Preprints, 6th IFAC Conference on Nonlinear Model Predictive Control
Madison, WI, USA, August 19-22, 2018

Data-driven Scenario Selection for Multistage Robust Model Predictive Control ⋆

Dinesh Krishnamoorthy, Mandar Thombre, Sigurd Skogestad, Johannes Jäschke

Department of Chemical Engineering, Norwegian University of Science & Technology , NO-7491 Trondheim , (e-mail: dinesh.krishnamoorthy@ntnu.no, mandar.thombre@ntnu.no, skoge@ntnu.no, johannes.jaschke@ntnu.no).

Abstract: A main assumption in many works considering multistage model predictive control (MPC) is that discrete realizations of the uncertainty are chosen a-priori and that the scenario tree is given. In this work, we focus on choosing the scenarios, which is an important practical aspect of scenario-based multistage MPC. In many applications, the distribution of the uncertain parameters is not available, but instead a finite set of data samples are available. Given this finite set of data samples, we present a data-driven approach to selecting the scenarios using principal component analysis (PCA). Using this approach, the scenarios are carefully selected such that the conservativeness of the solution can be reduced while still maintaining robustness towards constraint feasibility. The effectiveness of the proposed method is demonstrated using a simple example.

Keywords: Multistage MPC, Big data analysis, principal component analysis, MPC under uncertainty

1. INTRODUCTION

Model predictive control under uncertainty is an active research area that has received tremendous attention in the recent past, with developments in several different approaches to robust and stochastic MPC in the control literature. Many of these approaches solve an open loop optimization problem to determine the optimal control sequence, taking into account the uncertainty. However, this may not be optimal, since efficient handling of uncertainty requires feedback. In a recent review paper, Mayne (2014) notes that a better strategy would be to optimize over different control trajectories (closed-loop optimization) rather than a single control trajectory (open-loop optimization). One such closed-loop optimization strategy is the multistage scenario MPC also known as feedback min-max MPC or scenario-tree MPC (Scokaert and Mayne, 1998; Lucia et al., 2013).

In this approach, the evolution of the uncertainty in the prediction horizon is described by a scenario tree generated using discrete realizations of the uncertainty. By computing different control trajectories for the different scenarios, the notion of feedback, also known as recourse, is explicitly taken into account in the receding horizon implementation. This was later extended to nonlinear model predictive control by Lucia et al. (2013) in the framework of robust multistage MPC. The approach has since then received a lot of interest and has been applied to several chemical process systems (Lucia and Engell, 2013; Martí et al., 2015), autonomous vehicles (Klintberg et al., 2016), energy systems including power systems, oil and gas (Krishnamoorthy et al., 2016; Verheyleweghen and Jäschke, 2017), building climate control (Maiworm et al., 2015) etc. to name a few.

Most of these works assume that the uncertainty characteristics are known *a-priori* and that the discrete scenarios are given, for example, based on engineering insight before the MPC is designed. However, the issue of how to select the discrete realizations of the uncertainty for the scenario tree generation is an important practical aspect that has not been well studied in the control literature. Nevertheless, the problem has recently been considered in the operations research community under the topic of multistage stochastic optimization and is usually applied only for convex multistage optimization problems assuming full recourse. For example, Monte-Carlo sampling methods were considered in Shapiro (2003) and moment matching methods of the probability density functions (PDF) were used in Høyland et al. (2003). Lucia et al. (2013) also noted that the issue of how to generate the scenario tree for MPC applications is an important future research direction that must be addressed to enable practical implementation of such methods. Recently, a quadrature-based scenario tree generation was proposed using sparse grids by Leidereiter et al. (2014).

In many real applications, the probability distribution function (PDF) or the uncertainty set for the uncertain parameters is not readily available, but only a finite num-

⋆ D.K, S.S and J.J gratefully acknowledge the financial support from SFI SUBPRO. M.T, S.S and J.J gratefully acknowledge the financial support from FME HighEFF. Corresponding author: J.J.

ber of data samples may be available. Classical stochastic MPC frameworks make use of such data indirectly to infer the probability distribution of the uncertain problem parameters by means of statistical estimation methods. The estimated probability distribution function is then subsequently used in the optimization problem (Parys et al., 2016). Thus classical stochastic MPC problem is based on this two-step approach:

(1) estimate the PDF from the finite data samples
(2) use the estimated PDF in the optimization problem.

