CHEMICAL ENGINEERING TRANSACTIONS

VOL. 70, 2018

A publication of
AIDIC
The Italian Association
of Chemical Engineering
Online at www.aidic.it/cet

Guest Editors: Timothy G. Walmsley, Petar S. Varbanov, Rongxin Su, Jiří J. Klemeš
Copyright © 2018, AIDIC Servizi S.r.l.
ISBN 978-88-95608-67-9; **ISSN** 2283-9216

DOI: 10.3303/CET1870307

Hedging Against Uncertainty in Biomass Processing Network Design Using a Data-Driven Approach

Chao Ning[a], Daniel J. Garcia[b], Fengqi You[a,*]

[a]Cornell University, Ithaca, New York, 14853, USA
[b]Northwestern University, 2145 Sheridan Road, Evanston, Illinois, 60208, USA
fengqi.you@cornell.edu

Sustainability has become a key concern in process systems engineering. Renewable energy, such as biofuels, and renewable materials, such as bioproducts, could replace their non-renewable, petroleum-based counterparts. However, there remain many challenges in producing biofuels and bioproducts economically and efficiently. There are many different biomass feedstocks, processes to convert them, and many different possible biofuels and bioproducts to produce. Furthermore, prices and demands of biofuels and bioproducts are uncertain. The variation of price or demand of one bioproduct could influence price or demand of another, further complicating the problem. An approach that can identify economical, efficient, and sustainable biofuel and bioproduct production processes from the myriad possible options while also considering correlated and uncorrelated price and demand uncertainties of the final bioproducts is required. In this work, a data-driven decision-making framework is proposed for biomass processing network design that directly integrates machine learning with robust optimization. Principal component analysis (PCA) is used to identify latent uncertainties behind observed uncertainty data. A kernel density estimation approach captures probability distributions of the projected uncertainty data extracted from PCA. This uncertainty data analysis approach is applied to a bioconversion product and process network to identify cost-effective and environmentally-friendly biofuels and bioproducts production pathways. Our approach identifies a total annualized cost of $ 18.3M/y, 6 % lower than the cost found with conventional adaptive robust optimization.

1. Introduction

Production strategies for sustainable fuels and chemicals have advanced significantly in recent years (Yue et al., 2014). Multiple works aim to identify sustainable and cost-effective biomass conversion pathways from a host of possible feedstock, processing pathway, and final product options (Garcia and You, 2015). For biofuels and bioproducts to be truly sustainable alternatives to their non-renewable counterparts, their production must be sustainable in the face of constantly changing and uncertain market scenarios (Glenna and Cahoy, 2009). Static robust optimization (Bertsimas et al., 2011) and adaptive robust optimization (ARO) (Ben-Tal et al., 2004) have been employed to useful effect to consider uncertainty in decision-making, including for hydrogen refinery networks (Wei et al., 2017) or facility location problem (Hrabec et al., 2017). Some previous works consider uncertainty in select parameters, such as feedstock price and final product demand (Gong et al., 2016). However, price and demand data uncertainty of several different bioproducts may be correlated, suggesting previous ARO approaches may fall short in appropriately modelling uncertainties. Conventional robust optimization and ARO do not account for the structure and properties of uncertainty data (Ning and You, 2018). With increasing access to increasing amounts of uncertainty data, data-driven robust optimization has been proposed to begin tackling this shortcoming (Ning and You, 2017). However, there is still a need for a framework that can better analyse and utilise uncertainty datasets while leveraging their structure, including any correlation between uncertain parameters, to identify less conservative optimal solutions compared to conventional robust optimization.

In this work, machine learning techniques are integrated with ARO to better analyse uncertainty datasets. Uncertainty data must first be analysed by transforming the original correlated uncertain parameters into their uncorrelated components with a principal component analysis (PCA) (Wold et al., 1987). Next, the probability

Please cite this article as: Ning C., Garcia D.J., You F., 2018, Hedging against uncertainty in biomass processing network design using a data-driven approach , Chemical Engineering Transactions, 70, 1837-1842 DOI:10.3303/CET1870307

distributions of the transformed, uncorrelated uncertain parameters are defined with kernel density estimation (KDE) methods (Friedman et al., 2001) to arrive at more representative uncertainty sets than typical sets employed in conventional robust optimization approaches. Based on the uncertainty set using PCA and KDE, A data-driven adaptive robust biomass processing network design model is proposed and applied to a bioconversion product and process network to determine optimal biofuel and bioproduct production pathways under uncertain feedstock price and product demands.

