Proceedings of the 2015 Winter Simulation Conference
L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, eds.

QUANTIFYING STATISTICAL UNCERTAINTY FOR DEPENDENT INPUT MODELS WITH FACTOR STRUCTURE

Wei Xie

Department of Industrial and Systems Engineering
Rensselaer Polytechnic Institute
Troy, NY 12180-3590, USA

Cheng Li

Department of Statistical Science
Duke University
Durham, NC 27708-0251, USA

Hongtan Sun

Department of Industrial and Systems Engineering
Rensselaer Polytechnic Institute
Troy, NY 12180-3590, USA

ABSTRACT

Simulation used for the performance assessment of stochastic systems is usually driven by input models estimated from real-world data, which introduces both input and simulation uncertainty to the performance estimates. For many complex systems, because the components of input models are mutually dependent, an efficient estimation of dependence could improve the system performance assessment. Since the dependence could be caused by underlying common factors, we explore Gaussian copula factor models to characterize input models with dependence. We propose a Bayesian framework to quantify both input and simulation uncertainty. The input uncertainty is quantified by the posterior of input models and then propagated to output means by direct simulation, with the simulation estimation error characterized by the posterior distributions of system mean responses. This Bayesian framework delivers credible intervals that quantify the overall uncertainty of system performance estimates. Our approach is supported by both asymptotic theory and empirical study.

1 INTRODUCTION

When simulation is used to assess the behavior of complex stochastic systems, it is usually driven by the input models estimated from finite real-world data. There exist two sources of errors: the simulation and the input estimation errors. Ignoring either source of error could lead to unfounded confidence in the simulation assessment of system performance, see Xie, Nelson, and Barton (2014a). Therefore, it is necessary to have a statistical uncertainty analysis that could efficiently use the real-world data and the computational resource to quantify the overall uncertainty of system performance estimates, assuming the logic of the simulation model is valid.

The choice of input models has direct impact on the system performance estimates. Since it is typically difficult to construct a joint distribution and further generate samples from it to drive the simulations, especially as the dimension of the distribution increases, the prevalent practice is to have the input models composed of independent univariate distributions. However, this assumption does not always faithfully represent the underlying physical distributions. For example, in a project planning network, the activity durations for different tasks could be correlated if they are affected by the same nuisance factors, e.g., weather conditions. In the automobile industry, if the products are different models of cars and they are sold in several geographically-distinct markets, the joint distribution of their demands would likely be impacted by the local gas price and by other micro-economic conditions. In maintenance planning, the life

times of different components could be simultaneously impacted by the operation temperature. Therefore, to correctly assess the performance of stochastic systems, we should faithfully capture the dependence structure in the input models.

Considering the amount of information required to specify a joint distribution, most input modeling research focuses on methods that characterize the input models by some key properties, including marginals and dependence (Biller and Ghosh 2006). Given the partial information, various approaches were introduced to construct a full joint distribution, including NORmal-To-Anything (NORTA) and Gaussian copula. NORTA proposed by Cario and Nelson (1997) could represent and generate random vectors with flexible marginal distributions and a correlation matrix. Given finite real-world data, the estimation error of NORTA could be quantified by either frequentist approaches, e.g., the bootstrap, or Bayesian posterior distributions. Then, the impact of input uncertainty could be propagated to output means by either a metamodel or direct simulations; see Xie, Nelson, and Barton (2014b) and Akcay and Biller (2014).

Since NORTA is based on moment-matching, some feasible combination of marginals and a correlation matrix may not have a feasible NORTA representation, called *NORTA infeasible*. This issue could become worse as the dimension of joint input distribution increases (Ghosh and Henderson 2002). Biller and Corlu (2011) used the Gaussian copula to model dependent input models. For any feasible combination of continuous marginals and a correlation matrix, we can find a Gaussian copula representation. It could avoid NORTA infeasible issue.

In this paper, we assume that dependent input models are characterized by marginal distributions and correlation matrices. Given the partial information, we could build full joint distributions by Gaussian copula. Since for many practical problems, e.g., automobile and maintenance examples we mentioned above, the component-wise dependence is typically caused by underlying common factors, we further explore a factor structure to characterize the dependence. Motivated by Murray et al. (2013), we represent unknown input models by the *Gaussian copula factor models*. As the dimension of the input distribution increases, with finite amount of data, the correlation matrix in Gaussian copula can be better represented by a few important common factors plus some noise. In such situation, the Gaussian copula factor model is expected to produce more efficient estimation for the dependence in the input models.

We propose a unified Bayesian framework to quantify the input and simulation estimation uncertainty. Posterior distributions of the input models are used to quantify the input model and parameter estimation error, called *input uncertainty*. Then, we propagate the input uncertainty to outputs by direct simulation with the simulation estimation error quantified by the posterior of system mean responses. This Bayesian framework delivers a credible interval (CrI) that quantifies the overall uncertainty of the system mean performance estimate.

It is usually intractable to do fully Bayesian inference that characterizes both the marginal uncertainty and the dependence uncertainty simultaneously (Joe 2005). Therefore, *a two-stage estimation* is commonly employed in the copula literature, which is also used in our paper. This approximation quantifies the marginal and dependence estimation uncertainty separately. Specifically, in the first stage, the marginal uncertainty is estimated based on the data for each component of the input model. In the second stage, we only quantify the dependence uncertainty. According to the marginal likelihood approach in Section 8.3 Severini (2000), we could take marginal distributions as nuisance "parameters" and only make inference for the dependence based on a summary statistic that is independent of the marginals. Motivated by Hoff (2007) and Murray et al. (2013), the extended rank likelihood can be taken as such a summary statistic and we do Bayesian inference for the dependence structure with the estimation error quantified by a posterior distribution. Since our Bayesian inference for the dependence is based only on the extended rank likelihood instead of the full original data, the two-stage approach simplifies the estimation problem at the cost of only using partial information in the data. However, when the true marginals are all continuous, Hoff (2007) demonstrates that this estimation over the dependence is asymptotically efficient.

Our main focus in this paper is on quantifying the estimation error of the dependence structure. Following the idea of two-stage estimation, we take the empirical marginals as the true distributions and ignore the

marginal uncertainty. Developing Bayesian approaches to quantify both marginals and the dependence uncertainty is our on-going research.

The main contributions of this paper are as follows. First, since the component-wise dependence is typically caused by some underlying common factors in many situations, we explore the factor structure in the input models with dependence. In particular, we use the Gaussian copula factor model to represent flexible input joint distributions. Second, we make Bayesian inference on input models by the Gibbs sampling algorithm, prove the consistency of input distributions drawn from the posterior, and compare its empirical performance with the usual Gaussian copula model without factors. Third, a unified Bayesian framework is proposed to quantify the overall uncertainty of system performance estimates. Input uncertainty quantified by the posteriors of input models is propagated to outputs by direct simulation with the simulation estimation error quantified by posterior distributions of system mean responses.

The next section describes the problem statement and our objective. This is followed by a unified Bayesian framework supported by asymptotic study in Section 3. We then report results from an empirical study in Section 4 and conclude the paper in Section 5.

2 PROBLEM STATEMENT AND PROPOSED APPROACH

Suppose the simulation output is a function of input distributions F and random numbers. The output from the rth replication can be written as

$$Y_r(F) = \mu(F) + \varepsilon_r(F)$$

where, $\mu(F) = \mathrm{E}[Y_r(F)]$ denotes the unknown output mean and $\varepsilon_r(F)$ represents the simulation error with $\varepsilon_r(F) \sim N(0, \sigma_\varepsilon^2(F))$. Notice that the simulation outputs depend on the choice of input distributions F that could be composed of mutually independent univariate and multivariate joint distributions. For notation simplification, we focus on the case where there is only one multivariate joint distribution in F with the dimension, denoted by d.

We assume that the input distribution F is characterized by marginals and a correlation matrix. Suppose the marginals, denoted by $\{F_1, F_2, \ldots, F_d\}$, are continuous distributions. For an arbitrary feasible combination of marginals and a correlation matrix, there exists a Gaussian copula representation. Gaussian copula can be interpreted as a transformation from the domain of a $(d \times 1)$ random vector $\mathbf{X} \sim F$ to another domain where the dependence is easier to model, denoted by $\mathbf{Z}$, (Smith 2011)

$$\mathbf{X} \overset{U_j = F_j(X_j)}{\longrightarrow} \mathbf{U} \overset{Z_j = \Phi^{-1}(U_j)}{\longrightarrow} \mathbf{Z} \tag{1}$$

for $j = 1, 2, \ldots, d$, where $\mathbf{U}$ follows a multivariate uniform distribution and $\mathbf{Z}$ follows a multivariate normal distribution, $\mathbf{Z} \sim \mathbf{N}_d(\mathbf{0}, \mathbf{C})$ with $\mathbf{C}$ denoting the correlation matrix. The Gaussian copula representation for F could be written as

$$F(x_1, x_2, \ldots, x_d) = \Phi_d\Big(\Phi^{-1}\big[F_1(x_1)\big], \Phi^{-1}\big[F_2(x_2)\big], \ldots, \Phi^{-1}\big[F_d(x_d)\big]; \mathbf{C}\Big) \tag{2}$$

where, $\Phi_d(\cdot)$ and $\Phi(\cdot)$ denote the d-dimensional multivariate and univariate standard normal distributions. Therefore, F could be specified by marginals $F_1, F_2, \ldots, F_d$ and a correlation matrix $\mathbf{C}$. We have the unknown true input joint distribution, denoted by F^c, with the corresponding Gaussian copula representation specified by $(F_1^c, F_2^c, \ldots, F_d^c, \mathbf{C}^c)$.

Given m real-world data, denoted by a $(m \times d)$ matrix $\mathscr{X}_m^{(0)} \equiv \big(\mathbf{X}_1^{(0)}, \mathbf{X}_2^{(0)}, \ldots, \mathbf{X}_m^{(0)}\big)^\top$, with $\mathbf{X}_i^{(0)} \overset{i.i.d.}{\sim} F^c$ for $i = 1, 2, \ldots, m$, the input uncertainty can be characterized by the posterior distribution $\mathrm{P}(F|\mathscr{X}_m^{(0)})$. Since the input distribution could be specified by $(F_1, F_2, \ldots, F_d, \mathbf{C})$, the posterior distribution could be written as $\mathrm{P}\big(F_1, F_2, \ldots, F_d, \mathbf{C}|\mathscr{X}_m^{(0)}\big)$. We could generate B samples of input distribution from the posterior distribution to quantify the input uncertainty, denoted by $\{\tilde{F}^{(1)}, \tilde{F}^{(2)}, \ldots, \tilde{F}^{(B)}\}$, where $\tilde{F}^{(b)} \equiv \big(\tilde{F}_1^{(b)}, \tilde{F}_2^{(b)}, \ldots, \tilde{F}_d^{(b)}, \tilde{\mathbf{C}}^{(b)}\big)$ with $b = 1, 2, \ldots, B$. Notice that $\tilde{\cdot}$ denotes the sample from corresponding posterior distribution.

By the two-stage estimation, we quantify the marginal and dependence uncertainty separately. In this paper, we ignore the marginal uncertainty and mainly focus on estimating the uncertainty for dependence that is characterized by the posterior $\mathrm{P}\big(\mathbf{C}|\mathscr{X}_m^{(0)}\big)$. Let $\widehat{F}_j$ represent the empirical distribution for the jth marginal distribution and it is constructed based on data $\{X_{1j}^{(0)},X_{2j}^{(0)},\ldots,X_{mj}^{(0)}\}$. By plugging in the point estimator for marginals, we have the samples $\tilde{F}^{(b)}=\big(\widehat{F}_1,\widehat{F}_2,\ldots,\widehat{F}_d,\tilde{\mathbf{C}}^{(b)}\big)$ with $\tilde{\mathbf{C}}^{(b)}\overset{i.i.d.}{\sim}\mathrm{P}\big(\mathbf{C}|\mathscr{X}_m^{(0)}\big)$ for $b=1,2,\ldots,B$ quantifying the dependence estimation uncertainty.

To quantify the impact of input uncertainty on the system mean performance estimate, we need to propagate it to output means. If the mean response surface $\mu(\cdot)$ is known, a two-sided equal probability $(1-\alpha)100\%$ CrI, denoted by $[q_{\alpha/2},q_{1-\alpha/2}]$, could be used to quantify the impact of input uncertainty, where $q_\gamma\equiv\inf\{q:F_U(q)\geq\gamma\}$ and $F_U(t)\equiv\mathrm{P}(\mu(\tilde{F})\leq t|\mathscr{X}_m^{(0)})$ with $\gamma=\alpha/2,1-\alpha/2$. Since we typically do not have the closed form for the quantile q_γ, a Monte Carlo sampling approaches could be used to build a percentile CrI quantifying the impact of input uncertainty. To get a precise estimation on the quantile, B is recommended to be a few thousands.

However, the true mean response $\mu(\cdot)$ is typically unknown for many complex stochastic systems. At any F, we could estimate $\mu(F)$ by the simulation. Suppose each simulation run could be expensive. If the parametric families of marginal distributions are known, instead of running simulations at all samples of input distribution, $\tilde{F}^{(1)},\tilde{F}^{(2)},\ldots,\tilde{F}^{(B)}$, to estimate the system mean responses, we could construct a metamodel based on simulation outputs at a few samples of input distribution and then use it to propagate the input uncertainty to outputs, while reducing the simulation estimation uncertainty (Xie, Nelson, and Barton 2014a).

Since the parametric families of $F_1,F_2,\ldots,F_d$ are typically unknown, the input distribution F could not be specified by finite parameters. It is hard to build a metamodel for $\mu(F)$. In this paper, the direct simulation is used to propagate the input uncertainty to output means. Specifically, at each sample $\tilde{F}^{(b)}$, the direct simulation is used to estimate $\mu_b\equiv\mu(\tilde{F}^{(b)})$ and the estimation error is quantified by the posterior of system mean response estimate $\tilde{\mu}_b$. Suppose the number of replications allocated to $\tilde{F}^{(b)}$ is n_b and we have the simulation outputs, denoted by $\mathbf{Y}^{(b)}\equiv\{Y_1^{(b)},Y_2^{(b)},\ldots,Y_{n_b}^{(b)}\}$ with $Y_r^{(b)}|\tilde{F}^{(b)}\overset{i.i.d.}{\sim}\mathrm{N}(\mu_b,\sigma_\varepsilon^2(\tilde{F}^{(b)}))$. Then, the posterior $\mathrm{P}(\tilde{\mu}_b|\mathbf{Y}^{(b)},\tilde{F}^{(b)})$ could be used to quantify the simulation estimation uncertainty .

We are interested in $\mu^c\equiv\mu(F^c)$. Since $\mu(\cdot)$ and F^c are unknown, the input uncertainty is quantified by $\tilde{F}\sim P(F|\mathscr{X}_m^{(0)})$ and at any F, the simulation uncertainty is quantified by $\tilde{\mu}(F)\sim P(\mu(F)|\mathbf{Y}(F))$ with $\mathbf{Y}(F)$ representing the simulation outputs at F. Therefore, the estimation uncertainty for μ^c is quantified by the posterior of $\tilde{\mu}\equiv\tilde{\mu}(\tilde{F})$. Our goal is to efficiently use the information in real-world data and the simulation resource to accurately estimate the input model and reduce the estimation uncertainty of $\tilde{\mu}$.

3 A BAYESIAN FRAMEWORK

In this section, we propose a unified Bayesian framework to quantify both input and simulation uncertainty. In Section 3.1, we explore the factor structure for modeling the dependence. Our asymptotic study shows that as the amount of real-world data increases to infinity, the estimated input distribution converges to underlying true distribution F^c. In Section 3.2, the input uncertainty is propagated to output means by direct simulation with simulation estimation uncertainty quantified by the posteriors of mean responses. In Section 3.3, a procedure is constructed to account for both input and simulation uncertainty, which delivers a CrI quantifying the overall uncertainty for the system performance estimate. We show that as the amount of real-world data and the computational resource go to infinity, the system mean performance estimate converges to the true response μ^c.

3.1 Bayesian Quantification for Input Uncertainty

Given m real-world data, $\mathscr{X}_m^{(0)} = \left(\mathbf{X}_1^{(0)}, \mathbf{X}_2^{(0)}, \ldots, \mathbf{X}_m^{(0)}\right)^\top$, we would like to make inference about the dependence of the input distribution characterized by the correlation matrix $\mathbf{C}$. If the marginals $F_1, F_2, \ldots, F_d$ are known, we could have the corresponding data on latent variables $\mathbf{Z}$ by applying the transformation $Z_{ij}^{(0)} = \Phi^{-1}\left[F_j(X_{ij}^{(0)})\right]$ for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, d$. For continuous marginals $F_1, F_2, \ldots, F_d$, there exists a one-to-one mapping between $\mathbf{X}_i^{(0)}$ and $\mathbf{Z}_i^{(0)}$. Since $\mathbf{Z}_i^{(0)} \overset{i.i.d.}{\sim} \mathrm{N}_d(\mathbf{0}, \mathbf{C})$ for $i = 1, 2, \ldots, m$, the posterior distribution $\mathrm{P}\left(\mathbf{C} | \mathbf{Z}_1^{(0)}, \mathbf{Z}_2^{(0)}, \ldots, \mathbf{Z}_m^{(0)}\right)$ is straightforward to calculate; see Gelman et al. (2004). It is equivalent to the posterior $\mathrm{P}(\mathbf{C} | \mathscr{X}_m^{(0)})$ used to quantify the input uncertainty.

However, in general the marginal distributions $F_1, F_2, \ldots, F_d$ are unknown. The only information for the transformation $\Phi^{-1}\left[F_j(\cdot)\right]$ is an increasing function. Based on the marginal likelihood (Severini 2000), the extended rank likelihood proposed by Hoff (2007) could be used to generate a set of a $(m \times d)$ data matrix $\mathscr{Z}_m \equiv \left(\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_m\right)^\top$ that is consistent with real-world data in terms of the relative order. We denote this set by $D\left(\mathscr{X}_m^{(0)}\right)$,

$$D\left(\mathscr{X}_m^{(0)}\right) \equiv \{\mathscr{Z}_m : X_{ij}^{(0)} < X_{i'j}^{(0)} \Rightarrow Z_{ij} < Z_{i'j}\}.$$

Thus, $\mathscr{Z}_m \in D(\mathscr{X}_m^{(0)})$ is a statistic that is independent on the marginals $F_1, F_2, \ldots, F_d$ and only depends on the correlation $\mathbf{C}$. Given data $\mathscr{Z}_m \in D(\mathscr{X}_m^{(0)})$, the posterior $\mathrm{P}(\mathbf{C} | \mathscr{Z}_m \in D(\mathscr{X}_m^{(0)}))$ could be used to quantify the uncertainty of $\mathbf{C}$. Since the marginals distributions are unknown, $\mathscr{Z}_m$ consistent with $\mathscr{X}_m^{(0)}$ is not unique. To account for the impact from unknown marginals on the dependence estimation, we further generate samples of $\mathscr{Z}_m$ from $D\left(\mathscr{X}_m^{(0)}\right)$. Therefore, the uncertainty of correlation matrix could be characterized by $\mathrm{P}(\mathbf{C} | \mathscr{X}_m^{(0)}) = \mathrm{E}_{\mathscr{Z}_m \in D(\mathscr{X}_m^{(0)})}\left[\mathrm{P}\left(\mathbf{C} | \mathscr{Z}_m \in D(\mathscr{X}_m^{(0)})\right)\right]$.

Since the dependence between different components of $\mathbf{X}_i$ is often caused by some underlying common factors in many practical problems, we explore a factor model to characterize the dependence in the latent random vector $\mathbf{Z}_i$ so that we could efficiently estimate $\mathbf{C}$. Our modeling is motivated by Murray et al. (2013). Let $\mathbf{M}_i$ denote the scaled $(d \times 1)$ latent variables with $\mathbf{M}_i \sim \mathrm{N}_d(\mathbf{0}, \Sigma)$ and it has a simple factor model

$$\mathbf{M}_i = \Lambda \boldsymbol{\eta}_i + \mathbf{e}_i \tag{3}$$

where, Λ is a $(d \times k)$ loading matrix; $\boldsymbol{\eta}$ is a $(k \times 1)$ vector of common factors with $\boldsymbol{\eta} \sim \mathrm{N}_k(\mathbf{0}, \mathbf{I}_{k \times k})$; $\mathbf{e}$ is a $(d \times 1)$ vector of noise with $\mathbf{e} \sim \mathrm{N}_d(\mathbf{0}, \mathbf{I}_{d \times d})$; $\boldsymbol{\eta}$ and $\mathbf{e}$ are independent. Therefore, the covariance matrix of $\mathbf{M}$ has a factor structure, $\Sigma = \Lambda\Lambda^\top + \mathbf{I}_{d \times d}$. To make the factor loadings comparable to each other and also consistent with the Gaussian copula in Equation (2), we set $\mathbf{Z}_i$ to be equal to the scaled $\mathbf{M}_i$, $Z_{ij} = M_{ij} / \sqrt{\sum_{h=1}^k \lambda_{jh}^2 + 1}$, with $\mathbf{Z}_i \sim \mathrm{N}_d(\mathbf{0}, \mathbf{C})$, where λ_{ij} represents the element in the loading matrix Λ. Both the covariance matrix Σ and the correlation matrix $\mathbf{C}$ have factor structure.

Given $\mathscr{X}_m^{(0)}$, the input uncertainty is characterized by the posterior $\mathrm{P}(\mathbf{C} | \mathscr{X}_m^{(0)})$. Theorem 1 shows that as the amount of real-world data goes to infinity, the estimate of the input distribution $\tilde{F}$ converges to the true distribution F^c in probability.

Theorem 1 Suppose the prior of correlation matrix, denoted by $\mathrm{P}(\mathbf{C})$, has positive mass on the neighborhood of $\mathbf{C}^c$. Suppose $\mathbf{C}^c$ has a factor decomposition in k_0 factors. Let $\mathscr{X}_m \equiv \left(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m\right)^\top$ with $\mathbf{X}_i \overset{i.i.d.}{\sim} F^c$ for $i = 1, 2, \ldots, m$ and let $\widehat{F}_j$ be the empirical distribution for the jth marginal distribution F_j. Given a sample $\mathscr{Z}_m \in D(\mathscr{X}_m)$, $\tilde{\mathbf{C}}$ is a correlation matrix randomly drawn from the posterior $\mathrm{P}(\mathbf{C} | \mathscr{Z}_m \in D(\mathscr{X}_m))$, with a factor decomposition in k factors, where $d \geq k \geq k_0$. Let $\tilde{F}$ be a distribution with the Gaussian

copula factor representation

$$\tilde{F}(x_1,x_2,\ldots,x_d)=\Phi_d\Big(\Phi^{-1}\big[\widehat{F}_1(x_1)\big],\Phi^{-1}\big[\widehat{F}_2(x_2)\big],\ldots,\Phi^{-1}\big[\widehat{F}_d(x_d)\big];\tilde{\mathbf{C}}\Big).$$

Then, as $m\to\infty$, $\tilde{F}(x_1,x_2,\ldots,x_d)$ converges to $F^c(x_1,x_2,\ldots,x_d)$ uniformly for all $(x_1,x_2,\ldots,x_d)\in\mathbb{R}^d$ in probability, i.e. $\|\tilde{F}-F^c\|_\infty\xrightarrow{p}0$, where $\|\tilde{F}-F^c\|_\infty=\sup_{(x_1,x_2,\ldots,x_d)\in\mathbb{R}^d}\big|\tilde{F}(x_1,x_2,\ldots,x_d)-F^c(x_1,x_2,\ldots,x_d)\big|$.

Proof: For any generic matrix $\mathbf{A}$, let $\|\mathbf{A}\|=\sqrt{\mathrm{tr}(\mathbf{A}\mathbf{A}^\top)}$ be the Frobenius norm of $\mathbf{A}$. We first show that as $m\to\infty$, the correlation matrix estimate is consistent in the Frobenius norm, i.e. $\|\tilde{\mathbf{C}}-\mathbf{C}^c\|\xrightarrow{p}0$.

For any $\delta>0$, by Theorem 1 in Murray et al. (2013), we have

$$\lim_{m\to\infty}\mathrm{P}(\|\tilde{\mathbf{C}}-\mathbf{C}^c\|\le\delta|\mathscr{Z}_m\in D(\mathscr{X}_m))=1 \tag{4}$$

Then, by accounting for the finite sampling uncertainty for $\mathscr{X}_m$ and the impact of marginal uncertainty, we have

$$\begin{aligned}\lim_{m\to\infty}\mathrm{P}(\|\tilde{\mathbf{C}}-\mathbf{C}^c\|\le\delta)&=\lim_{m\to\infty}\Big(\mathrm{E}_{\mathscr{X}_m}\mathrm{E}_{\mathscr{Z}_m\in D(\mathscr{X}_m)|\mathscr{X}_m}\Big)\Big[\mathrm{P}(\|\tilde{\mathbf{C}}-\mathbf{C}^c\|\le\delta|\mathscr{Z}_m\in D(\mathscr{X}_m))\Big]\\&=\Big(\mathrm{E}_{\mathscr{X}_m}\mathrm{E}_{\mathscr{Z}_m\in D(\mathscr{X}_m)|\mathscr{X}_m}\Big)\Big[\lim_{m\to\infty}\mathrm{P}(\|\tilde{\mathbf{C}}-\mathbf{C}^c\|\le\delta|\mathscr{Z}_m\in D(\mathscr{X}_m))\Big]=1.\end{aligned}$$

The second step follows by applying the dominated convergence theorem, and the third step follows by applying Equation (4). Therefore, as $m\to\infty$, the correlation matrix estimate converges to the true correlation matrix in probability, $\tilde{\mathbf{C}}\xrightarrow{p}\mathbf{C}^c$.

By the Glivenko-Cantelli Theorem in Van Der Vaart (1998), we have

$$\|(\widehat{F}_1,\widehat{F}_2,\ldots,\widehat{F}_d)-(F_1^c,F_2^c,\ldots,F_d^c)\|_\infty\xrightarrow{a.s.}0\text{ as }m\to\infty.$$

Then, for any $(x_1,x_2,\ldots,x_d)$ in the domain of F^c, by applying the continuous mapping theorem (Van Der Vaart 1998), we have

$$\begin{aligned}\tilde{F}(x_1,x_2,\ldots,x_d)&=\Phi_d\Big(\Phi^{-1}\big[\widehat{F}_1(x_1)\big],\Phi^{-1}\big[\widehat{F}_2(x_2)\big],\ldots,\Phi^{-1}\big[\widehat{F}_d(x_d)\big];\tilde{\mathbf{C}}\Big)\\&\xrightarrow{p}\Phi_d\Big(\Phi^{-1}\big[F_1^c(x_1)\big],\Phi^{-1}\big[F_2^c(x_2)\big],\ldots,\Phi^{-1}\big[F_d^c(x_d)\big];\mathbf{C}^c\Big)=F^c(x_1,x_2,\ldots,x_d).\end{aligned}$$

Since both $\tilde{F}$ and F^c are distribution functions and $F^c(x_1,x_2,\ldots,x_d)$ is continuous for all $(x_1,x_2,\ldots,x_d)\in\mathbb{R}^d$, the input distribution $\tilde{F}$ uniformly converges to F^c in probability, i.e. $\|\tilde{F}-F^c\|_\infty\xrightarrow{p}0$, as $m\to\infty$. □

For finite samples, the posterior distribution $\mathrm{P}\big(\mathbf{C}|\mathscr{X}_m^{(0)}\big)$ could be used to characterize the input uncertainty. Since there is no closed form $\mathrm{P}\big(\mathbf{C}|\mathscr{X}_m^{(0)}\big)$, we use the Gibbs sampler to generate samples of $\tilde{\mathbf{C}}$ from $\mathrm{P}\big(\mathbf{C}|\mathscr{X}_m^{(0)}\big)$ to quantify the input uncertainty. Let $H=(\boldsymbol{\eta}_1,\boldsymbol{\eta}_2,\ldots,\boldsymbol{\eta}_m)$. Let $\boldsymbol{\lambda}_j$ denote the jth row of Λ with prior defined by $\mathrm{N}_k(\mathbf{0},\Psi_j)$ and $\mathscr{Z}_j$ denote the jth column of $\mathscr{Z}$. Since $\mathbf{e}_i\sim\mathrm{N}_d(\mathbf{0},\mathbf{I}_{d\times d})$, the conditional distributions involved in Gibbs sampling are

$$\boldsymbol{\lambda}_j|H,\mathscr{Z}_j\ \sim\ \mathrm{N}(\mathscr{Z}_jH^\top(HH^\top+\Psi_j^{-1})^{-1},(HH^\top+\Psi_j^{-1})^{-1})\text{ for }j=1,2,\ldots,d \tag{5}$$

$$\boldsymbol{\eta}_i|\Lambda,\mathbf{Z}_i\ \sim\ \mathrm{N}((\Lambda^\top\Lambda+\mathbf{I}_{k\times k})^{-1}\Lambda^\top\mathbf{Z}_i,(\Lambda^\top\Lambda+\mathbf{I}_{k\times k})^{-1})\text{ for }i=1,2,\ldots,m \tag{6}$$

$$Z_{ij}|\Lambda,\boldsymbol{\eta}_i\ \sim\ \mathrm{TN}\left(\sum_{h=1}^k\lambda_{jh}\eta_{hi},1,Z_{ij}^\ell,Z_{ij}^u\right)\text{ for }i=1,2,\ldots,m\text{ and }j=1,2,\ldots,d \tag{7}$$

$$\tilde{C}_{jj'}\ =\ \frac{\sum_{h=1}^k\lambda_{jh}\lambda_{hj'}}{\sqrt{1+\sum_{h=1}^k\lambda_{jh}^2}\sqrt{1+\sum_{h=1}^k\lambda_{hj'}^2}}\text{ for }j\neq j'\text{ and }j,j'=1,2,\ldots,d \tag{8}$$

where, $\mathrm{TN}(u,\sigma^2,a,b)$ denotes the normal with mean u, variance σ^2 and truncated to (a,b); $Z_{ij}^{\ell}=\max\{Z_{i'j}: X_{i'j}^{(0)}<X_{ij}^{(0)}\}$ and $Z_{ij}^{u}=\min\{Z_{i'j}: X_{i'j}^{(0)}>X_{ij}^{(0)}\}$. Based on Equations (5)-(8), we could use the Gibbs sampler to iteratively generate samples of $(\Lambda,H,\mathscr{Z},\tilde{\mathbf{C}})$; see Gelman et al. (2004) for the detailed description of the Gibbs sampling algorithm.

3.2 Bayesian Quantification for Simulation Estimation Uncertainty

By applying the Gibbs sampler, we could generate B samples of input distribution, $\tilde{F}^{(b)}=\left(\widehat{F}_1,\widehat{F}_2,\ldots,\widehat{F}_d,\tilde{\mathbf{C}}^{(b)}\right)$ with $b=1,2,\ldots,B$ to quantify the input uncertainty. At any $\tilde{F}^{(b)}$, the mean response $\mu(\tilde{F}^{(b)})$ is unknown and estimated by direct simulations. Denote the number of replications allocated to B samples of input distribution by $(n_1,n_2,\ldots,n_B)$. Let $\mathbf{Y}^{(b)}\equiv\left(Y_1(\tilde{F}^{(b)}),Y_2(\tilde{F}^{(b)}),\ldots,Y_{n_b}(\tilde{F}^{(b)})\right)$ denote the simulation outputs with $Y_r(\tilde{F}^{(b)})|\tilde{F}^{(b)} \overset{i.i.d.}{\sim} \mathrm{N}\left(\mu(\tilde{F}^{(b)}),\sigma_\varepsilon^2(\tilde{F}^{(b)})\right)$ for $r=1,2,\ldots,n_b$. Suppose there is no prior knowledge about $\mu(\cdot)$. Let $\tilde{\mu}_b\equiv\tilde{\mu}(\tilde{F}^{(b)})$. By DeGroot (1970), the posterior for $\tilde{\mu}_b|\mathbf{Y}^{(b)},\tilde{F}^{(b)}$ is $\mathrm{N}\left(\bar{Y}_b,\sigma_\varepsilon^2(\tilde{F}^{(b)})/n_b\right)$, where $\bar{Y}_b=\sum_{j=1}^{n_b}Y_j(\tilde{F}^{(b)})/n_b$. The unknown $\sigma_\varepsilon^2(\tilde{F}^{(b)})$ could be estimated by the sample variance of simulation outputs $\mathbf{Y}^{(b)}$, denoted by S_b^2. Therefore, the simulation estimation uncertainty for the mean response at $\tilde{F}^{(b)}$ could be characterized by the posterior $\mathrm{N}(\bar{Y}_b,S_b^2/n_b)$.

In the empirical study, when we propagate the input uncertainty to output mean, for simplification, we use the direct simulation with equal allocation, $n_1=n_2=\cdots=n_B$. The sequential experiment design proposed in Yi, Xie, and Zhou (2015) could assign more computational budget to the samples of input distributions that contribute most to estimating the CrI, $[q_{\alpha/2},q_{1-\alpha/2}]$. Therefore, it could efficiently use the simulation resource to propagate the input uncertainty to output mean and reduce the simulation estimation uncertainty.

3.3 Procedure to Construct a CrI

In this section, we propose a procedure to account for both input and simulation uncertainty and deliver a CrI quantifying the overall uncertainty of $\tilde{\mu}=\tilde{\mu}(\tilde{F})$. Then, we show this CrI is asymptotically consistent.

The procedure to construct the CrI mainly includes following steps.

(1) Specify the prior for Λ.
(2) Apply Gibbs sampling to generate B samples of $(\Lambda,H,\mathscr{Z},\tilde{\mathbf{C}})$ by using Equations (5)-(8).
(3) Allocate n_b replications to $\tilde{F}^{(b)}=(\widehat{F}_1,\widehat{F}_2,\ldots,\widehat{F}_d,\tilde{\mathbf{C}}^{(b)})$ for $b=1,2,\ldots,B$.
(4) Loop $b=1,2,\ldots,B$
   (a) Generate $\mathbf{Z}_i^{(b)} \overset{i.i.d.}{\sim} \mathrm{N}_d(\mathbf{0},\tilde{\mathbf{C}}^{(b)})$ and do transformation to obtain $\mathbf{X}_i^{(b)}$: $X_{ij}^{(b)}=\left(\widehat{F}_j\right)^{-1}\left[\Phi\left(Z_{ij}^{(b)}\right)\right]$.
   (b) Repeat Step 4.(a) to generate samples of $\mathbf{X}_i^{(b)}$. Use them to drive the simulations and obtain simulation outputs $Y_1^{(b)},Y_2^{(b)},\ldots,Y_{n_b}^{(b)}$. Calculate the sample mean $\bar{Y}_b$ and sample standard deviation S_b.
   (c) Generate $\tilde{\mu}_b\sim\mathrm{N}(\bar{Y}_b,S_b^2/n_b)$.
   End loop
(5) Report the $(1-\alpha)$ percentile CrI to quantify both input and simulation uncertainty:

$$\mathrm{CrI}=[\tilde{\mu}_{(\lceil B\alpha/2\rceil)},\tilde{\mu}_{(\lceil B(1-\alpha/2)\rceil)}]$$

where, $\alpha\in(0,1)$ and $\tilde{\mu}_{(b)}$ is the bth order statistic with $\tilde{\mu}_{(1)}\le\tilde{\mu}_{(2)}\le\cdots\le\tilde{\mu}_{(B)}$.

Theorem 2 describes the limiting behavior of the CrI obtained from the procedure above. As the number of real-world data and the computational resource increase to infinity, the CrI shrinks to the true mean response μ^c. This indicates that our CrI provides a valid Bayesian quantification of the uncertainty for the unknown system performance.

Theorem 2 Suppose the conditions in Theorem 1 and the following conditions hold. Let $n_{\min} = \min_b n_b$.

(1) The simulation variability is bounded around F^c: There exist finite constants $\varepsilon_0 > 0$ and $C > 0$, such that for any distribution F with $\|F - F^c\|_\infty \le \varepsilon_0$, $\sigma_\varepsilon^2(F) \le C$ holds.

(2) The mean response $\mu(\cdot)$ is continuous at F^c: For any $\delta_1 > 0$, there exists $\delta_2 > 0$ such that if $\|F - F^c\|_\infty \le \delta_2$, then $|\mu(F) - \mu(F^c)| \le \delta_1$.

Then, as $m \to \infty$ and $n_{\min} \to \infty$, the CrI $[\tilde{\mu}_{(\lceil B\alpha/2 \rceil)}, \tilde{\mu}_{(\lceil B(1-\alpha/2) \rceil)}]$ shrinks to $\mu^c = \mu(F^c)$ in probability.

Proof: By Theorem 1, for any $\tilde{F}$ randomly drawn from the posterior $\mathrm{P}(F|\mathscr{X}_m^{(0)})$, we have $\|\tilde{F} - F^c\|_\infty \xrightarrow{p} 0$ as $m \to \infty$. By applying Condition (2) and the continuous mapping theorem, we have $|\mu(\tilde{F}) - \mu(F^c)| \xrightarrow{p} 0$ as $m \to \infty$. Since $\bar{Y}_b|\tilde{F}^{(b)} \sim \mathrm{N}(\mu(\tilde{F}^{(b)}), \sigma_\varepsilon^2(\tilde{F}^{(b)})/n_b)$, for any $\delta > 0$,

$$\begin{aligned}
\mathrm{P}(|\bar{Y}_b - \mu(\tilde{F}^{(b)})| > \delta) &= \mathrm{E}_{\tilde{F}^{(b)}}\left[\mathrm{P}(|\bar{Y}_b - \mu(\tilde{F}^{(b)})| > \delta | \tilde{F}^{(b)})\right] \\
&= \mathrm{E}_{\tilde{F}^{(b)}}\left[2\left(1 - \Phi\left(\tfrac{\sqrt{n_b}\delta}{\sigma_\varepsilon(\tilde{F}^{(b)})}\right)\right)\right] \le 2\left[1 - \Phi\left(\tfrac{\sqrt{n_b}\delta}{\sqrt{C}}\right)\right] \to 0 \text{ as } n_b \to \infty
\end{aligned}$$

where, the third step follows by applying Condition (1). Therefore, $|\bar{Y}_b - \mu(\tilde{F}^{(b)})| \xrightarrow{p} 0$ as $n_b \to \infty$. Similarly, since $\tilde{\mu}_b|\mathbf{Y}_b, \tilde{F}^{(b)} \sim \mathrm{N}(\bar{Y}_b, S_b^2/n_b)$, for any $\delta > 0$,

$$\begin{aligned}
\mathrm{P}(|\tilde{\mu}_b - \bar{Y}_b| > \delta) &= \mathrm{E}_{\tilde{F}^{(b)}}\mathrm{E}_{\mathbf{Y}_b|\tilde{F}^{(b)}}\left[\mathrm{P}\left(\frac{|\tilde{\mu}_b - \bar{Y}_b|}{S_b/\sqrt{n_b}} > \frac{\delta\sqrt{n_b}}{S_b}\middle| \mathbf{Y}_b, \tilde{F}^{(b)}\right)\right] \\
&\le \mathrm{E}_{\tilde{F}^{(b)}}\mathrm{E}_{\mathbf{Y}_b|\tilde{F}^{(b)}}\left[2\mathrm{P}\left(T_{n_b-1} > \tfrac{\delta\sqrt{n_b}}{\sqrt{2C}}\middle| \mathbf{Y}_b, \tilde{F}^{(b)}\right)\right] \to 0 \text{ as } n_b \to \infty
\end{aligned}$$

where, T_{n_b-1} denotes a random variable having t-distribution with degrees of freedom $n_b - 1$, and the second step follows by applying $S_b \xrightarrow{a.s.} \sigma(\tilde{F}^{(b)})$ as $n_b \to \infty$, Condition (1) and the dominated convergence theorem. Therefore, we have $|\tilde{\mu}_b - \bar{Y}_b| \xrightarrow{p} 0$ as $n_b \to \infty$. Therefore, by triangular inequality,

$$|\tilde{\mu}_b - \mu(F^c)| \le |\tilde{\mu}_b - \bar{Y}_b| + |\bar{Y}_b - \mu(\tilde{F}^{(b)})| + |\mu(\tilde{F}^{(b)}) - \mu(F^c)|,$$

which implies that $|\tilde{\mu}_b - \mu^c| \xrightarrow{p} 0$ as $m \to \infty$ and $n_b \to \infty$. Since B is finite, this convergence holds uniformly over all $b \in \{1, 2, \ldots, B\}$ if $m \to \infty$ and $n_{\min} \to \infty$.

Finally we show that uniformly for all $b \in \{1, 2, \ldots, B\}$, the bth order statistic $\tilde{\mu}_{(b)}$ converges in probability to μ^c. For any $\delta > 0$ and any $b \in \{1, 2, \ldots, B\}$,

$$\begin{aligned}
P(|\tilde{\mu}_{(b)} - \mu^c| \ge \delta) &\le P(\tilde{\mu}_{(b)} \ge \mu^c + \delta) + P(\tilde{\mu}_{(b)} \le \mu^c - \delta) \\
&\le P(\tilde{\mu}_{(B)} \ge \mu^c + \delta) + P(\tilde{\mu}_{(1)} \le \mu^c - \delta) \\
&\le \sum_{i=1}^{B} P(\tilde{\mu}_i \ge \mu^c + \delta) + \sum_{i=1}^{B} P(\tilde{\mu}_i \le \mu^c - \delta) \to 0 \text{ as } m \to \infty, n_{\min} \to \infty.
\end{aligned}$$

Thus, for all $b \in \{1, 2, \ldots, B\}$, $|\tilde{\mu}_{(b)} - \mu^c| \xrightarrow{p} 0$ as $m \to \infty$ and $n_{\min} \to \infty$. Specifically, both $\tilde{\mu}_{(\lceil B\alpha/2 \rceil)}$ and $\tilde{\mu}_{(\lceil B(1-\alpha/2) \rceil)}$ converge in probability to μ^c. Therefore, the CrI $[\tilde{\mu}_{(\lceil B\alpha/2 \rceil)}, \tilde{\mu}_{(\lceil B(1-\alpha/2) \rceil)}]$ shrinks to $\mu^c = \mu(F^c)$ in probability as $m \to \infty$ and $n_{\min} \to \infty$. □

4 EMPIRICAL STUDY

In this section, we use a simple example to illustrate the finite sample performance of our Bayesian framework. A logistic company wants to estimate its utility cost, e.g., electricity, for next winter. The

company has $d/2$ warehouses at different sites located in the northeast part of the US. Suppose $d/2$ is an integer. The electricity is served by different companies. Let C_i and D_i denote the unit cost of the electricity and the amount of electricity demand at site i with $i = 1,2,\ldots,d/2$. Both C_i's and D_i's are affected by some underlying common factors. For example, C_i's are affected by the price of the fossil fuels and D_i's are affected by the climate. When the electricity demand becomes high, the electricity suppliers tend to increase the unit price to encourage energy saving. Therefore, C_i's and D_i's are also dependent. The cost function is $Y = \sum_{i=1}^{d/2} C_i D_i$. We are interested in the expected cost $\mathrm{E}(Y)$.

Let $\mathbf{X} \equiv (C_1, C_2, \ldots, C_{d/2}, D_1, D_2, \ldots, D_{d/2})^\top$ and $\mathbf{X} \sim F^c$. Suppose F^c has a Gaussian copula representation specified by $(F_1^c, F_2^c, \ldots, F_d^c, \mathbf{C}^c)$. The marginal distribution F_j^c is $\exp(2)$ for $j = 1,2,\ldots,d$. The correlation matrix $\mathbf{C}^c$ has $C^c_{j'j} = 0.6$ for $j' \neq j$ and $j', j = 1,2,\ldots,d$. Note that this correlation matrix can be decomposed into a factor structure as in Section 3.1 with the number of factors equal to 1. In the experiments we assume that all marginals and the correlation matrix are unknown. They are estimated from m real-world data $\mathscr{X}_m^{(0)} = (\mathbf{X}_1^{(0)}, \mathbf{X}_2^{(0)}, \ldots, \mathbf{X}_m^{(0)})^\top$ with $\mathbf{X}_i^{(0)} \overset{i.i.d.}{\sim} F^c$ for $i = 1,2,\ldots,m$.

Given finite real-world data, we compare the performance of Gaussian copula with and without the factor model on $\mathbf{C}^c$. Specifically, for the Gaussian copula without factor model, the Bayesian inference on the correlation matrix $\mathbf{C}$ is obtained by using the extended rank likelihood (Hoff 2007). We specify the prior on $\mathbf{C}$ to be an inverse-Wishart $(d+2, \mathbf{I}_{d\times d})$. For Gaussian copula factor model in Equation (3), we use the generalized double pareto prior $\mathrm{GDP}(\alpha, \beta)$ with $\alpha = 3$ and $\beta = 1$ on loading parameter λ_{ij} (Murray et al. 2013). This prior tends to shrink λ_{ij} to zero if its absolute value is very small and keep the large values of λ_{ij} unchanged.