The main issue with this two step approach is that the estimation step often aims to achieve maximum prediction accuracy without tailoring it to the optimization problem. Hence, the estimated probability distribution function itself may be uncertain as noted by Parys et al. (2016) (leading to recent developments in the so-called distributionally robust optimization). In multistage MPC, the scenario tree is generated using a finite number of uncertainty representations. Given finite data samples, the uncertainty representations may be chosen directly from this data set, thus releasing the assumption of the uncertainty having any particular distribution.

Therefore in this paper, we propose a data-driven multistage scenario MPC problem that avoids the estimation of probability distribution functions and selects discrete realizations of the uncertainty from the finite set of data samples using Big data analytics.

In the case of multi-dimensional parametric uncertainty, the scenario tree becomes large. In such cases, careful selection of scenarios becomes very important to reduce the conservativeness and keep the computation cost low. Given a finite set of data for the different parameters, the use of univariate statistical analysis may fail to detect the relationship between the different parameters. Consequently, this often leads us to choose the scenarios assuming that the parameters are independent of one another. The resulting scenario tree may then span over an unnecessarily large uncertainty space leading to conservative solutions. Big data analytics can examine such large and varied data sets to uncover hidden correlations and can help us choose the scenarios. Therefore in our approach, the relationship between the different parameters is exploited to carefully choose only those combinations of parameters that are likely to be the true realization of the uncertain parameters.

In this paper, we address the issue of how to choose the discrete scenarios from a finite number of data samples and propose the use of multivariate data mining tools such as principal component analysis (PCA) to judiciously choose the scenarios in order to reduce the conservativeness. We use an example to motivate and demonstrate the use of PCA in choosing the scenarios for the multistage MPC formulation.

Methods such as principal component analysis have long since been used together with model predictive control. Among these, the two main application areas combining MPC and PCA has been 1) online performance monitoring (Loquasto and Seborg, 2003; Qin and Yu, 2007; AlGhazzawi and Lennox, 2009) and 2) model reduction (Maurath et al., 1988; Wang et al., 2002; Drgoňa et al., 2018). In an another interesting approach by Liu et al. (2006), the MPC framework is used to control the score space of the PCA to reduce variations in product specifications.

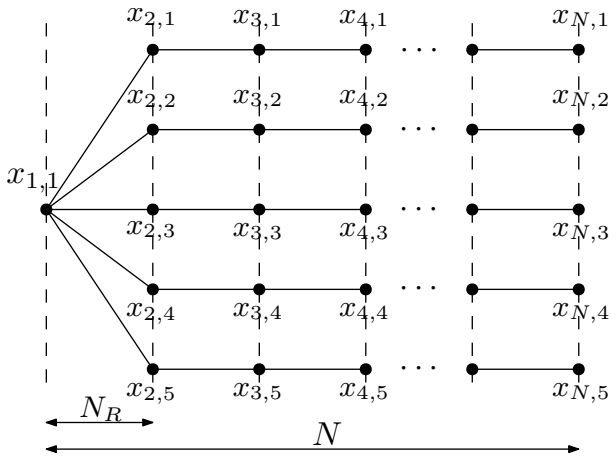

Fig. 1. Schematic representation of a scenario tree generated for $M = 5$ models and a robust horizon of $N_r = 1$.

The remainder of the paper is organized as follows. We introduce the multistage MPC problem in Section 2. Using a simple example, Section 3 motivates the need for data-mining techniques for choosing the discrete scenarios and describes the proposed data-driven multistage scenario MPC using principal component analysis (PCA). Simulation results for the corresponding multistage scenario MPC are provided in Section 4 as a proof-of-concept. Section 5 provides some useful discussions and future research directions towards Big data optimization with respect to multistage MPC before concluding the paper in Section 6.

2. MULTISTAGE MPC

Consider a discrete time nonlinear dynamic system

$$\mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k, \mathbf{p}) \tag{1}$$

where $\mathbf{x}_k \in \mathbb{R}^{n_x}$ and $\mathbf{u}_k \in \mathbb{R}^{n_u}$ denotes the states and inputs at time step k respectively and $\mathbf{p} \in \mathbb{R}^{n_p}$ denotes the vector of constant but uncertain parameters. The objective is to minimize a performance function $\mathbf{J}(\mathbf{x}_k, \mathbf{u}_k) : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \to \mathbb{R}$ while satisfying constraints $\mathbf{g}(\mathbf{x}_k, \mathbf{u}_k) \leq 0$ using an MPC with a prediction horizon of length N.

In multistage MPC, branching of the scenarios at each sample makes the problem size to grow exponentially over the prediction horizon. In order to curb the problem size, the scenario tree branching is stopped after a certain number of samples $N_r < N$ known as robust horizon as justified by Lucia et al. (2013).