2. Uncertainty set construction using machine learning techniques

In this section, a data-driven approach is presented to construct polyhedron uncertainty sets directly from uncertainty data following the work by Ning and You (2018). First, the latent uncertainty along each principal component is identified from observed uncertainty data using the PCA technique. The KDE method is employed to extract distributional information, which is then incorporated into a data-driven uncertainty set.

Consider an uncertainty data matrix X=[u$^{(1)}$,…, u$^{(N)}$]^T, in which each row represents an uncertainty data point in *m*-dimensional space. There is a total of *N* uncertainty data points. The PCA technique is capable of modelling high-dimensional uncertainty, and accurately extracts the first-order and second-order moment information from uncertainty data. PCA identifies the uncorrelated principal components via the eigenvalue decomposition of the sample covariance matrix (Wold et al., 1987). Data matrix *X* is scaled to zero-mean, which is shown in Eq(1).

$$\mathbf{X}_0 = \mathbf{X} - \mathbf{e}\boldsymbol{\mu}_0^T \tag{1}$$

where X$_0$ is an uncertainty data matrix after scaling, e denotes a column vector of all ones, and μ$_0$ represents the mean vector of uncertainty data.

The covariance matrix of uncertain parameters can be approximated with the sample covariance matrix S, as shown in Eq(2).

$$\mathbf{S} = \frac{1}{N-1}\mathbf{X}_0^T\mathbf{X}_0 \tag{2}$$

Eigenvalue decomposition leads to S=PΛP^T. The square matrix P consists of all the *m* eigenvectors, and Λ is a diagonal matrix consisting of all the eigenvalues. The proposed method then projects uncertainty data onto each principal component, which is shown below.

$$\mathbf{t}_k^{(i)} = \mathbf{p}_k^T\left[\mathbf{u}^{(i)} - \boldsymbol{\mu}_0\right] \tag{3}$$

where p$_k$ is the *k*-th principal component, and t$_k^{(i)}$ is the projection of uncertainty data point u$^{(i)}$ onto the *k*-th principal component. Let ξ$_k$ be the latent uncertainty along the *k*-th principal component. The estimated probability density function for ξ$_k$ by using the KDE method is given by

$$\hat{f}_{\text{KDE}}^{(k)}(\xi_k) = \frac{1}{N}\sum_{i=1}^{N} K_h\left(\xi_k, \mathbf{t}_k^{(i)}\right) \tag{4}$$

where *K$_h$* is a kernel function. In this work, the Gaussian kernel function is used, which is shown below.

$$K_h\left(\xi_k, \mathbf{t}_k^{(i)}\right) = \left(\frac{1}{\sqrt{2\pi}h}\right)\exp\left(-\left\|\xi_k - \mathbf{t}_k^{(i)}\right\|^2 \Big/ 2h^2\right) \tag{5}$$

Based on the cumulative density function of latent uncertainty ξ$_k$, the corresponding quantile function can be expressed as follows:

$$\hat{F}_{\text{KDE}}^{(k)\,-1}(\alpha) = \min\left\{\xi_k \in R \,\middle|\, \hat{F}_{\text{KDE}}^{(k)}(\xi_k) \ge \alpha\right\} \tag{6}$$

where α is a predefined parameter. Thus, the data-driven uncertainty set is presented as follows:

