Sparse Bayesian learning for data driven polynomial chaos expansion with application to chemical processes

Pham Luu Trung Duong¹, Le Quang Minh¹, Muhammad Abdul Qyyum, Moonyong Lee*  

School of Chemical Engineering, Yeungnam University, Gyeongsan, Republic of Korea

A R T I C L E   I N F O
Article history:
Received 11 June 2018
Received in revised form 31 July 2018
Accepted 3 August 2018
Available online 10 August 2018

A B S T R A C T
Uncertainties are ubiquitous in process system engineering. Hence, to develop a safe and profitable process, uncertainty quantification (UQ) is necessary in a reliability, availability, and maintainability (RAM) analysis. Generalized polynomial chaos expansions can be used as an efficient approach to UQ and work efficiently under the assumption of perfect knowledge with regard to the probability density distribution function of uncertainties. However, this assumption can hardly be satisfied in a real process scenario, mainly because of the limited knowledge regarding the probability density distribution function of uncertainties. To solve these issues, this study investigates the performance of orthogonal polynomial chaos in the UQ of chemical processes, including synthesis gas production and natural gas dehydration. Simultaneously, the limitations of orthogonal polynomial chaos were also investigated by an overwhelming sparse Bayesian learning approach considering a complicated nonlinear crude oil distillation unit with moderate uncertainty numbers. We found that the application of orthogonal polynomial chaos was limited to a small number of uncertainties, mainly because of using the polynomial’s tensor product. Finally, the orthogonal polynomial chaos and sparse Bayesian learning approach were rendered computationally effective in comparison with the conventional Monte Carlo method (approximately 96.5% improvement).

© 2018 Institution of Chemical Engineers. Published by Elsevier B.V. All rights reserved.

1. Introduction
To establish a successful, economic, and efficient chemical process, the best suitable chemical process should be selected before front end engineering design (FEED) takes place in the definition stage (Qyyum et al., 2018). Process synthesis is perceived as the best approach for engineers to select the best process from a set of promising design candidates. Recently, several researchers (Cremaschi, 2015; Tula et al., 2017) have presented excellent reviews with regard to state-of-the-art process synthesis problems. With the development of powerful computers, process synthesis activities will experience higher levels of data uncertainty and development of optimization approaches under uncertainty, such as robust optimization (Lappas and Gounaris, 2018), stochastic (Birge and Louveaux, 2011), and multi-parametric programming (Dominguez and Pistikopoulos, 2013).

Process design under uncertainty has attracted considerable attention, particularly with regard to safety, reliability, and economic decisions (Abubakar et al., 2015). Uncertainty quantification (UQ) investigates the effects of lack of knowledge or expertise error on the model output, and refers to the determination of uncertainty in the model outputs, which results from uncertainty in the model inputs. Additionally, sensitivity analysis (SA) is the assessment of the contribution of individual uncertain inputs to the uncertainty in the model outputs. Both of these analyses are necessary for model development and quality assurance (Song et al., 2015). The uncertainty is typically divided into two or three categories according to its qualities of aleatory uncertainty (i.e., inherent randomness and natural quality), epistemic uncertainty resulting from imperfect knowledge, and linguistic uncertainty (Uusitalo et al., 2015).

* Corresponding author.
E-mail address: mynlee@yu.ac.kr (M. Lee).
¹ These authors contributed equally to this work.
https://doi.org/10.1016/j.cherd.2018.08.006
0283-8762/© 2018 Institution of Chemical Engineers. Published by Elsevier B.V. All rights reserved.
A well-known approach to UQ is the Monte Carlo (MC) related approaches such as the Latin hypercube (Dunn and Shultis, 2012) and quasi Monte Carlo (QMC) (Binder, 1998), where the configurations of the model inputs are drawn randomly/quasi-randomly from their distributions. However, notwithstanding the simplicity of its implementation, the mean convergence rate of the MC approach is the inverse square root of the number of simulations, which makes it computationally expensive and occasionally infeasible for a complex process. This convergence is for the mean only. Subsequently, the different convergence rates will be applied to estimate the variance and failure probability. When considering a multi-fidelity deterministic model with a different level of accuracy, a multi-level Monte Carlo method (MLMC) (Kimiev and Ricardo-Sandoval, 2017) can be used to compute the expectations of the quantity of interest (QoI) with a reduced computational cost in comparison with the MC method. The key challenge for the MLMC is to sample a large number of the QoI by using low fidelity models, and then correct it with a few samples from the high-fidelity model. However, the method is restricted to computing the moment of the QoI only.

To address this issue, the original computational model can be replaced by a surrogate model that is convenient to evaluate and capable of tackling computationally expensive simulations (Celse et al., 2015; Schobi et al., 2015). In the case where the number of uncertainties is small, the polynomial chaos surrogate can be constructed by using a small number of the full model’s time-consuming evaluations. Uncertainty analysis using the generalized polynomial chaos (gPC) expansion has been investigated by numerous authors in various research areas, such as modeling, control, and optimization (Du et al., 2015; Duong and Lee, 2012, 2014; Shen and Bratza, 2016), and has been shown to be more computationally efficient for systems with a small number of random inputs in comparison to the MC method (Duong et al., 2016a; Ghanem and Spanos, 2003). However, in the case of a complex model, the propagation uncertainties through the MC simulation procedures usually require thousands of runs for the various input parameter values, which is not affordable even with modern high-performance computing architectures (Schobi et al., 2015). Simultaneously, when the number of model evaluations grows exponentially, the standard gPC approach becomes limited and may not be applicable to systems with a large number of uncertainties.