Given M discrete realizations of the uncertainty and a robust horizon of length N_r, we then have $S = M^{N_r}$ discrete scenarios in the scenario tree as shown in Fig. 1. The resulting multistage MPC problem can be formulated as,

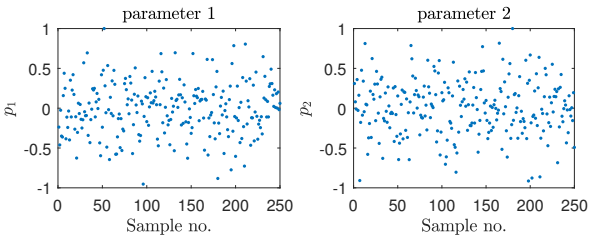

Fig. 2. Raw data of the two parameters p_1 (left subplot) and p_2 (right subplot).

$$\min_{\mathbf{x}_{k,j},\mathbf{u}_{k,j}} \quad \sum_{j=1}^{S} \omega_j \sum_{k=1}^{N} \mathbf{J}(\mathbf{x}_{k,j}, \mathbf{u}_{k,j}) \tag{2a}$$

s.t.

$$\mathbf{x}_{k+1,j} = \mathbf{f}(\mathbf{x}_{k,j}, \mathbf{u}_{k,j}, \mathbf{p}_j) \tag{2b}$$

$$\mathbf{g}(\mathbf{x}_{k,j}, \mathbf{u}_{k,j}) \leq 0 \tag{2c}$$

$$\sum_{j=1}^{S} \mathbf{E}_j \mathbf{u}_j = 0 \tag{2d}$$

$$\forall k \in \{1, \ldots, N\}, \forall j \in \{1, \ldots, S\}$$

where the subscript $(\cdot)_{k,j}$ denotes the time step k and scenario j and ω_j represents the weight given to each scenario. (2d) represents the non-anticipativity constraints with $\mathbf{u}_j = \left[\mathbf{u}_{0,j}^T \cdots \mathbf{u}_{N-1,j}^T\right]^T \in \mathbb{R}^{n_u N}$. The non-anticipativity constraints enforce the fact that all the decisions that branch at the same parent node are the same. This captures the real-time decision process correctly, since the control inputs cannot anticipate the future realization of the uncertainty, see Krishnamoorthy et al. (2018a,b) for more details on the structure of $\mathbf{E}_j$.

In this paper, we consider a constrained optimization problem under uncertainty, where the constraint feasibility must be ensured for any given realization of the uncertainty at the cost of conservativeness. Multistage MPC was shown to provide robust constraint feasible solutions whilst being less conservative than min-max approaches (Lucia et al., 2014).

In the next section, we will present a method for analyzing the data and choosing appropriate M discrete realizations of the uncertain parameters $\mathbf{p}$ given a finite set of data samples representing the uncertainty. Note that we require no knowledge on how the data is distributed, however, we assume that discrete historical data samples are available for the different uncertain parameters.

3. DATA-DRIVEN MULTISTAGE MPC

3.1 Motivating example

For the sake of simplicity, let us consider a system with two parameters ($n_p = 2$) and the finite data samples for each of the parameters are available as shown in Fig. 2. At first glance, the data samples tell us that each of the parameters vary in $[-1, 1]$. With no additional information, one often tends to assume that the parameters are uncorrelated, and assumes for example, a box uncertainty set. Consequently, the discrete realizations of the uncertainty from the four corners of the uncertainty set and the nominal value may be chosen, namely, $\mathbf{p}_j \in \{(-1,-1), (1,-1), (0,0), (-1,1), (1,1)\}$, to get a good representation of the uncertainty as shown in Fig. 3 (in gray shaded area). It is also often argued that a combination of extreme and nominal values of the all the parameters must be part of the scenario tree (Lucia et al., 2013).

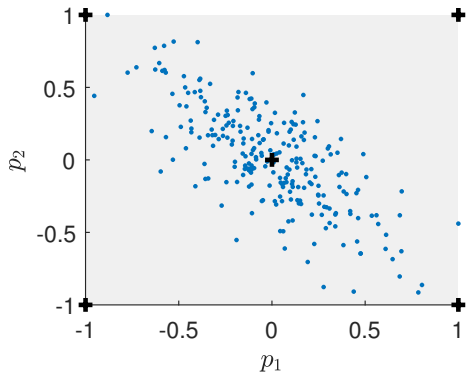

Fig. 3. Multivariate plot of the two parameters. The points $\mathbf{p}_j \in \{(-1,-1), (1,-1), (0,0), (-1,1), (1,1)\}$ are represented by the black '+' and the gray shaded area represents the univariate limits of the two parameters.