To quantify the input uncertainty characterized by $\mathrm{P}(\mathbf{C}|\mathscr{X}_m^{(0)})$, we use the Gibbs sampling algorithm to generate samples of $\tilde{\mathbf{C}}$ for Gaussian copula with and without factor model. In each Markov Chain Monte Carlo (MCMC) simulation, the warmup contains 5000 iterations. Then we continue the chain for another 10^5 iterations, and save one sample of $\tilde{\mathbf{C}}$ every 100 iterations, which results in 1000 sampled $\tilde{\mathbf{C}}$ from the chain. For the Gaussian copula factor model, a parameter-expanded approach is used to improve the MCMC mixing rate for Gaussian copula factor model; see Murray et al. (2013) and Liu and Wu (1999).

The mean and standard deviation of relative error of correlation matrix, defined as $\mathrm{err} = \|\tilde{\mathbf{C}} - \mathbf{C}^c\|/\|\mathbf{C}^c\|$, are used to measure the estimation efficiency of Gaussian copula with and without factor model. Here $\|\cdot\|$ denotes the Frobenius norm. Specifically, to estimate $\mathrm{E}[\mathrm{err}]$, we use 10 macro-replications and generate real-world data $\mathscr{X}_m^{(0)}$ in each macro-replication. Conditional on $\mathscr{X}_m^{(0)}$, we draw 1000 Markov chain samples of $\tilde{\mathbf{C}}$ from the posterior distribution $\mathrm{P}(\mathbf{C}|\mathscr{X}_m^{(0)})$ and calculate the relative error, denoted by err_{ij}, where i=1,2,...,10 and $j = 1,2,\ldots,1000$. We estimate the mean $\mathrm{E}[\mathrm{err}]$ and the standard deviation $\mathrm{SD}[\mathrm{err}]$ by

$$\begin{aligned}
\widehat{\mathrm{E}}[\mathrm{err}] &= \frac{1}{10}\sum_{i=1}^{10}\left[\frac{1}{1000}\sum_{j=1}^{1000}\mathrm{err}_{ij}\right] \\
\widehat{\mathrm{SD}}[\mathrm{err}] &= \frac{1}{10}\sum_{i=1}^{10}\frac{\mathrm{SD}[\mathrm{err}_i]}{\sqrt{999}}
\end{aligned}$$

where, $\mathrm{SD}[\mathrm{err}_i]$ represents the sample deviation for the samples from the ith chain.

When $d = 10, 20, 30$, the correlation matrix estimation errors are shown in Figure 1 and Table 1. Let err_F and err_G denote the estimation errors from Gaussian copula with and without factor model. In Figure 1, the horizontal axis is the size of real-world data m and the vertical axis gives the relative estimation errors of the correlation matrix. The dashed lines represent the results obtained by the usual Gaussian copula model without factor structure. The solid lines represent the results from Gaussian copula factor model. In general, as the sample size m increases, the estimation errors decrease. In all cases, the Gaussian copula model with factors has lower estimation errors on average than the model without factors, which indicates that the correlation matrix can be estimated more accurately by the factor model. This advantage becomes more obvious when the dimension d becomes large. When $d = 30$ and $m = 30$, the Gaussian copula model

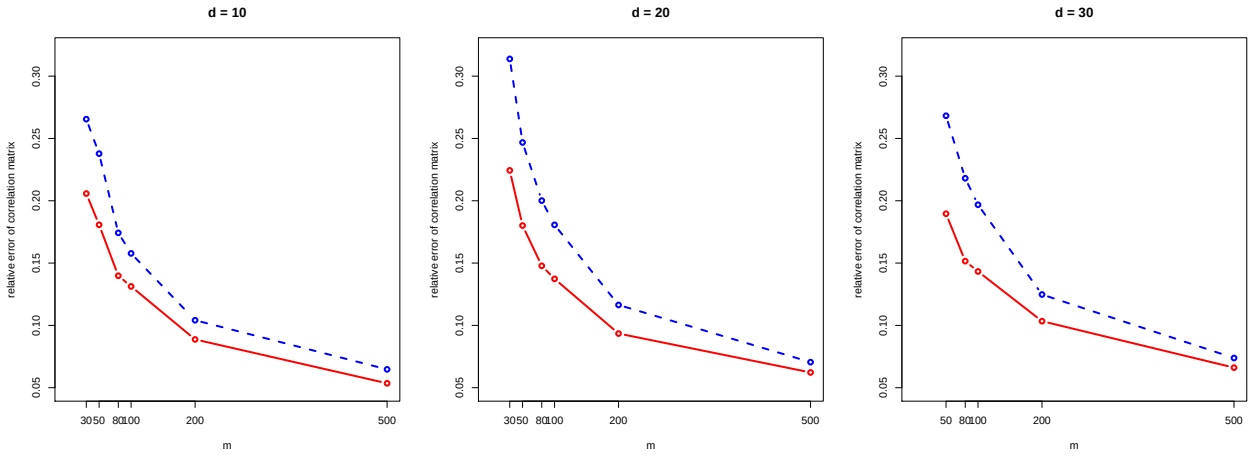

Figure 1: Relative error $\|\tilde{\mathbf{C}} - \mathbf{C}^c\|/\|\mathbf{C}^c\|$ obtained by Gaussian copula with and without factor model when $d = 10, 20, 30$. Solid lines give results from Gaussian copula factor model and dashed lines give results of Gaussian copula without factor model.

Table 1: The relative error for Gaussian copula with and without factor model when $d = 10, 20, 30$.

$d = 10$	$m = 30$	$m = 50$	$m = 80$	$m = 100$	$m = 200$	$m = 500$
$\widehat{\mathrm{E}}[\mathrm{err}_F]$	0.206	0.181	0.14	0.131	0.089	0.054
$\widehat{\mathrm{E}}[\mathrm{err}_G]$	0.266	0.238	0.174	0.158	0.104	0.065
$\widehat{\mathrm{E}}[\mathrm{err}_G] - \widehat{\mathrm{E}}[\mathrm{err}_F]$	0.06	0.057	0.034	0.027	0.015	0.011
$\widehat{\mathrm{SD}}[\mathrm{err}_F]$	0.002	0.0019	0.0015	0.0014	0.001	0.0007
$\widehat{\mathrm{SD}}[\mathrm{err}_G]$	0.002	0.0019	0.0013	0.0012	0.0009	0.0006
$d = 20$	$m = 30$	$m = 50$	$m = 80$	$m = 100$	$m = 200$	$m = 500$
$\widehat{\mathrm{E}}[\mathrm{err}_F]$	0.224	0.18	0.148	0.137	0.093	0.062
$\widehat{\mathrm{E}}[\mathrm{err}_G]$	0.314	0.247	0.2	0.181	0.116	0.071
$\widehat{\mathrm{E}}[\mathrm{err}_G] - \widehat{\mathrm{E}}[\mathrm{err}_F]$	0.089	0.067	0.052	0.043	0.023	0.008
$\widehat{\mathrm{SD}}[\mathrm{err}_F]$	0.002	0.0016	0.0016	0.0015	0.0004	0.001
$\widehat{\mathrm{SD}}[\mathrm{err}_G]$	0.002	0.002	0.0014	0.0012	0.0004	0.0006
$d = 30$	$m = 30$	$m = 50$	$m = 80$	$m = 100$	$m = 200$	$m = 500$
$\widehat{\mathrm{E}}[\mathrm{err}_F]$	0.245	0.19	0.15	0.143	0.103	0.066
$\widehat{\mathrm{E}}[\mathrm{err}_G]$	—	0.268	0.218	0.197	0.125	0.07
$\widehat{\mathrm{E}}[\mathrm{err}_G] - \widehat{\mathrm{E}}[\mathrm{err}_F]$	—	0.079	0.0665	0.054	0.021	0.008
$\widehat{\mathrm{SD}}[\mathrm{err}_F]$	0.001	0.0015	0.0015	0.0014	0.001	0.001
$\widehat{\mathrm{SD}}[\mathrm{err}_G]$	—	0.0018	0.0015	0.0013	0.0009	0.0006

without factors fails to produce results because $\tilde{\mathbf{C}}$ drawn from its posterior is singular, while the factor model still works even for cases with larger d since it has fewer number of parameters and does not have the singularity problem.

For $d = 20$ and $d = 30$, we further build a CrI to quantify both input and simulation estimation uncertainty with results shown in Table 2. These results are obtained based on 500 macro-replications. Each macro-replication uses m real-world data $\mathscr{X}_m^{(0)}$ and $B = 1000$ Markov chain samples of correlation matrices $\tilde{\mathbf{C}}$

from $\mathrm{P}(\mathbf{C}|\mathbf{X}_m^{(0)})$ to quantify the input uncertainty. Given the empirical marginals $\widehat{F}_j$ with $j = 1,2,\ldots,d$ and the sampled $\tilde{\mathbf{C}}$, we generate $\mathbf{X}$ by Equation (2) to drive the simulation. To control the simulation estimation error, we set the number of replications at each sampled input distribution to be $n = 10^5$. The results of CrI accounting for both input and simulation errors are shown in Table 2, where $q_{\alpha/2}$ and $q_{1-\alpha/2}$ denote the percentile bounds for the credible intervals obtained from either Gaussian copula (GC) or Gaussian copula with factor model (GCF). We also calculate the means and standard deviations of the width of CrIs, denoted by |CrI|, based on the results from 500 macro-replications. The marginal distributions typically have a large impact on the system performance estimate. Since we take the empirical marginals as the true distributions, there is no clear trend in the results.

Table 2: The credible intervals of system mean response with $\alpha = 0.1$.

$d = 20$		$m = 30$	$m = 50$	$m = 100$
GCF	$[q_{\alpha/2}, q_{1-\alpha/2}]$	[54.27, 60.81]	[55.76,62.54]	[57.20, 63.44]
	|CrI| mean	6.55	6.77	6.23
	|CrI| sd	0.086	0.128	0.052
GC	$[q_{\alpha/2}, q_{1-\alpha/2}]$	[52.88,61.73]	[55.09,62.07]	[57.25, 63.32]
	|CrI| mean	8.19	6.98	5.07
	|CrI| sd	0.113	0.132	0.038
$d = 30$		$m = 30$	$m = 50$	$m = 100$
GCF	$[q_{\alpha/2}, q_{1-\alpha/2}]$	[81.06, 89.78]	[82.71,91.68]	[84.88,90.09]
	|CrI| mean	8.71	8.97	9.20
	|CrI| sd	0.104	0.091	0.075
GC	$[q_{\alpha/2}, q_{1-\alpha/2}]$	—	[81.05,90.54]	[84.81,92.15]
	|CrI| mean	—	9.49	7.34
	|CrI| sd	—	0.101	0.056

5 CONCLUSION

We explore the Gaussian copula factor model for the input distribution with dependence, when the dependence between input components is caused by some underlying common factors. A unified Bayesian framework is proposed to quantify both the input and simulation estimation uncertainty by the posterior distributions with theoretical guarantee. Our empirical study shows that the Gaussian copula factor model provides more efficient estimation of the dependence structure and outperforms the usual Gaussian copula model in finite sample, which could be further used to reduce the overall uncertainty of system performance estimate.

REFERENCES

Akcay, A., and B. Biller. 2014. "Quantifying Input Uncertainty in an Assemble-to-order system Simulation with Correlated Input Variables of Mixed Types". In *Proceedings of the 2014 Winter Simulation Conference*: IEEE Computer Society, Washington, DC.

Biller, B., and C. G. Corlu. 2011. "Accounting for Parameter Uncertainty in Large-Scale Stochastic Simulations with Correlated Inputs". *Operations Research* 59:661–673.

Biller, B., and S. Ghosh. 2006. "Multivariate Input Processes". In *Handbooks in Operations Research and Management Science: Simulation*, edited by S. Henderson and B. L. Nelson, Chapter 5. Elsevier.

Cario, M., and B. L. Nelson. 1997. "Modeling and Generating Random Vectors with Arbitrary Marginal Distributions and Correlation Matrix". Technical report, Department of Industrial Engineering and Management Sciences, Northwestern University.

DeGroot, M. H. 1970. *Optimal Statistical Decisions*. McGraw-Hill, Inc.

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

Ghosh, S., and S. Henderson. 2002. "Properties of the NORTA Method in Higher Dimensions". In *Proceedings of the 2002 Winter Simulation Conference*, edited by E. Yűcesan, C. H. Chen, J. L. Snowdon, and J. M. Charnes, 263–269: IEEE Computer Society, Washington, DC.

Hoff, P. D. 2007. "Extending the Rank Likelihood for Semiparametric Copula Estimation". *The Annals of Applied Statistics* 1:265–283.

Joe, H. 2005. "Asymptotic Efficiency of the Two-stage Estimation Method for Copula-based Models". *Journal of Multivariate Analysis* 94:401419.

Liu, J. S., and Y. N. Wu. 1999. "Parameter Expansion for Data Augmentation". *Journal of the American Statistical Association* 94:1264.

Murray, J. S., D. B. Dunson, L. Carin, and J. E. Lucas. 2013. "Bayesian Gaussian Copula Factor Models for Mixed Data". *Journal of the American Statistical Association* 108:656–665.

Severini, T. A. 2000. *Likelihood Methods in Statistic*. 22. Oxford Statistical Science Series.

Smith, M. S. 2011. "Bayesian Approaches to Copula Modeling". Technical report, University of Melbourne.

Van Der Vaart, A. W. 1998. *Asymptotic Statistics*. Cambridge, UK: Cambridge University Press.

Xie, W., B. L. Nelson, and R. R. Barton. 2014a. "A Bayesian Framework for Quantifying Uncertainty in Stochastic Simulation". *Operations Research* 62 (6): 1439–1452.

Xie, W., B. L. Nelson, and R. R. Barton. 2014b. "Statistical Uncertainty Analysis for Stochastic Simulation with Dependent Input Models". In *Proceedings of the 2014 Winter Simulation Conference*: IEEE Computer Society, Washington, DC.

Yi, Y., W. Xie, and E. Zhou. 2015. "A Sequential Experiment Design for Input Uncertainty Quantification in Stochastic Simulation". In *Proceedings of the 2015 Winter Simulation Conference*: IEEE Computer Society, Washington, DC.

AUTHOR BIOGRAPHIES

WEI XIE is an assistant professor in the Department of Industrial and Systems Engineering at Rensselaer Polytechnic Institute. Her research interests are in computer simulation, risk management and data analytics. Her email address is xiew3@rpi.edu and her web page is http://homepages.rpi.edu/~xiew3/.

CHENG LI is now a postdoctoral associate at Department of Statistical Science, Duke University. His research interests are in Bayesian modeling and theory for massive datasets, Bayesian nonparametric methods, and Bayesian model selection. His email address is cl332@stat.duke.edu.

HONGTAN SUN is a PhD student in the Department of Industrial and Systems Engineering at Rensselaer Polytechnic Institute. Her research centers on the simulation input modeling and uncertainty quantification. Her email address is sunh6@rpi.edu.