$$U_{\text{PCA+KDE}} = \left\{\mathbf{u} \,\middle|\, \begin{array}{l} \mathbf{u} = \boldsymbol{\mu}_0 + \mathbf{P}\xi,\ \xi = \underline{\xi} \circ \mathbf{z}^- + \overline{\xi} \circ \mathbf{z}^+, \\ \mathbf{0} \le \mathbf{z}^-,\ \mathbf{z}^+ \le \mathbf{e},\ \mathbf{z}^- + \mathbf{z}^+ \le \mathbf{e},\ \mathbf{e}^T\left(\mathbf{z}^- + \mathbf{z}^+\right) \le \Phi \\ \underline{\xi} = \left[\hat{F}_{\text{KDE}}^{(1)\,-1}(\alpha), \ldots,\ \hat{F}_{\text{KDE}}^{(m)\,-1}(\alpha)\right]^T \\ \overline{\xi} = \left[\hat{F}_{\text{KDE}}^{(1)\,-1}(1-\alpha), \ldots,\ \hat{F}_{\text{KDE}}^{(m)\,-1}(1-\alpha)\right]^T \end{array}\right\} \tag{7}$$

where $U_{\text{PCA+KDE}}$ is the data-driven uncertainty set using PCA and KDE, z^- is a backward deviation vector, z^+ is a forward deviation vector, e is a column vector of all ones, and Φ is an uncertainty budget. The uncertainty budget can be interpreted as the maximum number of latent uncertainties allowed to deviate from their mean values. Note that the notation $\circ$ denotes the Hadamard product. ξ is the latent uncertainty vector. The lower and upper bound vectors describe the minimum and maximum levels of latent uncertainties, respectively. They jointly define the confidence interval of latent uncertainties according to the probability density estimation obtained using the KDE method. α is a parameter that specifies the size of intervals according to confidence level of (1-2α). The confidence interval becomes smaller as α increases.

The uncertainty set in (7) is essentially a polytope. A novel feature of this uncertainty set is that it effectively incorporates the correlations and distributional information embedded within the uncertainty data. Additionally, the uncertainty set $U_{\text{PCA+KDE}}$ is not necessarily symmetric, as it accounts for the forward and backward deviations separately. As a result, it flexibly adapts to the intrinsic structure and complexity of uncertainty data.

3. Data-driven adaptive robust biomass processing network design model

In this section, a data-driven adaptive robust biomass processing network design model is presented using the uncertainty set in the previous section. In a production conversion network, biomass feedstocks, such as corn and switchgrass, are converted into a variety of biofuels and bioproducts via different processing and upgrading technologies (Garcia and You, 2015). One needs to make decisions on the selection of technology pathway, capacity and operating level of each technology, purchase amounts of feedstocks and quantities of products to sell. The goal is to minimize the total annualized cost. The first-stage decision variables are decisions on the selection and the capacity of technology. The second-stage decisions include production levels, quantity of biomass feedstock to purchase and amounts of products to sell.

The data-driven ARO model for bioconversion network design can be cast as a multi-level mixed-integer program. All the decision variables are separated into the first-stage decisions that are made before uncertainty is revealed, and the second-stage decisions that are adjustable to uncertainty realization. The objective function of the bioconversion network design is shown in Eq(8). The constraints include technology capacity constraint Eq(9), production level constraint Eq(10), mass balance constraint Eq(11), biomass feedstock availability constraint Eq(12), biofuel product demand satisfaction constraint Eq(13), non-negativity and integrity constraints Eq(14)- Eq(15). The data-driven uncertainty sets for product demand and feedstock prices are shown in Eq(16) and Eq(17).

$$\min_{Y_i, Q_i} \sum_{i \in I} c_{1,i} E_i + \max_{d \in U_1,\, fp \in U_2} \min_{W_i, P_j, S_j} \sum_{i \in I} c_{2,i} W_i + \sum_{j \in J} fp_j P_j \tag{8}$$

$$\text{s.t.}\ \ q_i^L \cdot Y_i \le Q_i \le q_i^U \cdot Y_i, \quad \forall i \tag{9}$$

$$W_i \le Q_i, \quad \forall i \tag{10}$$

$$P_j - \sum_i \kappa_{ij} \cdot W_i - S_j = 0, \quad \forall j \tag{11}$$

$$P_j \le b_j, \quad \forall j \tag{12}$$

$$S_j \ge d_j, \quad \forall j \tag{13}$$

$$Q_i,\ P_j,\ S_j,\ W_i \ge 0 \quad \forall i, j \tag{14}$$

$$Y_i \in \{0,1\}, \quad \forall i \tag{15}$$

$$U_1 = \left\{ d_j \left| \begin{array}{l} d_j = d_j^0 + \sum_k p_{jk} \left(\xi_k^L \cdot z_k^- + \xi_k^U \cdot z_k^+ \right),\ \forall j \\ \sum_k \left(z_k^- + z_k^+ \right) \le \Phi^{\text{dem}},\ z_k^- + z_k^+ \le 1,\ 0 \le z_k^-,\ z_k^+ \le 1 \end{array} \right. \right\} \tag{16}$$

$$U_2 = \left\{ fp_j \left| \begin{array}{l} fp_j = fp_j^0 + \sum_l r_{jl} \left(\beta_l^L \cdot \delta_l^- + \beta_l^U \cdot \delta_l^+ \right),\ \forall j \\ \sum_l \left(\delta_l^- + \delta_l^+ \right) \le \Phi^{pri},\ \delta_l^- + \delta_l^+ \le 1,\ 0 \le \delta_l^-,\ \delta_l^+ \le 1 \end{array} \right. \right\} \tag{17}$$

where Q_i is a decision on total capacity of technology *i*, Y_i is a binary decision variable to reflect whether technology *i* is selected in the pathway, W_i denotes the production level of technology *i*, P_j is the purchase quantity of compound *j*, and S_j is the sale amount of product *j*. $c_{1,i}$ and $c_{2,i}$ represent the coefficients for economic evaluation. fp_j denotes the feedstock price of compound *j*, κ_{ij} is a mass balance coefficient, b_j is the availabilities of feedstocks, and d_j is the demand for compound *j*. q_i^L and q_i^U are upper and lower bounds of process capacity. Note that E_i is a piecewise linear function.

U_1 is the data-driven uncertainty set for demand. d_j^0 is the mean value of demand d_j, p_{jk} is the j-th element of the k-th principal component for demand, the backward deviation is z_k^-, and the forward deviation is z_k^+. ξ_k^L is the lower bound of the k-th latent demand uncertainty, and ξ_k^U is the upper bound of the k-th latent demand uncertainty. Φ^{dem} is the demand uncertainty budget enforcing the maximum deviations. U_2 is the data-driven uncertainty set for price, fp_j^0 is the mean value of product price, and r_{jl} is the j-th element of the l-th principal component for price. δ_l^- and δ_l^+ are backward and forward deviation vectors for feedstock price. β_l^L and β_l^U are the lower and upper bounds of latent price uncertainty. Φ^{pri} is the price uncertainty budget, which is used to specify the maximum deviations allowed in latent price uncertainties.

The resulting biomass processing network design problem is cast as a multi-level mixed-integer program. It explicitly takes advantage of machine learning methods to decipher the uncertainty data for the decision making on the biomass processing network design.

4. Case study

This section considers a large-scale bioconversion process network taken from (Garcia and You, 2015). In this comprehensive network, there are 142 compounds and 197 process technologies. Because of the market fluctuations, all feedstock prices and biofuel product demands are subject to uncertainty. Since biofuel product demand and biomass feedstock price are well-documented, these two types of uncertainties are considered following the literature (Gong et al., 2016). For the feedstock price uncertainty, a set of 20,000 uncertainty data are used for constructing the uncertainty set. Each data point has eight dimensions for a combination of all feedstocks. Regarding the product demand uncertainty, 1,000 uncertainty data points are utilized, and each data point represents a combination of all the three biofuel products, namely biodiesel, gasoline and ethanol.

In this case study, the conventional ARO method with a gamma uncertainty set is also used in addition to the proposed data-driven ARO approach to demonstrate the advantages. All optimization problems are modelled in GAMS 24.7.3 (Rosenthal, 2008), solved with CPLEX 12.6.3. The optimality gap for CPLEX 12.6.3 is set to be 0, and the relative optimality gap for the decomposition algorithm is 10^{-6}. The uncertainty budget of feedstock prices is set to be two, and demand uncertainty budget is set to be one for both methods. The number of pieces in the piecewise linear function E_i is 50.

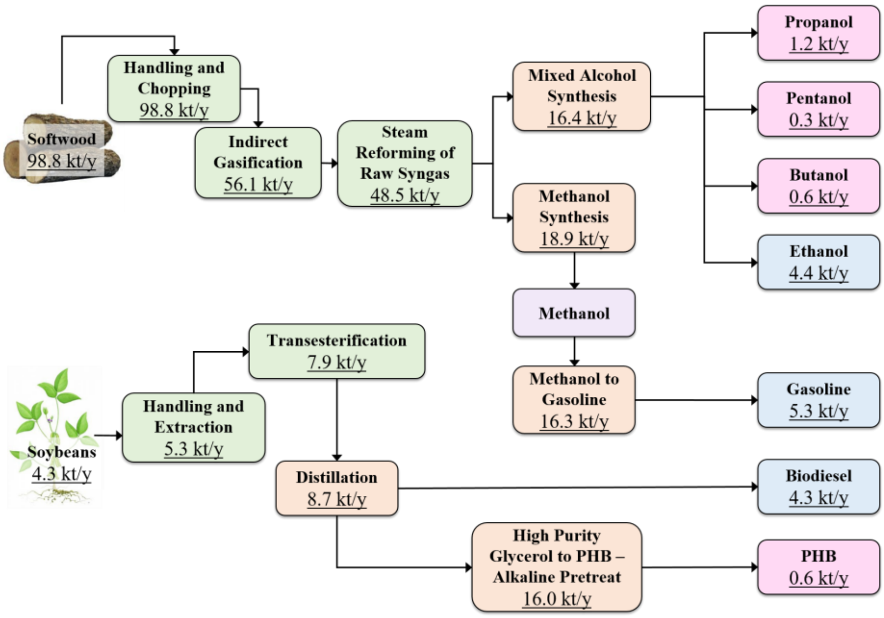

Figure 1: The optimal network design of the conventional ARO method with a gamma uncertainty set.

As for the objective values, the total annualized cost determined by the conventional ARO method is $ 19.6 MM/y, whereas the annualized cost determined by the proposed data-driven ARO approach using PCA and KDE is only $ 18.3 MM/y. The proposed approach with PCA and KDE generates a less conservative robust solution by lowering the annualized cost by 6.6 %.

The biomass process network designs determined by the conventional ARO method and the proposed data-driven approach using KDE are presented in Figure 1 and Figure 2. For both methods, the feedstock of soybeans is selected to produce biodiesel through technologies of handling and extraction, transesterification and distillation. Polyhydroxybutyrate (PHB – a biodegradable plastic) is produced as a byproduct. Soybeans are chosen as the biodiesel process produces glycerol, which can be used to synthesize to PHB. Indirect gasification of softwood is also selected to make ethanol and gasoline. Indirect gasification of softwood is chosen because the process produces syngas, a product that can be transformed into a number of products, boosting the process's flexibility. By comparing the optimal processing pathways in Figure 1 and Figure 2, technologies of acetic acid synthesis and acetic acid hydrogenation are selected only in the optimal pathway determined by the proposed data-driven approach. The optimal production level of technology determined by the proposed approach using KDE is shown in Figure 3. The production level of acetic acid hydrogenation is almost seven times higher than that of mixed alcohol synthesis, indicating that the proposed approach produces ethanol mainly by acetic acid hydrogenation under the worst-case uncertainty realization. Since the operating levels are adaptive decisions, the production levels of these two technologies and their ratio could vary with the realized price and demand uncertainties.

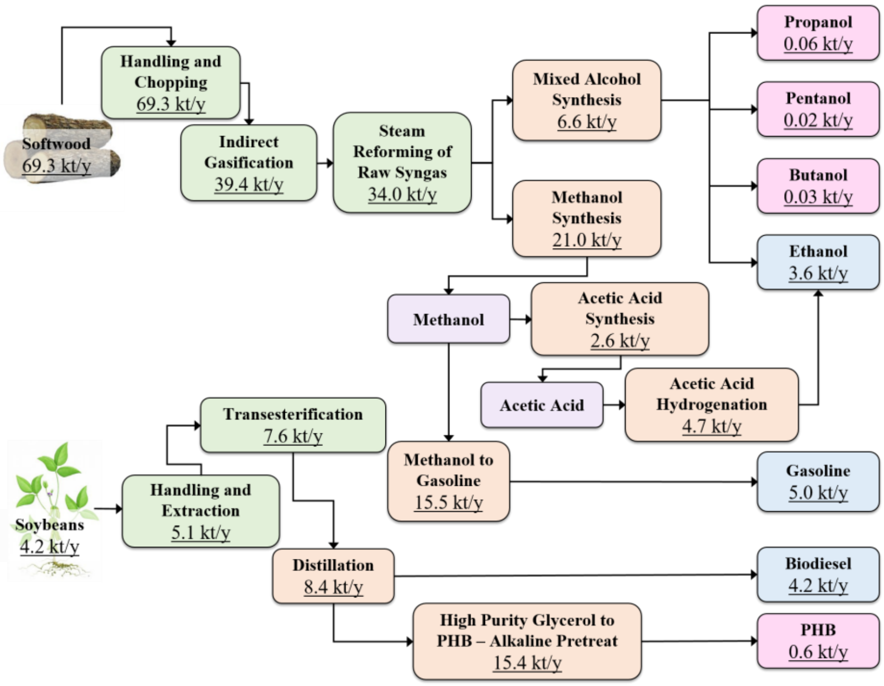
Handling and Chopping 69.3 kt/y
Softwood 69.3 kt/y
Indirect Gasification 39.4 kt/y
Steam Reforming of Raw Syngas 34.0 kt/y
Mixed Alcohol Synthesis 6.6 kt/y
Methanol Synthesis 21.0 kt/y
Propanol 0.06 kt/y
Pentanol 0.02 kt/y
Butanol 0.03 kt/y
Ethanol 3.6 kt/y
Methanol
Acetic Acid Synthesis 2.6 kt/y
Acetic Acid
Acetic Acid Hydrogenation 4.7 kt/y
Transesterification 7.6 kt/y
Soybeans 4.2 kt/y
Handling and Extraction 5.1 kt/y
Methanol to Gasoline 15.5 kt/y
Gasoline 5.0 kt/y
Distillation 8.4 kt/y
Biodiesel 4.2 kt/y
High Purity Glycerol to PHB – Alkaline Pretreat 15.4 kt/y
PHB 0.6 kt/y

Figure 2: The optimal network design decisions determined by the proposed data-driven approach.

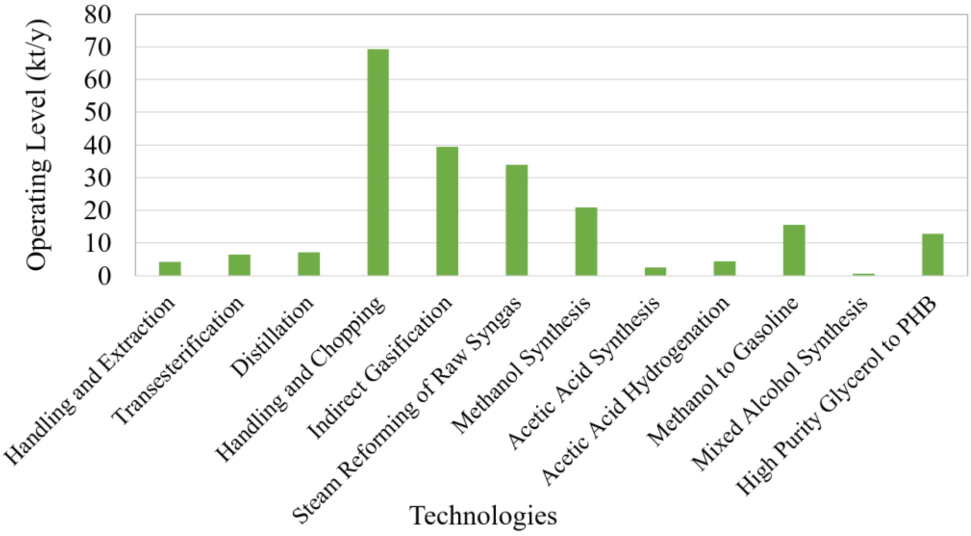

Figure 3: The optimal production levels determined by the proposed data-driven approach.

5. Conclusions

This work developed a new approach to analyse uncertainty data with machine learning techniques to more appropriately consider correlations and asymmetries between uncertain parameters in ARO models. The proposed approach minimizes the total annualised cost of a bioconversion process network design. The final bioconversion process network design converted softwood to ethanol, gasoline, propanol, pentanol, and butanol via indirect gasification of softwood to mixed alcohol synthesis, acetic acid synthesis and hydrogenation, and traditional conversion of soybean to biodiesel as well as glycerine conversion to PHB. Compared to the solution of a conventional ARO approach, the identified pathway led to greater reliance on acetic acid synthesis and hydrogenation rather than mixed alcohol synthesis to produce fuel ethanol. The total annualised cost of the solution found with the proposed data-driven approach was 6.6 % lower than the solution of the conventional ARO approach.

References

Ben-Tal A., Goryashko A., Guslitzer E., Nemirovski A., 2004, Adjustable robust solutions of uncertain linear programs, Mathematical Programming, 99, 351-376.

Bertsimas D., Brown D. B., Caramanis C., 2011, Theory and applications of robust optimization, SIAM Review, 53, 464-501.

Friedman J., Hastie T., Tibshirani R., 2001, The elements of statistical learning, Vol. 1, Springer Series in Statistics, Springer, Berlin, Germany.

Garcia D. J., You F., 2015, Multiobjective optimization of product and process networks: General modeling framework, efficient global optimization algorithm, and case studies on bioconversion, AIChE Journal, 61, 530-554.

Garcia D. J., You F., 2015, Network-Based Life Cycle Optimization of the Net Atmospheric CO2-eq Ratio (NACR) of Fuels and Chemicals Production from Biomass. ACS Sustainable Chemistry & Engineering, 3, 1732-1744.

Glenna L. L., Cahoy D. R., 2009, Agribusiness concentration, intellectual property, and the prospects for rural economic benefits from the emerging biofuel economy. Southern Rural Sociology, 24(2), 111.

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

Gong J., You F., 2015, Sustainable design and synthesis of energy systems. Current Opinion in Chemical Engineering, 10, 77-86.

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

Hrabec D., Šomplák R., Nevrlý V., Janošťák F., Rosecký M., Kůdela J., 2017, Robust facility location problem for bio-waste transportation, Chemical Engineering Transactions, 61, 1093-1098.

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

Ning C., You F., 2017, A data-driven multistage adaptive robust optimization framework for planning and scheduling under uncertainty, AIChE Journal, 63, 4343-4369.

Ning C., You F., 2018, Data-driven decision making under uncertainty integrating robust optimization with principal component analysis and kernel smoothing methods. Computers & Chemical Engineering, 112, 190-210.

Ning C., You F., 2018, Data-driven stochastic robust optimization: General computational framework and algorithm leveraging machine learning for optimization under uncertainty in the big data era. Computers & Chemical Engineering, 111, 115-133.

Rosenthal E., 2008, GAMS-a user's guide, GAMS Development Corporation, Washington, DC, United States.

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

Wei L., Liao Z., Jiang B., Wang J., Yang Y., 2017, Robust optimization of refinery hydrogen networks using worst-case conditional value-at-risk concept, Chemical Engineering Transactions, 61, 685-690.

Wold S., Esbensen K., Geladi P., 1987, Principal component analysis, Chemometrics and Intelligent Laboratory Systems, 2, 37-52.

Yue D., You F., Snyder S. W., 2014, Biomass-to-bioenergy and biofuel supply chain optimization: Overview, key issues and challenges. Computers & Chemical Engineering, 66, 36-56.