However, plotting the two parameters against each other gives us more information about the relationship between the two parameters, as shown in Fig. 3. One can immediately see that $\mathbf{p}_j$ selected using independent parameter variations includes parameter combinations that are unlikely to be the true realization of the uncertainty. Seeking robustness against parameter combinations that are not likely can lead to very conservative and hence suboptimal operation. The information from the simple multivariate plot in Fig. 3 gives us more information into the data's hidden structure which can be exploited to choose the different uncertainty realizations that are more likely. This simple two parameter example already motivates the need for multivariate data analysis methods when choosing the discrete scenarios for multistage MPC.

When the number of parameters and the number of data points increases, it can be cumbersome and time consuming, if not impossible, to plot two parameters at a time to find out the hidden structures in the data simply due to information overload and the effort required to make each plot. Multivariate data mining approaches that attempt to find the hidden structure in big data sets can thus lead to more information that would not have been otherwise discovered. This can directly be exploited in the scenario tree generation to select and include only those parameter combinations that are likely to be the true realization of the uncertainty.

3.2 Data mining using principal component analysis

Principal component analysis (PCA) is a universal data mining tool for extracting useful information hidden in massive amounts of data (Seber, 1984). Principal component analysis attempts to explain the variability in a given set of data by separating the data into so-called *principal components* (PC) where each PC contributes to explaining the total variability of the data. More specifically, PCA uses an orthogonal transformation to convert

a set of (possibly correlated) data into a set of linearly uncorrelated principal components. This transformation is such that the first principal component explains the largest variance in the data set and the other PCs are ordered in decreasing component variance. A principal component therefore points out which variables contribute most to the observed variability in the data and finds the relationship between the different variables (Rao, 1964). In simplest terms, PCA can be thought of as fitting a multi-dimensional ellipsoid to the data, where each axis of the ellipsoid represents a principal component (Hotelling, 1933).

Consider a data set with n_o number of observations for each parameter and the data is represented by a data matrix $\mathbf{P} \in \mathbb{R}^{n_o \times n_p}$. It is important to note that PCA is sensitive to the scaling of the variables and hence the data must be scaled. In addition, mean-centering is also necessary to ensure that the first principal component describes the direction of maximum variance (Jolliffe, 1986).

Therefore, let $\mathbf{P}_0 \in \mathbb{R}^{n_o \times n_p}$ be the scaled and mean-centered data matrix corresponding to $\mathbf{P}$. PCA returns the bilinear model

$$\mathbf{P}_0 = \mathbf{\Lambda}\mathbf{C}^T \tag{3}$$

where the matrix $\mathbf{\Lambda} \in \mathbb{R}^{n_o \times n_p}$ contains the so-called scores (left-hand eigenvectors). The scores represent the distance of the different data points from the mean along the direction of the principal components. The matrix $\mathbf{C} \in \mathbb{R}^{n_p \times n_p}$ contains the coefficients of the principal components which represents the weight by which each original data point should be multiplied to get the component score.

The principal components, scores and coefficients are useful means of understanding the correlation between the different parameters. This information can be exploited in choosing the scenarios as explained in the section below.

In the following subsection, we show how PCA can be used to select scenarios for the multistage robust MPC framework, which to the best of our knowledge has not been used before.

3.3 Scenario generation using data

We now describe how the scores and the coefficients from the principal component analysis can be used to select the discrete realizations of the uncertain parameters. The variance in the scores along the different principal components can be used to describe the uncertainty set instead of using the univariate parameter data. To do this, we pick the data points corresponding to the maximum and minimum scores along the directions of the different principal components that explains the variability with sufficient component variance. Using the coefficients of the principal components, we can then transform this to the original parameter space. These points now form the discrete realizations of the uncertainty that represents the uncertainty space.

This is further illustrated using the data set for two parameters shown in Fig. 2 and Fig. 3. The score plot for this data set is shown in Fig. 4, where the data points corresponding to the maximum and minimum scores along first and second principal component directions are shown

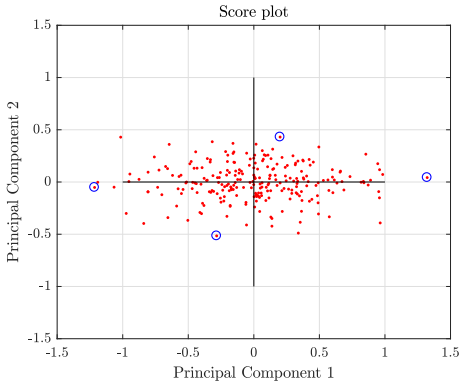

Fig. 4. Score plot along the two principal component directions. The data points corresponding to the maximum and minimum scores along the two PC directions are shown in blue circles.

