Bayesian Network Based Risk and Sensitivity Analysis for Production Process Stability Control

Wei Xie[a,*], Bo Wang[a], Cheng Li[b], Jared Auclair[a], Peter Baker[1]

[a]*Northeastern University*
[b]*National University of Singapore*

Abstract

The biomanufacturing industry is growing rapidly and becoming one of the key drivers of personalized medicine and life science. However, biopharmaceutical production faces critical challenges, including complexity, high variability, long lead time and rapid changes in technologies, processes, and regulatory environment. Driven by these challenges, we explore the bio-technology domain knowledge and propose a rigorous risk and sensitivity analysis framework for biomanufacturing innovation. Built on the causal relationships of raw material quality attributes, production process, and bio-drug properties in safety and efficacy, we develop a Bayesian Network (BN) to model the complex probabilistic interdependence between process parameters and quality attributes of raw materials/in-process materials/drug substance. It integrates various sources of data and leads to an interpretable probabilistic knowledge graph of the end-to-end production process. Then, we introduce a systematic risk analysis to assess the criticality of process parameters and quality attributes. The complex production processes often involve many process parameters and quality attributes impacting on the product quality variability. However, the real-world (batch) data are often limited, especially for customized and personalized bio-drugs. We propose uncertainty quantification and sensitivity analysis to analyze the impact of model risk. Given very limited process data, the empirical results show that we can provide reliable and inter-

*Corresponding author
Email addresses: w.xie@northeastern.edu (Wei Xie), wang.bo2@husky.neu.edu (Bo Wang), stalic@nus.edu.sg (Cheng Li), j.auclair@northeastern.edu (Jared Auclair), pebake64@gmail.com (Peter Baker)

pretable risk and sensitivity analysis. Thus, the proposed framework can provide the science- and risk-based guidance on the process monitoring, data collection, and process parameters specifications to facilitate the production process learning and stability control.
Keywords: Decision analysis, biomanufacturing, Bayesian network, production process risk analysis, sensitivity analysis
2017 MSC: 00-01, 99-00

---

1. Introduction

In the past decades, pharmaceutical companies have invested billions of dollars in the research and development (R&D) of new bio-medicines for the treatment of many severe illnesses, including cancer cells and adult blindness. More than 40 percent of the overall pharmaceutical industry R&D and products in the development pipeline are biopharmaceuticals and this percentage is expected to continuously increase.

Compared to the classical pharmaceutical manufacturing, biopharmaceutical production faces several challenges, including complexity, high variability, long lead time and rapid changes in technologies, processes, and regulatory environment (Kaminsky & Wang, 2015). Biotechnology products are produced in living organisms, which induces a lot of uncertainty in the production process. Compared to small molecule drugs, biopharmaceuticals are considerably larger and it is challenging to get the complete evaluation of large-molecule protein drugs. Thus, *the (production) process is the product.* The interactions of hundreds of factors impact drug critical quality attributes. Besides, since the biologic life cycle is short and production lead time is long, there are limited amounts of historical batch data, in particular for drugs in early development stages. Therefore, the development of stable and efficient biomanufacturing requires us to fully utilize the data, provide a better understanding of production process, correctly assess the criticality of process parameters and in-process/final product quality attributes, and control their specifications.

Kaminsky & Wang (2015) provide the review of Operations research (OR) methodologies proposed for biomanufacturing. State-of-the-art studies in the field have several

key limitations. First, as far as we know, the existing OR approaches introduced for biomanufacturing still focus on developing *general* methodologies, and they do not fully explore the pharmaceutical biotechnology domain knowledge (e.g., the underlying physical mechanics causing the interdependence of raw material quality, production process, and bio-drug properties in safety and efficacy). For complex biopharmaceutical production process which often has tight batch data, this limits OR methodology performance and interpretability, as well as its adoption in real applications. Second, the existing approaches tend to focus on a certain (limited) part of the biomanufacturing system and there is no appropriate and reliable end-to-end risk analysis and control framework. Some recent works, e.g., Martagan et al. (2016, 2017, 2018), incorporate physical-chemical characteristics and develop Markov decision models to find the optimal operational policies with the scope limited in fermentation and chromatography.

For complex systems, Bayesian network (BN) can be used to combine expert knowledge and data. It can facilitate the end-to-end data integration and analysis in various applications, including additive manufacturing (Wang et al., 2018), gene co-expression (Troyanskaya et al., 2003), information system security (Feng et al., 2014) and etc. Besides, global probabilistic sensitivity analysis can evaluate the contribution of each random input to the variation of output; see the review in Borgonovo & Plischke (2016). Since the commonly used variance-based sensitivity measures (i.e., first-order effects and total effects) fail to adequately account for probabilistic dependence and structural interactions, Owen (2014) and Song et al. (2016) introduce and study a new sensitivity measure, called *Shapley value* (SV), which is motivated by game theory.

Driven by the critical challenges in biomanufacturing, in this paper, we propose a new BN based risk and sensitivity analysis framework for end-to-end production process risk management. It can overcome the key limitations of existing OR methodologies. By exploring the causal relationships of raw material quality, production process, and bio-drug properties in safety and efficacy, we first introduce a relational graph that can meaningfully integrate all the data collected from the production processes. Then, we develop a BN modeling the underlying probabilistic interdependence of critical process parameters (CPPs) and critical quality attributes (CQAs) of raw materials/in-process materials/Active Pharmaceutical Ingredients (API). After that, we propose a

production process risk analysis to study the contribution of each CPP/CQA to the output variance and assess their criticality. Since the BN is estimated from limited real-world batch data, there exists the model risk (MR). We further introduce uncertainty quantification (UQ) and sensitivity analysis (SA) to study the impact of MR on the production process risk analysis. Thus, the proposed framework can facilitate the systematic learning and guide the science- and risk-based process monitoring and CPPs/CQAs specification to improve the production process stability.

This paper is organized as follows. In Section 2, we present the problem description. Then, we propose a production process risk analysis in Section 3, and further introduce BN based uncertainty quantification and sensitivity analysis in Section 4. We conduct the empirical study in Section 5 and conclude this paper in Section 6.

2. Problem Description and Proposed Framework

We propose a systematic BN-SV-MR based risk and sensitivity analysis framework to facilitate the understanding and stability control for end-to-end biopharmaceutical production processes. Since biopharmaceuticals are produced in living organisms, the production process involves up to hundreds of factors determining the product quality; see a simplified illustration in Figure 1. The production process includes the main unit operations: (1) media preparation, (2) inoculum fermentation, (3) main fermentation, (4) centrifugation(s), (5) chromatography/purification, (6) filtration, (7) fill and finish, and (8) quality control. Steps (1)–(3) belong to upstream cell culture, Steps (4)–(6) belong to downstream target protein purification, and Steps (7)–(8) are for finished drug filling/formulation and final product quality control testing.

The *interactions* of many factors impact the variability of drug quality; see the fishbone representation of the production process in Figure 1. In general, these factors can be divided into CPPs and CQAs; see the definitions of CPPs/CQAs in ICH Q8(R2).

CPP: At each process unit operation, CPPs are defined as those parameters whose *variability* has an impact on critical quality and therefore should be monitored and controlled to ensure the process produces the desired quality.

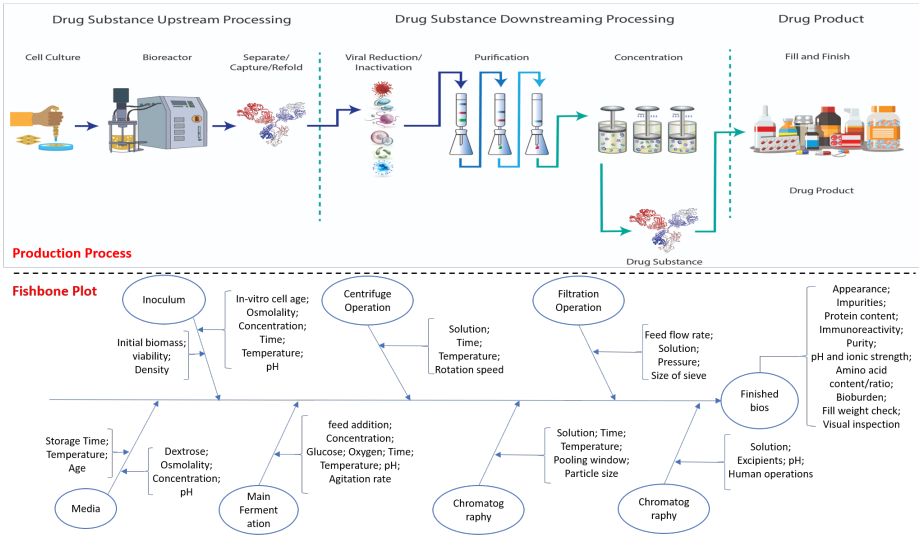

Figure 1: A general biomanufacturing process (Walsh, 2013).

CQA: A physical, chemical, biological, or microbiological property or characteristic that should be within an appropriate limit, range, or distribution to ensure the desired product quality.

Since the raw material attributes are outputs of release materials, they should be considered along with CPPs as impacting process variability.

We represent the variability of any bio-drug critical quality attribute with a random variable (r.v.), denoted by Y. It depends on the random inputs, denoted by $\mathbf{X}$, including selected CPPs/CQAs and other variables introduced in the production process operations (e.g., contamination), denoted by $\mathbf{e}$. Thus, we model the complex interactions of inputs $(\mathbf{X}, \mathbf{e})$ and production process impacting on the response by $Y = g(\mathbf{X}, \mathbf{e}|\boldsymbol{\theta})$, where $g(\cdot)$ is a unknown function specified by parameters $\boldsymbol{\theta}$. The proposed framework can be naturally extended to a vector of responses.

There are two types of uncertainty: *stochastic uncertainty* (inherent variability of biopharmaceutical production process, i.e., random inputs $\mathbf{X}$ and $\mathbf{e}$) and *model risk* (i.e., limited knowledge on statistical models for $(\mathbf{X}, \mathbf{e})$ and function $g(\cdot|\boldsymbol{\theta})$ representing the complex interactions). We can control the impact of stochastic uncertainty by improv-

ing the production process CPPs/CQAs specifications/control and reduce the model risk through process monitoring and data collection.

To produce a successful batch, there could exist more than 100 factors that need to be considered (Otto et al., 2014). *We explore the causal relationships of underlying physical mechanics causing the interdependence of raw material quality, production process, and bio-drug properties in safety and efficacy.* We further develop a BN with parameters $\boldsymbol{\theta}$ modeling the complex probabilistic interdependence, $Y = g(\mathbf{X}, \mathbf{e}|\boldsymbol{\theta})$. We measure the variability of product quality by the output variance, $\mathrm{Var}(Y)$. The contribution of any random input from $\mathbf{X}$ and $\mathbf{e}$ to $\mathrm{Var}(Y)$ can be quantified by variance-based sensitivity measures. In Section 2.1, we review relevant sensitivity measures, especially focusing on the SV utilized in this paper. Then, before providing the detailed presentation in Sections 3 and 4, we summarize and provide the insights of the proposed production process risk and sensitivity analysis framework accounting for both stochastic uncertainty and model risk in Section 2.2.

2.1. Variance-based Sensitivity Measures - Shapley Value

Since hundreds of factors could impact on the product critical attributes, we want to identify those inputs that can reduce the output variance $\mathrm{Var}(Y)$ the most. The contribution of each input, W_k in $\mathbf{W} = (\mathbf{X}, \mathbf{e})$ with $k \in \mathscr{K}$, to the output variance relies on *probabilistic dependence* and *structural interactions*, where $\mathscr{K}$ represents the index set of all inputs in $\mathbf{W}$. Here, the probabilistic dependence represents the underlying interdependence among different inputs (e.g., the CQAs of raw materials and in-process products) and the structural interactions are induced by the complex production process logic. Two most commonly used variance-based sensitivity measures are: (1) the first-order effect $V_k \equiv \mathrm{Var}(Y) - \mathrm{E}[\mathrm{Var}(Y|W_k)]$ that considers the variance reduction when we fix W_k; and (2) the total effect $T_k \equiv \mathrm{E}[\mathrm{Var}(Y|\mathbf{W}_{-k})]$ that considers the expected remaining variance when all other inputs, denoted by $\mathbf{W}_{-k}$, are fixed. However, both measures fail to appropriately quantify the sensitivity when there exist probabilistic dependence and/or structural interaction among inputs (Song et al., 2016).

In this paper, we consider the SV, a new variance-based sensitivity measure introduced by Owen (2014). It can overcome the limitations of first-order effect and total

effect measures. SV was originally introduced to evaluate the performance of a player in a cooperative game from game theory (Shapley, 1953). In the proposed framework, it is used to quantify the contribution of each random input W_k to the output variance $\mathrm{Var}(Y)$, defined as

$$\mathrm{Sh}(W_k) = \sum_{\mathscr{J} \subset \mathscr{K}/\{k\}} \frac{(K - |\mathscr{J}| - 1)!|\mathscr{J}|!}{K!} \left[c(\mathscr{J} \cup \{k\}) - c(\mathscr{J}) \right], \tag{1}$$

where $K = |\mathbf{W}|$ denotes the total number of random inputs and $|\mathscr{J}|$ is the size of index subset $\mathscr{J}$ from $\mathscr{K}/\{k\}$; see the description and motivation of SV in Owen (2014) and Song et al. (2016). Since $\frac{(K-s-1)!s!}{K!} = \frac{1}{K}\binom{K-1}{s}^{-1}$, we put equal weight to all possible sizes of subsets $(s = 0, 1, \ldots, K-1)$ and also equal weight to all possible subsets of size s. The cost function $c(\mathscr{J})$ represents the variance of Y induced by random inputs in any subset $\mathbf{W}_{\mathscr{J}}$. Here, we set the cost function $c(\mathscr{J})$ to be the expected remaining variance of response Y when all other input factors $\mathbf{W}_{-\mathscr{J}} = \mathbf{W}_{\mathscr{K}\setminus\mathscr{J}}$ are fixed,

$$c(\mathscr{J}) = \mathrm{E}[\mathrm{Var}[Y|\boldsymbol{W}_{-\mathscr{J}}]].$$

Then, $c(\mathscr{J} \cup \{k\}) - c(\mathscr{J})$ is the expected increment of $\mathrm{Var}[Y|\boldsymbol{W}_{-\mathscr{J}}]$ induced by including W_k into the set $\mathbf{W}_{\mathscr{J}}$. *Therefore, $Sh(W_k)$ defined in Equation (1) can be interpreted as the average variance $Var(Y)$ increment induced by including W_k into all possible subsets $\mathbf{W}_{\mathscr{J}}$, which allows us to appropriately account for probabilistic dependence and structural interaction.*

The main benefits of SV over first-order and total effect sensitivity measures include: (1) the uncertainty contributions sum up to total variance of output; (2) SV can automatically account for probabilistic dependence and structural interactions occurring in the complex production process; and (3) combing SV with BN (represented by BN-SV) can facilitate the appropriate and interpretable risk and sensitivity analysis since BN is built based on underlying physical mechanics causing the interdependence of raw material quality, production process, and bio-drug properties.

2.2. Summary of Proposed Process Risk and Sensitivity Framework

In Section 3, we first explore the causal relationships of CPPs/CQAs of raw materials/in-process materials/API and develop a BN specified by parameters $\boldsymbol{\theta}$. Since we often have limited batch data, it can efficiently model the complex probabilistic interdependence of production process and improve the interpretability. Then, we derive the variance decomposition, $\mathrm{Var}(Y|\boldsymbol{\theta}) = \sum_{X_k} \mathrm{Sh}(X_k|\boldsymbol{\theta}) + \sum_{e_k} \mathrm{Sh}(e_k|\boldsymbol{\theta})$, where the SVs, $\mathrm{Sh}(X_k|\boldsymbol{\theta})$ and $\mathrm{Sh}(e_k|\boldsymbol{\theta})$, measure the contribution of any CPP/CQA $X_k \in \mathbf{X}$ and other factor $e_k \in \mathbf{e}$ to $\mathrm{Var}(Y|\boldsymbol{\theta})$. Then, the *criticality*, defined by $p_{W_k,Y}(\boldsymbol{\theta}) \equiv \mathrm{Sh}(W_k|\boldsymbol{\theta})/\mathrm{Var}(Y|\boldsymbol{\theta})$, can be used to identify those inputs W_k (i.e., X_k or e_k) that can reduce $\mathrm{Var}(Y)$ the most. Thus, the proposed *BN-SV based risk analysis* can improve our understanding of the end-to-end production process, and facilitate the CPPs/CQAs specification and control to efficiently reduce the output variation.

The "correct" BN parameters, denoted by $\boldsymbol{\theta}^c$, characterizing the underlying probabilistic interdependence of the production process is unknown. We estimate it by using the real-world data, denoted by $\mathscr{X}$, collected from the production process, and quantify the model risk (MR) with the posterior distribution $p(\boldsymbol{\theta}|\mathscr{X})$. In Section 4, we first characterize the estimation uncertainty of variance contribution from any random input W_k with the posterior distribution of $\mathrm{Sh}(W_k|\widetilde{\boldsymbol{\theta}})$ with $\widetilde{\boldsymbol{\theta}} \sim p(\boldsymbol{\theta}|\mathscr{X})$. In this paper, we use $\widetilde{\cdot}$ to denote any posterior sample. We can use the posterior variance to quantify the overall estimation uncertainty, $\mathrm{Var}^*[\mathrm{Sh}(W_k)|\mathscr{X}] \equiv \mathrm{Var}^*_{p(\boldsymbol{\theta}|\mathscr{X})}[\mathrm{Sh}(W_k|\widetilde{\boldsymbol{\theta}})|\mathscr{X}]$. The subscript "$*$" represents any measure calculated based on the posterior $p(\boldsymbol{\theta}|\mathscr{X})$.

Since the BN is built on underlying physical mechanics of the production process, we can find the subset of BN parameters determining $\mathrm{Sh}(W_k)$, denoted by $\boldsymbol{\theta}(W_k,Y)$. We introduce the *BN-SV-MR based sensitivity analysis* to provide the comprehensive study over the impact of model risk. We derive the posterior variance decomposition,

$$\mathrm{Var}^*[\mathrm{Sh}(W_k)|\mathscr{X}] = \sum_{\theta_\ell \in \boldsymbol{\theta}(W_k,Y)} \mathrm{Sh}^*_{\theta_\ell}\left[\mathrm{Sh}\left(W_k\left|\widetilde{\boldsymbol{\theta}}(W_k,Y)\right.\right)\right|\mathscr{X}\right] = \sum_{\theta_\ell \in \boldsymbol{\theta}(W_k,Y)} \mathrm{Sh}^*_{\theta_\ell}[\mathrm{Sh}(W_k)|\mathscr{X}], \quad (2)$$

where $\mathrm{Sh}^*_{\theta_\ell}[\cdot|\mathscr{X}]$ measures the contribution from parameter uncertainty on $\theta_\ell \in \boldsymbol{\theta}(W_k,Y)$. Since θ_ℓ represents a certain part of production process, the decomposition in (2) provides the detailed information on how the model risk over the end-to-end

production process influences the estimation uncertainty of $\mathrm{Sh}(W_k)$. It can be used to guide the knowledge improvement through process monitoring and data collection.

In sum, the proposed framework can provide systematic production process risk analysis and assess the criticality or contribution of each CPP/CQA to the output variance. It accounts for the impact from model risk. Thus, the framework can provide reliable guidance on CPPs/CQAs monitoring and specifications, as well as quality control for raw materials, intermediate and final products, so that we can efficiently improve the stability of integrated biopharmaceutical production process.

3. Integrated Production Process Risk Analysis

Here we introduce an end-to-end production process risk analysis. In Section 3.1, we explore the causal relationships and introduce a relational graph to meaningfully connect all sources of data collected from various process unit operations. Then, in Section 3.2, we develop a probabilistic knowledge graph or Bayesian network modeling the complex interdependence between raw material quality, production process, and bio-drug properties in safety and efficacy. In Section 3.3, we propose the BN-SV risk analysis, assess the criticality of each CPP/CQA or any other factor, and identify the main sources of uncertainty contributing the most to the product quality variation; see the algorithm for production process risk analysis in Appendix D.

3.1. Relational Graphical Model for Production Process

Based on the interactions of CPPs/CQAs and other factors in each process unit operation and also connections among production steps, we develop a relational graphical model for biopharmaceutical production process from raw materials to finished drug substance; see Figure 2 for illustration. In the graph, nodes represent factors (i.e., CPPs and CQAs) impacting on the product quality. The directed edges model the input-output dependence in each process unit and also the interdependence among different production steps. Each big dashed box in Figure 2 illustrates one process unit operation. The shaded nodes represent the variables with real-world observations, including the testing and sensor monitoring data of CPPs/CQAs for raw materials, intermediate

and final drug products. The unshaded and dashed nodes represent variables without observations, including the complete quality status of intermediate and final drug products, and other uncontrollable factors (e.g., contamination) introduced during the process unit operations. Since protein drugs have very complex structures, we can not observe the underlying complete quality status and the monitoring of CQAs can carry partial information.

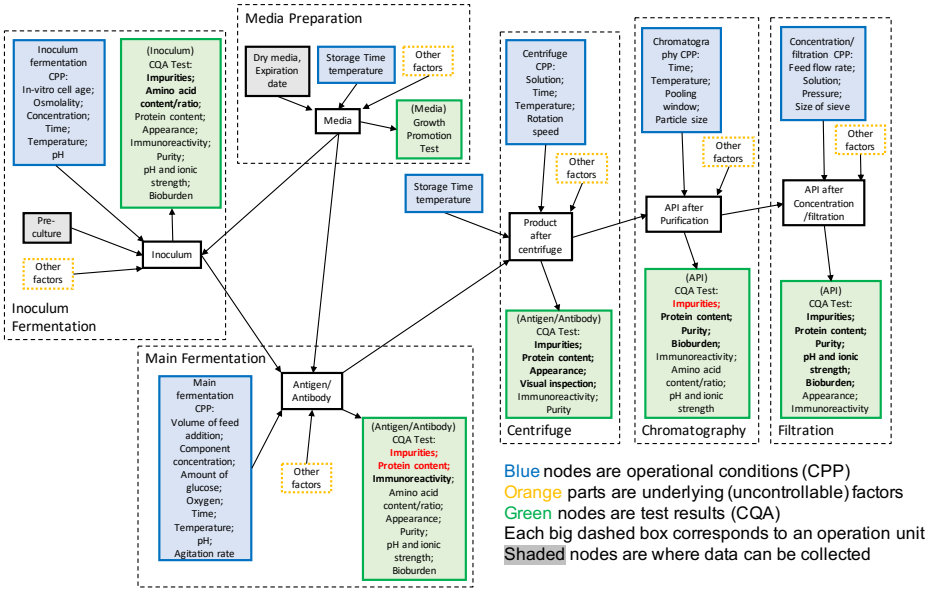

Figure 2: Relational graph for biopharmaceutical production process.

3.2. Bayesian Network Development for Production Process

Based on the relational graph introduced in Section 3.1, we develop a BN probabilistic graphical model composing of random variables and their conditional dependencies via directed edges. For biopharmaceutical production processes, it can characterize the probabilistic interdependence between CPPs and CQAs of raw materials, intermediate product/materials, and API.

3.2.1. Bayesian Network Modeling Development Illustrated with An Example

Without loss of generality, we use a simple example including two production steps (say media preparation and main fermentation) in Figure 3 to illustrate the probabilis-

tic graphical model development. It is based on the causal relationships and the interactions between CPPs and CQAs. Each node represents a CPP/CQA with a r.v. X modeling its variability. Each directed edge with parameter β_{ij} represents the impact of parent node X_i on child node X_j. The pattern-fill nodes (X_1, X_2, X_3) represent the CPPs. The solid fill nodes (X_6, X_7) represent the monitored CQAs of intermediate materials and drug products. The nodes X_4 and X_5 represent the underlying complete quality status of media and drug product after main fermentation. The CQAs X_6 and X_7 represent the partial information of quality variables X_4 and X_5. Except the CPPs X_1, X_2, X_3, the impacts from other factors (i.e., uncontrollable factors) introduced during media preparation and main fermentation are modeled through e_4' and e_5'.

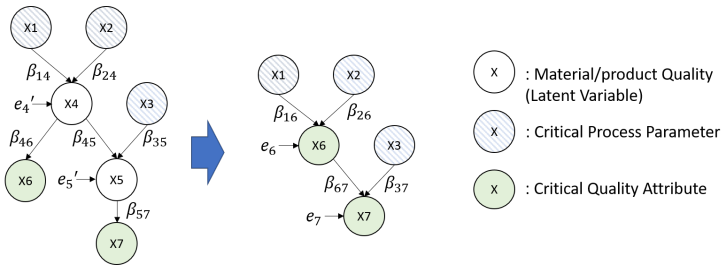

Figure 3: Left: knowledge relational graph; Right: simplified knowledge graph.

Since X_4 and X_5 are hidden and also it is hard to uniquely specify the underlying properties of media and product, they could lead to the identification issue. Thus, we simplify and transform the relational graphical model to a graph without hidden nodes, depicted in the right panel of Figure 3. The CQA X_6 only carries the partial information of the underlying media quality X_4, and e_6 accounts for the impact from factor e_4' on X_6. Combining with other factor e_5', the impact of remaining media properties on the CQA X_7 is modeled through e_7.

According to the right plot in Figure 3, the sources of uncertainty impacting on the variability of X_7 include the process parameters X_1, X_2, X_3 and other factors e_6, e_7. In each process unit operation, we have CPPs X_1, X_2 as inputs and CQA X_6 as output for the first step, and have CQA X_6 and CPP X_3 as inputs and X_7 as output for the second step. To study the impact of each CPP on the CQA of interest (e.g., X_6 and X_7), we can decompose the variance of X_6 and X_7 into the contributions from X_1, X_2 and X_3, and

remaining parts coming from e_6 and e_7. In this way, we can identify the main sources of uncertainty and quantify their impacts, which can guide the CPPs/CQAs specifications and the quality control to improve the product quality stability.

Since there are often limited batch data, in Section 3.2.2, we develop a Gaussian Bayesian network modeling the probabilistic interdependence of production process. Then, we derive a variance decomposition and propose a systematic BN-SV based risk analysis to estimate the criticality of each CPP/CQA in Section 3.3.

3.2.2. Gaussian Bayesian Network Model Development

We consider a probabilistic graphical model with $m+1$ nodes representing the interdependence of CPPs/CQAs ($\mathbf{X}$), bio-product critical quality attribute (Y or X_{m+1}) and other factors ($\mathbf{e}$). Let the first m^p nodes representing CPPs $\mathbf{X}^p = \{X_1, X_2, \ldots, X_{m^p}\}$, the next m^a nodes representing CQAs $\mathbf{X}^a = \{X_{m^p+1}, X_{m^p+2}, \ldots, X_m\}$, and the last node representing the response $Y \triangleq X_{m+1}$ with $m = m^p + m^a$. We construct the linear Gaussian model of the marginal and conditional distributions for each node as follows:

$$X_k \sim \mathcal{N}(\mu_k, v_k^2) \text{ for CPP } X_k \text{ with } k = 1, 2, \ldots, m^p, \tag{3}$$

$$X_k = \mu_k + \sum_{X_j \in Pa(X_k)} \beta_{jk}(X_j - \mu_j) + e_k \text{ for CQA } X_k \text{ with } k = m^p+1, \ldots, m+1 \tag{4}$$

where $Pa(X_k)$ denotes the parent nodes of X_k, and we assume $e_k \sim \mathcal{N}(0, v_k^2)$ with the conditional variance $v_k^2 = \mathrm{Var}[X_k | Pa(X_k)]$.

Given the BN parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{v}^2, \boldsymbol{\beta})$ with mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_{m+1})^\top$, conditional variance $\boldsymbol{v}^2 = (v_1^2, \ldots, v_{m+1}^2)^\top$, and linear coefficients $\boldsymbol{\beta} = \{\beta_{jk}; k = m^p+1, \ldots, m+1 \text{ and } X_j \in Pa(X_k)\}$, the conditional distribution for each CQA node X_k,

$$p(X_k | Pa(X_k)) = \mathcal{N}\left(\mu_k + \sum_{X_j \in Pa(X_k)} \beta_{jk}(X_j - \mu_j), v_k^2\right) \text{ for } k = m^p + 1 \ldots, m+1. \tag{5}$$

For any CPP node X_k without parent nodes, $Pa(X_k)$ is an empty set and $P(X_k | Pa(X_k))$ is just the marginal distribution $P(X_k)$ in (3). Let $Z_k \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and (5) becomes $X_k = \mu_k + \sum_{X_j \in Pa(X_k)} \beta_{jk}(X_j - \mu_j) + v_k Z_k$. Thus, the joint distribution characterizing the

interdependence of CPPs and CQAs involved in the production process can be written, $p(X_1, X_2, \ldots, X_{m+1}) = \prod_{k=1}^{m+1} p(X_k | Pa(X_k))$.

3.3. CPPs and CQAs Criticality Assessment

Given the BN parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{v}^2, \boldsymbol{\beta})$, we present the BN-SV based risk analysis for the end-to-end production process and further quantify the criticality of each random input through measuring its contribution to the product quality variance $\mathrm{Var}(X_{m+1})$.

3.3.1. CPPs Impact on CQAs of Intermediate and Final Products

We derive the SV quantifying the contribution of each random input from CPPs $\mathbf{X}^p$ and other factors $\mathbf{e}$ to $\mathrm{Var}(X_{m+1})$. Suppose $\mathbf{X}^p$ and $\mathbf{e}$ are independent with each other. According to the Gaussian BN model presented in (3) and (4), we can derive

$$X_{m+1} = \mu_{m+1} + \sum_{k=1}^{m^p} \gamma_{k,m+1}(X_k - \mu_k) + \sum_{k=m^p+1}^{m+1} \gamma_{k,m+1} e_k, \tag{6}$$

where the weight coefficient of any CPP X_k to CQA X_n with $k \leq m^p < n \leq m+1$,

$$\gamma_{kn} = \beta_{kn} + \sum_{m^p < \ell < n} \beta_{k\ell}\beta_{\ell n} + \sum_{m^p < \ell_1 < \ell_2 < n} \beta_{k\ell_1}\beta_{\ell_1\ell_2}\beta_{\ell_2 n} + \ldots + \beta_{k,m^p+1}\beta_{m^p+1,m^p+2}\ldots\beta_{n-1,n}, \tag{7}$$

the weight coefficient of any e_k to a CQA node X_n with $m^p < k < n \leq m+1$,

$$\gamma_{kn} = \beta_{kn} + \sum_{k < \ell < n} \beta_{k\ell}\beta_{\ell n} + \sum_{k < \ell_1 < \ell_2 < n} \beta_{k\ell_1}\beta_{\ell_1\ell_2}\beta_{\ell_2 n} + \ldots + \beta_{k,k+1}\beta_{k+1,k+2}\ldots\beta_{n-1,n}; \tag{8}$$

and $\gamma_{nn} = 1$ for any n; see the derivation for (6) in Appendix A. The weight coefficient γ_{kn} accounts for all paths from node X_k to node X_n in the BN.

Let $\boldsymbol{W} = \{X_1, \ldots, X_{m^p}, e_{m^p+1}, \ldots, e_{m+1}\} \triangleq \{W_1, W_2, \ldots, W_{m+1}\}$ represent all random inputs, including the variability from raw materials and production process. Denote the index set $\mathscr{K} = \{1, 2, \ldots, m+1\}$. Then, the SV for the k-th input factor W_k is,

$$\mathrm{Sh}(W_k) = \sum_{\mathscr{J} \subset \mathscr{K}/\{k\}} \frac{(m - |\mathscr{J}|)!|\mathscr{J}|!}{(m+1)!} \left[c(\mathscr{J} \cup \{k\}) - c(\mathscr{J})\right].$$

Based on (6), we compute the cost function $c(\mathscr{J}) = \mathrm{E}[\mathrm{Var}[X_{m+1}|\boldsymbol{W}_{-\mathscr{J}}]] = \sum_{k\in\mathscr{J}} \gamma_{k,m+1}^2 v_k^2$, and obtain $c(\mathscr{J}\cup\{k\}) - c(\mathscr{J}) = \gamma_{k,m+1}^2 v_k^2$. Given the BN parameters $\boldsymbol{\theta}$, we derive $\mathrm{Sh}(X_k|\boldsymbol{\theta})$ for any CPP X_k with $k = 1,\ldots,m^p$,

$$\mathrm{Sh}(X_k|\boldsymbol{\theta}) = \sum_{\mathscr{J}\subset\mathscr{K}/\{k\}} \frac{(m-|\mathscr{J}|)!|\mathscr{J}|!}{(m+1)!}\gamma_{k,m+1}^2 v_k^2 = \gamma_{k,m+1}^2 v_k^2 \sum_{s=0}^{m} \frac{(m-s)!s!}{(m+1)!}\binom{m}{s} = \gamma_{k,m+1}^2 v_k^2,$$

and $\mathrm{Sh}(e_k|\boldsymbol{\theta}) = \gamma_{k,m+1}^2 v_k^2$ for any other factor e_k with $k = m^p+1,\ldots,m+1$. Therefore, we can decompose the variance of drug substance attribute X_{m+1} and estimate the contribution from each random input from $\mathbf{X}^p$ and $\mathbf{e}$,

$$\mathrm{Var}(X_{m+1}|\boldsymbol{\theta}) = \sum_{k=1}^{m^p} \mathrm{Sh}(X_k|\boldsymbol{\theta}) + \sum_{k=m^p+1}^{m+1} \mathrm{Sh}(e_k|\boldsymbol{\theta}) = \sum_{k=1}^{m^p} \gamma_{k,m+1}^2 v_k^2 + \sum_{k=m^p+1}^{m+1} \gamma_{k,m+1}^2 v_k^2. \quad (9)$$

Equation (9) can be used to identify the dominant factors in $\mathbf{X}^p$ and $\mathbf{e}$ contributing the most to the product quality variance, which can guide the parameters specification to improve the production process stability. As a result, given the BN parameters $\boldsymbol{\theta}$, the *criticality* of any input W_k can be quantified by $p_{W_k,X_{m+1}}(\boldsymbol{\theta}) \equiv \mathrm{Sh}(W_k|\boldsymbol{\theta})/\mathrm{Var}(X_{m+1}|\boldsymbol{\theta})$.

In addition, we can assess the impact of inputs on any intermediate product CQA, say X_i with $m^p < i < m+1$. Let $\mathbf{X}^p(X_i)$ denote a set containing all CPPs from previous process unit operations impacting on $\mathrm{Var}(X_i|\boldsymbol{\theta})$. For example in Figure 3(b), we consider the subgraph $\{X_1,X_2,X_6\}$ and then $\mathbf{X}^p(X_6) = \{X_1,X_2\}$. Following the similar derivation as that for (6), we have

$$X_i = \mu_i + \sum_{X_k\in\mathbf{X}^p(X_i)} \gamma_{ki}(X_k-\mu_k) + \sum_{k=m^p+1}^{i} \gamma_{ki}e_k. \quad (10)$$

We can compute the SV measuring the contribution from each factor X_k or $e_{k'}$ on the variance of CQA X_i: $\mathrm{Sh}_i(X_k|\boldsymbol{\theta}) = \gamma_{ki}^2 v_k^2$ and $\mathrm{Sh}_i(e_{k'}|\boldsymbol{\theta}) = \gamma_{k'i}^2 v_{k'}^2$ for $\{k : X_k \in \mathbf{X}^p(X_i)\}$ and $k' = m^p+1,\ldots,i$. Thus, we can decompose the variance of CQA X_i to the contribution from each source of uncertainty as

$$\mathrm{Var}(X_i|\boldsymbol{\theta}) = \sum_{X_k\in\mathbf{X}^p(X_i)} \mathrm{Sh}_i(X_k|\boldsymbol{\theta}) + \sum_{k=m^p+1}^{i} \mathrm{Sh}_i(e_k|\boldsymbol{\theta}) = \sum_{X_k\in\mathbf{X}^p(X_i)} \gamma_{ki}^2 v_k^2 + \sum_{k=m^p+1}^{i} \gamma_{ki}^2 v_k^2.$$

3.3.2. CQAs and CPPs Impact on Final Product or Intermediate CQA

Here we consider the partial production process including one or multiple process unit operations. Suppose that we start from certain operation unit and want to estimate the impact of starting CQAs and remaining steps CPPs on the intermediate or final product CQA of interest. For example, in Figure 3(b), we consider the subgraph $\{X_3, X_6, X_7\}$ with the starting CQA X_6 carrying some information from previous operation step. We study the impacts of X_6 and CPP X_3 on the variability of CQA X_7.

Define the complete BN model with parameters $\boldsymbol{\theta}$ as $G(\mathbf{N}|\boldsymbol{\theta})$, where the set $\mathbf{N} = \{X_1, \ldots, X_{m+1}\}$ includes all nodes. Suppose the part of production process of interest is represented by a subgraph $G(\mathbf{N}'|\boldsymbol{\theta}(\mathbf{N}'))$ having the node set denoted by $\mathbf{N}' \subseteq \mathbf{N}$ with size $n' = |\mathbf{N}'|$ and the index set denoted by $\mathscr{K}'$. We define the set including the starting CQAs as $\mathbf{X}^a(\mathbf{N}') = \{X_k \in (\mathbf{X}^a \cap \mathbf{N}') : Pa(X_k) \cap \mathbf{N}' = \emptyset\}$, i.e., any CQA node in $\mathbf{N}'$ having parent nodes located out of $\mathbf{N}'$. Define the set of CPP nodes as $\mathbf{X}^p(\mathbf{N}') = \{\mathbf{X}^p \cap \mathbf{N}'\}$. Then each remaining CQA node $X_j \in \mathbf{N}'/(\mathbf{X}^a(\mathbf{N}') \cup \mathbf{X}^p(\mathbf{N}'))$ has the corresponding variation factor e_j. Denote the set of e_j's in the subgraph as $\mathbf{e}(\mathbf{N}')$. Without loss of generality, suppose there is a node $X_i \in \mathbf{N}'$ ($m^p < i \leq m+1$) without succeeding nodes in $\mathbf{N}'$.

We quantify the contributions of random inputs for the subgraph, including starting CQAs and CPPs, to $\mathrm{Var}(X_i|\boldsymbol{\theta})$. Following the idea similar with (6), we obtain that

$$X_i = \mu_i + \sum_{X_k \in \mathbf{X}^p(\mathbf{N}')} \gamma_{ki}(X_k - \mu_k) + \sum_{X_k \in \mathbf{X}^a(\mathbf{N}')} \gamma_{ki}(X_k - \mu_k) + \sum_{e_k \in \mathbf{e}(\mathbf{N}')} \gamma_{ki} e_k. \tag{11}$$

Notice that CQAs in $\mathbf{X}^a(\mathbf{N}')$ could be dependent on each other. *The Shapley value used for risk analysis can correctly account for both probabilistic dependence and structural interaction of input factors.* Let $\boldsymbol{W}' = \{\mathbf{X}^p(\mathbf{N}') \cup \mathbf{X}^a(\mathbf{N}') \cup \mathbf{e}(\mathbf{N}')\} \triangleq \{W_k : k \in \mathscr{K}'\}$ represent all random inputs in the subgraph. For any index subset $\mathscr{J} \subset \mathscr{K}'$, by applying (11), we get

$$c(\mathscr{J}) = \mathrm{E}[\mathrm{Var}[X_i|\boldsymbol{W}'_{-\mathscr{J}}]] = \sum_{k \in \mathscr{J}} \gamma_{ki}^2 \mathrm{Var}(W_k) + 2 \sum_{k_1 < k_2 \in \mathscr{J}} \gamma_{k_1 i} \gamma_{k_2 i} \mathrm{Cov}(W_{k_1}, W_{k_2}).$$

Suppose any CPP $X_k \in \mathbf{X}^p(\mathbf{N}')$ or factor $e_k \in \mathbf{e}(\mathbf{N}')$ is independent with other inputs in

$\boldsymbol{W}'$. It implies $\text{Cov}(W_{k_1}, W_{k_2}) = 0$ if either $W_{k_1} \in \{\mathbf{X}^p(\mathbf{N}') \cup \mathbf{e}(\mathbf{N}')\}$ or $W_{k_2} \in \{\mathbf{X}^p(\mathbf{N}') \cup \mathbf{e}(\mathbf{N}')\}$. When both $W_{k_1} = X_{k_1}$ and $W_{k_2} = X_{k_2}$ are CQAs belonging to $\mathbf{X}^a(\mathbf{N}')$, the covariance can be computed by using the linear representation in (10),

$$\text{Cov}(X_{k_1}, X_{k_2}) = \sum_{X_\ell \in \{\mathbf{X}^p(X_{k_1}) \cap \mathbf{X}^p(X_{k_2})\}} \gamma_{\ell k_1} \gamma_{\ell k_2} v_\ell^2 + \sum_{\ell = m^p+1}^{\min(k_1,k_2)} \gamma_{\ell k_1} \gamma_{\ell k_2} v_\ell^2; \tag{12}$$

see the derivation in Appendix B. Then, for each W_k and $\mathscr{J} \subset \mathscr{K}'/\{k\}$, we can obtain

$$\begin{aligned} c(\mathscr{J} \cup \{k\}) - c(\mathscr{J}) &= \gamma_{ki}^2 \text{Var}(W_k) + 2 \sum_{\ell \in \mathscr{J}} \gamma_{ki} \gamma_{\ell i} \text{Cov}(W_k, W_\ell), \\ &= \begin{cases} \gamma_{ki}^2 v_k^2, \ W_k = X_k \in \mathbf{X}^p(\mathbf{N}') \ or \ W_k = e_k \in \mathbf{e}(\mathbf{N}') \\ \gamma_{ki}^2 \text{Var}(X_k|\boldsymbol{\theta}) + 2 \sum_{X_\ell \in \mathbf{X}^a(\mathbf{N}'), \ell \in \mathscr{J}} \gamma_{ki} \gamma_{\ell i} \text{Cov}(X_k, X_\ell|\boldsymbol{\theta}), \ W_k = X_k \in \mathbf{X}^a(\mathbf{N}') \end{cases} \end{aligned}$$

where $\text{Var}(X_k|\boldsymbol{\theta})$ is the variance of starting CQAs X_k and it can be calculated by applying (11) or estimated by using the real-world data. Then, the SV of each input W_k on the drug quality CQA X_i with $m^p < i \leq m+1$ can be calculated as follows:

$$\begin{aligned} \text{Sh}_i(W_k|\boldsymbol{\theta}) &= \sum_{\mathscr{J} \subset \mathscr{K}/\{k\}} \frac{(n' - |\mathscr{J}| - 1)!|\mathscr{J}|!}{n'!} [c(\mathscr{J} \cup \{k\}) - c(\mathscr{J})] \\ &= \gamma_{ki}^2 \text{Var}(W_k) + \sum_{\ell \neq j} \gamma_{ki} \gamma_{\ell i} \text{Cov}(W_k, W_\ell) \\ &= \begin{cases} \gamma_{ki}^2 v_k^2, \ \ W_k = X_k \in \mathbf{X}^p(\mathbf{N}') \ or \ W_k = e_k \in \mathbf{e}(\mathbf{N}') \\ \gamma_{ki}^2 \text{Var}(X_k|\boldsymbol{\theta}) + \sum_{X_\ell \in \mathbf{X}^a(\mathbf{N}')/\{X_k\}} \Big(\sum_{X_h \in \{\mathbf{X}^p(X_k) \cap \mathbf{X}^p(X_\ell)\}} \gamma_{hk} \gamma_{ki} \gamma_{h\ell} \gamma_{\ell i} v_h^2 \\ \quad + \sum_{h=m^p+1}^{\min(k,\ell)} \gamma_{hk} \gamma_{ki} \gamma_{h\ell} \gamma_{\ell i} v_h^2 \Big), \ \ W_k = X_k \in \mathbf{X}^a(\mathbf{N}'). \end{cases} \end{aligned} \tag{13}$$

The derivation of (13) is provided in Appendix C. Thus, we have the variance decomposition:

$$\begin{aligned} \text{Var}(X_i|\boldsymbol{\theta}) &= \sum_{X_k \in \mathbf{X}^p(\mathbf{N}')} \text{Sh}_i(X_k|\boldsymbol{\theta}) + \sum_{X_k \in \mathbf{X}^a(\mathbf{N}')} \text{Sh}_i(X_k|\boldsymbol{\theta}) + \sum_{e_k \in \mathbf{e}(\mathbf{N}')} \text{Sh}_i(e_k|\boldsymbol{\theta}) \\ &= \sum_{X_k \in \mathbf{X}^p(\mathbf{N}')} \gamma_{ki}^2 v_k^2 + \sum_{X_k \in \mathbf{X}^a(\mathbf{N}')} \Bigg[\gamma_{ki}^2 \text{Var}(X_k|\boldsymbol{\theta}) + \sum_{X_\ell \in \mathbf{X}^a(\mathbf{N}')/\{X_k\}} \end{aligned}$$

$$\Big(\sum_{X_h\in\{\mathbf{X}^p(X_k)\cap\mathbf{X}^p(X_\ell)\}}\gamma_{hk}\gamma_{ki}\gamma_{h\ell}\gamma_{\ell i}v_h^2+\sum_{h=m^p+1}^{\min(k,\ell)}\gamma_{hk}\gamma_{ki}\gamma_{h\ell}\gamma_{\ell i}v_h^2\Big)\Big]+\sum_{e_k\in\mathbf{e}(\mathbf{N}')}\gamma_{ki}^2v_k^2. \tag{14}$$

The criticality of W_k on X_i can be measured by $p_{W_k,X_i}(\boldsymbol{\theta})=\mathrm{Sh}_i(W_k|\boldsymbol{\theta})/\mathrm{Var}(X_i|\boldsymbol{\theta})$.

4. Uncertainty Quantification and Sensitivity Analysis

Since the "correct" BN parameters $\boldsymbol{\theta}^c$ are unknown, given finite real-world data $\mathscr{X}$, there exists the *model risk* (MR) characterizing our limited knowledge on the underlying probabilistic interdependence of raw material quality, production process, and bio-drug safety and efficacy. To study the impact of MR on the production process risk analysis and CPPs/CQAs criticality assessment, we propose the *BN-SV-MR based UQ and SA* which can guide the risk- and science-based process monitoring and data collection. In Section 4.1, we develop the Bayesian learning, derive the posterior $p(\boldsymbol{\theta}|\mathscr{X})$ and provide a Gibbs sampler to generate posterior samples, $\widetilde{\boldsymbol{\theta}}^{(b)}\sim p(\boldsymbol{\theta}|\mathscr{X})$ with $b=1,2,\ldots,B$, quantifying the model risk. Then, in Section 4.2, we present UQ to quantify the estimation uncertainty of variance contribution and criticality of any random input W_k. In Section 4.3, we propose the BN-SV-MR based SA that can provide the comprehensive study on how the MR over each part of the end-to-end production process impacts on the risk analysis and the CPPs/CQAs criticality assessment.

4.1. Bayesian Learning for Model Risk Quantification

We first study the case with R batches of complete production process data, denoted as $\mathscr{X}=\{(x_1^{(r)},x_2^{(r)},\ldots,x_{m+1}^{(r)}),r=1,2,\ldots,R\}$. Without strong prior information, we consider the following non-informative prior,

$$p(\boldsymbol{\mu},\boldsymbol{v}^2,\boldsymbol{\beta})=\prod_{i=1}^{m+1}p(\mu_i)p(v_i^2)\cdot\prod_{i\neq j}p(\beta_{ij}), \tag{15}$$

with $p(\mu_i)=\mathscr{N}(\mu_i^{(0)},\sigma_i^{(0)2}),p(v_i^2)=\text{Inv-}\Gamma\left(\frac{\kappa_i^{(0)}}{2},\frac{\lambda_i^{(0)}}{2}\right)$ and $p(\beta_{ij})=\mathscr{N}(\theta_{ij}^{(0)},\tau_{ij}^{(0)2})$, where Inv-$\Gamma$ denotes the inverse-gamma distribution. Given the data $\mathscr{X}$, by applying

the Bayes' rule, we can obtain the posterior distribution

$$p(\boldsymbol{\mu},\boldsymbol{v}^2,\boldsymbol{\beta}|\mathscr{X}) \propto \prod_{r=1}^{R}\left[\prod_{i=1}^{m+1} p(x_i^{(r)}|x_{Pa(X_i)}^{(r)})\right] p(\boldsymbol{\mu},\boldsymbol{v}^2,\boldsymbol{\beta}), \tag{16}$$

quantifying the estimation uncertainty of the BN model characterizing the probabilistic interdependence of production process.

To develop a Gibbs sampler for the posterior distribution in (16), we derive the conditional posterior for each parameter in $(\boldsymbol{\mu},\boldsymbol{v}^2,\boldsymbol{\beta})$. Let $\boldsymbol{\mu}_{-i}$, $\boldsymbol{v}^2_{-i}$ and $\boldsymbol{\beta}_{-ij}$ denote the collection of parameters $\boldsymbol{\mu},\boldsymbol{v},\boldsymbol{\beta}$ excluding the i-th or (i,j)-th element. Let $S(X_i)$ denote the set of direct succeeding or child nodes of node X_i. We first derive the conditional posterior for the coefficient β_{ij},

$$p(\beta_{ij}|\mathscr{X},\boldsymbol{\mu},\boldsymbol{v}^2,\boldsymbol{\beta}_{-ij}) = \mathscr{N}(\theta_{ij}^{(R)},\tau_{ij}^{(R)2}), \tag{17}$$

where $\theta_{ij}^{(R)} = \dfrac{\tau_{ij}^{(0)2}\sum_{r=1}^{R}\alpha_i^{(r)}m_{ij}^{(r)} + v_j^2\theta_{ij}^{(0)}}{\tau_{ij}^{(0)2}\sum_{r=1}^{R}\alpha_i^{(r)2} + v_j^2}$ and $\tau_{ij}^{(R)2} = \dfrac{\tau_{ij}^{(0)2}v_j^2}{\tau_{ij}^{(0)2}\sum_{r=1}^{R}\alpha_i^{(r)2} + v_j^2}$ with $\alpha_i^{(r)} = x_i^{(r)} - \mu_i$ and $m_{ij}^{(r)} = (x_j^{(r)} - \mu_j) - \sum_{X_k\in Pa(X_j)/\{X_i\}}\beta_{kj}(x_k^{(r)} - \mu_k)$. Then, we derive the conditional posterior for $v_i^2 = \mathrm{Var}[X_i|Pa(X_i)]$ with $i = 1,2,\ldots,m+1$,

$$p(v_i^2|\mathscr{X},\boldsymbol{\mu},\boldsymbol{v}^2_{-i},\boldsymbol{\beta}) = \text{Inv-}\Gamma\left(\frac{\kappa_i^{(R)}}{2},\frac{\lambda_i^{(R)}}{2}\right), \tag{18}$$

where $\kappa_i^{(R)} = \kappa_i^{(0)} + R$, $\lambda_i^{(R)} = \lambda_i^{(0)} + \sum_{r=1}^{R}u_i^{(r)2}$ and $u_i^{(r)} = (x_i^{(r)} - \mu_i) - \sum_{X_k\in Pa(X_i)}\beta_{ki}(x_k^{(r)} - \mu_k)$. After that, we derive the conditional posterior for the mean parameter μ_i with $i = 1,2,\ldots,m+1$ for any CPP/CQA,

$$p(\mu_i|\mathscr{X},\boldsymbol{\mu}_{-i},\boldsymbol{v}^2,\boldsymbol{\beta}) \propto p(\mu_i)\prod_{r=1}^{R}\left[p(x_i^{(r)}|x_{Pa(X_i)}^{(r)})\prod_{j\in\mathscr{S}(X_i)}p(x_j^{(r)}|x_{Pa(X_j)}^{(r)})\right] = \mathscr{N}(\mu_i^{(R)},\sigma_i^{(R)2}), \tag{19}$$

where $\mu_i^{(R)} = \sigma_i^{(R)2}\left[\dfrac{\mu_i^{(0)}}{\sigma_i^{(0)2}} + \sum_{r=1}^{R}\dfrac{a_i^{(r)}}{v_i^2} + \sum_{r=1}^{R}\sum_{X_j\in S(X_i)}\dfrac{\beta_{ij}c_{ij}^{(r)}}{v_j^2}\right]$ and $\dfrac{1}{\sigma_i^{(R)2}} = \dfrac{1}{\sigma_i^{(0)2}} + \dfrac{R}{v_i^2} + \sum_{X_j\in S(X_i)}\dfrac{R\beta_{ij}^2}{v_j^2}$ with $a_i^{(r)} = x_i^{(r)} - \sum_{X_k\in Pa(X_i)}\beta_{kj}(x_k^{(r)} - \mu_k)$ and $c_{ij}^{(r)} = \beta_{ij}x_i^{(r)} - (x_j^{(r)} - \mu_j) + \sum_{X_k\in Pa(X_j)/\{X_i\}}\beta_{kj}(x_k^{(r)} - \mu_k)$. The Gibbs sampler iteratively draws the posterior

samples of $(\boldsymbol{\mu}, \boldsymbol{v}^2, \boldsymbol{\beta})$ by applying the conditional posterior distributions given in (17), (18), and (19) until convergence (Gelman et al. 2004).

Besides the case with complete production data, we often have additional incomplete batch data. Since the lead time for biopharmaceutical production is lengthy (Otto et al., 2014), we can have some batches in the middle of production. In addition, the bio-drug quality requirements are restricted, especially for human drugs. Following the quality control, we could discard some batches after main fermentation or even in the middle of downstream purification. Thus, we provide the description, derivation and Gibbs sampler (see Algorithm 3) for both cases with complete or mixing data in online Appendix E.1.2.

4.2. Uncertainty Quantification Accounting for Model Risk

Here we present the UQ to study the overall impact of BN model risk on the production process risk analysis and CPPs/CQAs criticality assessment. Based on Section 3.3, the contribution from any random input W_k in $\boldsymbol{W} \equiv \{\mathbf{X}^p \cup \mathbf{X}^a \cup \mathbf{e}\}$ to the variance of product quality attribute X_{m+1} is measured by the Shapley value, $\mathrm{Sh}(W_k|\boldsymbol{\theta}^c)$. The unknown parameters $\boldsymbol{\theta}^c$ specifying the underlying probabilistic interdependence of production process are estimated by using limited real-world data $\mathscr{X}$. Thus, the estimation uncertainty of the contribution from input W_k can be quantified by the posterior distribution, $\mathrm{Sh}(W_k|\widetilde{\boldsymbol{\theta}})$ with $\widetilde{\boldsymbol{\theta}} \sim p(\boldsymbol{\theta}|\mathscr{X})$. We can use the posterior mean to estimate the expected variance contribution and criticality, $\mathrm{E}^*[\mathrm{Sh}(W_k)|\mathscr{X}] \equiv \mathrm{E}^*_{p(\boldsymbol{\theta}|\mathscr{X})}[\mathrm{Sh}(W_k|\widetilde{\boldsymbol{\theta}})|\mathscr{X}]$ and $\mathrm{E}^*[p_{W_k,X_{m+1}}|\mathscr{X}] \equiv \mathrm{E}^*_{p(\boldsymbol{\theta}|\mathscr{X})}[p_{W_k,X_{m+1}}(\widetilde{\boldsymbol{\theta}})|\mathscr{X}]$, where $p_{W_k,X_{m+1}}(\widetilde{\boldsymbol{\theta}}) = \dfrac{\mathrm{Sh}(W_k|\widetilde{\boldsymbol{\theta}})}{\mathrm{Var}(X_{m+1}|\widetilde{\boldsymbol{\theta}})}$. The posterior variance is used to quantify the overall estimation uncertainty induced by MR, $\mathrm{Var}^*[\mathrm{Sh}(W_k)|\mathscr{X}] \equiv \mathrm{Var}^*_{p(\boldsymbol{\theta}|\mathscr{X})}[\mathrm{Sh}(W_k|\widetilde{\boldsymbol{\theta}})|\mathscr{X}]$ and $\mathrm{Var}^*[p_{W_k,X_{m+1}}|\mathscr{X}] \equiv \mathrm{Var}^*_{p(\boldsymbol{\theta}|\mathscr{X})}[p_{W_k,X_{m+1}}(\widetilde{\boldsymbol{\theta}})|\mathscr{X}]$.

Since we do not have the closed form solutions, we can estimate the posterior mean and variance of $\mathrm{Sh}(W_k)$ and $p_{W_k,X_{m+1}}$ through the sampling approach. By applying the Gibbs sampler in online Appendix E, we can generate posterior samples $\widetilde{\boldsymbol{\theta}}^{(b)} \sim p(\boldsymbol{\theta}|\mathscr{X})$ with $b = 1, 2, \ldots, B$. At any $\widetilde{\boldsymbol{\theta}}^{(b)}$, we can compute $\mathrm{Sh}(W_k|\widetilde{\boldsymbol{\theta}}^{(b)})$ following the description in Section 3.3. The expected contribution from W_k to the variance of X_{m+1}

is estimated by $\widehat{\mathrm{E}}^{*}[\mathrm{Sh}(W_k)|\mathscr{X}] = \bar{\mathrm{Sh}}(W_k|\mathscr{X}) = \frac{1}{B}\sum_{b=1}^{B}\mathrm{Sh}(W_k|\widetilde{\boldsymbol{\theta}}^{(b)})$. And the overall estimation uncertainty can be estimated by sample variance,

$$\widehat{\mathrm{Var}}^{*}[\mathrm{Sh}(W_k)|\mathscr{X}] = \frac{1}{B-1}\sum_{b=1}^{B}\left[\mathrm{Sh}(W_k|\widetilde{\boldsymbol{\theta}}^{(b)}) - \bar{\mathrm{Sh}}(W_k|\mathscr{X})\right]^2. \tag{20}$$

Similarly, we can estimate the expected criticality by $\widehat{\mathrm{E}}^{*}[p_{W_k,X_{m+1}}|\mathscr{X}] = \bar{p}_{W_k,X_{m+1}} = \frac{1}{B}\sum_{b=1}^{B} p_{W_k,X_{m+1}}(\widetilde{\boldsymbol{\theta}}^{(b)})$ and estimate the overall estimation uncertainty by

$$\widehat{\mathrm{Var}}^{*}[p_{W_k,X_{m+1}}|\mathscr{X}] = \frac{1}{B-1}\sum_{b=1}^{B}\left[p_{W_k,X_{m+1}}(\widetilde{\boldsymbol{\theta}}^{(b)}) - \bar{p}_{W_k,X_{m+1}}\right]^2. \tag{21}$$

4.3. Sensitivity Study of Risk Analysis to Model Risk

In this section, we propose the BN-SV-MR based sensitivity analysis, and also provide the UQ and SA procedure in Algorithm 1. It allows us to analyze the effect of model risk on the production process risk analysis and criticality assessment for each CPP/CQA. Steps (1)–(3) estimate $\mathrm{Var}^{*}[\mathrm{Sh}(W_k)|\mathscr{X}]$ to quantify the overall impact of model risk on $\mathrm{Sh}(W_k)$ estimation. Steps (4)–(14) further study its sensitivity to the estimation uncertainty induced from each relevant SN parameter.

Here we use $\mathrm{Var}^{*}[\mathrm{Sh}(W_k)|\mathscr{X}]$ for illustration and the similar idea can be applied to CPPs/CQAs criticality assessment $\mathrm{Var}^{*}[p_{W_k,X_{m+1}}|\mathscr{X}]$. Let $\boldsymbol{\theta}(W_k,X_{m+1}) \subset \boldsymbol{\theta}$ represent the subset of BN parameters impacting $\mathrm{Sh}(W_k|\boldsymbol{\theta})$. We set $\boldsymbol{\theta}(W_k,X_{m+1}) = \{v_i^2 : X_i \in \boldsymbol{N}(W_k,X_{m+1})\} \cup \{\beta_{ij}^2 : X_i,X_j \in \boldsymbol{N}(W_k,X_{m+1}), X_i \in Pa(X_j)\}$, where $\boldsymbol{N}(W_k,X_{m+1})$ is the set containing nodes located along paths from W_k to X_{m+1}; see Section 3.3. Notice that $\boldsymbol{\mu}$ has no impact on $\mathrm{Sh}(W_k|\boldsymbol{\theta})$. Since SV can appropriately account for the probabilistic dependence characterized by $p(\boldsymbol{\theta}|\mathscr{X})$ and structural interactions, we can measure the contribution from any parameter $\theta_\ell \in \boldsymbol{\theta}(W_k,X_{m+1})$ through the posterior variance decomposition,

$$\mathrm{Var}^{*}[\mathrm{Sh}(W_k)|\mathscr{X}] = \sum_{\theta_\ell \in \boldsymbol{\theta}(W_k,X_{m+1})} \mathrm{Sh}^{*}_{\theta_\ell}\left[\mathrm{Sh}\left(W_k\middle|\widetilde{\boldsymbol{\theta}}\right)\middle|\mathscr{X}\right] = \sum_{\theta_\ell \in \boldsymbol{\theta}(W_k,X_{m+1})} \mathrm{Sh}^{*}_{\theta_\ell}\left[\mathrm{Sh}(W_k)|\mathscr{X}\right].$$

The proposed BN-SV-MR sensitivity analysis can provide the comprehensive and interpretable understanding on how the model risk or knowledge limitation over each

related part of production process (i.e., represented by parameter $\theta_\ell \in \boldsymbol{\theta}(W_k, X_{m+1})$) impacts on the system risk analysis.

Algorithm 1: Procedure for the BN-SV-MR Based UQ and SA

Input: BN structure $G(\mathbf{N}|\boldsymbol{\theta})$, data $\mathscr{X}$, number of samples N_π, B, B_O and B_I, index subset $\mathscr{L}_k$.

Output: Return $\widehat{\widehat{\mathrm{Sh}}}^*_{\theta_\ell}[\mathrm{Sh}(W_k)|\mathscr{X}]$ and $\widehat{\mathrm{Sh}}^*_{\theta_\ell}\left[p_{W_k,X_{m+1}}\middle|\mathscr{X}\right]$ for any $W_k \in \boldsymbol{W} = \{\mathbf{X}^p \cup \mathbf{X}^a \cup \mathbf{e}\}$.

(1) Call Algorithm 3 in Appendix E.1.3 to obtain the posterior samples $\widetilde{\boldsymbol{\theta}}^{(b)} = (\widetilde{\boldsymbol{\mu}}^{(b)}, \widetilde{\boldsymbol{v}}^{(b)2}, \widetilde{\boldsymbol{\beta}}^{(b)})$ with $b = 1,2,\ldots,B$ for UQ and $\widetilde{\boldsymbol{\theta}}^{(b_O)} = (\widetilde{\boldsymbol{\mu}}^{(bo)}, \widetilde{\boldsymbol{v}}^{(b_O)2}, \widetilde{\boldsymbol{\beta}}^{(bo)})$ with $b_O = 1,2,\ldots,B_O$ for SA;

(2) Call Algorithm 2 in Appendix D to compute $\mathrm{Sh}(W_k|\widetilde{\boldsymbol{\theta}}^{(b)})$ and criticality $p_{W_k,X_{m+1}}(\widetilde{\boldsymbol{\theta}}^{(b)})$ for $b = 1,2,\ldots,B$;

(3) Calculate the overall estimation uncertainty by using $\widehat{\mathrm{Var}}^*[\mathrm{Sh}(W_k)|\mathscr{X}]$ and $\widehat{\mathrm{Var}}^*[p_{W_k,X_{m+1}}|\mathscr{X}]$ in Equations (20) and (21);

(4) Randomly generate N_π permutations, $\pi_n \sim \Pi(\mathscr{L}_k)$ with $n = 1,\ldots,N_\pi$;

for *Each* π_n **do**

(5) Set $\widehat{c}(P_{\pi_n(1)}(\pi_n)) = 0$;

for $\ell = 1,\ldots,L_k$ **do**

if $\ell < L_k$ **then**

for $b_O = 1,\ldots,B_O$ **do**

(7) Set initial value $\boldsymbol{\theta}^{(b_O,0)}_{\mathscr{J}} = \widetilde{\boldsymbol{\theta}}^{(b_O)}_{\mathscr{J}}$ with $\mathscr{J} = P_{\pi_n(\ell+1)}(\pi_n)$;

for $t = 1,\ldots,T$ **do**

(8) For each $\theta_{\mathscr{J}(\ell)} \in \boldsymbol{\theta}_{\mathscr{J}}$, generate $\theta^{(b_O,t)}_{\mathscr{J}(\ell)} \sim p(\theta_{\mathscr{J}(\ell)}|\mathscr{X}, \widetilde{\boldsymbol{\theta}}^{(b_O)}_{\mathscr{L}_k-\mathscr{J}}, \theta^{(b_O,t)}_{\mathscr{J}(1)}, \ldots, \theta^{(b_O,t)}_{\mathscr{J}(\ell-1)}, \theta^{(b_O,t-1)}_{\mathscr{J}(\ell+1)}, \ldots, \theta^{(b_O,t-1)}_{\mathscr{J}(J)})$ by applying Equations (17)/(18)/(19) for the case with complete data or Equations (E.1)/(E.2)/(E.3) for cases with mixing data (see the online Appendix). Obtain the new sample $\boldsymbol{\theta}^{(b_O,t)}_{\mathscr{J}}$;

(9) Set $\widetilde{\boldsymbol{\theta}}^{(b_O,b_I)}_{\mathscr{J}} = \boldsymbol{\theta}^{(b_O,(b_I-1)h+1)}_{\mathscr{J}}$ with some constant integer h to reduce the correlation between consecutive samples;

(10) Compute $\widehat{c}(P_{\pi_n(\ell+1)}(\pi_n))$ by Equations (25) and (27);

else

(11) Set $\widehat{c}(P_{\pi_n(\ell+1)}(\pi_n)) = \widehat{\mathrm{Var}}^*[\mathrm{Sh}(W_k)|\mathscr{X}]$ and $\widehat{\mathrm{Var}}^*[p_{W_k,X_{m+1}}|\mathscr{X}]$;

(12) Compute $\Delta_{\pi_n(\ell)} c(\pi_n) = \widehat{c}(P_{\pi_n(\ell+1)}(\pi_n)) - \widehat{c}(P_{\pi_n(\ell)}(\pi_n))$;

(13) Estimate $\widehat{\widehat{\mathrm{Sh}}}^*_{\theta_\ell}[\mathrm{Sh}(W_k)|\mathscr{X}]$ and $\widehat{\mathrm{Sh}}^*_{\theta_\ell}\left[p_{W_k,X_{m+1}}\middle|\mathscr{X}\right]$ by using Equations (26) and (28).

Then, we derive SV, $\mathrm{Sh}^*_{\theta_\ell}[\mathrm{Sh}(W_k)|\mathscr{X}] = \sum_{\mathscr{J}\subset\mathscr{L}_k/\{\ell\}} \frac{(L_k - |\mathscr{J}| - 1)!|\mathscr{J}|!}{L_k!}$ $[c(\mathscr{J}\cup\{\ell\}) - c(\mathscr{J})]$, to measure the contribution from the estimation uncertainty

of θ_ℓ. Denote the size of relevant BN parameters by $L_k = |\boldsymbol{\theta}(W_k, X_{m+1})|$ and denote the index set by $\mathscr{L}_k$. Thus, $\boldsymbol{\theta}(W_k, X_{m+1}) = \boldsymbol{\theta}_{\mathscr{L}_k}$. We further denote any subset by $\boldsymbol{\theta}_{\mathscr{J}} \subset \boldsymbol{\theta}(W_k, X_{m+1})$ with size $J = |\boldsymbol{\theta}_{\mathscr{J}}|$ and the corresponding index set $\mathscr{J} = \{\mathscr{J}(1), \mathscr{J}(2), \ldots, \mathscr{J}(J)\} \subset \mathscr{L}_k$. For any $\mathscr{J} \subset \mathscr{L}_k$, the cost function is given as,

$$c(\mathscr{J}) = \mathrm{E}^*_{p(\boldsymbol{\theta}_{\mathscr{L}_k - \mathscr{J}} | \mathscr{X})}[\mathrm{Var}^*_{p(\boldsymbol{\theta}_{\mathscr{J}} | \boldsymbol{\theta}_{\mathscr{L}_k - \mathscr{J}}, \mathscr{X})}[\mathrm{Sh}(W_k) | \widetilde{\boldsymbol{\theta}}_{\mathscr{L}_k - \mathscr{J}}]], \tag{22}$$

where $\boldsymbol{\theta}_{\mathscr{L}_k - \mathscr{J}} = \boldsymbol{\theta}_{\mathscr{L}_k \setminus \mathscr{J}}$. Denote a permutation of $\mathscr{L}_k$ as π and define the set $P_\ell(\pi)$ as the index set preceding ℓ in π. The SV can be rewritten as,

$$\mathrm{Sh}^*_{\theta_\ell}[\mathrm{Sh}(W_k) | \mathscr{X}] = \sum_{\pi \in \Pi(\mathscr{L}_k)} \frac{1}{L_k!} [c(P_\ell(\pi) \cup \{\ell\}) - c(P_\ell(\pi))], \tag{23}$$

where $\Pi(\mathscr{L}_k)$ denotes the set of all $L_k!$ permutations of $\mathscr{L}_k$.

The number of all possible subsets $\mathscr{J}$ could grow exponentially as L_k increase. To address this computational issue, we use the Monte Carlo sampling approach, ApproShapley, suggested by Song et al. (2016) and Castro et al. (2009), by estimating the Shapley value in (23) by

$$\widehat{\mathrm{Sh}}^*_{\theta_\ell}[\mathrm{Sh}(W_k) | \mathscr{X}] = \frac{1}{N_\pi} \sum_{n=1}^{N_\pi} [c(P_\ell(\pi_n) \cup \{\ell\}) - c(P_\ell(\pi_n))] \triangleq \frac{1}{N_\pi} \sum_{n=1}^{N_\pi} \Delta_\ell c(\pi_n), \tag{24}$$

where N_π denotes the number of permutations $\pi_1, \ldots, \pi_{N_\pi}$ randomly generated from $\Pi(\mathscr{L}_k)$ and $\Delta_\ell c(\pi_n) = c(P_\ell(\pi_n) \cup \{\ell\}) - c(P_\ell(\pi_n))$ is the incremental posterior variance $\mathrm{Var}^*[\mathrm{Sh}(W_k) | \mathscr{X}]$ induced by including the ℓ-th BN parameter input in $P_\ell(\pi_n)$.

To efficiently compute $\Delta_\ell c(\pi_n)$, according to Song et al. (2016), we compute $c(\cdot)$ in an order of exact location in permutation π_n. Assume the permutation index set $\pi_n = \{\pi_n(1), \pi_n(2), \ldots, \pi_n(L_k)\}$, we compute $c(P_{\pi_n(\ell)}(\pi_n))$ as the order of $\pi_n(\ell)$ for $\ell = 1, \ldots, L$, where $c(P_{\pi_n(1)}(\pi_n)) = c(\emptyset) = 0$, $c(P_{\pi_n(L_k+1)}(\pi_n)) = c(\pi_n) = \mathrm{Var}^*[\mathrm{Sh}(W_k) | \mathscr{X}]$, and $P_{\pi(j+1)}$ denotes $P_{\pi(j)}(\pi) \cup \{\pi(j)\}$ for $0 < j < L_k$. Then we can compute the incremental variance related to each $\theta_{\pi_n(\ell)}$ for $\ell = 1, \ldots, L_k$: $\Delta_{\pi_n(\ell)} c(\pi_n) = c(P_{\pi_n(\ell+1)}(\pi_n)) - c(P_{\pi_n(\ell)}(\pi_n))$. Since $\{\theta_{\pi_n(1)}, \theta_{\pi_n(2)} \ldots, \theta_{\pi_n(L_k)}\} \triangleq \boldsymbol{\theta}_{\pi_n}$ is just a permutation of $\boldsymbol{\theta}_{\mathscr{L}_k}$, it is equivalent as computing $\Delta_\ell c(\pi_n)$ for all θ_ℓ under permutation π_n.

We can not analytically compute $c(\mathscr{J})$ in (22) and a Monte Carlo sampling approach is developed to estimate it. Since the posterior samples obtained from the Gibbs sampler in online Appendix E.1.3 can not be directly used to estimate $\mathrm{E}^*_{p(\boldsymbol{\theta}_{\mathscr{L}_k-\mathscr{J}}|\mathscr{X})}[\mathrm{Var}^*_{p(\boldsymbol{\theta}_{\mathscr{J}}|\boldsymbol{\theta}_{\mathscr{L}_k-\mathscr{J}},\mathscr{X})}[\mathrm{Sh}(W_k)|\widetilde{\boldsymbol{\theta}}_{\mathscr{L}_k-\mathscr{J}}]]$, we introduce a *nested Gibbs sampling approach*. The posterior sample $\widetilde{\boldsymbol{\theta}}^{(b_O)}_{\mathscr{L}_k-\mathscr{J}}$ can be directly obtained by applying the Gibbs sampling in Appendix E.1.3 to generate $\widetilde{\boldsymbol{\theta}}^{(b_O)} \sim p(\boldsymbol{\theta}|\mathscr{X})$ and keeping components with index $\mathscr{L}_k - \mathscr{J}$ for $b_O = 1,\ldots,B_O$. They are called "outer" samples and we use them to estimate $\mathrm{E}^*_{p(\boldsymbol{\theta}_{\mathscr{L}_k-\mathscr{J}}|\mathscr{X})}$. At each $\widetilde{\boldsymbol{\theta}}^{(b_O)}_{\mathscr{L}_k-\mathscr{J}}$, a conditional sampling is further developed to generate samples from $p(\boldsymbol{\theta}_{\mathscr{J}}|\widetilde{\boldsymbol{\theta}}^{(b_O)}_{\mathscr{L}_k-\mathscr{J}},\mathscr{X})$. More specifically, we set the initial value $\boldsymbol{\theta}^{(b_O,0)}_{\mathscr{J}} = \widetilde{\boldsymbol{\theta}}^{(b_O)}_{\mathscr{J}}$. In each t-th MCMC iteration, given the previous sample $\boldsymbol{\theta}^{(b_O,t-1)}_{\mathscr{J}}$, we apply the Gibbs sampling to sequentially generate one sample from the conditional posterior for each $\theta_{\mathscr{J}(\ell)} \in \boldsymbol{\theta}_{\mathscr{J}}$ with $\ell = 1,\ldots,|\mathscr{J}|$,

$$\theta^{(b_O,t)}_{\mathscr{J}(\ell)} \sim p\left(\theta_{\mathscr{J}(\ell)} \,\middle|\, \mathscr{X}, \widetilde{\boldsymbol{\theta}}^{(b_O)}_{\mathscr{L}_k-\mathscr{J}}, \theta^{(b_O,t)}_{\mathscr{J}(1)},\ldots,\theta^{(b_O,t)}_{\mathscr{J}(\ell-1)}, \theta^{(b_O,t-1)}_{\mathscr{J}(\ell+1)},\ldots,\theta^{(b_O,t-1)}_{\mathscr{J}(|\mathscr{J}|)}\right).$$

By repeating this procedure, we can get samples $\boldsymbol{\theta}^{(b_O,t)}_{\mathscr{J}}$ with $t = 0,\ldots,T$. We keep one for every h samples to reduce the correlations between consecutive samples. Consequently, we obtain "inner" samples $\widetilde{\boldsymbol{\theta}}^{(b_O,b_I)}_{\mathscr{J}}$ with $b_I = 1,\ldots,B_I$.

Thus, this nested Gibbs sampling can generate $B_O \cdot B_I$ samples $\{(\widetilde{\boldsymbol{\theta}}^{(b_O,b_I)}_{\mathscr{J}}, \widetilde{\boldsymbol{\theta}}^{(b_O)}_{\mathscr{L}_k-\mathscr{J}}) : b_O = 1,\ldots,B_O \text{ and } b_I = 1,\ldots,B_I\}$. For any $\mathscr{J} \subset \mathscr{L}_k$, the cost function can be estimated as,

$$\widehat{c}(\mathscr{J}) = \frac{1}{B_O}\sum_{b_O=1}^{B_O}\left\{\frac{1}{B_I-1}\sum_{b_I=1}^{B_I}\left[\mathrm{Sh}(W_k|\widetilde{\boldsymbol{\theta}}^{(b_O,b_I)}_{\mathscr{J}}, \widetilde{\boldsymbol{\theta}}^{(b_O)}_{\mathscr{L}_k-\mathscr{J}}) - \bar{\mathrm{Sh}}(W_k|\widetilde{\boldsymbol{\theta}}^{(b_O)}_{\mathscr{L}_k-\mathscr{J}})\right]^2\right\}, \tag{25}$$

where $\bar{\mathrm{Sh}}(W_k|\widetilde{\boldsymbol{\theta}}^{(b_O)}_{\mathscr{L}_k-\mathscr{J}}) = \sum_{b_I=1}^{B_I}\mathrm{Sh}(W_k|\widetilde{\boldsymbol{\theta}}^{(b_O,b_I)}_{\mathscr{J}}, \widetilde{\boldsymbol{\theta}}^{(b_O)}_{\mathscr{L}_k-\mathscr{J}})/B_I$. By plugging $\widehat{c}(\mathscr{J})$ into Equation (24), we can quantify the contribution from the estimation uncertainty on each BN parameter $\theta_\ell \in \boldsymbol{\theta}_{\mathscr{L}_k}$,

$$\widehat{\widehat{\mathrm{Sh}}}^*_{\theta_\ell}[\mathrm{Sh}(W_k)|\,\mathscr{X}] = \frac{1}{N_\pi}\sum_{n=1}^{N_\pi}\Delta_\ell\widehat{c}(\pi_n), \tag{26}$$

where $\Delta_\ell \widehat{c}(\pi_n) = \widehat{c}(P_\ell(\pi_n) \cup \{\ell\}) - \widehat{c}(P_\ell(\pi_n))$ and it is equivalent with computing $\Delta_{\pi_n(\ell)} \widehat{c}(\pi_n) = \widehat{c}(P_{\pi_n(\ell+1)}(\pi_n)) - \widehat{c}(P_{\pi_n(\ell)}(\pi_n))$ for all $\ell = 1, \ldots, L_k$. Similarly, for CPP/CQA criticality assessment, we can estimate the cost function,

$$\widehat{c}'(\mathscr{J}) = \frac{1}{B_O} \sum_{b_O=1}^{B_O} \left\{ \frac{1}{B_I - 1} \sum_{b_I=1}^{B_I} \left[p_{W_k, X_{m+1}}(\widetilde{\boldsymbol{\theta}}_{\mathscr{J}}^{(b_O, b_I)}, \widetilde{\boldsymbol{\theta}}_{\mathscr{L}_k - \mathscr{J}}^{(b_O)}) - \bar{p}_{W_k, X_{m+1}}(\widetilde{\boldsymbol{\theta}}_{\mathscr{L}_k - \mathscr{J}}^{(b_O)}) \right]^2 \right\}, \tag{27}$$

where $\bar{p}_{W_k, X_{m+1}}(\widetilde{\boldsymbol{\theta}}_{\mathscr{L}_k - \mathscr{J}}^{(b_O)}) = \sum_{b_I=1}^{B_I} p_{W_k, X_{m+1}}(\widetilde{\boldsymbol{\theta}}_{\mathscr{J}}^{(b_O, b_I)}, \widetilde{\boldsymbol{\theta}}_{\mathscr{L}_k - \mathscr{J}}^{(b_O)})/B_I$. Then, we estimate the contribution from the estimation uncertainty of θ_ℓ on the criticality assessement,

$$\widehat{\mathrm{Sh}}_{\theta_\ell}^* \left[p_{W_k, X_{m+1}} \middle| \mathscr{X} \right] = \frac{1}{N_\pi} \sum_{n=1}^{N_\pi} \Delta_\ell \widehat{c}'(\pi_n). \tag{28}$$

5. Empirical Study

To study the performance of the proposed framework, we consider the production process example for a monoclonal antibody (mAbs) bio-drug. The production procedure starts from main fermentation to API, including the main operation steps: (1) main fermentation, (2) centrifugation, (3) chromatography/purification, and (4) filtration. The simplified relational graph for this production procedure is provided in Figure 4. Here, we only consider the dominant CPPs/CQAs in each step, and the impacts of remaining factors are included in $\mathbf{e}$. To provide a clear illustration, CPPs and CQAs are grouped in Figure 4. In total, this BN has 21 nodes, consisting of 10 CPPs ($\mathbf{X}^p$) and 8 CQAs ($\mathbf{X}^a$) for intermediate product and 3 CQAs ($\mathbf{Y}$) for the final drug substance. The size of parameters $\boldsymbol{\theta}$ is 91, including 21 μ_i's, 21 v_i's, and 49 β_{ij}'s parameters. Even though there are many factors impacting on the bio-drug quality, the amount of real-world batch data could be very limited. *Thus, it is important to explore the causal relationships of the biopharmaceutical production process, which can reduce the model risk and also increase the interpretability for process risk and sensitivity analysis.* Given very limited process batch data, the empirical results show that we can provide reliable and interpretable risk analysis conclusions.

To study the performance of the proposed framework, we generate the simulated production process data $\mathscr{X}$, which mimics the "real-world data collection." The BN

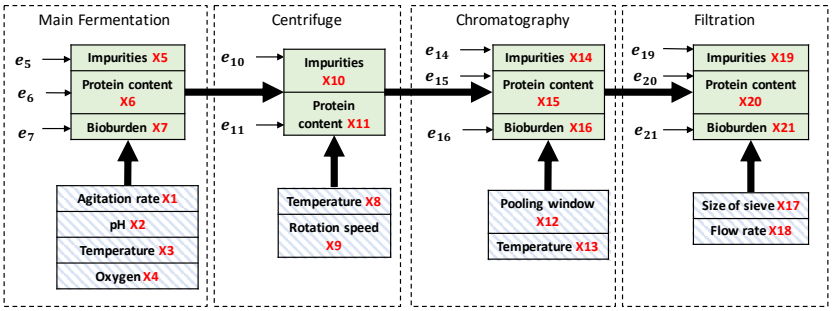

Figure 4: Relational graph for an antibody bio-drug API production process.

with parameters $\boldsymbol{\theta}^c = (\boldsymbol{\mu}^c, (\boldsymbol{\nu}^2)^c, \boldsymbol{\beta}^c)$ characterizing the underlying risk and interdependence is used for data generation, which is built on the biomanufacturing domain knowledge; see the detailed setting in online Appendix F. To assess the performance of the proposed framework, we assume that the true parameter values are unknown. We empirically study the convergence of BN parameter inference in Appendix G.

5.1. Risk Analysis and Criticality Assessment of CPPs/CQAs

In the real applications, the amount of biopharmaceutical production batch data could be limited. In the following empirical study, we generate the data $\mathscr{X}$ with size $R = 30$ to study the performance of the proposed risk and sensitivity analysis framework. For any CQA X_i of interest, at each posterior sample $\widetilde{\boldsymbol{\theta}}$, we follow Algorithm 2 in Appendix D to assess the criticality of CPPs/CQAs and other factors by applying Equation (14). Then, we can estimate the expected criticality for any input $W_k \in \{\mathbf{X}^p(\mathbf{N}') \cup \mathbf{X}^a(\mathbf{N}') \cup \mathbf{e}(\mathbf{N}')\}$. Specifically, in the k-th macro-replication, we first generate the "real-world" batch data $\mathscr{X}^{(k)}$. Let $p_{W_k,X_i}(\widetilde{\boldsymbol{\theta}}) = \mathrm{Sh}_i(W_k|\widetilde{\boldsymbol{\theta}})/\mathrm{Var}(X_i|\widetilde{\boldsymbol{\theta}})$ denote the criticality of W_k to the variance of X_i. We estimate the expected criticality $\mathrm{E}[p_{W_k,X_i}] = \iint p_{W_k,X_i}(\boldsymbol{\theta}) dP(\boldsymbol{\theta}|\mathscr{X}) dP(\mathscr{X}|\boldsymbol{\theta}^c) \times 100\%$ by using $\widehat{\mathrm{E}}[p_{W_k,X_i}] = \frac{1}{KB}\sum_{k=1}^{K}\sum_{b=1}^{B} p_{W_k,X_i}(\widetilde{\boldsymbol{\theta}}^{(k,b)}) \times 100\%$ with $\widetilde{\boldsymbol{\theta}}^{(k,b)} \sim p(\boldsymbol{\theta}|\mathscr{X}^{(k)})$ for $k = 1,\ldots,K$ and $b = 1,\ldots,B$, with $K = 20$ and $B = 1000$, and then record the results in terms of percentage (%) in Tables 1 and 2.

The BN model risk is characterized by the posterior $p(\boldsymbol{\theta}|\mathscr{X})$ and its impact on the CPPs/CQAs criticality assessment can be quantified by the posterior standard deviation

Table 1: The estimated criticality level $\widehat{\mathrm{E}}[p_{W_k,X_i}]$ and standard deviation $\widehat{\mathrm{SD}}[p_{W_k,X_i}]$ (in %) of any input CPP or other factor W_k impacting on the variance of intermediate or final product CQA X_i.

$\widehat{\mathrm{E}}[p_{W_k,X_i}]$	$X_i=X_5$	X_6	X_7	X_{10}	X_{11}	X_{14}	X_{15}	X_{16}	X_{19}	X_{20}	X_{21}
$W_k=X_1$	8.91(3.09)	8.93(3.11)	9.42(3.59)	8.51(2.87)	8.51(2.87)	5.87(1.88)	5.87(1.88)	5.87(1.88)	5.52(1.74)	5.52(1.74)	5.52(1.74)
X_2	0.82(0.38)	0.76(0.35)	0.96(0.75)	0.75(0.32)	0.75(0.32)	0.52(0.21)	0.52(0.21)	0.52(0.2)	0.49(0.19)	0.49(0.19)	0.49(0.19)
X_3	4.28(1.6)	4.33(1.61)	4.22(1.97)	4.05(1.46)	4(1.44)	2.75(0.93)	2.75(0.93)	2.75(0.93)	2.59(0.86)	2.59(0.86)	2.59(0.86)
X_4	85.75(4.02)	85.73(4.03)	83.29(4.84)	81.52(4.55)	81.6(4.53)	58.2(7.32)	58.22(7.32)	58.2(7.32)	55.09(7.21)	55.09(7.21)	55.06(7.21)
e_5	0.23(0.1)			0.05(0.04)	0.04(0.04)	0.03(0.02)	0.03(0.02)	0.03(0.02)	0.03(0.02)	0.03(0.02)	0.03(0.02)
e_6		0.24(0.1)		0.04(0.03)	0.05(0.04)	0.03(0.02)	0.03(0.02)	0.03(0.02)	0.03(0.02)	0.03(0.02)	0.03(0.02)
e_7			2.11(0.86)	0.05(0.04)	0.06(0.04)	0.03(0.02)	0.03(0.02)	0.03(0.02)	0.03(0.02)	0.03(0.02)	0.03(0.02)
X_8				1.02(0.41)	1.01(0.41)	0.68(0.24)	0.68(0.25)	0.69(0.25)	0.64(0.23)	0.64(0.23)	0.64(0.23)
X_9				3.91(1.42)	3.86(1.41)	2.66(0.9)	2.68(0.91)	2.67(0.9)	2.51(0.84)	2.51(0.84)	2.51(0.84)
e_{10}				0.1(0.04)		0.02(0.01)	0.02(0.01)	0.02(0.01)	0.02(0.01)	0.02(0.01)	0.02(0.01)
e_{11}					0.12(0.05)	0.02(0.01)	0.02(0.02)	0.02(0.01)	0.02(0.01)	0.02(0.01)	0.02(0.01)
X_{12}						1.86(0.63)	1.9(0.66)	1.86(0.63)	1.76(0.59)	1.76(0.59)	1.76(0.59)
X_{13}						27.31(6.46)	27.18(6.45)	27.3(6.46)	25.72(6.09)	25.73(6.09)	25.71(6.09)
e_{14}						0.02(0.01)			<0.01(<0.01)	<0.01(<0.01)	<0.01(<0.01)
e_{15}							0.06(0.02)		<0.01(<0.01)	<0.01(<0.01)	<0.01(<0.01)
e_{16}								0.02(0.01)	<0.01(<0.01)	<0.01(<0.01)	<0.01(<0.01)
X_{17}									1.27(0.45)	1.27(0.43)	1.27(0.43)
X_{18}									4.23(1.39)	4.26(1.39)	4.3(1.41)
e_{19}									0.04(0.01)		
e_{20}										0.01(<0.01)	
e_{21}											0.01(<0.01)

Table 2: The estimated criticality level $\widehat{\mathrm{E}}[p_{W_k,X_i}]$ and standard deviation $\widehat{\mathrm{SD}}[p_{W_k,X_i}]$ (in %) of any input CQA W_k on the variance of intermediate or final product CQA X_i.

$\widehat{\mathrm{E}}[p_{W_k,X_i}]$	$X_i=X_{10}$	X_{11}	X_{14}	X_{15}	X_{16}	X_{19}	X_{20}	X_{21}
$W_k=X_5$	43.1(11.18)	38.05(11.44)	28.68(6.48)	28.46(6.68)	28.44(6.48)	27.05(6.14)	26.99(6.11)	27.01(6.07)
X_6	37.97(10.79)	42.44(11.27)	28.64(6.32)	28.8(6.5)	28.88(6.28)	27.13(5.98)	27.18(5.95)	27.14(5.91)
X_7	13.91(4.42)	14.52(4.92)	10.11(2.62)	10.2(2.73)	10.11(2.67)	9.59(2.49)	9.6(2.49)	9.59(2.48)
X_{10}			37.17(6.55)	33.62(9.62)	34.59(6.7)	33.52(5.69)	33.01(5.16)	33.3(5.02)
X_{11}			33.64(6.42)	37.24(9.74)	36.23(6.72)	33.44(5.69)	33.95(5.2)	33.63(5.03)
X_{14}						32.65(12.69)	31.91(7.8)	33.74(6.69)
X_{15}						21.49(12.22)	25.1(6.9)	25.25(5.52)
X_{16}						40.31(14.51)	37.45(7.64)	35.42(7.85)

(SD), $\mathrm{SD}^*[p_{W_k,X_i}(\widetilde{\boldsymbol{\theta}})|\mathscr{X}]$. Based on the results from K macro-replications, we compute the expected SD for criticality estimation, $\mathrm{SD}[p_{W_k,X_i}] = \sqrt{\mathrm{E}[\mathrm{Var}^*(p_{W_k,X_i}(\widetilde{\boldsymbol{\theta}})|\mathscr{X})]} \times 100\%$, with the estimate,

$$\widehat{\mathrm{SD}}[p_{W_k,X_i}] = \sqrt{\frac{1}{K(B-1)}\sum_{k=1}^{K}\sum_{b=1}^{B}\left[p_{W_k,X_i}\left(\widetilde{\boldsymbol{\theta}}^{(k,b)}\right) - \bar{p}^{(k)}_{W_k,X_i}\right]^2} \times 100\%$$

where $\bar{p}^{(k)}_{W_k,X_i} = \frac{1}{B}\sum_{b=1}^{B} p_{W_k,X_i}(\widetilde{\boldsymbol{\theta}}^{(k,b)})$. In Tables 1 and 2, we record the results of SD in terms of percentage (%) in the bracket.

For the example shown in Figure 4, the CQAs of intermediate products include nodes $\{X_5,X_6,X_7,X_{10},X_{11},X_{14},X_{15},X_{16}\}$ and the final API CQAs are represented by nodes $\{X_{19},X_{20},X_{21}\}$. For each CQA output X_i, we record the criticality with the estimated mean $\widehat{\mathrm{E}}[p_{W_k,X_i}]$ and SD $\widehat{\mathrm{SD}}[p_{W_k,X_i}]$ from any CPP or other factor $W_k \in \{\mathbf{X}^p(\mathbf{N}') \cup \mathbf{e}(\mathbf{N}'))\}$ in Table 1. Then, we record the criticality of CQAs W_k to each following CQA X_i in Table 2. Under the example setting, we can see that the variations

in X_4 (oxygen in main fermentation) and X_{13} (temperature in chromatography) have the dominant impact on both intermediate and final product CQAs' variance. Compared with main fermentation and chromatography, the other two operation units (i.e., centrifuge and filtration) have relatively small impact on the final product quality variation.

The CQAs after main fermentation, i.e., $\{X_5, X_6, X_7\}$, together account for about 50% of final product CQA variance; and CQAs after chromatography, i.e., $\{X_{14}, X_{15}, X_{16}\}$ together account for about 90% of final CQA variation. *Thus, the CQAs of intermediate product close to the end of production process provides better explanation of the CQAs variation of final API and we can predict more accurate on the API quality.* This information can be used to guide the production process quality control. In addition, according to Table 2, compared with bioburden X_7, the CQAs impurities and protein X_5 and X_6 contribute relatively higher to the following intermediate product and final product CQAs variation.

5.2. Uncertainty Quantification and Sensitivity Analysis

Here we consider the product protein content X_{20} in Figure 4 to illustrate the performance of the proposed sensitivity analysis. Based on the results in Table 1, the CPPs X_4 and X_{13} have the dominant contributions to the variance of output X_{20} and they also have the relatively high estimation uncertainty. Thus, we conduct the BN-SV-MR sensitivity analysis to get the comprehensive information on how the BN model risk impacts on the criticality assessment for $p_{X_4,X_{20}}$ and $p_{X_{13},X_{20}}$.

Given the data $\mathscr{X}$, we have the posterior variance decomposition studying the criticality estimation error induced by the model risk, $\mathrm{Var}^*_{p(\boldsymbol{\theta}|\mathscr{X})}[p_{W_k,X_i}(\widetilde{\boldsymbol{\theta}})|\mathscr{X}] = \sum_{\theta_\ell \in \boldsymbol{\theta}(W_k,X_i)} \mathrm{Sh}^*_{\theta_\ell}\left[p_{W_k,X_i}(\widetilde{\boldsymbol{\theta}})\middle|\mathscr{X}\right]$. Then, we can estimate the relative contribution from each BN parameter $\theta_\ell \in \boldsymbol{\theta}(W_k,X_i)$ with $\mathrm{EP}_{\theta_\ell}(p_{W_k,X_i}) \equiv \mathrm{E}\left[\dfrac{\mathrm{Sh}^*_{\theta_\ell}\left[p_{W_k,X_i}(\widetilde{\boldsymbol{\theta}})\middle|\mathscr{X}\right]}{\mathrm{Var}^*_{p(\boldsymbol{\theta}|\mathscr{X})}\left[p_{W_k,X_i}(\widetilde{\boldsymbol{\theta}})\middle|\mathscr{X}\right]}\right]$. In the k-th macro-replication, given the data $\mathscr{X}^{(k)}$, we can estimate the contribution from each θ_ℓ by using $\widehat{\widehat{\mathrm{Sh}}}^*_{\theta_\ell}\left[p_{W_k,X_i}(\widetilde{\boldsymbol{\theta}})\middle|\mathscr{X}^{(k)}\right]$ and $\widehat{\mathrm{Var}}^*_{p(\boldsymbol{\theta}|\mathscr{X})}\left[p_{W_k,X_i}(\widetilde{\boldsymbol{\theta}})\middle|\mathscr{X}^{(k)}\right]$ which is estimated by using $N_\pi = 500$, $B_O = 5$ and $B_I = 20$; see Song et al. (2016) for the selection of sampling parameter setting. Thus, we have $\widehat{\mathrm{EP}}_{\theta_\ell}(p_{W_k,X_i}) \equiv \frac{1}{K}\sum_{k=1}^{K} \dfrac{\widehat{\widehat{\mathrm{Sh}}}^*_{\theta_\ell}\left[p_{W_k,X_i}(\widetilde{\boldsymbol{\theta}})\middle|\mathscr{X}^{(k)}\right]}{\widehat{\mathrm{Var}}^*_{p(\boldsymbol{\theta}|\mathscr{X})}\left[p_{W_k,X_i}(\widetilde{\boldsymbol{\theta}})\middle|\mathscr{X}^{(k)}\right]}$ with $K = 20$.

The parameters contributing to the estimation of $\mathrm{Sh}_{20}(X_4)$ include v_4^2 and 18 linear coefficients $\boldsymbol{\beta}$ on the paths from node X_4 to node X_{20}. The parameters contributing to the estimation of $\mathrm{Sh}_{20}(X_{13})$ include v_{13}^2 and 6 linear coefficients $\boldsymbol{\beta}$ located on the paths from X_{13} to X_{20}. Due to the space limit, we only present the top five parameters contributing most to the estimation uncertainty of criticality $p_{X_4,X_{20}}$ and $p_{X_{13},X_{20}}$, and aggregate the results for remaining parameters. The sensitivity analysis results, $\widehat{\mathrm{EP}}_{\theta_\ell}(p_{W_k,X_i}) \pm \mathrm{SE}\left[\widehat{\mathrm{EP}}_{\theta_\ell}(p_{W_k,X_i})\right]$, for $p_{X_4,X_{20}}$ and $p_{X_{13},X_{20}}$ are shown in Table 3, where SE stands for the standard error (SE). Notice that the parameters that contribute the most to the estimation error of the criticality $p_{X_4,X_{20}}$ and $p_{X_{13},X_{20}}$ are the variance parameters of CPPs (v_4^2 and v_{13}^2). The estimation errors of linear coefficients have similar and relatively small contributions. This information can guide the production process monitoring and data collection to efficiently reduce the estimation uncertainty of criticality assessment and improve our knowledge on the underlying probabilistic interdependence of the biopharmaceutical production process.

Table 3: The estimated relative contribution of each BN parameter estimation uncertainty (in terms of %) on criticality assessment $\widehat{\mathrm{EP}}_{\theta_\ell}(p_{W_k,X_i}) \pm \mathrm{SE}\left[\widehat{\mathrm{EP}}_{\theta_\ell}(p_{W_k,X_i})\right]$ for $p_{X_4,X_{20}}$ and $p_{X_{13},X_{20}}$.

$\theta_\ell \in \boldsymbol{\theta}(X_4,X_{20})$	v_4^2	$\beta_{11,15}$	$\beta_{10,14}$	$\beta_{15,20}$	$\beta_{14,20}$	rest
$\widehat{\mathrm{EP}}_{\theta_\ell}(p_{X_4,X_{20}})$	73.75±1.96	1.59±0.13	1.58±0.19	1.57±0.21	1.56±0.13	19.95±4.70
$\theta_\ell \in \boldsymbol{\theta}(X_{13},X_{20})$	v_{13}^2	$\beta_{13,15}$	$\beta_{16,20}$	$\beta_{15,20}$	$\beta_{13,16}$	rest
$\widehat{\mathrm{EP}}_{\theta_\ell}(p_{X_{13},X_{20}})$	68.16±6.40	5.57±0.50	5.54±0.49	5.42±0.44	5.23±0.46	10.08±8.87

6. Conclusions

Driven by the critical challenges in biomanufacturing, we propose a new BN based risk and sensitivity analysis framework to facilitate the production process learning and stability control. Since hundreds of factors could impact on the product quality and the amount of real-world batch data is often very limited, we explore the causal relationships and develop a BN characterizing the production process probabilistic interdependence. Considering SV can correctly account for probabilistic dependence and structural interactions, we propose the BN-SV based risk analysis to assess the criticality of each random input on the variance of product quality attributes. Given limited real-world batch data, there exists the model risk. We further introduce the

BN-SV-MR based UQ and SA that can provide the comprehensive understanding of how the model risk impacts on the end-to-end production process risk analysis and CPPs/CQAs criticality assessment. An antibody bio-drug production case is used to study the performance of the proposed framework.

There are several directions worth further investigation. In this paper, we ignore the risk induced by the model family selection error. How to quantify and incorporate this source of uncertainty will be considered in the future research. In addition, built on this study, we could develop new methodologies that can facilitate the production process forward prediction and backward problem detection.

References

Borgonovo, E., & Plischke, E. (2016). Sensitivity analysis: A review of recent advances. *European Journal of Operational Research*, *248*, 869 – 887.

Castro, J., Gómez, D., & Tejada, J. (2009). Polynomial calculation of the shapley value based on sampling. *Computers & Operations Research*, *36*, 1726–1730.

Feng, N., Wang, H. J., & Li, M. (2014). A security risk analysis model for information systems: Causal relationships of risk factors and vulnerability propagation analysis. *Information sciences*, *256*, 57–73.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian Data Analysis*. (2nd ed.). New York: Taylor and Francis Group, LLC.

Kaminsky, P., & Wang, Y. (2015). Analytical models for biopharmaceutical operations and supply chain management: A survery of research literature, pharmaceutical bioprocess. *Pharmaceutical Bioprocess*, *2*, 61–73.

Martagan, T., Krishnamurthy, A., & Leland, P. (2018). Managing trade-offs in protein manufacturing: how much to waste? *Manufacturing & Service Operations Management*, .

Martagan, T., Krishnamurthy, A., Leland, P. A., & Maravelias, C. T. (2017). Performance guarantees and optimal purification decisions for engineered proteins. *Operations Research*, *66*, 18–41.

Martagan, T., Krishnamurthy, A., & Maravelias, C. T. (2016). Optimal condition-based harvesting policies for biomanufacturing operations with failure risks. *IIE Transactions*, *48*, 440–461.

Mitchell, M. (2013). Determining criticality-process parameters and quality attributes part i: criticality as a continuum. *BioPharm International*, *26*.

Otto, R., Santagostino, A., & Schrader, U. (2014). From science to operations: Questions, choices, and strategies for success in biopharma. *McKinsey & Company*, .

Owen, A. B. (2014). Sobol'indices and shapley value. *SIAM/ASA Journal on Uncertainty Quantification*, *2*, 245–251.

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, *2*, 307–317.

Song, E., Nelson, B. L., & Staum, J. (2016). Shapley effects for global sensitivity analysis: Theory and computation. *SIAM/ASA Journal on Uncertainty Quantification*, *4*, 1060–1083.

Troyanskaya, O. G., Dolinski, K., Owen, A. B., Altman, R. B., & Botstein, D. (2003). A bayesian framework for combining heterogeneous data sources for gene function prediction (in saccharomyces cerevisiae). *Proceedings of the National Academy of Sciences*, *100*, 8348–8353.

Walsh, G. (2013). *Pharmaceutical biotechnology: concepts and applications*. John Wiley & Sons.

Wang, Y., Blache, R., Zheng, P., & Xu, X. (2018). A knowledge management system to support design for additive manufacturing using bayesian networks. *Journal of Mechanical Design*, *140*, 051701.

Appendix A. Detailed Derivation of Equation (6)

In order to show Equation (6), we consider more general results as following,

$$X_n = \mu_n + \sum_{k=1}^{m^p} \gamma_{k,n}(X_k - \mu_k) + \sum_{k=m^p+1}^{n} \gamma_{k,n} e_k, \tag{A.1}$$

for $n = m^p + 1, \ldots, m+1$, where $\gamma_{k,n}$ is given as Equations (7) and (8). Notice according to linear Gaussian model (4), we can write $X_{m^p+1} = \mu_{m^p+1} + \sum_{k=1}^{m^p} \beta_{k,m^p+1}(X_k - \mu_k) + e_{m^p+1}$, where $\beta_{k,m^p+1} = 0$ for $k \notin Pa(X_{m^p+1})$. Suppose Equation (A.1) holds for all $n = m^p + 1, \ldots, n_0$. For $n = n_0 + 1$, by applying linear Gaussian model, we have

$$\begin{aligned}
X_{n_0+1} &= \mu_{n_0+1} + \sum_{k=1}^{n_0} \beta_{k,n_0+1}(X_k - \mu_k) + e_{n_0+1}, \\
&= \mu_{n_0+1} + \sum_{k=1}^{m^p} \beta_{k,n_0+1}(X_k - \mu_k) + \sum_{\ell=m^p+1}^{n_0} \beta_{\ell,n_0+1} \left[\sum_{k=1}^{m^p} \gamma_{k,\ell}(X_k - \mu_k) + \sum_{k=m^p+1}^{\ell} \gamma_{k,\ell} e_k \right] \\
&\quad + e_{n_0+1}, \qquad (A.2) \\
&= \mu_{n_0+1} + \sum_{k=1}^{m^p} \left[\beta_{k,n_0+1} + \sum_{\ell=m^p+1}^{n_0} \gamma_{k,\ell}\beta_{\ell,n_0+1} \right] (X_k - \mu_k) + \sum_{k=m^p+1}^{n_0} \left[\sum_{\ell=k}^{n_0} \gamma_{k,\ell}\beta_{\ell,n_0+1} \right] e_k \\
&\quad + e_{n_0+1}, \\
&= \mu_{n_0+1} + \sum_{k=1}^{m^p} \gamma_{k,n_0+1}(X_k - \mu_k) + \sum_{k=m^p+1}^{n_0+1} \gamma_{k,n_0+1} e_k. \qquad (A.3)
\end{aligned}$$

Step (A.2) follows by applying (A.1). Step (A.3) follows by applying Equations (7) and (8). By mathematical induction, we can conclude that Equation (A.1) holds for all $n = m^p + 1, \ldots, m+1$.

Appendix B. Detailed Derivation of Equation (12)

According to the linear representation (10), for any $X_{k_1}, X_{k_2} \in \mathbf{X}^a(\mathbf{N}')$, the covariance,

$$\mathrm{Cov}(X_{k_1}, X_{k_2}) = \mathrm{Cov}\left[\sum_{X_\ell \in \mathbf{X}^p(X_{k_1})} \gamma_{\ell k_1} X_\ell + \sum_{\ell=m^p+1}^{k_1} \gamma_{\ell k_1} e_\ell, \sum_{X_\ell \in \mathbf{X}^p(X_{k_2})} \gamma_{\ell k_2} X_\ell + \sum_{\ell=m^p+1}^{k_2} \gamma_{\ell k_1} e_\ell \right]$$

$$= \sum_{X_\ell \in \{\mathbf{X}^p(X_{k_1}) \cap \mathbf{X}^p(X_{k_2})\}} \gamma_{\ell k_1}\gamma_{\ell k_2}\mathrm{Var}[X_\ell] + \sum_{\ell=m^p+1}^{\min(k_1,k_2)} \gamma_{\ell k_1}\gamma_{\ell k_2}\mathrm{Var}[e_\ell]$$

$$= \sum_{X_\ell \in \{\mathbf{X}^p(X_{k_1}) \cap \mathbf{X}^p(X_{k_2})\}} \gamma_{\ell k_1}\gamma_{\ell k_2}v_\ell^2 + \sum_{\ell=m^p+1}^{\min(k_1,k_2)} \gamma_{\ell k_1}\gamma_{\ell k_2}v_\ell^2,$$

since CPPs $X_\ell \in \mathbf{X}^p$ and error terms e_ℓ are mutually independent with each other.

Appendix C. Detailed Derivation of Equation (13)

We consider W_k and $\mathscr{J} \subset \mathscr{K}'/\{k\}$. For $\mathscr{J} = \emptyset$, we have

$$\frac{(n' - |\mathscr{J}| - 1)!|\mathscr{J}|!}{n'!}[c(\mathscr{J} \cup \{k\}) - c(\mathscr{J})] = \frac{1}{n'}\gamma_{ki}^2\mathrm{Var}(W_k).$$

For $|\mathscr{J}| = m'$ with $m' = 1, \ldots, n'-1$, we have

$$\begin{aligned}
&\sum_{\{\mathscr{J}:|\mathscr{J}|=m'\}} \frac{(n' - |\mathscr{J}| - 1)!|\mathscr{J}|!}{n'!}[c(\mathscr{J} \cup \{k\}) - c(\mathscr{J})] \\
&= \sum_{\{\mathscr{J}:|\mathscr{J}|=m'\}} \frac{(n'-m'-1)!m'!}{n'!}\left[\gamma_{ki}^2\mathrm{Var}(W_k) + 2\sum_{\ell \in \mathscr{J}} \gamma_{ki}\gamma_{\ell i}\mathrm{Cov}(W_k, W_\ell)\right] \\
&= \frac{(n'-m'-1)!m'!}{n'!}\left\{\binom{n'-1}{m'}\gamma_{ki}^2\mathrm{Var}(W_k)\right. \qquad \text{(C.1)} \\
&\quad \left. + 2\sum_{\ell \in \mathscr{K}'/\{k\}}\left[\sum_{\{\mathscr{J}:|\mathscr{J}|=m' \text{ and } \ell \in \mathscr{J}\}} \gamma_{ki}\gamma_{\ell i}\mathrm{Cov}(W_k, W_\ell)\right]\right\} \qquad \text{(C.2)} \\
&= \frac{1}{n'}\gamma_{ki}^2\mathrm{Var}(W_k) + 2\frac{(n'-m'-1)!m'!}{n'!}\binom{n'-2}{m'-1}\sum_{\ell \in \mathscr{K}'/\{k\}} \gamma_{ki}\gamma_{\ell i}\mathrm{Cov}(W_k, W_\ell) \qquad \text{(C.3)} \\
&= \frac{1}{n'}\gamma_{ki}^2\mathrm{Var}(W_k) + \frac{2m'}{n'(n'-1)}\sum_{\ell \in \mathscr{K}'/\{k\}} \gamma_{ki}\gamma_{\ell i}\mathrm{Cov}(W_k, W_\ell).
\end{aligned}$$

Step (C.1) holds because the number of all subsets $\mathscr{J}$ with size m' is $\binom{n'-1}{m'}$. In Step (C.2), we shift the order of sums over $\mathscr{J}$ and ℓ. Then, Step (C.3) holds because given W_ℓ, the number of subset $\{\mathscr{J} : |\mathscr{J}| = m' \text{ and } \ell \in \mathscr{J}\}$ is $\binom{n'-2}{m'-1}$. So, we

get the Shapley value,

$$
\begin{aligned}
\mathrm{Sh}_i(W_k|\boldsymbol{\theta}) &= \sum_{\mathscr{J}\subset\mathscr{K}/\{k\}} \frac{(n'-|\mathscr{J}|-1)!|\mathscr{J}|!}{n'!}[c(\mathscr{J}\cup\{k\})-c(\mathscr{J})] \\
&= \sum_{m'=0}^{n'-1}\left[\frac{1}{n'}\gamma_{ki}^2\mathrm{Var}(W_k)+\frac{2m'}{n'(n'-1)}\sum_{\ell\in\mathscr{K}'/\{k\}}\gamma_{ki}\gamma_{\ell i}\mathrm{Cov}(W_k,W_\ell)\right] \\
&= \gamma_{ki}^2\mathrm{Var}(W_k)+\sum_{\ell\in\mathscr{K}'/\{k\}}\gamma_{ki}\gamma_{\ell i}\mathrm{Cov}(W_k,W_\ell), \\
&= \begin{cases}
\gamma_{ki}^2 v_k^2, \;\; W_k = X_k \in \mathbf{X}^p(\mathbf{N}') \; or \; W_k = e_k \in \mathbf{e}(\mathbf{N}') \\
\gamma_{ki}^2\mathrm{Var}(X_k|\boldsymbol{\theta})+\sum_{X_\ell\in\mathbf{X}^a(\mathbf{N}')/\{X_k\}}\Big(\sum_{X_h\in\{\mathbf{X}^p(X_k)\cap\mathbf{X}^p(X_\ell)\}}\gamma_{hk}\gamma_{ki}\gamma_{h\ell}\gamma_{\ell i}v_h^2 \\
\quad +\sum_{h=m^p+1}^{\min(k,\ell)}\gamma_{hk}\gamma_{ki}\gamma_{h\ell}\gamma_{\ell i}v_h^2\Big), \;\; W_k = X_k \in \mathbf{X}^a(\mathbf{N}').
\end{cases}
\end{aligned}
$$

The last step is obtained by applying Equation (12).

Appendix D. Procedure for Production Process Risk Analysis

Given the BN parameters $\boldsymbol{\theta}$, we summarize the procedure for production process BN-SV based risk analysis in Algorithm 2. Suppose that we consider several consecutive operation steps corresponding to the (sub)graph $G(\mathbf{N}'|\boldsymbol{\theta}(\mathbf{N}'))$ with $\mathbf{N}'\subseteq\mathbf{N}$, and we are interested in a single response X_i with $m^p < i \le m+1$. Our objective is to quantify the contribution of each random input in $G(\mathbf{N}'|\boldsymbol{\theta}(\mathbf{N}'))$ to $\mathrm{Var}(X_i|\boldsymbol{\theta})$. For the complete production process, we have $\mathbf{N}'=\mathbf{N}$ and $X_i = X_{m+1}$.

Algorithm 2: Procedure for Production Process BN-SV Based Risk Analysis

Input: BN parameters $\boldsymbol{\theta}$, subgraph $G(\mathbf{N}'|\boldsymbol{\theta}(\mathbf{N}'))$, response node X_i.
Output: Variance decomposition of X_i in terms of all random inputs within $G(\mathbf{N}'|\boldsymbol{\theta}(\mathbf{N}'))$.
(1) Identify node sets $\mathbf{X}^a(\mathbf{N}')$, $\mathbf{X}^p(\mathbf{N}')$ and $\mathbf{e}(\mathbf{N}')$ representing the random inputs in the subgraph $G(\mathbf{N}'|\boldsymbol{\theta}(\mathbf{N}'))$; see Section 3.3.2;
(2) Calculate the Shapley value $\mathrm{Sh}_i(W_k|\boldsymbol{\theta})$ with $W_k = X_k$ or e_k by using Equation (13), which measures the contribution from W_k to the variance of response CQA X_i;
(3) Provide the variance decomposition of $\mathrm{Var}(X_i|\boldsymbol{\theta})$ by using Equation (14), and obtain the criticality of W_k on the variance of X_i: $p_{W_k,X_i}(\boldsymbol{\theta}) = \mathrm{Sh}_i(W_k|\boldsymbol{\theta})/\mathrm{Var}(X_i|\boldsymbol{\theta})$.

Appendix E. Derivation and Procedure for BN Learning and Gibbs Sampler

Appendix E.1. Bayesian Learning for BN Model Risk Quantification

We derive the posterior distribution of BN model parameters $p(\boldsymbol{\theta}|\mathscr{X})$ and introduce a Gibbs sampling approach to generate the posterior samples, $\widetilde{\boldsymbol{\theta}}^{(b)} \sim p(\boldsymbol{\theta}|\mathscr{X})$ with $b = 1,2,\ldots,B$ quantifying the model risk. In Section Appendix E.1.1, we first provide the derivation for conditional posterior distribution with complete production process data described in Section 4.1. Considering the situations where we could have some additional incomplete batch data (e.g., batches in the middle of production or thrown away at certain production step based on the quality control strategy), we further extend the Bayesian learning approach to cases with mixing data in Section Appendix E.1.2. Then, we provide the Gibbs sampling procedure to generate the posterior samples $\widetilde{\boldsymbol{\theta}}^{(b)}$ with $b = 1,2,\ldots,B$ in Section Appendix E.1.3.

Appendix E.1.1. Knowledge Learning for Cases with Complete Production Process Data

Following Section 4.1, we first derive the conditional posterior distribution for the weight coefficient β_{ij},

$$
\begin{aligned}
p(\beta_{ij}|\mathscr{X},\boldsymbol{\mu},\boldsymbol{v}^2,\boldsymbol{\beta}_{-ij}) &\propto \left[\prod_{r=1}^{R} p(x_j^{(r)}|x_{Pa(X_j)}^{(r)})\right] p(\beta_{ij}),\\
&\propto \exp\left\{-\sum_{r=1}^{R}\frac{1}{2v_j^2}\left[(x_j^{(r)}-\mu_j)-\beta_{ij}(x_i^{(r)}-\mu_i)-\sum_{k\in Pa(j)/\{i\}}\beta_{kj}(x_k^{(r)}-\mu_k)\right]^2\right.\\
&\quad \left.-\frac{1}{2\tau_{ij}^{(0)2}}\left(\beta_{ij}-\theta_{ij}^{(0)}\right)^2\right\},\\
&\propto \exp\left\{-\frac{1}{2v_j^2}\sum_{r=1}^{R}\left(\alpha_i^{(r)}\beta_{ij}-m_{ij}^{(r)}\right)^2-\frac{1}{2\tau_{ij}^{(0)2}}\left(\beta_{ij}-\theta_{ij}^{(0)}\right)^2\right\},\\
&\propto \exp\left\{-\frac{\beta_{ij}^2}{2}\left(\sum_{r=1}^{R}\frac{\alpha_i^{(r)2}}{v_j^2}+\frac{1}{\tau_{ij}^{(0)2}}\right)+\beta_{ij}\left(\sum_{r=1}^{R}\frac{\alpha_i^{(r)}m_{ij}^{(r)}}{v_j^2}+\frac{\theta_{ij}^{(0)}}{\tau_{ij}^{(0)2}}\right)\right\} = \mathscr{N}(\theta_{ij}^{(R)},\tau_{ij}^{(R)2}),
\end{aligned}
$$

where $\theta_{ij}^{(R)} = \dfrac{\tau_{ij}^{(0)2}\sum_{r=1}^{R}\alpha_i^{(r)}m_{ij}^{(r)} + v_j^2\theta_{ij}^{(0)}}{\tau_{ij}^{(0)2}\sum_{r=1}^{R}\alpha_i^{(r)2} + v_j^2}$ and $\tau_{ij}^{(R)2} = \dfrac{\tau_{ij}^{(0)2}v_j^2}{\tau_{ij}^{(0)2}\sum_{r=1}^{R}\alpha_i^{(r)2} + v_j^2}$ with $\alpha_i^{(r)} = x_i^{(r)} - \mu_i$, and $m_{ij}^{(r)} = (x_j^{(r)} - \mu_j) - \sum_{X_k \in Pa(X_j)/\{X_i\}} \beta_{kj}(x_k^{(r)} - \mu_k)$.

Second, we derive the conditional posterior distribution for the variance parameter $v_i^2 = \mathrm{Var}[X_i|Pa(X_i)]$ with $i = 1, 2, \ldots, m+1$,

$$
\begin{aligned}
p(v_i^2|\mathscr{X}, \boldsymbol{\mu}, \boldsymbol{v}_{-i}^2, \boldsymbol{\beta}) &\propto \left[\prod_{r=1}^{R} p(x_i^{(r)}|x_{Pa(X_i)}^{(r)})\right] p(v_i^2) \\
&\propto (v_i^2)^{-R/2-\kappa_i^{(0)}/2-1} \exp\left\{-\frac{1}{2v_j^2}\sum_{r=1}^{R}\left[(x_i^{(r)} - \mu_i) - \sum_{X_k \in Pa(X_i)} \beta_{ki}(x_k^{(r)} - \mu_k)\right]^2\right\} \\
&\propto (v_i^2)^{-R/2-\kappa_i^{(0)}/2-1} \exp\left\{-\frac{1}{2v_j^2}\sum_{r=1}^{R} u_i^{(r)2} - \frac{\lambda_i^{(0)}}{2v_j^2}\right\} = \text{Inv-}\Gamma\left(\frac{\kappa_i^{(R)}}{2}, \frac{\lambda_i^{(R)}}{2}\right),
\end{aligned}
$$

where $\kappa_i^{(R)} = \kappa_i^{(0)} + R$, $\lambda_i^{(R)} = \lambda_i^{(0)} + \sum_{r=1}^{R} u_i^{(r)2}$ and $u_i^{(r)} = (x_i^{(r)} - \mu_i) - \sum_{X_k \in Pa(X_i)} \beta_{ki}(x_k^{(r)} - \mu_k)$.

Third, we derive the conditional posterior distribution of mean parameter μ_i with $i = 1, 2, \ldots, m+1$ for any CPP and CQA,

$$
\begin{aligned}
&p(\mu_i|\mathscr{X}, \boldsymbol{\mu}_{-i}, \boldsymbol{v}^2, \boldsymbol{\beta}) \propto p(\mu_i)\prod_{r=1}^{R}\left[p(x_i^{(r)}|x_{Pa(X_i)}^{(r)}) \prod_{j \in \mathscr{S}(X_i)} p(x_j^{(r)}|x_{Pa(X_j)}^{(r)})\right] \\
&\propto \exp\left\{-\frac{1}{2v_i^2}\sum_{r=1}^{R}\left[(x_i^{(r)} - \mu_i) - \sum_{X_k \in Pa(X_i)} \beta_{ki}(x_k^{(r)} - \mu_k)\right]^2\right. \\
&\quad \left. -\sum_{r=1}^{R}\sum_{X_j \in \mathscr{S}(X_i)} \frac{1}{2v_j^2}\left[(x_j^{(r)} - \mu_j) - \sum_{X_k \in Pa(X_j)} \beta_{kj}(x_k^{(r)} - \mu_k)\right]^2 - \frac{1}{2\sigma_i^{(0)2}}\left(\mu_i - \mu_i^{(0)}\right)^2\right\}, \\
&\propto \exp\left\{-\frac{1}{2v_i^2}\sum_{r=1}^{R}\left(\mu_i - a_i^{(r)}\right)^2 - \sum_{r=1}^{R}\sum_{X_j \in \mathscr{S}(X_i)} -\frac{1}{2v_j^2}\left(\beta_{ij}\mu_i - c_{ij}^{(r)}\right)^2\right. \\
&\quad \left. -\frac{1}{2\sigma_i^{(0)2}}\left(\mu_i - \mu_i^{(0)}\right)^2\right\}, \\
&\propto \exp\left\{-\frac{\mu_i^2}{2}\left(\frac{R}{v_i^2} + \sum_{X_j \in S(X_i)} \frac{R\beta_{ij}^2}{v_j^2} + \frac{1}{\sigma_i^{(0)2}}\right) + \mu_i\left(\sum_{r=1}^{R}\frac{a_i^{(r)}}{v_i^2} + \sum_{r=1}^{R}\sum_{X_j \in \mathscr{S}(X_i)} \frac{\beta_{ij}c_{ij}^{(r)}}{v_j^2}\right.\right.
\end{aligned}
$$

$$+\frac{\mu_i^{(0)}}{\sigma_i^{(0)2}}\Bigg)\Bigg\} = \mathscr{N}(\mu_i^{(R)}, \sigma_i^{(R)2}),$$

where $\mu_i^{(R)} = \sigma_i^{(R)2}\left[\frac{\mu_i^{(0)}}{\sigma_i^{(0)2}} + \sum_{r=1}^{R}\frac{a_i^{(r)}}{v_i^2} + \sum_{r=1}^{R}\sum_{X_j\in S(X_i)}\frac{\beta_{ij}c_{ij}^{(r)}}{v_j^2}\right]$ and $\frac{1}{\sigma_i^{(R)2}} = \frac{1}{\sigma_i^{(0)2}} + \frac{R}{v_i^2} + \sum_{X_j\in S(X_i)}\frac{R\beta_{ij}^2}{v_j^2}$, with $a_i^{(r)} = x_i^{(r)} - \sum_{X_k\in Pa(X_i)}\beta_{kj}(x_k^{(r)} - \mu_k)$ and $c_{ij}^{(r)} = \beta_{ij}x_i^{(r)} - (x_j^{(r)} - \mu_j) + \sum_{X_k\in Pa(X_j)/\{X_i\}}\beta_{kj}(x_k^{(r)} - \mu_k)$.

Appendix E.1.2. Knowledge Learning for Cases with Mixing Data

Except the case with complete production data discussed in Section Appendix E.1.1, we consider the cases with additional incomplete data corresponding to certain "Top Sub-Graph", denoted by $G(\mathbf{N}'|\boldsymbol{\theta}(\mathbf{N}'))$ with $\mathbf{N}' \subseteq \mathbf{N}$, such that any CQA node $X_j \in \mathbf{N}'$ has $Pa(X_j) \subset \mathbf{N}'$. Since batch data collected from biopharmaceutical production process are usually limited, we want to fully utilize both complete and incomplete data to estimate the BN model parameters and improve our knowledge of production process.

Without loss of generality, we consider the real-world data including two data sets $\mathscr{X} = \{\mathscr{X}_1, \mathscr{X}_2\}$ with the complete data $\mathscr{X}_1 = \{(x_1^{(r_1)}, x_2^{(r_1)}, \ldots, x_{m+1}^{(r_1)})$ for $r_1 = 1,2,\ldots,R_1\}$ and the incomplete data $\mathscr{X}_2 = \{(x_i^{(r_2)} : X_i \in \mathbf{N}')$ for $r_2 = R_1+1, R_1+2,\ldots,R\}$, where $R = R_1 + R_2$. Our approach can be easily extended to cases with multiple incomplete data sets. We use the same prior distribution $p(\boldsymbol{\mu}, \boldsymbol{v}^2, \boldsymbol{\beta})$ as shown in Equation (15). Given the mixing data $\mathscr{X} = \{\mathscr{X}_1, \mathscr{X}_2\}$, we can derive the posterior distribution of $\boldsymbol{\theta}$,

$$p(\boldsymbol{\mu}, \boldsymbol{v}^2, \boldsymbol{\beta}|\mathscr{X}) \propto \prod_{r_1=1}^{R_1}\left[\prod_{i=1}^{m+1} p(x_i^{(r_1)}|x_{Pa(X_i)}^{(r_1)})\right]\prod_{r_2=R_1+1}^{R}\left[\prod_{X_i\in\mathbf{N}'} p(x_i^{(r_2)}|x_{Pa(X_i)}^{(r_2)})\right] p(\boldsymbol{\mu}, \boldsymbol{v}^2, \boldsymbol{\beta}).$$

For β_{ij} with $X_j \notin \mathbf{N}'$ or v_i^2 and μ_i with node $X_i \notin \mathbf{N}'$, the conditional posterior is the same as complete data case and we can utilize Equations (17), (18) and (19) by replacing $\mathscr{X}$ with $\mathscr{X}_1$.

Thus, to derive the full Gibbs sampler, we only need to provide the updated conditional posterior accounting for those nodes included in the incomplete data set $\mathscr{X}_2$. We

first derive the conditional posterior distribution for weight parameter β_{ij} with $X_j \in \mathbf{N}'$.

$$
\begin{aligned}
p(\beta_{ij}|\mathscr{X},\boldsymbol{\mu},\boldsymbol{v}^2,\boldsymbol{\beta}_{-ij}) &\propto \left[\prod_{r=1}^{R_1+R_2} p(x_j^{(r)}|x_{Pa(X_j)}^{(r)})\right] p(\beta_{ij}), \\
&\propto \exp\left\{-\sum_{r=1}^{R_1+R_2}\frac{1}{2v_j^2}\left[(x_j^{(r)}-\mu_j)-\beta_{ij}(x_i^{(r)}-\mu_i)-\sum_{k\in Pa(j)/\{i\}}\beta_{kj}(x_k^{(r)}-\mu_k)\right]^2\right. \\
&\qquad \left.-\frac{1}{2\tau_{ij}^{(0)2}}\left(\beta_{ij}-\theta_{ij}^{(0)}\right)^2\right\}, \\
&\propto \exp\left\{-\frac{1}{2v_j^2}\sum_{r=1}^{R_1+R_2}\left(\alpha_i^{(r)}\beta_{ij}-m_{ij}^{(r)}\right)^2-\frac{1}{2\tau_{ij}^{(0)2}}\left(\beta_{ij}-\theta_{ij}^{(0)}\right)^2\right\}, \\
&\propto \exp\left\{-\frac{\beta_{ij}^2}{2}\left(\sum_{r=1}^{R_1+R_2}\frac{\alpha_i^{(r)2}}{v_j^2}+\frac{1}{\tau_{ij}^{(0)2}}\right)+\beta_{ij}\left(\sum_{r=1}^{R_1+R_2}\frac{\alpha_i^{(r)}m_{ij}^{(r)}}{v_j^2}+\frac{\theta_{ij}^{(0)}}{\tau_{ij}^{(0)2}}\right)\right\} \\
&= \mathscr{N}(\theta_{ij}^{(R_1+R_2)},\tau_{ij}^{(R_1+R_2)2}),
\end{aligned}
\tag{E.1}
$$

where $\theta_{ij}^{(R_1+R_2)}=\dfrac{\tau_{ij}^{(0)2}\sum_{r=1}^{R_1+R_2}\alpha_i^{(r)}m_{ij}^{(r)}+v_j^2\theta_{ij}^{(0)}}{\tau_{ij}^{(0)2}\sum_{r=1}^{R_1+R_2}\alpha_i^{(r)2}+v_j^2}$ and $\tau_{ij}^{(R_1+R_2)2}=\dfrac{\tau_{ij}^{(0)2}v_j^2}{\tau_{ij}^{(0)2}\sum_{r=1}^{R_1+R_2}\alpha_i^{(r)2}+v_j^2}$ with $\alpha_i^{(r)}=x_i^{(r)}-\mu_i$ and $m_{ij}^{(r)}=(x_j^{(r)}-\mu_j)-\sum_{X_k\in Pa(X_j)/\{X_i\}}\beta_{kj}(x_k^{(r)}-\mu_k)$ for $r=1,2,\ldots,R$.

Then, we derive the conditional posterior distribution for v_i^2 with $X_i \in \mathbf{N}'$,

$$
\begin{aligned}
p(v_i^2|\mathscr{X},\boldsymbol{\mu},\boldsymbol{v}_{-i}^2,\boldsymbol{\beta}) &\propto \left[\prod_{r=1}^{R_1+R_2} p(x_i^{(r)}|x_{Pa(X_i)}^{(r)})\right] p(v_i^2), \\
&\propto (v_i^2)^{-(R_1+R_2)/2-\kappa_i^{(0)}/2-1}\exp\left\{-\frac{1}{2v_j^2}\sum_{r=1}^{R_1+R_2}\left[(x_i^{(r)}-\mu_i)-\sum_{X_k\in Pa(X_i)}\beta_{ki}(x_k^{(r)}-\mu_k)\right]^2\right\}, \\
&\propto (v_i^2)^{-(R_1+R_2)/2-\kappa_i^{(0)}/2-1}\exp\left\{-\frac{1}{2v_j^2}\sum_{r=1}^{R_1+R_2}u_i^{(r)2}-\frac{\lambda_i^{(0)}}{2v_j^2}\right\} \\
&= \text{Inv-}\Gamma\left(\frac{\kappa_i^{(R_1+R_2)}}{2},\frac{\lambda_i^{(R_1+R_2)}}{2}\right),
\end{aligned}
\tag{E.2}
$$

where $\kappa_i^{(R_1+R_2)}=\kappa_i^{(0)}+R$ and $\lambda_i^{(R_1+R_2)}=\lambda_i^{(0)}+\sum_{r=1}^{R}u_i^{(r)2}$ with $u_i^{(r)}=(x_i^{(r)}-\mu_i)-\sum_{X_k\in Pa(X_i)}\beta_{ki}(x_k^{(r)}-\mu_k)$ for $r=1,2,\ldots,R$.

After that, we derive the conditional posterior for mean parameter μ_i with $X_i \in \mathbf{N}'$,

$$
\begin{aligned}
p(\mu_i|\mathscr{X},\boldsymbol{\mu}_{-i},\boldsymbol{\nu}^2,\boldsymbol{\beta}) &\propto p(\mu_i)\prod_{r=1}^{R_1+R_2} p(x_i^{(r)}|x_{Pa(X_i)}^{(r)})\prod_{r_1=1}^{R_1}\prod_{X_j\in\mathscr{S}(X_i)} p(x_j^{(r_1)}|x_{Pa(X_j)}^{(r_1)}) \\
&\quad\cdot\prod_{r_2=R_1+1}^{R_1+R_2}\prod_{X_j\in S(X_i)\cap\mathbf{N}'} p(x_j^{(r_2)}|x_{Pa(X_j)}^{(r_2)}), \\
&\propto \exp\Bigg\{-\frac{1}{2v_i^2}\sum_{r=1}^{R_1+R_2}\left[(x_i^{(r)}-\mu_i)-\sum_{X_k\in Pa(X_i)}\beta_{ki}(x_k^{(r)}-\mu_k)\right]^2 \\
&\quad-\sum_{r_1=1}^{R_1}\sum_{X_j\in\mathscr{S}(X_i)}\frac{1}{2v_j^2}\left[(x_j^{(r_1)}-\mu_j)-\sum_{X_k\in Pa(X_j)}\beta_{kj}(x_k^{(r_1)}-\mu_k)\right]^2 \\
&\quad-\sum_{r_2=R_1+1}^{R_1+R_2}\sum_{X_j\in S(X_i)\cap\mathbf{N}'}\frac{1}{2v_j^2}\left[(x_j^{(r_2)}-\mu_j)-\sum_{X_k\in Pa(X_j)}\beta_{kj}(x_k^{(r_2)}-\mu_k)\right]^2 \\
&\quad-\frac{1}{2\sigma_i^{(0)2}}\left(\mu_i-\mu_i^{(0)}\right)^2\Bigg\}, \\
&\propto \exp\Bigg\{-\frac{\mu_i^2}{2}\left(\frac{R_1+R_2}{v_i^2}+\sum_{X_j\in\mathscr{S}(X_i)}\frac{R_1\beta_{ij}^2}{v_j^2}+\sum_{X_j\in S(X_i)\cap\mathbf{N}'}\frac{R_2\beta_{ij}^2}{v_j^2}\frac{1}{\sigma_i^{(0)2}}\right) \\
&\quad+\mu_i\left(\sum_{r=1}^{R_1+R_2}\frac{a_i^{(r)}}{v_i^2}+\sum_{r_1=1}^{R_1}\sum_{X_j\in\mathscr{S}(X_i)}\frac{\beta_{ij}c_{ij}^{(r_1)}}{v_j^2}+\sum_{r_2=R_1+1}^{R}\sum_{X_j\in\mathscr{S}(X_i)\cap\mathbf{N}'}\frac{\beta_{ij}c_{ij}^{(r_2)}}{v_j^2}+\frac{\mu_i^{(0)}}{\sigma_i^{(0)2}}\right)\Bigg\}, \\
&= \mathscr{N}\left(\mu_i^{(R_1+R_2)},\sigma_i^{(R_1+R_2)2}\right),
\end{aligned}
\tag{E.3}
$$

$$
\mu_i^{(R_1+R_2)}=\sigma_i^{(R_1+R_2)2}\left[\frac{\mu_i^{(0)}}{\sigma_i^{(0)2}}+\sum_{r=1}^{R_1+R_2}\frac{a_i^{(r)}}{v_i^2}+\sum_{r_1=1}^{R_1}\sum_{X_j\in\mathscr{S}(X_i)}\frac{\beta_{ij}c_{ij}^{(r_1)}}{v_j^2}+\sum_{r_2=R_1+1}^{R}\sum_{X_j\in\mathscr{S}(X_i)\cap\mathbf{N}'}\frac{\beta_{ij}c_{ij}^{(r_2)}}{v_j^2}\right],
$$

and

$$
\frac{1}{\sigma_i^{(R_1+R_2)2}}=\frac{1}{\sigma_i^{(0)2}}+\frac{R_1+R_2}{v_i^2}+\sum_{X_j\in\mathscr{S}(X_i)}\frac{R_1\beta_{ij}^2}{v_j^2}+\sum_{X_j\in\mathscr{S}(X_i)\cap\mathbf{N}'}\frac{R_2\beta_{ij}^2}{v_j^2}
$$

with $a_i^{(r)}=x_i^{(r)}-\sum_{X_k\in Pa(X_i)}\beta_{kj}(x_k^{(r)}-\mu_k)$ and $c_{ij}^{(r)}=\beta_{ij}x_i^{(r)}-(x_j^{(r)}-\mu_j)+\sum_{X_k\in Pa(X_j)/\{X_i\}}\beta_{kj}(x_k^{(r)}-\mu_k)$ for $r=1,2,\ldots,R$. Here for illustration, we have only provided the conditional posteriors with two datasets $\mathscr{X}_1$ and $\mathscr{X}_2$. These derivations can be easily extended to similar cases with multiple datasets collected from complete graph and different top sub-graphs.

Appendix E.1.3. Gibbs Sampling Procedure for BN Model Bayesian Inference

Based on the derived conditional posterior distributions in Sections Appendix E.1.1 and Appendix E.1.2, we provide the Gibbs sampling procedure in Algorithm 3 to generate posterior samples $\widetilde{\boldsymbol{\theta}}^{(b)} \sim p(\boldsymbol{\theta}|\mathscr{X})$ with $\widetilde{\boldsymbol{\theta}}^{(b)} = (\widetilde{\boldsymbol{\mu}}^{(b)}, \widetilde{\boldsymbol{v}}^{(b)2}, \widetilde{\boldsymbol{\beta}}^{(b)})$ and $b = 1, \ldots, B$. We first set the non-informative prior $p(\boldsymbol{\theta}) = p(\boldsymbol{\mu}, \boldsymbol{v}^2, \boldsymbol{\beta})$ as Equation (15), and generate the initial point $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\mu}^{(0)}, \boldsymbol{v}^{(0)2}, \boldsymbol{\beta}^{(0)})$ by sampling from the prior. Within each t-th iteration of Gibbs sampling, given the previous sample $\boldsymbol{\theta}^{(t-1)} = (\boldsymbol{\mu}^{(t-1)}, \boldsymbol{v}^{(t-1)2}, \boldsymbol{\beta}^{(t-1)})$, we sequentially compute and generate one sample from the conditional posterior distribution for each parameter β_{ij}, v_i^2 and μ_i. By repeating this procedure, we can get samples $\boldsymbol{\theta}^{(t)} = (\boldsymbol{\mu}^{(t)}, \boldsymbol{v}^{(t)2}, \boldsymbol{\beta}^{(t)})$ with $t = 1, \ldots, T$. To reduce the initial bias and correlations between consecutive samples, we remove the first T_0 samples and keep one for every h samples. Consequently, we obtain the posterior samples $\widetilde{\boldsymbol{\theta}}^{(b)} \sim p(\boldsymbol{\theta}|\mathscr{X})$ with $b = 1, \ldots, B$.

Algorithm 3: Gibbs Sampling Procedure for BN Model Risk Quantification

Input: the prior $p(\boldsymbol{\theta})$ and real-world data $\mathscr{X}$.

Output: Posterior samples $\widetilde{\boldsymbol{\theta}}^{(b)} = (\widetilde{\boldsymbol{\mu}}^{(b)}, \widetilde{\boldsymbol{v}}^{(b)2}, \widetilde{\boldsymbol{\beta}}^{(b)}) \sim p(\boldsymbol{\theta}|\mathscr{X})$ with $b = 1, \ldots, B$.

(1) Set the initial value $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\mu}^{(0)}, \boldsymbol{v}^{(0)2}, \boldsymbol{\beta}^{(0)})$ by sampling from prior $p(\boldsymbol{\theta})$;

for $t = 1, 2, \ldots, T$ **do**

(2) Given the previous sample $\boldsymbol{\theta}^{(t-1)} = (\boldsymbol{\mu}^{(t-1)}, \boldsymbol{v}^{(t-1)2}, \boldsymbol{\beta}^{(t-1)})$;

(3) For each β_{ij}, generate $\beta_{ij}^{(t)} \sim p(\beta_{ij}|\mathscr{X}, \beta_{12}^{(t)}, \ldots, \beta_{i,j-1}^{(t)}, \beta_{i,j+1}^{(t-1)}, \ldots, \beta_{m,m+1}^{(t-1)}, \boldsymbol{\mu}^{(t-1)}, \boldsymbol{v}^{(t-1)2})$ through Equation (17) for complete data or (E.1) for mixing data;

(4) For each v_i^2, generate $v_i^{(t)2} \sim p(v_i^2|\mathscr{X}, \boldsymbol{\beta}^{(t)}, v_1^{(t)2}, \ldots, v_{i-1}^{(t)2}, v_{i+1}^{(t-1)2}, \ldots, v_{m+1}^{(t-1)2}, \boldsymbol{\mu}^{(t-1)})$ through Equation (18) for complete data or (E.2) for mixing data;

(5) For each μ_i, generate $\mu_i^{(t)} \sim p(\mu_i|\mathscr{X}, \boldsymbol{\beta}^{(t)}, \boldsymbol{v}^{(t),2}, \mu_1^{(t)}, \ldots, \mu_{i-1}^{(t)}, \mu_{i+1}^{(t-1)}, \ldots, \mu_n^{(t-1)})$ through Equation (19) for complete data or (E.3) for mixing data;

(6) Obtain a new posterior sample $\boldsymbol{\theta}^{(t)} = (\boldsymbol{\mu}^{(t)}, \boldsymbol{v}^{(t)2}, \boldsymbol{\beta}^{(t)})$;

(7) Set $\widetilde{\boldsymbol{\theta}}^{(b)} = \boldsymbol{\theta}^{(T_0+(b-1)h+1)}$ with some constant integer T_0 and h, to reduce the initial bias and correlation between consecutive samples.

Appendix F. Simulated Biopharmaceutical Production Data

To study the performance of proposed framework, we generate the simulated production process data $\mathscr{X}$, which mimics the "real-world data collection." The BN with parameters $\boldsymbol{\theta}^c$ characterizing the underlying production process interdependence is used for data generation, which is built according to the biomanufacturing domain knowledge. The ranges of CPPs/CQAs are listed Table F.4. For each CPP $X_j \in \mathbf{X}^p$ with range $[x_j^{low}, x_j^{up}]$, we can specify the marginal distribution $X_j \sim \mathcal{N}(\mu_j^c, (v_j^c)^2)$ with mean $\mu_j^c = (x_j^{low} + x_j^{up})/2$ and standard deviation $v_j^c = (x_j^{up} - x_j^{low})/4$. For each CQA $X_i \in \{\mathbf{X}^a \cup \mathbf{Y}\}$ with range $[x_i^{low}, x_i^{up}]$, we have mean $\mu_i^c = (x_i^{low} + x_i^{up})/2$ and marginal variance $\text{Var}(X_i) = [(x_i^{up} - x_i^{low})/4]^2$. Based on Equation (14), the corresponding parameter v_i^c can be computed through back-engineering. For the complex interdependence, Table F.5 provides the relative associations with levels (i.e., high, median, low) between input CPPs/CQAs with output CQAs in each operation unit, which is built based on the "cause-and-effect matrix" in Mitchell (2013). For the high, median and low association between X_i to X_j, we set the coefficient $\beta_{ij}^c = 0.9, 0.6, 0.3$ respectively. Thus, we can specify the underlying true parameters $\boldsymbol{\theta}^c = (\boldsymbol{\mu}^c, (\boldsymbol{v}^2)^c, \boldsymbol{\beta}^c)$. To mimic the "real-world" data collection, we generate the production batch data $\mathscr{X}$ using the BN model with $\boldsymbol{\theta}^c$. Then, to assess the performance of proposed framework, we assume that the true parameter values are unknown.

Table F.4: Range of CPPs/CQAs in the production procedure.

Process Unit Operation	CPP	Range	CQA	Range
Main Fermentation	pH	6.8-7.2	impurities	3-11 pl
	temperature	20-30 C	protein content	1-5 g/L
	Oxygen	2.5-7.5%	bioburden	5-15 CFU/100mL
	agitation rate	1.1-2.5 m/s		
Centrifuge	temperature	20 to 30 C	impurities	3-11 pl
	rotation speed	3-5K RPM	protein content	5-15 CFU/100mL
Chromatography	pooling window	10-30 min	impurities	3-11 pl
	temperature	2-10 C	protein content	1-5 g/L
			bioburden	5-15 CFU/100mL
Filtration	size of sieve	0.1-0.5 um	impurities	3-11 pl
	flow rate	25-100 mL/min	protein content	1-5 g/L
			bioburden	5-15 CFU/100mL

Table F.5: Relative association between input CPPs/CQAs with output CQAs in each process unit operation.

Process Unit Operation	Input CPPs/CQAs	Output CQAs		
		impurities	protein content	bioburden
Main Fermentation	pH	high	high	low
	temperature	high	high	low
	Oxygen	high	high	low
	agitation rate	high	high	low
Centrifuge	temperature	medium	medium	—
	rotation speed	medium	medium	—
	impurities (main fermentation)	medium	medium	—
	protein content (main fermentation)	medium	medium	—
	bioburden (main fermentation)	medium	medium	—
Chromatography	pooling window	high	medium	high
	temperature	high	medium	high
	impurities (centrifuge)	high	medium	high
	protein content (centrifuge)	high	medium	high
Filtration	size of sieve	low	medium	medium
	flow rate	low	medium	medium
	impurities (chromatography)	low	medium	medium
	protein content (chromatography)	low	medium	medium
	bioburden (chromatography)	low	medium	medium

Appendix G. Study the Bayesian Learning and Inference

To study the performance of proposed framework, we generate the simulated production process data $\mathscr{X}$, which mimics the "real-world data collection." The BN with parameters $\boldsymbol{\theta}^c = (\boldsymbol{\mu}^c, (\boldsymbol{v}^2)^c, \boldsymbol{\beta}^c)$ characterizing the underlying risk and interdependence is used for data generation, which is built on the biomanufacturing domain knowledge; see the detailed setting in online Appendix Appendix F.

To assess the performance of proposed framework, we assume that the true parameter values are unknown. We empirically study the convergence of BN parameter inference. In each k-th macro-replication, we first mimic the "real-world" production batch data collection through generating $\mathscr{X}^{(k)} = \{\mathbf{X}_1^{(k)}, \ldots, \mathbf{X}_R^{(k)}\}$ with $\mathbf{X}_i^{(k)} \sim F(\mathbf{X}|\boldsymbol{\theta}^c)$ for $i = 1, \ldots, R$ and $k = 1, \ldots, K$. Then, we generate B posterior samples $\widetilde{\boldsymbol{\theta}}^{(k,b)} \sim p(\boldsymbol{\theta}|\mathscr{X}^{(k)})$ with $b = 1, 2, \ldots, B$. For the Gibbs sampler in Algorithm 3 provided in online Appendix Appendix E.1.3, we set the initial warm-up length $T_0 = 500$ and step-size $h = 10$. With different size of complete "real-world" batch data $R = 30, 100, 500$, we compute the mean squared error (MSE) for each parameter $\theta_\ell \in \boldsymbol{\theta}$: $\text{MSE}(\theta_\ell) = \iint \left(\theta_\ell - \theta_\ell^c\right)^2 dP(\theta_\ell|\mathscr{X}) dP(\mathscr{X}|\boldsymbol{\theta}^c)$. Based on $K = 20$ macro-replications and $B = 1000$ posterior samples of BN parameters,

we estimate $\mathrm{MSE}(\theta_\ell)$ with $\widehat{\mathrm{MSE}}(\theta_\ell) = \frac{1}{KB}\sum_{k=1}^{K}\sum_{b=1}^{B}\left(\widetilde{\theta}_\ell^{(k,b)} - \theta_\ell^c\right)^2$. Since the total number of parameters is large, we further group parameters by mean $\boldsymbol{\mu}$, conditional variance $\boldsymbol{v}^2$ and linear coefficients $\boldsymbol{\beta}$, and take average of the sample MSE respectively: $\widehat{\mathrm{MSE}}(\boldsymbol{\mu}) = \frac{1}{|\boldsymbol{\mu}|}\sum_{\theta_\ell\in\boldsymbol{\mu}}\widehat{\mathrm{MSE}}(\theta_\ell)$, $\widehat{\mathrm{MSE}}(\boldsymbol{v}^2) = \frac{1}{|\boldsymbol{v}^2|}\sum_{\theta_\ell\in\boldsymbol{v}^2}\widehat{\mathrm{MSE}}(\theta_\ell)$, and $\widehat{\mathrm{MSE}}(\boldsymbol{\beta}) = \frac{1}{|\boldsymbol{\beta}|}\sum_{\theta_\ell\in\boldsymbol{\beta}}\widehat{\mathrm{MSE}}(\theta_\ell)$. The corresponding results are reported in Table G.6. As the size of real-world data R increases, the average MSE decreases, which implies the posterior samples obtained by Gibbs sampling procedure can converge to the true parameters.

Table G.6: The MSE of $\boldsymbol{\mu}$, $\boldsymbol{v}^2$ and $\boldsymbol{\beta}$ esimated by using the Gibbs sampling.

Batch Data Size	$\widehat{\mathrm{MSE}}(\boldsymbol{\mu})$	$\widehat{\mathrm{MSE}}(\boldsymbol{v}^2)$	$\widehat{\mathrm{MSE}}(\boldsymbol{\beta})$
$R = 30$	0.122±0.032	0.276±0.029	0.0225±0.0013
$R = 100$	0.075±0.023	0.061±0.006	0.0063±0.0004
$R = 500$	0.009±0.003	0.013±0.001	0.0011±0.00004