Moreover, the accuracy and effectiveness of the gPC approach rely on the exact availability of the uncertainties’ probability density function (PDF). In real-world applications, the underlying probability density/distribution of uncertainties may be intrinsically complicated and vary under different circumstances. When dealing with high-dimensional uncertainties, it is rather challenging to select the type of uncertainties by prior knowledge and further evaluate its convergence/divergence with regard to UQ (Shang et al., 2017). For decision-making under uncertainties, big data methods that exploit massive amounts of available data (Bertsimas et al., 2011; Calaf et al., 2015; Qin, 2014) can be used. Oladyshkin and Nowak (2012) reported that polynomial chaos could be directly constructed from the available data by using uncertainty moments. They performed UQ analysis by using the data driven polynomial chaos (DD-gPC) for CO2 stored in the geological formation, and concluded that the DD-gPC was significantly better than any other polynomial chaos expansions based on the fitted PDF. However, their method utilized the tensor product to construct the polynomial basis dictionary, which limited the use of the method only to problems with a small number of uncertainties (Ahlfeldt et al., 2016).

To analyze the limitations of DD-gPC to small number of uncertainties, Ahlfeldt et al. (2016) employed a sparse grid approach to construct sparse polynomial chaos expansions and tested several multivariate functions. The main advantage of this method was that it could tackle problems with limited information about the input distributions.

Another possible approach for handling a large number of inputs is to utilize the low rank approximation. Gujjar et al. (2017) developed a sparse principal component analysis (SPCA) by combining a generic algorithm to improve the interpretability of variable loadings on principal components and to address the difficulty in the selection of the number of non-zero loadings in SPC. Kieslich et al. (2016) presented an accurate prediction of protein secondary structure based on support vector machines, which provides exceptional accuracy for the prediction results. Sidhu et al. (2018) applied a sparse proper orthogonal decomposition (SPOD) — Galerkin methodology to build a reduced-order model for a nonlinear parabolic partial differential equation system with moving boundaries for describing the fracture propagation in a hydraulic fracturing process. In general, the SPOD/SPCA methods search for an approximation of QoI, which is known as the PARAFAC/CANDECOMP decomposition (Gold and Bader, 2009). The standard approach for finding PARAFAC decomposition is alternative least square (Hyun Phan and Cichocki, 2011).

Different from above approaches, a variety of practical basis selection methods such as the LASSO (Tibshirani, 1996), iteratively reweighted LASSO (Candes et al., 2008), sparse Bayesian learning (Wipf and Rao, 2004), sparse regression (Brunton et al., 2016), best subset regression (Wilson and Sahinidis, 2017), alternating direction method of multipliers (Duong et al., 2016b), and the two-stage gPC method (Mish et al., 2017) have been introduced to find the maximally sparse representations from overcomplete dictionaries. The effectiveness of the sparse Bayesian learning (SBL) approach has been consistently demonstrated with regard to the selection of sparse presentation and in comparison with other methods (Wipf and Rao, 2004). In this study, the SBL was combined with data driven polynomial chaos to obtain a sparse data driven polynomial chaos (SDD-gPC) surrogate model for systems with moderate stochastic inputs/parameters. Note that these basis selection methods are linear sparse regression methods which are different from nonlinear basis selection methods in the series by Tran et al. (2017, 2016a, 2016b), where sparse nonlinear regression and Bayesian approach in combination with the data-driven model for the design of steam methane reforming.

This paper is organized as follows: Section 2 briefly explains the data driven polynomial chaos methods. It also explains the limitation of DD-gPC based on the tensor product and how the proposed method (SDD-gPC) can improve computational efficiency by using the sparse Bayesian learning approach. Additionally, at the end of this section the SA with DD-gPC/SDD-gPC will be discussed in brief. With regard to the UQ and SA, Section 3 demonstrates and validates the applica-
bility of the DD-gPCE of complex chemical processes, such as synthesis
gas (syngas), natural gas dehydration, and the SDD-gPCE of a crude oil
distillation unit with a large amount of uncertainties. For the crude oil
distillation unit, the standard DD-gPCE could not be computed effect-
tively because of the exponential growth of the required simulations.
The computational efficiency of the SDD-gPCE method was successfully
demonstrated as superior to that of the conventional MC. The com-
putational requirements and accuracy of the results obtained by the
DD-gPCE/SDD-gPCE method were compared with those of the standard
MC method. Furthermore, the Sobol sensitivity indices were obtained
directly from the DD-gPCE/SDD-gPCE surrogate model, and were in turn
used to detect the influential inputs beyond the uncertainty propa-
gation. Finally, the conclusions drawn from this study are presented in
Section 4.

2. Data driven polynomial chaos and
sparse data driven polynomial chaos

2.1. Uncertainty quantification and sensitivity
analysis with data driven polynomial chaos
expansions

Let us consider a computational intensive model for a chemi-
atical process with n independent uncertain inputs \(Z = (z_1, ..., z_n)\),
for which raw samples are available and their output is
\[
y = \mathcal{M}(Z)
\]
(1)

The objective of the UQ is to obtain a density function of
output \(\rho_y\) and associated statistical metrics (mean, variances,
etc.).

When the probability density functions \(\rho_i\) are available, the
stochastic model output can be expanded with a total degree
of \(P = \sum_{i=1}^{n} d_i\), as follows:
\[
y \sim \sum_{i=1}^{n} c_i \psi_i(Z) = C^T \Psi(Z)
\]
(2)
\[
M + 1 = \frac{(n + P)!}{n!}
\]
where \(\Psi(Z) = [\psi_1(Z), ..., \psi_M(Z)]^T\) is a collection of basic
functions, the basic function \(\psi_i(Z)\) is obtained by the tensoriza-
tion of one dimensional (univariate) basis, and \(M\) is the number of
expansion terms.

However, exact knowledge of the probability density function
is hardly available in reality. The main idea of data driven
polynomial chaos (DD-gPC) (Ahfeldt et al., 2016; Oladyshkin
and Nowak, 2012) is to construct the univariate polynomial
basis \(\psi_i(Z)\) directly from the given raw samples. This relies on
the knowledge that one can construct an orthogonal univari-
ate polynomial and associated quadrature from the statistical
moment of the underlying distribution (Gautschi, 2004). Its
algorithm is described briefly below.

A \(k\)th order raw statistical moment for a continuous ran-
don variable is defined as follows:
\[
\mu_{ik} = \int_{\Omega_i} z_i^k \, \rho_i(z_i) \, dz_i
\]
(3)
where \(\Omega_i\) is the support of the random input \(z_i\).

With a given set of \(N\) samples, its density function can be
estimated numerically by using the diffusion method (Botev
et al., 2010). With the calculated numerical value of the density
function \(\rho_i(z_i)\), the raw statistical moments can be computed
up to order \(2d\) in Eq. (3) by using the composite trapezoid
method.

Then, the positive definite Hankel matrix of moments can be
defined as follows:
\[
M = \begin{bmatrix}
\mu_{i0} & \mu_{i1} & \cdots & \mu_{id} \\
\mu_{i1} & \mu_{i2} & \cdots & \mu_{i(d+1)} \\
\vdots & \vdots & \ddots & \vdots \\
\mu_{id} & \mu_{i(d+1)} & \cdots & \mu_{i(2d)}
\end{bmatrix}
\]
(4)

By using the Cholesky decomposition, the Hankel matrix
\(M\) can be decomposed into:
\[
M = R^T R
\]
(5)
\[
R = \begin{bmatrix}
r_{11} & r_{12} & \cdots & r_{1(d+1)} \\
r_{22} & r_{23} & \cdots & r_{2(d+1)} \\
\vdots & \vdots & \ddots & \vdots \\
r_{(d+1)(d+1)} & r_{(d+2)(d+1)} & \cdots & r_{(2d)(d+1)}
\end{bmatrix}
\]

The quadrature points \(z_{inm}\) can be computed as eigenval-
ues of the Jacobian matrix \(J\), and the weights are as follows:
\[
\omega_{im} = (v_{1m})^2 \cdot m = 1, ..., d_i
\]
(6)

where \(v_{1m}\) is the first component of the normalized eigenvec-
tor corresponding to the \(m\) eigenvalue.

Because it has a univariate polynomial (constructed from
the data), associated points, and weight of quadrature, the
coefficients \(c_i\) of the polynomial chaos expansion in Eq. (1)
can be calculated with tensorized cubature as follows:
\[
c_i = \sum_{i=1}^{d_i} \sum_{m=1}^{d_i} \lambda M(z_{1i}, ..., z_{ni}) \psi(z_{1i}, ..., z_{ni}) (\omega_{1i} \cdots \omega_{ni})
\]
(7)

With the computed coefficients of the polynomial chaos
expansions, one can have an approximate analytical model for
the relationship between the uncertain inputs and the out-
put in a practical way. For UQ purposes, this approximate and
computationally inexpensive analytical model can be evaluated with the raw samples of uncertain inputs instead of the computational model. Moreover, the mean and variance of the output can be computed easily as follows:

\[ m_y = c_1 \cdot D_y = \sum_{i=2}^{M} c_i^2 \quad (10) \]

The orthogonal polynomials in the expansion offer closed-form formulae for the Sobol sensitivity indices, which are essential in identifying the uncertain input that is relatively more important in comparison with the others (Haro Sandovai et al., 2012). More details on the Sobol indices for sensitivity analysis can be found in Haro Sandovai et al. (2012) and Saltelli et al. (2008).

Let us define a set of multi indices as follows:

\[ I_{i_1,...,i_k} = \{(i_1, ..., i_k): 0 \leq j \leq M, \gamma_k^j = 0, k \in \{1, ..., n\}\} \quad (11) \]

where \( \gamma_k^j \) is the degree of the univariate polynomial with respect to the sub set of inputs \( k \in \{1, ..., n\}\setminus\{i_1, ..., i_k\} \).

Without considering the joint effect with other inputs, the primary Sobol indices \( S_i \) that determine the main effect of a random input \( z_i \) on the variance of the output \( y \) is approximated as follows:

\[ S_i = \frac{\sum_{j \in I_i} c_j^2}{D_y} \quad (12) \]

The joint sensitivity indices for each sub set of inputs \( \{z_{i_1}, ..., z_{i_k}\} \), which quantify the joint effect of this sub set input on the output, can be computed as follows:

\[ S_{i_1,...,i_k} = \frac{\sum_{j \in I_{i_1,...,i_k}} c_j^2}{D_y} \quad (13) \]

The total sensitivity index for an input \( z_i \), can be computed by summing all of the sensitivity indices that involve that input, as follows:

\[ T_i = \sum_{j \in I_i} S_j \quad (14) \]

This index quantifies the total impact of the uncertain input \( z_i \) on the output, including all of the interactions with other inputs.

**Remarks.** All sensitivity indices were normalized quantities, i.e., they ranged between \([0,1]\). If a total sensitivity index \( T_i \) was close to zero, the contribution of input \( z_i \) to the model output was not significant and could be neglected. The joint sensitivity index \( S_{i_1,...,i_k} \) characterized the interaction between inputs \( \{z_{i_1}, ..., z_{i_k}\} \).

### 2.2 Sparse data driven polynomial chaos expansions with sparse Bayesian learning

From Eq. (9), it can be seen that the development of the DD-gPC using tensorized cubature, which was described in the previous subsection, requires a total of \( \prod_{i=1}^{n} d_i \) computational model evaluations. Thus, the number of evaluations increased exponentially with respect to the number of uncertain inputs. For a model with a moderate/large number of uncertain inputs, the method described previously may require an even larger computational effort in comparison to the effort required by the MC/QMC methods for UQ. Typically, in engineering applications, high order interactions between the inputs have trivial contributions to the outputs and most coefficients in the expansion (2) are negligible. Regarding this observation, a method for selecting sparse polynomial chaos expansion can be used, which includes the least angle regression (LARS) (Hastie et al., 2015; Efron et al., 2004), alternating direction multiplier method (ADMM) (Boyd et al., 2011), and so on, because only the basis functions and coefficients that make significant contributions to the model response of interest are retained. Among these sparse basis selection methods, sparse Bayesian learning (SBL) (Wipf and Rao, 2004) has demonstrated its effectiveness and accuracy. This subsection will describe how the SBL framework can be used to construct the sparse data driven polynomial chaos with a moderate/large number of uncertain inputs.

With the sparsity assumption, the coefficients of polynomial chaos expansions can be determined by solving the optimization approximately, as follows:

\[ \hat{C} = \arg \min_{C} ||AC - Y||^2 + \lambda||C||_0 \]

\[ A_{ij} = \psi_j(Z^0) \]

\[ Y = [y_1, ..., y_Q]^T \]

where \( Z^0 \) is an element of experimental design set \( \{Z^{(1)}, ..., Z^{(Q)}\} \), and \( y_i \) is the corresponding model output obtained from \( y_i = M(Z^{(0)}) \).

The first term in (15) is used to penalize the least square error between the surrogate (polynomial chaos) and the real model. The number of nonzero coefficients \( ||C||_0 \) is a penalty to model complexity and induces a sparse polynomial expansion, while \( \lambda \) is the penalty constant.

It is difficult to obtain a solution to the original problem (15) because of the problem’s combinatorial nature. A variety of methods including ADMM and LARS, where the zero-norm of the coefficient vector \( ||C||_0 \) is replaced with its \( l_1 \) norm \( ||C||_1 \), has been proposed with the convex objective of

\[ \min_{C} ||AC - Y||^2 + \lambda||C||_1 \]

Wipf and Rao (2004) demonstrated that the SBL with an iterative reweight method can enhance the accuracy of the recovered coefficients in comparison with the \( l_1 \) norm based methods. This is because the \( l_1 \) norm solution does not always correspond to the solution of (15), despite being the globally minimum solution. In other words, as in Wipf and Nagarajan (2010), the SBL finds a coefficient vector \( C \) as the solution of

\[ \min_{C} ||A - YC||^2 + g_{SBL}(C, y, u) \]

where

\[ g_{SBL}(C, y, u) = \min_{\gamma \in \mathbb{R}^D} \sum_{i=1}^{M} \frac{c_i^2}{y_i} + u_i \gamma_i - h^*(u) \]

(17)
and
\[
h^*(u) = \min_{x \geq 0} \|y - \log|I + A^T A|^T x\|
\]
(18)
\[
\Gamma = \text{diag}(\gamma); \gamma = [\gamma_1, \ldots, \gamma_N]; u = [u_1, \ldots, u_N]
\]
Using the block coordinate descent, problems (16)–(18) can be solved as follows:
For a fixed \(y\) and \(u\), problem (16) corresponds to a weighted \(l_1\) regularized least squares problem that can be solved for \(C\). With the resulting \(C\) and fixed \(u\), Eq. (17) can be solved for optimum \(\gamma\) as
\[
\gamma_i = u_i^{-1/2}|c_i|
\]
(19)
By substitution the value of \(\gamma\) from (19) to (17), one can solve (17) for \(u\) as
\[
u = \text{diag}(A^T(I + A^T A)^{-1} A)
\]
(20)
The SBL algorithm for the sparse data expansion of the driven polynomial chaos is thus summarized as follows:
- For an experimental set (a small portion from the available raw inputs data) \(\{Z^{(1)}, \ldots, Z^{(M)}\}\), run the simulations \(Q < M\) times and form the information matrix \(A_j = \psi_j(Z^{(k)})\) and the corresponding output vector \(Y = [y_1, \ldots, y_Q]^T\)
- Initialize \(u_i^{(0)} = 1, \text{ } j = 1, \ldots, M\)
- For \(k = 1:\) maximum number of iteration
  - Solve the weighted \(l_1\) regularized least squared:
  \[
  C^{(k+1)} = \arg \min_{C} \|A - YC\|_2^2 + \sum_{i=1}^{M} |u_i^{(k)}|^{1/2}|c_i|
  \]
  (21)
  - Obtain \(\gamma^{(k+1)}\) at step \(k + 1:\)
  \[
  \gamma^{(k+1)} = (u_i^{(k)})^{-1/2}|c_i^{(k+1)}|
  \]
  (22)
  - Solve \(u^{(k+1)}\) at step \(k + 1:\)
  \[
  u^{(k+1)} = \text{diag}(A^T(I + A^T A)^{-1} A)
  \]
  (23)
End
- The resulting \(C\) in the previous step will be used as the coefficients for constructing the SDD-gPC
\[
\psi = C \psi(Z)
\]
(24)
The analytical approximation (24) can be easily sampled to obtain the density function of output. The coefficients can be sum up to obtain SA indices.

Remark. The number of available raw data samples for inputs will decide the order of 1-D polynomial basis. When a system is highly nonlinear, high order expansions are required. For all our current case studies, one dimensional basis could be constructed by using 10000 raw samples with the order up to 7. The simulation study results confirmed that the 1-D basis with the order up to 7 is adequate for a representation of the problems considered in this study. However, for other cases, if the higher order expansions are required, the required raw data of input should be larger.

3. Case studies

The following case studies were considered:
3.1 Case-I: syngas production plant
3.2. Case-II: natural gas dehydration
3.3. Case-III: crude oil distillation column

In this study, the UQ and SA for complex chemical processes were estimated based on the data driven based gPC approach. The major challenge was that the amount of computation that was required to evaluate the cost-to-go function was often high. Therefore, emphasis was placed on conducting a large number of simulations off-line to generate the simulation data and meta-model (termed approximate cost-to-go function), which is proposed to approximate the relationship between the process inputs and the cost-to-go function.

3.1. Case-I: syngas production plant

Syngas, which contains carbon monoxide, carbon dioxide, and hydrogen, is a crucial intermediate industrial product. Typically, the raw feed (mainly methane), steam, and air (including \(N_2\) (79%) and \(O_2\) (21%)) are used to form syngas. Moreover, the formation is endothermic and requires high temperature during operation. As shown in Fig. 1, the syngas production chain includes four chemical reactions: two conversions, one combustion, and one equilibrium. The reactions are demonstrated below (Aasbjerg-Petersen et al., 2011):

Conversion reaction:
\[
CH_4 + H_2O \rightarrow CO + 3H_2
\]
\[
CH_4 + 2H_2O \rightarrow CO_2 + 4H_2
\]
Combustion reaction:
\[
CH_4 + 2O_2 \rightarrow CO_2 + 2H_2O
\]
Equilibrium reaction:
\[
CO + H_2O \leftrightarrow CO_2 + H_2
\]
According to HYSYS, two conversion (a reformer and a combustor) and three equilibrium reactors in series constitute the syngas process (HYSYS, 2007). In the reformer, the steam reacts with a large amount of methane to produce a mixture of syngas with \(CO_2\). Effluent is introduced to the combustor (partial auto-thermal reforming) with air and reforming steam. In the exothermic reaction, methane co-reacts with oxygen, while the combustion steam is used to maintain the reactor temperature and ensure the complete conversion of the excess methane from the natural gas. Then, the ratio of \(H_2:N_2\) in the syngas is controlled in the last three equilibrium reactors when the water gas shift reaction occurs. Here, syngas is taken as feedstock for use at the ammonia production plant. In the equilibrium reaction, the operating temperature increases gradually on the side of the product. To facilitate the production of ammonia, the stoichiometric condition (i.e., the molar ratio of hydrogen over nitrogen) is necessarily controlled at approximately 3:1 in the syngas. The set operations are used to specify the particular pressure of the steam and air streams. The Peng Robinson state equation, which is widely used in the hydrocarbon system, was selected to predict the vapor–liquid equilibrium behavior.

In this study, MATLAB code was developed to generate the data passed to the HYSYS rigorous model (see Fig. 1).
The feed flow rate ($F_a$), temperature ($T_a$), and reformer steam flow rate ($F_s$) were arbitrarily chosen as main variables to affect the process outputs and assumed to be uncertain. These three variables were randomized without prior knowledge of the density functions in the ranges of (320, 420) °C; (80, 120) kg mol h⁻¹; (200, 280) kg mol h⁻¹. Moreover, a data set of 10,000 synthesis raw samples of uncertain inputs was available, as shown in Fig. 2. In this case, the impact of uncertainties on the hydrogen–nitrogen ratio (HNr) was monitored and recorded to carry out the UQ.

By using the theory described in Section 2, the DD-gPC basis was derived from the generated data (set of 10,000 synthesis raw samples) of the UQ with HNr output. The resulting surrogate model for this case study was described in the S1 of the Supplementary Material. The results obtained by the proposed method were compared using the MC method (10,000 samples). Note that a large amount of samples is required to obtain an accurate result with the MC method. Table 1 lists the statistical properties of the HNr obtained by the gPC approach. Fig. 3 shows the density functions of HNr in the syngas model by the DD-gPC when all three factors of the feed flow rate, natural gas temperature, and reformer steam varied. For a comparison purpose, the density functions by the MC method are also illustrated in this figure for one random input and all three factors of the feed flow rate, natural gas temperature, and reformer steam. It can be seen that the results obtained by the DD-gPC closely match the results obtained by the traditional MC method. The Sobol sensitivity indices that were derived by the proposed method correctly identified the influential parameters. Furthermore, the number of simulations required by the proposed method was significantly lower in comparison with that required by the MC method. The computational costs were significantly reduced by the proposed method. As can be seen in Fig. 3, these density results suggested that the HNr can potentially vary between any values in the range of 2.35–4.25, with a mean value of 3.42. Moreover, the process was assumed to have failed if the HNr was less than 3.0 and the probability $P[HNr \leq 3.0] = 0.1208$. This means that the syngas process had an expected reliability of 87.92%.

Another advantage was that the single and total sensitivity indices corresponding with the three uncertain inputs (feed flow, temperature of natural gas, and reformer steam) could be obtained. Table 2 lists the Sobol sensitivity indices. It can be seen, that only the natural gas flow rate was critical, while the other indices, such as the feed temperature and reformer steam, were non-influential and could remain constant at the nominal value. It is worthwhile to note that to calculate the sensitivity indices with the MC method, a large computational effort (10,000 × (3 + 2) simulations) was required. Thus, we did not use the MC approach to validate the SA indices with the DD-gPC method. However, Fig. 3 demonstrates that the DD-gPC correctly identified the critical input (natural gas flow rate).

3.2. Case-II: natural gas dehydration

Fig. 4 shows a simplified schematic diagram of the conventional triethylene glycol (TEG) dehydration unit (HYSYS, 2007). In this case study, the wet natural gas contained a mixture of: water, nitrogen, carbon dioxide, hydrogen sulfur, methane, ethane, propane, i-butane, n-butane, i-pentane, and n-pentane. The wet gas entered the absorption tower (contactor) from the bottom, while the lean solvent entered from the top of the tower. Within the absorber, the water in the gas phase was absorbed by the TEG such that the rich glycol and dry gas streams exited from the bottom and top tower, respectively. Different amounts of water were removed from the raw feed to obtain the lean and dry sale gas that satisfied the dew point requirement and production rate of the lean dry natural gas. Increased pressure drops, reduced pipeline capacity, and potential compressor damages can be avoided by avoiding the liquid dropout in the pipe lines (Märch et al., 2006; Skyalogianni et al., 2016). Subsequently, the rich glycol flowed into the middle section of the regeneration column, and then passed through a rich-lean exchanger to exchange heat with a stream of hot lean glycol. Then, the lean glycol that was generated from the bottom was fed into a pump, where its pressure was increased to the required pressure of the glycol contactor and was sub-

![Fig. 1 – Synthesis gas production process.](image-url)
Fig. 2 – Histogram of input data distributions for syngas production process.

<table>
<thead>
<tr>
<th>Method</th>
<th>Synthesis gas</th>
<th>Natural gas dehydration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of simulations</td>
<td>Runtime (s)</td>
</tr>
<tr>
<td>MC (three inputs)</td>
<td>10,000</td>
<td>5272</td>
</tr>
<tr>
<td>MC (one input)</td>
<td>10,000</td>
<td>5196</td>
</tr>
<tr>
<td>MC (two inputs)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Proposed (DD-gPC)</td>
<td>343</td>
<td>178</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2 – Sobol indices obtained by DD-gPC method.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sobol sensitivity indices ( (S_n, T_n) )</td>
</tr>
<tr>
<td>S1</td>
</tr>
<tr>
<td>Synthesis gas</td>
</tr>
<tr>
<td>Natural gas dehydration</td>
</tr>
</tbody>
</table>

Subsequently recycled into the absorber after passing through a heat exchanger to exchange heat with the dry gas. The Glycol package was selected in this simulation.

As in the previous case study, a stochastic module was integrated into this process (Fig. 4) to obtain the required performance indices. As can be seen in Fig. 5, the random inputs associated with the natural gas feed flow rate \( F_n \), water content in natural gas \( W \), and solvent/feed flow rate ratio \( S \), were assumed to be uncertain in the ranges of \( F_n = [(448.3, 547.9)] \) kg mol/h, \( W = (0.449, 0.549) \) kg mol/h, \( S = [0.0148, 0.0181] \). The precise knowledge of density functions with uncertain inputs was not available. Other parameters,
such as the vessel volume, column pressure, and number of theoretical stages, were assumed to be deterministic. Table 2 also lists the Sobol indices derived from the surrogate DD-gPC model, which can be found in the S2 of the Supplementary Material. Consequently, one may predict that the two random inputs (\(F_M\) and \(S\)) were the most influential. The remaining factor (\(W\)) was fixed at its mean value. Fig. 6 shows the density functions of the reboiler duty uncertainty with respect to all three random inputs; namely, the natural gas feed flow rate (\(F_M\)), water content in natural gas (\(W\)), and solvent/feed flow rate ratio (\(S\)). The red line denotes the proposed DD-gPC, while the blue one is the MC method (using 10,000 samples). Additionally, the density function of the reboiler duty from the MC method with two uncertainties (using 10,000 simulations) was compared in Fig. 6 (green line). The results from the DD-gPC and MC methods are in good agreement, which indicates that the derived sensitivity indices correctly indicated the influential inputs for the UQ. The statistical properties of the reboiler duty and simulation parameters of the proposed method (DD-gPC) and MC method are listed in Table 1. Because the precise estimation of the process output was not accessible, the sensitivity functions of the proposed method were compared with those of the MC approach with a sufficiently large amount of simulations. To accurately estimate the probability with the MC method, a standard number of samples was selected from the Chernoff bound (Tempo et al., 2012). It is worthwhile to note that the MC method was computationally expensive and required a large number of simulations to accurately estimate the expected values, variances, and density functions.

3.3. Case-III: crude oil distillation column using SDD-gPC approach

When adequate smoothness conditions are provided, the DD-gPC becomes efficient for UQ and SA, as was previously shown by two examples. However, in most engineering applications, a current limitation of the DD-gPC, where the coefficients are estimated by the tensor product, is that it may not be applicable to systems with a moderate/large number of uncertainties because the number of model evaluations grows exponentially. For example, the DD-gPC requires \(5^8 = 15,625\) simulations for Case-III problem. Additionally, to tackle the
Fig. 5 – Histogram of input data distribution for natural gas dehydration process.

Fig. 6 – Density profiles of natural gas dehydration process. (For interpretation of the references to color in the text, the reader is referred to the web version of this article.)
practical and time-consuming problems of the UQ and SA, the Bayesian learning technique was used in the proposed SDD-gPC to build a sparse gPC expansion with moments. In this study, the six input variables of the crude oil distillation unit were varied for UQ and SA by using the proposed method. Working under the assumption that the model prediction produces a sparse representation, a smaller number of sampling points was required and the process was less computationally expensive in comparison to the conventional MC method.

A crude distillation unit (CDU) consumes approximately 20–30% of the total energy required to distill a given crude material into products (Mahalec and Sanchez, 2012). Therefore, minimizing the operating costs on crude distillation systems while imposing constraints on the quality of the products has recently been considered (Mahalec and Sanchez, 2012). In a CDUs system, the large number of degrees of freedom and uncertainties in the feed mixture and process present a complex problem, and thereby their optimization becomes a highly challenging task. Sensitivity analysis can identify the uncertainties that are not influential and can be discarded during the design optimization of the CDUs system. Moreover, the UQ investigation can measure the comparative influences of the important inputs on the output.

The study was carried out with Aspen HYSYS v10, and the Peng–Robinson package with the state equation option was used for the binary interaction parameters and enthalpy calculations. The crude oil composition and boiling point profiles can be found in Minh et al. (2018). The CDU system was chosen to obtain five products including naphtha, kerosene, diesel, atmospheric gas oil (AGO), and their respective residues. The feed flowrate was assumed to be 100,000 barrels/day at the temperature of 232.2 °C.

Fig. 7 demonstrates a rigorous CDU simulation model (HYSYS, 2007), whose structure included a main shell consisting of 29 trays and three side strippers associated with three pump-around of the middle products. Because of the high temperature at the flash zone (bottom section), a main stripping steam was inserted at the bottom to prevent the excessive thermal cracking of the crude oil. A heat source (Q-trim) was introduced onto this tray to “trim” the feed temperature as required by the initial solution trials, and facilitate a convergence. Additionally, a pre-flash separator was used to enhance the efficiency of the light component separation from the liquid phase (Errico et al., 2009).

In this study, an interface between MATLAB and HYSYS was developed by using the COM functionality to transfer the data required for the global SA and the UQ with a moderate/large number of uncertainties, as was described in Section 2.2. This was established by using the client-server application automation provided by MATLAB R2017b. For UQ and SA, the variation of six inputs, such as the flow rates of the main stripping steam and the three product stripping streams, crude oil temperature, and column reflux ratio, were assumed to follow the unknown independent distribution. They are presented by the 10,000 elements of the raw data shown in Fig. 8, and to individually affect the process performance. The objective of this study was to obtain the specific operating cost (OP), which is the operating cost per unit of crude oil processed (Minh et al., 2018). Table 3 lists all the random inputs with their confidence boundaries.

According to the SDD-gPC, a set of 334 samples from the raw data was selected and passed to the HYSYS simulation to obtain the SA and UQ results. The value $\lambda = 0.00001$ was selected in SBL to minimize leave one out (LOO) error. Note that LOO can be computed efficiently for this sparse linear regression (Duong et al., 2016b), which is different from sparse nonlinear regression in Tran et al. (2018a, 2018b). Furthermore, the sequential thresholded least-squares algorithm (Brunton et al., 2016) was applied for finding the coefficient values of SDD-gPC in comparison with SBL. Unfortunately, for this particular case study, i.e., Case-III, we could not find $\lambda$ to make the method converge. It is because the sparse regression used in Brunton et al. (2016) is for studies where the information matrix A is a tall matrix ($Q \gg M$: the number of observation is larger than the number of basis function), meanwhile, the matrix A is a fat matrix with $Q < M$ in this case study. Table 4 lists the values of the SA indices obtained by the SDD-gPC. Consequently, for the OP, the total SA indices demonstrated that one random variable, such as the crude oil temperature,

---

**Table 3 – Random input boundaries of CDU distillation process.**

<table>
<thead>
<tr>
<th>Variables</th>
<th>Base</th>
<th>Lower</th>
<th>Upper</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main steam</td>
<td>188.8</td>
<td>170.0</td>
<td>207.7</td>
<td>kmol/h</td>
</tr>
<tr>
<td>Kerosene steam</td>
<td>75.6</td>
<td>68.0</td>
<td>83.1</td>
<td>kmol/h</td>
</tr>
<tr>
<td>Diesel steam</td>
<td>75.6</td>
<td>68.0</td>
<td>83.1</td>
<td>kmol/h</td>
</tr>
<tr>
<td>AGO steam</td>
<td>62.9</td>
<td>56.7</td>
<td>69.2</td>
<td>kmol/h</td>
</tr>
<tr>
<td>Crude oil temperature</td>
<td>232.2</td>
<td>209.0</td>
<td>255.4</td>
<td>°C</td>
</tr>
<tr>
<td>Reflux ratio</td>
<td>0.55</td>
<td>0.5</td>
<td>0.6</td>
<td></td>
</tr>
</tbody>
</table>

---

**Table 4 – Sobol indices for the CDU obtained by the SDD-gPC method.**

<table>
<thead>
<tr>
<th>Sobol sensitivity indices ($S_i, T_j$)</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
<th>S6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crude oil distillation unit</td>
<td>1.86E – 5</td>
<td>2.76E – 5</td>
<td>6.22E – 6</td>
<td>3.0E – 7</td>
<td>0.9736</td>
<td>0.0206</td>
</tr>
<tr>
<td></td>
<td>T1</td>
<td>T2</td>
<td>T3</td>
<td>T4</td>
<td>T5</td>
<td>T6</td>
</tr>
<tr>
<td></td>
<td>2.59E – 5</td>
<td>3.88E – 5</td>
<td>1.63E – 5</td>
<td>9.52E – 6</td>
<td>0.9737</td>
<td>0.0262</td>
</tr>
</tbody>
</table>
could be identified as critical. The others parameters were non-influential and could be ignored in the subsequent UQ stage (Duong et al., 2016a).

The detected stage of the non-influential variables from the SA step enabled the simplification of the surrogate model. The UQ was considered for the OP. Because a precise estimation of the process output did not exist, Fig. 9 compares the density functions of the OP, which were obtained by the proposed method (SDD-gPC) and by the MC approach (about 10,000 simulations) (Tempo et al., 2012) associated with all six random inputs. Moreover, these density functions were compared with those obtained by the MC method with the important input (crude oil temperature) identified in the SA step. As seen in Table 5, the statistical properties in the proposed (SDD-gPC) and MC methods were comparable. In other words, the computational cost of the SDD-gPC method was reduced up to 96.5% in comparison with the computational cost of the conventional MC method. The computational time of the SDD-gPC consisted of the computational time to run 334 simulations and the development of the surrogate

Fig. 8 – Histogram of input data distribution for CDU process.

Fig. 9 – Probability distribution profiles of CDU process.
model for UQ/SA with the SBL. The MC method was computationally expensive because it required 10,000 simulations to accurately estimate the expected values, variances, and densities. Moreover, as can be observed in Fig. 9, the SDD-gPC method achieved acceptable results (by matching the density functions) with less computational cost in comparison to the conventional MC method. A description of the SDD-gPC surrogate model for this case study was defined in the Analysis of the Supplementary Material.

4. Conclusions

This study applied sparse Bayesian learning to a DD-gPC approach to tackle the challenging task of UQ and SA with regard to complex chemical processes with moderate/large uncertain inputs. For a small number of uncertainties, the advantage of DD-gPC over the conventional gPC was investigated in the context of syngas production and natural gas dehydration processes. By using sparse Bayesian learning to find the sparse polynomial expansion, the SDD-gPC became superior for a moderate/large number of uncertainties. The proposed methods did not require prior knowledge of the distribution function of uncertain inputs, and used only the raw available input samples. In these case studies, the results of the DD-gPC/SDD-gPC were compared with those obtained by the traditional MC method. The computational efficiency of the DD-gPC/SDD-gPC was significantly higher than that of the MC (up to 96.5%). Moreover, DD-gPC/SDD-gPC allowed the use of a direct/inexpensive function to calculate the sensitivity indices, which were used to detect the non-influential UQ inputs. In future work, the SDD-gPC method could be applied to other complex chemical processes for the purpose of carrying out UQ and SA.

Conflict of interest

The authors declare that there is no conflict of interest.

Acknowledgements

This study was supported by the Basic Science Research Program of the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2018R1A2B6001566), Priority Research Centers Program of the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2014R1A6A1031189), and the R&D Center for Reduction of Non-CO2 Greenhouse Gases (201700240008) funded by the Ministry of Environment as a ‘Global Top Environment R&D Program’.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.cej.2018.08.006.

References

Duong, P.L.T., Minh, L.Q., Pham, T.N., Gocalves, J., Kwok, E., Lee, M., 2016b. Uncertainty quantification and global sensitivity analysis of complex chemical processes with a large number