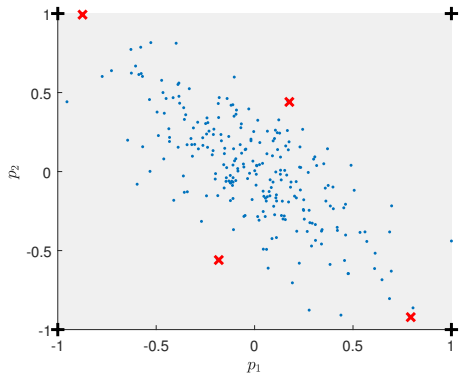

Fig. 5. Data plot in the original space, with realizations corresponding to maximum scores on first and second principal components marked by red '×'. The black '×' correspond to realizations picked by simply taking the combination of extreme values of the parameters.

in blue circles. These are then transformed into the original parameter space as shown in Fig. 5 using a red '×'. We can see that the discrete realizations selected using principal component analysis captures the parameter variations more tightly than the ones chosen by looking at the parameter variations independently.

The proposed approach can thus be summarized by the following steps,

(1) Scale and mean center the data set $\mathbf{P}$ to obtain $\mathbf{P}_0$.
(2) Perform PCA to compute the principal components and the scores for each of the data points $\mathbf{\Lambda}$ and the corresponding co-efficient matrix $\mathbf{C}$ of the principal components.
(3) Pick out the maximum and minimum scores along the direction of the different principal components that sufficiently explain the total variance of the data.
(4) Using the coefficient matrix, re-transform the selected scores from step 3 to the original data space.
(5) Generate the scenario tree based on the discrete realizations of the uncertainty selected in step 4.

4. ILLUSTRATIVE EXAMPLE

In this section, we now compare the effect of the discrete realizations of the uncertainty on the performance of the multistage scenario MPC. We consider a simple example, where the system is given by a model with two states $\mathbf{x} = [x_1, x_2]^T$ and one control input $\mathbf{u} = u$ and two uncertain parameters $\mathbf{p} = [p_1, p_2]^T$ as shown below,

$$\dot{x}_1 = \frac{1}{\tau}\left(-3.5u^2 + 30u - x_1\right) \tag{4}$$
$$\dot{x}_2 = \frac{1}{\tau}\left(4u + 2p_1 + 4p_2 + 10 - x_2\right)$$

with $\tau = 5s$ being the time constant.

The objective is to maximize x_1 while satisfying constraints on x_2 despite the uncertainty in p_1 and p_2. We apply the multistage scenario MPC approach (2) with,

- stage cost $\mathbf{J}(\mathbf{x}_{k,j}, \mathbf{u}_{k,j}) = -x_1$,
- system model (4) discretized using third order direct collocation,
- inequality constraint (2c): $x_2 \leq 20$,
- uncertain parameters discretized into $M = 5$ realizations of the uncertain parameter
- non-anticipativity constraints (2d).

We choose a prediction horizon of T=1 min divided equally into $N = 60$ samples and a robust horizon of $N_r = 2$ samples (25 scenarios). The true realization of the parameters for the simulation was chosen to be at its nominal value $(0, 0)$. The resulting multistage MPC was implemented in `MATLAB` using `CasADi` algorithmic differentiation tool (Andersson, 2013) version 3.1.0, and `IPOPT` solver (Wächter and Biegler, 2006) was used to solve the resulting nonlinear programming problem.

We first simulate the multistage scenario MPC using $\mathbf{p}_j$ selected using the parameters variations independently, as shown in Table.1 (Simulation 1) and in Fig. 5 using black '+'. This corresponds to using the corner points of the box $[-1 \quad 1] \times [-1 \quad 1]$. The points at the boundaries that constitute a combination of the minimum and maximum values of the uncertain parameters along with the nominal point (0,0) have been selected to get a good representation of the uncertainty based on the time series (univariate) data in Fig.2. This simulation is used as a benchmark.

We then solve the same problem but by replacing $\mathbf{p}_j$ which is now selected using principal component analysis as shown in Table.1 (Simulation 2) and Fig. 5 using red '×'. The simulation results are compared in Fig. 6. The left subplot shows x_1 which has to be maximized, the right subplot shows x_2 which must be maintained below its maximum value of 20. It can be clearly seen from the simulation results that the scenarios chosen using principal component analysis is much less conservative than the scenarios chosen using the parameter data independently. This is because, in the proposed approach, we do not consider scenarios in the scenario tree that are not likely to be the true realization of the uncertainty.

We then simulated the system for 30 runs with different randomly chosen realizations of the uncertain parameters in the plant simulator as shown in Fig. 7 (right subplot). To evaluate the performance, we also plot the integrated objective (left subplot), which is the objective function J

Table 1. Discrete realizations of $\mathbf{p}$ used in the simulations

j	Simulation 1		Simulation 2	
	p_1	p_2	p_1	p_2
1	1	1	0.18	0.44
2	1	-1	0.79	-0.92
3	0	0	0	0
4	-1	1	-0.87	0.99
5	-1	-1	-0.18	-0.56

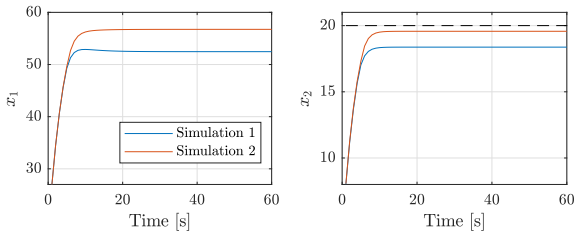

Fig. 6. Simulations results with two different set of scenarios.

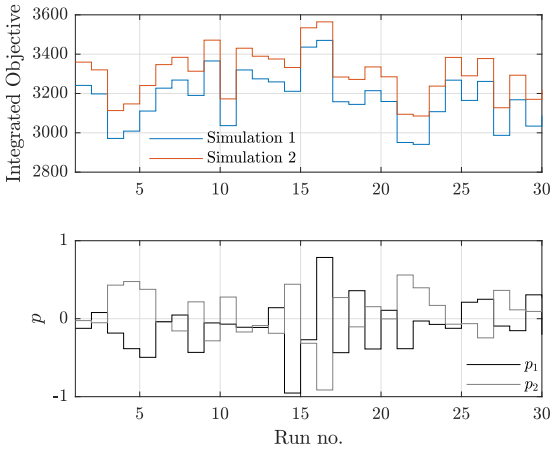

Fig. 7. Monte Carlo Simulations results with different realizations of the uncertain parameters.

integrated over the entire simulation time of $t = 60$ min for each simulation, i.e. integrated objective

$$J_{int} = \int_{t=0}^{t=60} J(t)\,\mathrm{d}t.$$

It can be clearly seen that by using the scenarios selected using the PCA method, we are able to improve the performance for different realizations of the uncertainty from the given data set whilst being robust feasible.

5. DISCUSSION AND FUTURE WORK

In this paper, we proposed to use data mining approaches to select the scenarios based on a finite set of data samples. Note that we have purposefully used a simple example with two uncertain parameters to clearly demonstrate the concept to readers of any level of expertise with such methods. Indeed, the full potential of such data-mining techniques is realized for large data sets with multidimensional parameters, where it may be difficult to select the scenarios purely based on engineering intuition and univariate analysis. For example, consider the building

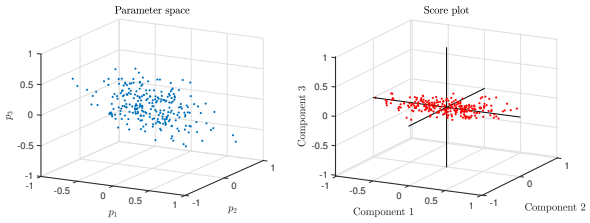

Fig. 8. Data samples for three parameters and the corresponding score plot.

climate control problem, where many uncertain parameters such as temperature, humidity, cloud cover, solar radiation, building occupancy level etc. affect the building climate control problem. Multistage MPC for the building climate control was shown to be a promising approach in Maiworm et al. (2015). One or more of these parameters affecting the climate control problem may be correlated, such as cloud cover, solar radiation and temperature. Using historical data of the weather conditions and building occupancy levels, multivariate data mining techniques can be used to appropriately choose the scenarios to reduce the conservativeness, instead of picking the scenarios based on a combination of maximum and minimum values of the different parameters. By doing so, one can potentially reduce the number of scenarios or the span of the scenario tree to be used in the multistage MPC problem. The application of the proposed data-based scenario selection approach for a building climate control problem is an ongoing work.

5.1 Scenario reduction using variability explanation

Methods such as principal component analysis also provides the percentage of variability explained by the different principal components. This information can in addition, be used to discard scenarios that do not sufficiently explain the variability in the data, hence reducing the number of scenarios that must be considered in the multistage scenario MPC problem. This can help reduce the problem size. For example, consider a different data set for three parameters as shown in Fig. 8. The PCA for this data set returns three principle components, where the first principle component explains 72.5% and the second principal component explains 26.9% of the variability in the data. The third principal component explains only 0.48% of the variability. Based on this, we can then select the maximum and minimum scores along the direction of the first and second principle components and discard the scenario combinations along principal component 3, since it does not sufficiently explain the variability of the data. This helps in reducing the number of scenarios to be included in the scenario tree.

5.2 Weighting in the MPC cost function

The different scenarios can be weighted in the optimization problem as shown in (2a). The results from principal component analysis can not only be used to select the scenarios from the data, but also provide a weight for the selected scenario.

As mentioned earlier, the scores provided by the PCA represent how far a data point is from the mean along the direction of the principal components. Since the data matrix is mean-centered, the data points with large scores are far away from the mean and the vice versa. The weight given to a data point that is far away from the mean (i.e. large score) must be low, compared to the weight given to a data point that is closer to the mean (i.e. low score). Therefore, the weights for the discrete scenarios selected by the PCA method are chosen to be inversely proportional to its score.

5.3 Online update of scenarios

In this paper, we assumed that a finite set of data samples are available which was used to select the scenarios offline using principal component analysis. As more data points become available, PCA can also be used online to continuously adapt the scenarios to reflect the most recent data points. This can be especially useful when the uncertain parameters are time varying in nature. As more data points become available, this information can be included to update the different scenarios in the multistage MPC formulation.

5.4 Other data analytic methods

It must be noted here that PCA does have its limitations, although it works well with the example considered in this work. PCA aims to find hidden linear correlations within the data set, and is thus lacking when data has inherently nonlinear correlations. Further, it only finds PCs that are orthogonal to each other, whereas the projections within the data with highest variance may be nonorthogonal in nature.

For data that is not linearly separable, other data classifiers such as the nonlinear support vector machines (SVM) may be used. The nonlinear SVM maps the given data into a higher-dimensional space using so-called kernel functions, and the transformed data is then linearly separable. Another avenue for further research in improving upon the proposed methodology would be to use advanced data mining techniques for outlier detection. This would be helpful in eliminating the selection of parameters that correspond to 'unlikely' scenarios and help reducing the conservativeness of the solution.

In the previous section, we selected the scenarios corresponding to the maximum scores along the different PC directions. This was done in order to ensure robust constraint feasibility for any realization of the uncertain parameters from the given data set. Alternatively, the scenarios can be chosen based on the scores that falls within some user-defined percentile along the different PC directions to further reduce the conservativeness by trading off on the constraint satisfaction. For example, the scenarios can be chosen using the scores that fall within the 90th percentile in order to reduce the conservativeness to ensure constraint satisfaction with a given probability (analogous to using a chance constrained MPC formulation). Alternatively, one may also use the associated probabilities of the data points to appropriately choose the scenarios. Note that more rigourous analysis must be carried out to get an equivalent performance as using a chance constrained optimization, which is another useful future research direction.

6. CONCLUSION

To conclude, we have motivated the need to develop methods to appropriately choose the scenarios based on a finite sample of historical data. Using a simple example we have demonstrated the concept of how data mining techniques such as principal component analysis can be used to uncover hidden structures in the data, which can then be exploited in choosing the necessary scenarios and discarding the scenarios that need not be considered in the optimization problem. This leads to a less conservative solution as demonstrated in the simulation example. We have also provided some discussions on possible research avenues towards using data mining techniques for scenario selection and hope to stimulate further research in this direction.

REFERENCES

AlGhazzawi, A. and Lennox, B. (2009). Model predictive control monitoring using multivariate statistics. *Journal of Process Control*, 19(2), 314–327.

Andersson, J. (2013). *A General-Purpose Software Framework for Dynamic Optimization.* PhD thesis, KU Leuven.

Drgoňa, J., Picard, D., Kvasnica, M., and Helsen, L. (2018). Approximate model predictive building control via machine learning. *Applied Energy*, 218, 199–216.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6), 417.

Høyland, K., Kaut, M., and Wallace, S.W. (2003). A heuristic for moment-matching scenario generation. *Computational optimization and applications*, 24(2-3), 169–185.

Jolliffe, I.T. (1986). Principal component analysis and factor analysis. In *Principal component analysis*, 115–128. Springer.

Klintberg, E., Dahl, J., Fredriksson, J., and Gros, S. (2016). An improved dual newton strategy for scenario-tree mpc. In *IEEE 55th Conference on Decision and Control (CDC), 2016*, 3675–3681. IEEE.

Krishnamoorthy, D., Foss, B., and Skogestad, S. (2016). Real time optimization under uncertainty - applied to gas lifted wells. *Processes*, 4(4). doi:10.3390/pr4040052.

Krishnamoorthy, D., Foss, B., and Skogestad, S. (2018a). A distributed algorithm for scenario-base model predictive control using primal decomposition. *IFAC ADCHEM 2018 (In-Press)*.

Krishnamoorthy, D., Suwartadi, E., Foss, B., Skogestad, S., and Jäschke, J. (2018b). Improving scenario decomposition for multistage mpc using a sensitivity-based path-following algorithm. *IEEE Control Systems Letters*, 2(4). doi:10.1109/LCSYS.2018.2845108.

Leidereiter, C., Potschka, A., and Bock, H.G. (2014). Quadrature-based scenario tree generation for nonlinear model predictive control. *IFAC Proceedings Volumes*, 47(3), 11087–11092.

Liu, X., Chen, X., Wu, W., and Zhang, Y. (2006). Process control based on principal component analysis for maize drying. *Food control*, 17(11), 894–899.

Loquasto, F. and Seborg, D.E. (2003). Model predictive controller monitoring based on pattern classification and pca. In *American Control Conference, 2003. Proceedings of the 2003*, volume 3, 1968–1973. IEEE.

Lucia, S. and Engell, S. (2013). Robust nonlinear model predictive control of a batch bioreactor using multi-stage stochastic programming. In *Control Conference (ECC), 2013 European*, 4124–4129. IEEE.

Lucia, S., Finkler, T., and Engell, S. (2013). Multi-stage nonlinear model predictive control applied to a semi-batch polymerization reactor under uncertainty. *Journal of Process Control*, 23(9), 1306–1319.

Lucia, S., Paulen, R., and Engell, S. (2014). Multi-stage nonlinear model predictive control with verified robust constraint satisfaction. In *IEEE 53rd Annual Conference on Decision and Control (CDC), 2014*, 2816–2821. IEEE.

Maiworm, M., Bäthge, T., and Findeisen, R. (2015). Scenario-based model predictive control: Recursive feasibility and stability. *IFAC-PapersOnLine*, 48(8), 50–56.

Martí, R., Lucia, S., Sarabia, D., Paulen, R., Engell, S., and de Prada, C. (2015). Improving scenario decomposition algorithms for robust nonlinear model predictive control. *Computers & Chemical Engineering*, 79, 30–45.

Maurath, P.R., Laub, A.J., Seborg, D.E., and Mellichamp, D.A. (1988). Predictive controller design by principal components analysis. *Industrial & engineering chemistry research*, 27(7), 1204–1212.

Mayne, D.Q. (2014). Model predictive control: Recent developments and future promise. *Automatica*, 50(12), 2967–2986.

Parys, B.P.G.V., Kuhn, D., Goulart, P.J., and Morari, M. (2016). Distributionally robust control of constrained stochastic systems. *IEEE Transactions on Automatic Control*, 61(2), 430–442. doi: 10.1109/TAC.2015.2444134.

Qin, S.J. and Yu, J. (2007). Recent developments in multivariable controller performance monitoring. *Journal of Process Control*, 17(3), 221–227.

Rao, C.R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhyā: The Indian Journal of Statistics, Series A*, 329–358.

Scokaert, P. and Mayne, D. (1998). Min-max feedback model predictive control for constrained linear systems. *IEEE Transactions on Automatic Control*, 43(8), 1136–1142.

Seber, G. (1984). Multivariate analysis of variance and covariance. *Multivariate observations*, 433–495.

Shapiro, A. (2003). Monte carlo sampling methods. *Handbooks in operations research and management science*, 10, 353–425.

Verheyleweghen, A. and Jäschke, J. (2017). Framework for combined diagnostics, prognostics and optimal operation of a subsea gas compression system. *IFAC-PapersOnLine*, 50(1), 15916–15921.

Wächter, A. and Biegler, L.T. (2006). On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical programming*, 106(1), 25–57.

Wang, P., Litvak, M., and Aziz, K. (2002). Optimization of production operations in petroleum fields. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers.