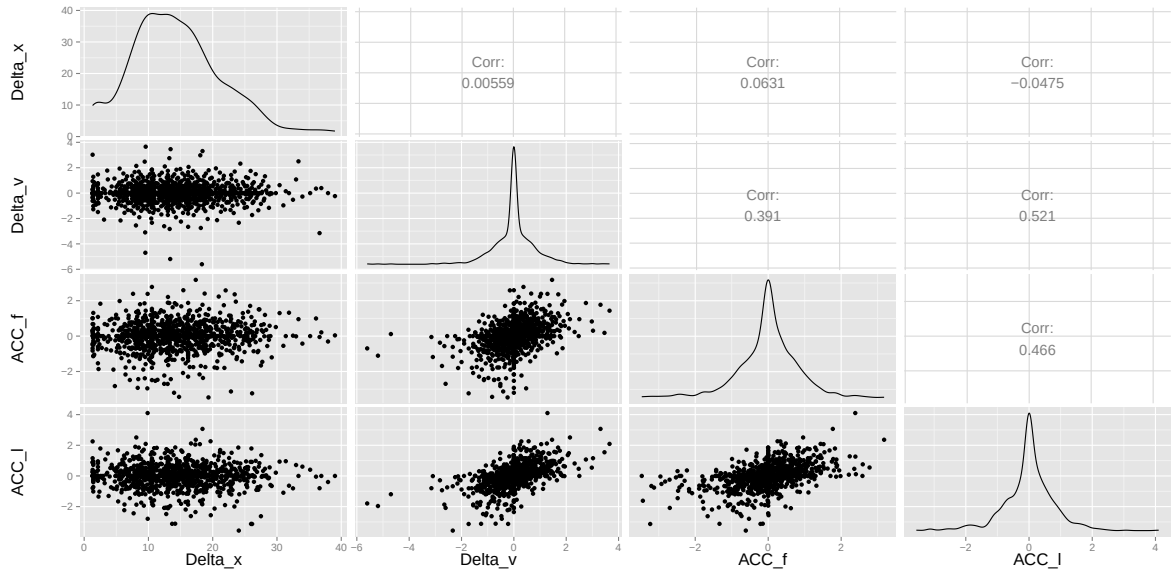


Sampling dependent parameters in traffic simulation models with Gaussian copula

Qiao Ge

Monica Menendez

ETH Zurich IVT

April 2015

STRC

15th Swiss Transport Research Conference

Monte Verità / Ascona, April 15 – 17, 2015

ETH Zurich IVT

Sampling dependent parameters in traffic simulation models with Gaussian copula

Qiao Ge
ETH Zurich
Institute for Transport Planning and Systems
HIL F 37.3, Stefano-Franscini-Platz 15
8093 Zurich, Switzerland
phone: +41-44-633 32 49
fax: +41-44-633 10 57
qiao.ge@ivt.baug.ethz.ch

Monica Menendez
ETH Zurich
Institute for Transport Planning and Systems
HIL F 37.2, Stefano-Franscini-Platz 15
8093 Zurich, Switzerland
phone: +41-44-633 66 95
fax: +41-44-633 10 57
monica.menendez@ivt.baug.ethz.ch

April 2015

Abstract

Making samples with certain marginal distributions and dependence structures is an essential but difficult step to perform sampling-based SA for traffic simulation models with dependent parameters. In this paper, we present a general approach for generating samples for dependent parameters. It utilizes the Gaussian copula in the sampling process, which makes it attractive for sampling parameters from any arbitrary marginal distribution. Furthermore, the Spearman's rank correlation coefficient is employed instead of the traditional linear correlation coefficient, so that the dependence structure of the empirical data can be retained throughout the non-linear transform of the Gaussian copula.

A case study that generates samples for the kinematic parameters of Wiedemann-74 car-following model is included to demonstrate the application of this approach. It has shown that the marginal distributions and correlation coefficients of the generated samples are comparable with that of the empirical data. Specifically, the 1,024 samples, which are generated by employing the Sobol sequence in the sampling process, also present consistent marginal distributions and correlation coefficients as the empirical data. This has demonstrated that the proposed sampling approach is also useful for making proper samples of computationally expensive models, for which a big number of model runs are not always affordable.

Keywords

Data Sampling, Dependent Parameters, Gaussian Copula, Traffic Model

1 Introduction

Traffic simulation has become a major resource in the field of traffic engineering. Along with the development of computational techniques, microscopic traffic simulation models are more advanced and realistic nowadays. On the other hand, the complexity of the model also significantly increases, especially due to the fact that there are more and more parameters contained in the model. To help model users to better understand the model, and manage the uncertainties in the simulation result, it is necessary to investigate the relationship between the model inputs (i.e., the parameters of the model) and outputs (i.e., the simulation results), especially when the model itself behaves like a blackbox. One important and widely used tool for such task is Sensitivity Analysis (SA).

SA studies the relationship between the inputs and outputs of a model. Many SA applications in traffic simulation models have shown that it is capable to provide both qualitative and quantitative sensitivity information in an efficient way. For instance, in (Ge *et al.*, 2014c) and (Ge and Menendez, 2014), SA was used to rank the parameters of simulators VISSIM and Aimsun based on their impacts on the variation of the simulation results. However, to the authors' knowledge, most SA methods applied for microscopic traffic simulations found from literature are only suitable for models with independent parameters. When performing SA for a complex and/or computationally expensive model, practitioner tend to assume that all parameters are independent beforehand, or simply group the dependent parameters as one independent parameter (Ge *et al.*, 2014b).

Applying the SA approaches which are dedicated to independent parameters to the model with dependent parameters may actually provide wrong sensitivity information. For instance, with the same SA approach, some of the model inputs, which are considered as unimportant if they are independent inputs, can also be considered as important if they are highly correlated with the most important inputs (a simple example is given in Section 2). In addition, given the fact that there are usually many dependent parameters contained in a microscopic traffic model (e.g., the speed and the acceleration rate of a vehicle), the research of an efficient and accurate SA approach dedicated to complex models with dependent parameters is very important.

In (Kucherenko *et al.*, 2012), the authors extended the Sobol's formula (Sobol, 1993) for the model with independent parameters to the model with dependent parameters. These formulas are very helpful as they allow to quantify the model sensitivity by the first-order and total sensitivity indexes. However, these sensitivity indexes can not be analytically derived when the model is a blackbox (this is especially true for most commercial simulators such as VISSIM and Aimsun). As an alternative, we propose to utilize the sampling-based SA to approximate these sensitivity

indexes. Furthermore, to correctly generate the samples based on the distribution and correlation as the measured data, the Gaussian copula and the Spearman's rank correlation coefficient andare employed in this study (Iman and Conover, 1982; Mara and Tarantola, 2012). A case study is included to demonstrate the proposed sampling approach by making samples of the kinematic parameters of the Wiedemann-74 car-following model.

The paper is organized as follows: a brief introduction of the sampling-based SA is given in Section 2; the proposed sampling method for dependent parameters is described in Section 3; the application of the proposed sampling algorithm for dependent parameters is illustrated with a case study in Section 4; the conclusions and suggestions for future work are included in Section 5.

2 Sampling-based SA

The sampling-based SA uses Monte Carlo simulation or other approaches such as Latin Hypecube Sampling (LHS, see Saltelli *et al.* (2007)), quasi-random sampling (e.g., Sobol sequence, see Sobol (1976)) to generate random samples. Suppose a model M has n parameters, i.e., Z_1, Z_2, $\cdots$, Z_n, and m random samples (i.e., each sample is a combination of certain values of all inputs based on their distributions) are generated. These samples can be described in the following matrix:

$$\tilde{Z} = \begin{bmatrix} z_1^{(1)} & z_2^{(1)} & \cdots & z_n^{(1)} \\ z_1^{(2)} & z_2^{(2)} & \cdots & z_n^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ z_1^{(m)} & z_2^{(m)} & \cdots & z_n^{(m)} \end{bmatrix}, \tag{1}$$

where $z_i^{(d)}$ ($i \in [1, n], d \in [1, m]$) is the d-th sample of input parameter Z_i.

The model is then executed consecutively by taking values from each row of $\tilde{Z}$ as the model inputs (Helton *et al.*, 2006). The model output $Y = [y^{(1)}, y^{(2)}, \cdots, y^{(m)}]^T$ with respect to $\tilde{Z}$ is obtained accordingly:

$$Y = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{pmatrix} = \begin{pmatrix} M(z_1^{(1)}, z_2^{(1)}, \cdots, z_n^{(1)}) \\ M(z_1^{(2)}, z_2^{(2)}, \cdots, z_n^{(2)}) \\ \vdots \\ M(z_1^{(m)}, z_2^{(m)}, \cdots, z_n^{(m)}) \end{pmatrix}. \tag{2}$$

For each parameter Z_i, we can plot m points at coordinates ($z_i^{(j)}, y^{(j)}, j \in [1, m]$) in a scatter plot. The shape of the points cloud in the scatter plot represents the sensitivity of the output with respect to Z_i, and it can be visually analyzed. For example, considering a simple linear model with two independent parameters (i.e., Z_1, Z_2) that are normally distributed:

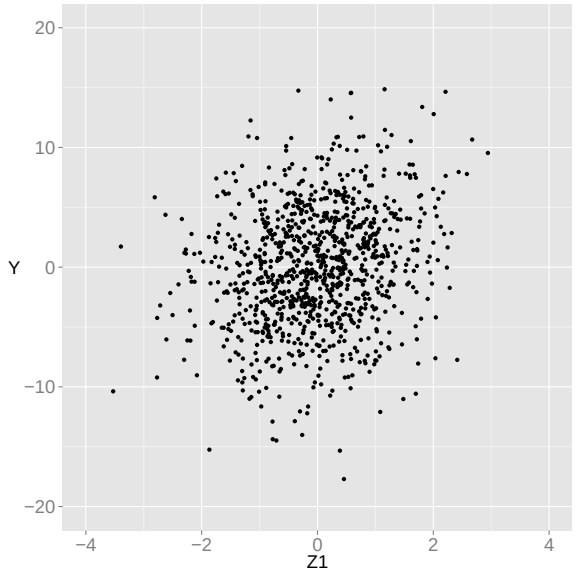
$$Y = Z_1 + Z_2, \tag{3}$$

in which $Z_1 \sim \mathcal{N}(0, 1)$, $Z_2 \sim \mathcal{N}(0, 5)$.

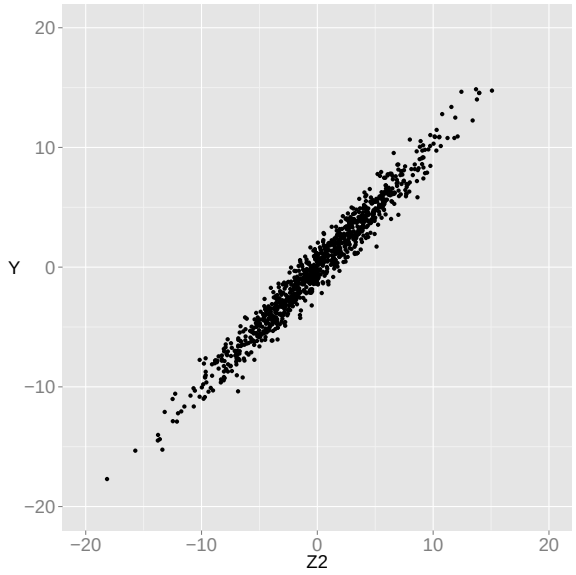
By generating a size of 1000 random samples (i.e., $m = 1000$) for Z_1 and Z_2 each using the normal distribution, we can plot the corresponding input-output (i.e., $Z_1 - Y$ and $Z_2 - Y$) in the scatter plots (Figure 1). In Fig. 1(a), the cloud of Z_1 is more or less uniformly distributed across different values of Z_1 (i.e., it looks like a circle). On the contrary, in Fig. 1(b) the cloud of Z_2 has a much wider dispersion along with different values of Z_2, and a clear linearly relationship between Z_2 and Y can be determined. It obviously shows that input Z_2 has much higher impacts on the variation of model output than input Z_1.

Figure 1: Scatter plot for a linear model with two independent inputs $Y = Z_1 + Z_2$

(a) Scatter plot of input Z_1 and the output Y



(b) Scatter plot of input Z_2 and the output Y



Furthermore, if enough samples are generated from the input space, the sampling-based SA can also be used to derive the quantitative sensitivity measures such as the first order and the total sensitivity indexes (Saltelli *et al.*, 2007; Kucherenko *et al.*, 2012) [1]. Therefore, the sampling-based SA is one of the simplest ways for performing global SA. This method is also very useful for the SA of "blackbox" models, in which the global sensitivity indexes can not be analytically derived. On the other hand, as the sampling-based SA typically requires to run the model with certain amount of random samples generated, when the SA is performed for models that have many inputs and/or are computationally expensive, it would be less attractive due to its low efficiency. In those cases, the computational cost required by the sampling-based SA is just too high, and the SA can be even infeasible.

Another issue for the sampling-based SA comes from sampling of dependent parameters. As

[1] The details about how to derive the quantitative sensitivity measures using sampling-based SA will be included in our next paper. In this paper we focus on the sampling approach.

mentioned in Section 1, many SA practices found in the literature were performed based on the assumption that the parameters of the model are independent. However, this assumption will very likely lead bias or wrong SA results when the input parameters are actually dependent. To make a simple example, let us recall the above linear model (Eq. (3)). The inputs Z_1 and Z_2 have the same marginal distribution as before, but their correlation coefficient ρ_{Z_1,Z_2} (Eq. (4)) is 0.8 this time.

$$\rho_{Z_1,Z_2} = \frac{E[(Z_1 - \mu_{Z_1})(Z_2 - \mu_{Z_2})]}{\sigma_{Z_1}\sigma_{Z_2}}, \tag{4}$$

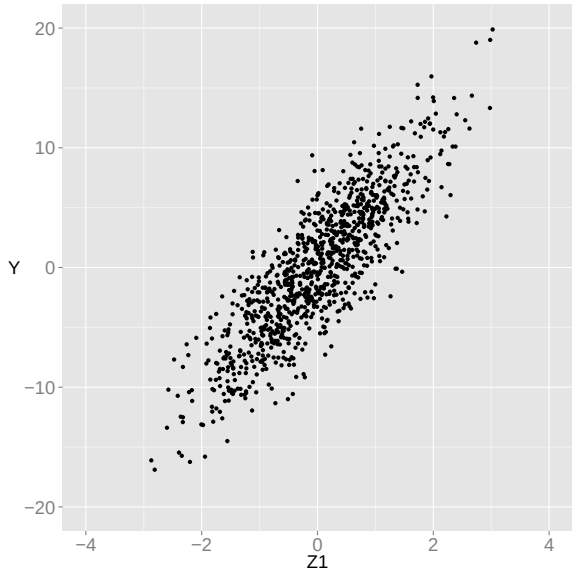
where $E[\cdot]$ stands for the expectation.

We also generate 1000 samples for both Z_1 and Z_2 using the bivariate normal distribution with the joint Probability Density Function (PDF) shown in Eq. (5). Then we plot the scatter plots for both $Z_1 - Y$ and $Z_2 - Y$ in Figure 2.

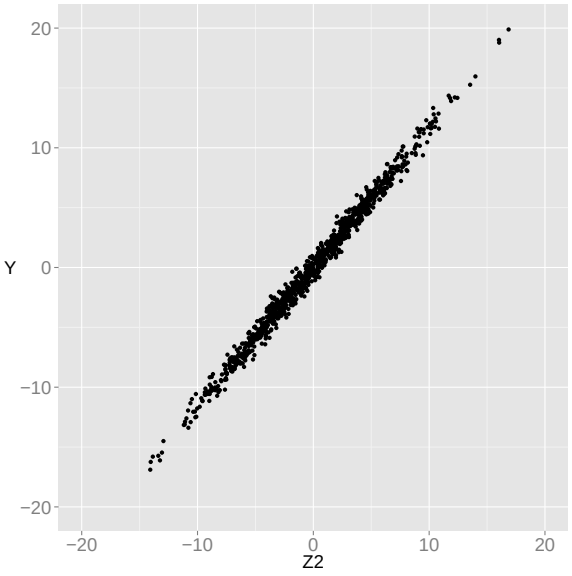
$$f(Z_1, Z_2) = \frac{1}{2\pi\sigma_{Z_1}\sigma_{Z_2}\sqrt{1-\rho^2_{Z_1,Z_2}}} e^{-\frac{1}{2(1-\rho^2_{Z_1,Z_2})}\left[\frac{(Z_1-\mu_{Z_1})^2}{\sigma^2_{Z_1}} + \frac{(Z_2-\mu_{Z_2})^2}{\sigma^2_{Z_2}} - \frac{2\rho_{Z_1,Z_2}(Z_1-\mu_{Z_1})(Z_2-\mu_{Z_2})}{\sigma_{Z_1}\sigma_{Z_2}}\right]}. \tag{5}$$

Figure 2: Scatter plot for a linear model with two dependent inputs $Y = Z_1 + Z_2$, $\rho_{Z_1,Z_2} = 0.8$

(a) Scatter plot of input Z_1 and the output Y



(b) Scatter plot of input Z_2 and the output Y



Comparing the shape of the clouds of the corresponding parameters in Fig. 1 and Fig. 2, it is

clear that the shape of the cloud of Z_1 in Fig. 2(a) is quite different with that in Fig. 1(a), while the difference between Fig. 2(b) and Fig. 1(b) is not very significant. This is because Z_1 now has a strong positive correlation with Z_2, and since Z_2 is a very influential parameter, Z_1 also becomes an important parameter in this case.

Therefore,to avoid wrong conclusions by the sampling-based SA, the interdependency among the parameters should not be ignored when generating samples. If all inputs have distributions that are from a known standard multivariate distribution, it is straightforward to use the corresponding joint PDF such as Eq. (5) for generating the samples. However, this can be very challenging for many microscopic traffic models, as the inputs are actually not following any standard distribution, and/or they are from different distributions. Therefore, a general sampling approach that can cope with any arbitrary distribution and correctly represent different the dependence structure is very important. For this purpose, we propose to use the idea of Gaussian copula to generate dependent samples, and the details are given in the next section.

3 Methodology

To draw a sample (i.e., z) of a random parameter Z with a given Cumulative Distribution Function (CDF) $F_Z(z)$ based on Monte Carlo simulation, we can first draw a sample u from the standard uniform distribution $\mathcal{U}[0,1]$, and apply the inverse CDF to u as $z = F_Z^{-1}(u)$, in which $F_Z^{-1}(\cdot)$ is the inverse CDF of Z. Similarly, for the model M with n independent parameters $\{Z_1, \cdots, Z_n\}$ and corresponding marginal CDFs $\{F_{Z_1}(z_1), \cdots, F_{Z_n}(z_n)\}$, the model output Y with respect to the sample matrix $\tilde{Z}$ (see Eq. (2)) can be derived as:

$$Y = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{pmatrix} = \begin{pmatrix} M(z_1^{(1)}, z_2^{(1)}, \cdots, z_n^{(1)}) \\ M(z_1^{(2)}, z_2^{(2)}, \cdots, z_n^{(2)}) \\ \vdots \\ M(z_1^{(m)}, z_2^{(m)}, \cdots, z_n^{(m)}) \end{pmatrix} = \begin{pmatrix} M\left(F_{Z_1}^{-1}(u_1^{(1)}), \cdots, F_{Z_n}^{-1}(u_n^{(1)})\right) \\ M\left(F_{Z_1}^{-1}(u_1^{(2)}), \cdots, F_{Z_n}^{-1}(u_n^{(2)})\right) \\ \vdots \\ M\left(F_{Z_1}^{-1}(u_1^{(m)}), \cdots, F_{Z_n}^{-1}(u_n^{(m)})\right) \end{pmatrix}. \tag{6}$$

where $u_i^{(d)} (i \in [1,n], d \in [1,m])$ is the d-th sample drawn from the uniform distribution $\mathcal{U}[0,1]$ for parameter Z_i, and $F_{Z_i}^{-1}(\cdot)$ stands for the inverse CDF of parameter Z_i.

In the above case, since $Z_1, \cdots, Z_n$ are independent parameters, the joint CDF of $\{Z_1, \cdots, Z_n\}$ is just the product function of the marginal CDFs of all parameters:

$$\begin{aligned} F_{Z_1,\cdots,Z_n}(z_1, \cdots, z_n) &= \mathbb{P}\Big(\{Z_1 \le z_1\} \cap \cdots \cap \{Z_n \le z_n\} \Big) \\ &= \mathbb{P}(Z_1 \le z_1) \times \cdots \times \mathbb{P}(Z_n \le z_n) \\ &= \prod_{i=1}^{n} F_{Z_i}(z_i) \end{aligned}, \tag{7}$$

where $\mathbb{P}(\cdot)$ stands for the probability.

In the case of dependent parameters, the joint CDF of $\{Z_1, \cdots, Z_n\}$ can also be written as a function (obviously, it is not a simple product function here due to the correlation of the parameters) of the marginal CDFs:

$$\begin{aligned} F_{Z_1,\cdots,Z_n}(z_1, \cdots, z_n) &= C(F_{Z_1}(z_1), \cdots, F_{Z_n}(z_n)) \\ &= C(u_1, \cdots, u_n), \end{aligned} \tag{8}$$

where $u_i = F_{Z_i}(z_i), u_i \sim \mathcal{U}[0,1]$ for $\forall i \in [1,n]$.

The function $C(\cdot)$ is known as the copula (for details see Nelsen (1999)). It is defined as the joint CDF of random variables $\{U_1, \cdots, U_n\}$ that have uniform marginal distributions in $\mathcal{U}[0, 1]$:

$$C(u_1, u_2, \ldots, u_n) = \mathbb{P}\Big(\{U_1 \leq u_1\} \cap \cdots \cap \{U_n \leq u_n\} \Big). \tag{9}$$

Copula has been quite popular in the fields such as risk management, quantitative finance, civil engineering recently. It can be used to generate multivariate distributions for modelling the dependence structure of correlated multivariate data. There are many different types of copulas, and one commonly used copula for sampling dependent parameters is the Gaussian copula:

$$C_n^{\text{Gauss}}(u_1, \ldots, u_n) = \Phi_n \left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n) \right) = \Phi_n(x_1, \ldots, x_n), \tag{10}$$

where $\Phi^{-1}(\cdot)$ is the inverse CDF of the univariate standard normal distribution, $\Phi_n(\cdot)$ is the joint CDF of the multivariate standard normal distribution for random variables $\{X_1, \cdots, X_n\}$:

$$\begin{aligned} \Phi_n(x_1, \ldots, x_n) &= \mathbb{P}\Big(\{X_1 \leq x_1\} \cap \cdots \cap \{X_n \leq x_n\} \Big) \\ &= \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_2} \int_{-\infty}^{x_1} \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}} \mathrm{d}x_1 \mathrm{d}x_2 \cdots \mathrm{d}x_n, \\ \mathbf{x} &= (x_1, \ldots, x_n). \end{aligned} \tag{11}$$

$\boldsymbol{\Sigma}$ is the covariance matrix for $\{X_1, \cdots, X_n\}$, i.e., $x_i \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ for $\forall i \in [1, n]$:

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho_{X_1,X_2} & \cdots & \rho_{X_1,X_n} \\ \rho_{X_2,X_1} & 1 & \cdots & \rho_{X_2,X_n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{X_n,X_1} & \rho_{X_n,X_2} & \cdots & 1 \end{bmatrix}, \tag{12}$$

and $|\boldsymbol{\Sigma}|$ is the determinant of matrix $\boldsymbol{\Sigma}$. The correlation coefficient ρ_{X_i,X_j} for $i \neq j$ can be calculated using Eq. (4).

If combining Eq. (8) and Eq. (10), we can get:

$$\begin{aligned} F_{Z_1,\cdots,Z_n}(z_1,\cdots,z_n) &= C_n^{\text{Gauss}}(F_{Z_1}(z_1),\cdots,F_{Z_n}(z_n)) \\ &= \Phi_n\left(\Phi^{-1}(F_{Z_1}(z_1)),\ldots,\Phi^{-1}(F_{Z_n}(z_n))\right) \\ &= \Phi_n(x_1,\ldots,x_n). \end{aligned} \tag{13}$$

Accordingly, $x_i = \Phi^{-1}(u_i) = \Phi^{-1}(F_{Z_i}(z_i))$ for $\forall i \in [1,n]$. Then by applying the inverse transform, the following equation can be obtained:

$$z_i = F_{Z_i}^{-1}\left(\Phi(x_i)\right). \tag{14}$$

Hence, the model parameters $\{Z_1,\cdots,Z_n\}$ with arbitrary marginal CDFs $\{F_{Z_1}(z_1),\cdots,F_{Z_n}(z_n)\}$ can be represented by $\{X_1,\cdots,X_n\}$ with n-variate standard normal distribution $\mathcal{N}_n(0,\boldsymbol{\Sigma})$. However, it should be noted that in most cases, the transform from X_i to Z_i is not linear, hence the linear correlation coefficient ρ (Eq. (4)) will tend to be different for X_i and Z_i, i.e., $\rho_{X_i,X_j} \neq \rho_{Z_i,Z_j}$. To solve this problem, the rank correlation coefficient is considered. In this study, we use the Spearman's correlation coefficient:

$$\rho^s_{Z_i,Z_j} = 1 - \frac{6\sum_{d=1}^{m}\left(r_{z_i^{(d)}} - r_{z_j^{(d)}}\right)^2}{m(m^2-1)}, \tag{15}$$

where $r_{z_i^{(d)}}$ stands for the rank (in the ascending order) of the d-th ($d \in [1,m]$) sample of input parameter Z_i. For instance, if Z_1 has three samples $[z_1^{(1)}, z_1^{(2)}, z_1^{(3)}]^{\mathrm{T}} = [3.5, 2.1, 4.8]^{\mathrm{T}}$, then the corresponding rank vectors are $[r_{z_1^{(1)}}, r_{z_1^{(2)}}, r_{z_1^{(3)}}]^{\mathrm{T}} = [2,1,3]^{\mathrm{T}}$.

The benefit of the rank correlation coefficient is its invariance through the monotonic transform, regardless of the linearity of the transform. In the above Gaussian copula, as $F_{Z_i}^{-1}(\cdot)$ and $\Phi(\cdot)$ are both monotonic functions, the rank vector of X_i should be the same as the rank vector of Z_i. Therefore, the rank correlation coefficients are also the same, i.e., $\rho^s_{X_i,X_j} = \rho^s_{Z_i,Z_j}$. Specifically, with the Gaussian copula, the linear correlation coefficient and Spearman's rank correlation coefficient have the following relationship (Hotelling and Pabst, 1936):

$$\rho_{X_i,X_j} = 2\sin\left(\frac{\pi}{6}\rho^s_{X_i,X_j}\right). \tag{16}$$

The proposed algorithm for generating samples of dependent parameters $\{Z_1, \cdots, Z_n\}$ with arbitrary marginal CDFs $\{F_{Z_1}(z_1), \cdots, F_{Z_n}(z_n)\}$ is developed based on the approach in (Iman and Conover, 1982). The details are described below.

Step 1 Generate m samples for each of the n random parameters (i.e., $V_1, \cdots, V_n$) from the uniform distribution $\mathcal{U}(0, 1)$. The samples are presented as $\tilde{V}$. The samples of V_i, i.e., the i-th column vector $\tilde{V}_i$, can be produced by using pseudo-random number generators or low discrepancy series (e.g., Sobol' sequence (Sobol, 1976)). As a results, $\tilde{V}_1, \cdots, \tilde{V}_n$ will be independent with each other.

$$\tilde{V} = \begin{bmatrix} v_1^{(1)} & v_2^{(1)} & \cdots & v_n^{(1)} \\ v_1^{(2)} & v_2^{(2)} & \cdots & v_n^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ v_1^{(m)} & v_2^{(m)} & \cdots & v_n^{(m)} \end{bmatrix} \tag{17}$$

Step 2 Apply the inverse CDF of the standard normal distribution, i.e., $\Phi^{-1}(\cdot)$, to every element in $\tilde{V}$. The resulting matrix is $\tilde{X}$. Obviously, any column vector $\tilde{X}_i$ has a standard normal distribution, and it is independent with the other column vectors.

$$\tilde{X} = \begin{bmatrix} \Phi^{-1}(v_1^{(1)}) & \Phi^{-1}(v_2^{(1)}) & \cdots & \Phi^{-1}(v_n^{(1)}) \\ \Phi^{-1}(v_1^{(2)}) & \Phi^{-1}(v_2^{(2)}) & \cdots & \Phi^{-1}(v_n^{(2)}) \\ \vdots & \vdots & \ddots & \vdots \\ \Phi^{-1}(v_1^{(m)}) & \Phi^{-1}(v_2^{(m)}) & \cdots & \Phi^{-1}(v_n^{(m)}) \end{bmatrix} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_n^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_n^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(m)} & x_2^{(m)} & \cdots & x_n^{(m)} \end{bmatrix} \tag{18}$$

Step 3 Compute the Spearman's rank correlation coefficient $\rho^s_{Z_i,Z_j}$ for all pairs of $\{Z_i, Z_j\}$ from the empirical data using Eq. (15). Due to the symmetry of the covariance matrix, we only need to compute the correlation coefficients for the cases when $1 \le i < j \le n$. Since $\rho^s_{X_i,X_j} = \rho^s_{Z_i,Z_j}$, we can compute the covariance matrix, i.e., $\mathbf{\Sigma}$ (see Eq. (12)), based on Eq. (16):

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & 2\sin\left(\frac{\pi}{6}\rho^s_{Z_1,Z_2}\right) & \cdots & 2\sin\left(\frac{\pi}{6}\rho^s_{Z_1,Z_n}\right) \\ 2\sin\left(\frac{\pi}{6}\rho^s_{Z_1,Z_2}\right) & 1 & \cdots & 2\sin\left(\frac{\pi}{6}\rho^s_{Z_2,Z_n}\right) \\ \vdots & \vdots & \ddots & \vdots \\ 2\sin\left(\frac{\pi}{6}\rho^s_{Z_1,Z_n}\right) & 2\sin\left(\frac{\pi}{6}\rho^s_{Z_2,Z_n}\right) & \cdots & 1 \end{bmatrix}. \tag{19}$$

Step 4 Since $\mathbf{\Sigma}$ is a symmetric positive definite matrix, it can be decomposed as the product of a lower triangular matrix L and the corresponding transpose matrix L^{T} by using the Cholesky decomposition:

$$\mathbf{\Sigma} = L \cdot L^{\mathrm{T}} = \begin{bmatrix} l_{1,1} & 0 & \cdots & 0 \\ l_{1,2} & l_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{1,n} & l_{2,n} & \cdots & l_{n,n} \end{bmatrix} \cdot \begin{bmatrix} l_{1,1} & l_{1,2} & \cdots & l_{1,n} \\ 0 & l_{2,2} & \cdots & l_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & l_{n,n} \end{bmatrix}. \tag{20}$$

Step 5 The normally distributed correlated sample matrix is obtained as:

$$\tilde{X}^c = \tilde{X} \cdot L^{\mathrm{T}} = \begin{bmatrix} x_1^{c(1)} & x_2^{c(1)} & \cdots & x_n^{c(1)} \\ x_1^{c(2)} & x_2^{c(2)} & \cdots & x_n^{c(2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{c(m)} & x_2^{c(m)} & \cdots & x_n^{c(m)} \end{bmatrix}. \tag{21}$$

Step 6 Finally, the desired sample matrix $\tilde{Z}$ for parameters $\{Z_1, \cdots, Z_n\}$ is obtained though applying the transform of $F_{Z_i}^{-1}(\Phi(\cdot))$ to the corresponding element in $\tilde{X}^c$:

$$\tilde{Z} = \begin{bmatrix} F_{Z_1}^{-1}\left(\Phi(x_1^{c(1)})\right) & F_{Z_2}^{-1}\left(\Phi(x_2^{c(1)})\right) & \cdots & F_{Z_n}^{-1}\left(\Phi(x_n^{c(1)})\right) \\ F_{Z_1}^{-1}\left(\Phi(x_1^{c(2)})\right) & F_{Z_2}^{-1}\left(\Phi(x_2^{c(2)})\right) & \cdots & F_{Z_n}^{-1}\left(\Phi(x_n^{c(2)})\right) \\ \vdots & \vdots & \ddots & \vdots \\ F_{Z_1}^{-1}\left(\Phi(x_1^{c(m)})\right) & F_{Z_2}^{-1}\left(\Phi(x_2^{c(m)})\right) & \cdots & F_{Z_n}^{-1}\left(\Phi(x_n^{c(m)})\right) \end{bmatrix}. \tag{22}$$

In the next section, we will use a case study to demonstrate the application of the proposed sampling approach.

4 Case Study

The Wiedemann-74 car-following model (Wiedemann, 1974) is a well-known car-following model in microscopic traffic simulation. It has been implemented with the commercial microscopic traffic simulator VISSIM for modeling the car-following behavior in the urban area. This model contains 31 parameters, including 7 kinematic parameters of the vehicles in the car-following process, i.e., positions (x_f and x_l, the subscripts l and f indicate the leading vehicle and the following vehicle respectively), speeds (v_f and v_l), acceleration rates (acc_f and acc_l), as well as the length of the leading vehicle (L_l). Interested reader may refer to Ge *et al.* (2014b)) for a more detailed review of the parameters in the Wiedemann-74 car-following model.

A SA study was performed for this model in (Ge *et al.*, 2014b). In this paper, the authors analyzed the sensitivity of all parameters through sequentially applying the quasi-OTEE approach and the Kriging-based SA approach (Ge and Menendez, 2014; Ge *et al.*, 2014a,c). Due to the fact that the 7 kinematic variables are highly correlated, they cannot be sampled independently in the sequential SA. As a result, they were grouped as one single input Kin in the SA, i.e., any value assigned to Kin represents a combination of the 7 kinematic parameters. The sequential SA in (Ge *et al.*, 2014b) shows that Kin is the most influential parameter, which accounts for 50% of the variations of the resulting acceleration rate. However, as the 7 kinematic parameters are jointly sampled by groups, it is very hard to tell which parameters among them are the most important ones. Therefore, to further investigate the impacts of individual kinematic parameter, it is reasonable to generate samples separately for each parameter, under the condition that the samples have similar marginal distributions and dependence structure as the empirical data. In this case study we will use the aforementioned sampling approach for this task.

In this paper, we adopt the same empirical data as those in (Ge *et al.*, 2014b) for generating the samples. The empirical data came from 6 car-following experiments under different road and traffic conditions, hence it is expected that they are good representatives of different combinations of the 7 kinematic parameters. The data collection was carried out on the roads in Naples, Italy, area under real traffic conditions between October 2002 and July 2003. In the experiments, four vehicles were driven along urban and interurban roads under various traffic conditions without any lane changing. All vehicles were equipped with kinematic GPS receivers, and the position of each vehicle was recorded at an interval of 0.1s. Post data processing included differential correction of raw GPS coordinates, which utilized data gathered by a fifth stationary receiver (i.e., a base station) and an elaborate filtering procedure. Details on the car-following experiments are provided in (Punzo *et al.*, 2005; Punzo and Simonelli, 2005), and the car-following data are available on request to public from the MULTITUDE project (MULTITUDE, 2015). In this paper, a total of 16,425 car-following observations are included, i.e., for each of the 7

kinematic parameters there are 16,425 observations which can be used to derive the corresponding distributions and dependence structure.

In addition, since only the position difference and speed difference between the leading and following vehicles are used to derive the acceleration rate in the Wiedemann-74 model (see Wiedemann (1974)), in this case study we introduce two new parameters Δx ($\Delta x = x_l - x_f - L_l$) and Δv ($\Delta v = v_l - v_f$) in the sampling. The 5 kinematic parameters in the original model, i.e., x_l, x_f, L_l, v_l, and v_f, are not included in the following analysis. The marginal distributions of parameters Δx, Δv, acc_f, and acc_l, as well as the linear correlation coefficients and the scatter plots of any two parameters are illustrated in Figure 3. Moreover, the corresponding linear correlation coefficients and the Spearman's rank correlation coefficients are reported in Table 1. It is found in Table 1 that there are certain correlations for three parameter pairs Δv-acc_f, Δv-acc_l, and acc_f-acc_l, while the correlations for the rest parameter pairs are not significant.

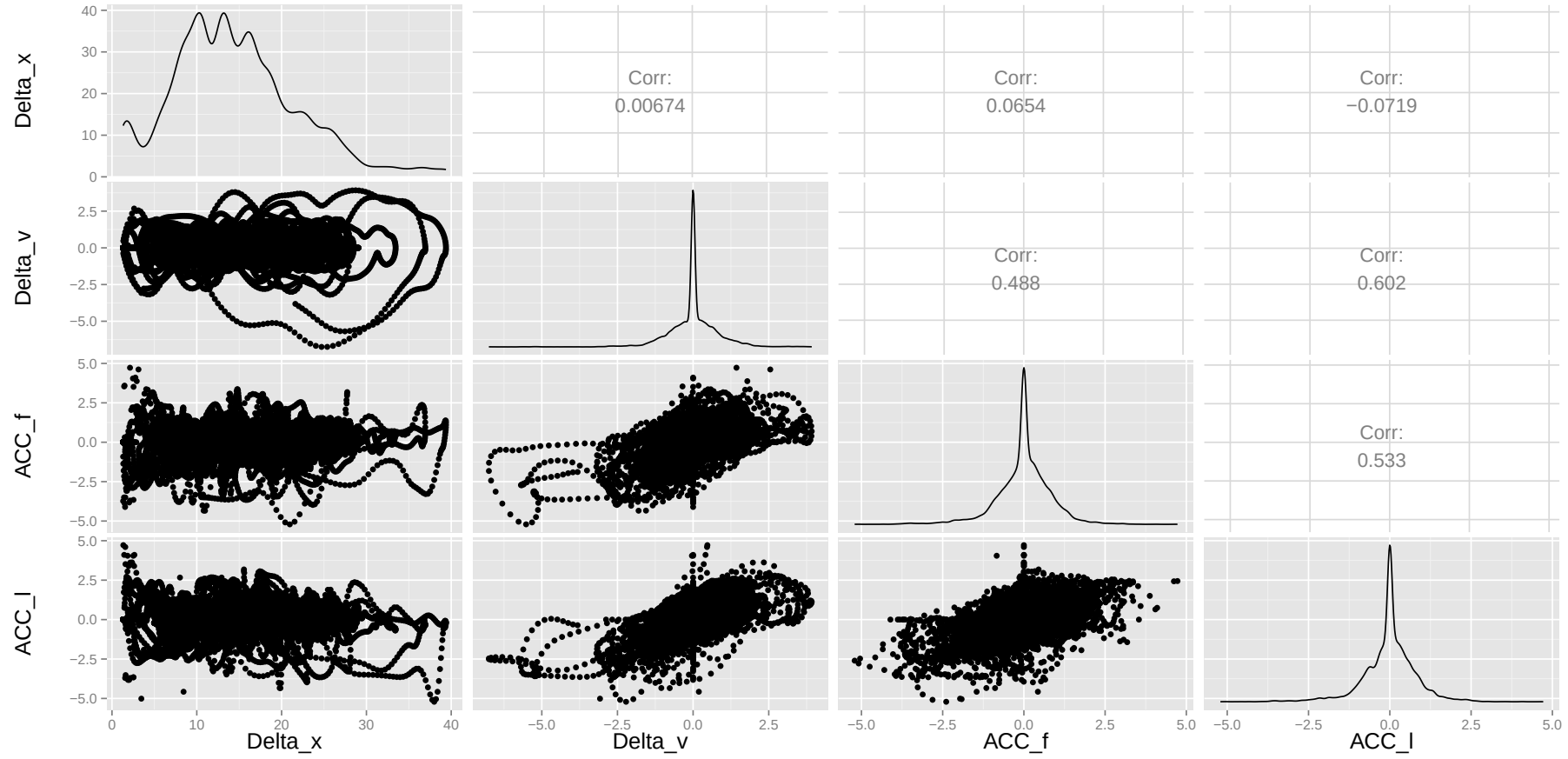
Table 1: Correlation coefficients from the empirical data

Parameters	Linear correlation coefficient ρ	Spearman's rank correlation coefficient ρ^s
Δx and Δv	0.0067	0.0176
Δx and acc_f	0.0654	0.0711
Δx and acc_l	-0.0719	-0.0298
Δv and acc_f	0.4884	0.4203
Δv and acc_l	0.6022	0.5417
acc_f and acc_l	0.5333	0.4877

To generate the desired samples using the approach proposed in Section 3, the first step is to generate four independent samples that are uniformly distributed in the open interval $\mathcal{U}(0, 1)$, i.e., $\tilde{V}_{\Delta x}$, $\tilde{V}_{\Delta v}$, $\tilde{V}_{acc_f}$, and $\tilde{V}_{acc_l}$. Then the inverse CDF of the standard normal distribution is applied to $\tilde{V}_{\Delta x}$, $\tilde{V}_{\Delta v}$, $\tilde{V}_{acc_f}$, and $\tilde{V}_{acc_l}$ in order to transform them into samples with standard normal distribution, i.e., $\tilde{X}_{\Delta x}$, $\tilde{X}_{\Delta v}$, $\tilde{X}_{acc_f}$, and $\tilde{X}_{acc_l}$. Note that they are still independent with each other at this step. In the next step, we transform the Spearman's rank correlation coefficient ρ^s into the linear correlation coefficient ρ for the Gaussian copula. This is done by applying Eq. (19). The covariance matrix $\mathbf{\Sigma}$ required for the Gaussian copula is shown below:

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & 0.0185 & 0.0745 & -0.0312 \\ 0.0185 & 1 & 0.4366 & 0.5597 \\ 0.0745 & 0.4366 & 1 & 0.5052 \\ -0.0312 & 0.5597 & 0.5052 & 1 \end{bmatrix}.$$

Figure 3: Marginal distribution, scatter plot and the linear correlation coefficient of the empirical data. Plots in the diagonal: marginal distribution of the corresponding parameter. Plots below the diagonal: scatter plots of the two parameters from the corresponding row and column. Plots above the diagonal: linear correlation coefficients of the two parameters from the corresponding row and column.



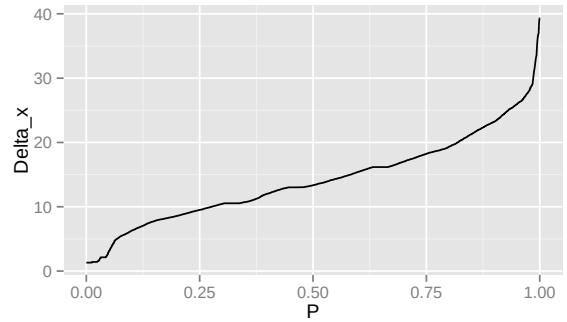
The covariance matrix can be further decomposed as the product of a lower triangular matrix and its transpose matrix by Cholesky decomposition. The transpose matrix L^{T} is shown below:

$$L^{\mathrm{T}} = \begin{bmatrix} 1.0000 & 0.0185 & 0.0745 & -0.0312 \\ 0 & 0.9998 & 0.4353 & 0.5604 \\ 0 & 0 & 0.8972 & 0.2938 \\ 0 & 0 & 0 & 0.7737 \end{bmatrix}.$$

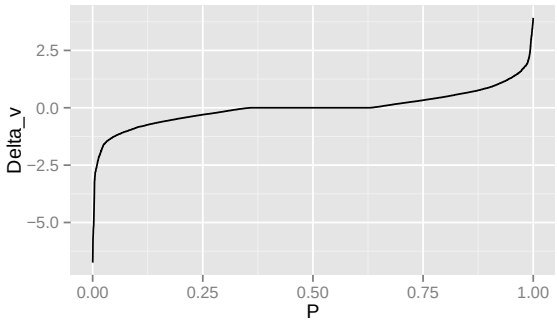
By multiplying $\tilde{X}_{\Delta x}$, $\tilde{X}_{\Delta v}$, $\tilde{X}_{acc_f}$, and $\tilde{X}_{acc_l}$ with L^{T} (Eq. (21)), the independent samples are transformed into correlated samples, i.e., $\tilde{X}^c_{\Delta x}$, $\tilde{X}^c_{\Delta v}$, $\tilde{X}^c_{acc_f}$, and $\tilde{X}^c_{acc_l}$. The final transform will be performed by applying the CDF of the standard normal distribution to the correlated samples to derive $\Phi(\tilde{X}^c_{\Delta x})$, $\Phi(\tilde{X}^c_{\Delta v})$, $\Phi(\tilde{X}^c_{acc_f})$, and $\Phi(\tilde{X}^c_{acc_l})$. Then the inverse CDF of each parameter (see Fig. 4) is applied to obtain the final samples. For example, if $\Phi(\tilde{X}^c_{\Delta v})$ is 0.5, then the corresponding sample for Δv is 0 according to the inverse CDF in Fig. 4(b).

Figure 4: Inverse cumulative distribution functions of Δx, Δv, acc_f, and acc_l obtained from the empirical data. Horizontal axle represents the probability, vertical axle represents the value of the corresponding parameter.

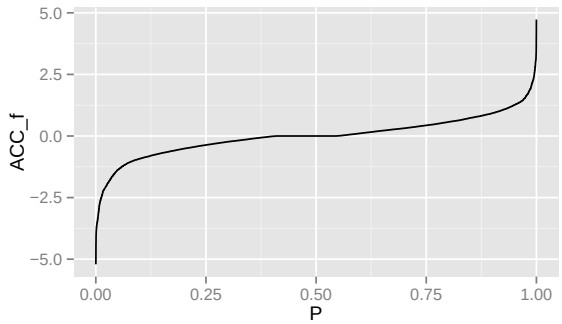
(a) Inverse CDF of Δx



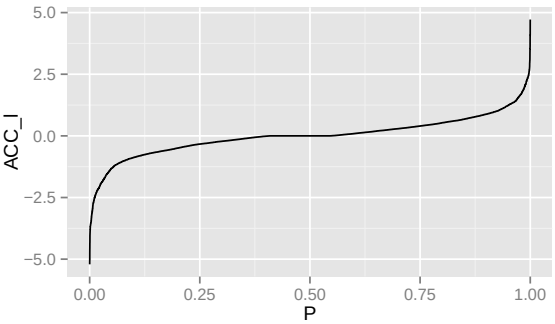
(b) Inverse CDF of Δv



(c) Inverse CDF of acc_f



(d) Inverse CDF of acc_l



For the demonstration purpose, we have conducted two experiments with two different sample sizes. The first experiment employs a sample size of 1,024 based on the Sobol sequence. This

one simulates the sampling process for the computationally expensive models without a big number of model runs. The other experiment has a sample size of 10,000 using random sampling, which simulates the sampling process for the computationally cheap models. The marginal distributions, scatter plots and the linear correlation coefficients of the final samples in these two experiments are shown in Fig. 5 and Fig. 6 respectively. The corresponding linear correlation coefficients and the Spearman's rank correlation coefficients are reported in Table 2.

Table 2: Correlation coefficients from the two experiments with sample size 1024 and 10000.

	Sample size = 1024		Sample size = 10000	
Parameters	ρ	ρ^s	ρ	ρ^s
Δx and Δv	0.0055	0.0149	0.0116	0.0190
Δx and acc_f	0.0631	0.0663	0.0756	0.0679
Δx and acc_l	-0.0475	-0.0339	-0.0292	-0.0298
Δv and acc_f	0.3907	0.4161	0.4073	0.4155
Δv and acc_l	0.5209	0.5367	0.5312	0.5365
acc_f and acc_l	0.4659	0.4916	0.4791	0.4913

Comparing Figs. 5 and 6 with Fig. 3, it is obvious that the marginal distributions in the second experiment (Fig. 6) are more similar to that of the empirical data. The main reason for such difference in the two experiments is because it has much more samples, which cannot always be achieved for computationally expensive models. On the other hand, the marginal distributions in the first experiment with 1,024 pseudo-random samples (Fig. 5) also show satisfactory similarities with that of the empirical data, especially for parameters Δv, acc_f, and acc_l.

We further make pair-wise comparison of the empirical data and the two experiments in terms of the linear correlation coefficients and the Spearman's rank correlation coefficients in Table 2 and Table 1. It is found that the dependence structures of the samples in the two experiments are also quite similar to that of the empirical data. Specifically, the orders of both linear and rank correlation coefficients for all parameter pairs in the two experiments are exactly the same with the that of the empirical data. For example, Δv-acc_l has the strongest correlation, followed by acc_f-acc_l and Δv-acc_f. Therefore, the accuracy of both experiments are acceptable, and the samples can be used for the sampling-based SA in the next step.

Figure 5: Marginal distribution, scatter plot and the linear correlation coefficient of the experiment with a sample size of 1024. Plots in the diagonal: marginal distribution of the corresponding parameter. Plots below the diagonal: scatter plots of the two parameters from the corresponding row and column. Plots above the diagonal: linear correlation coefficients of the two parameters from the corresponding row and column.

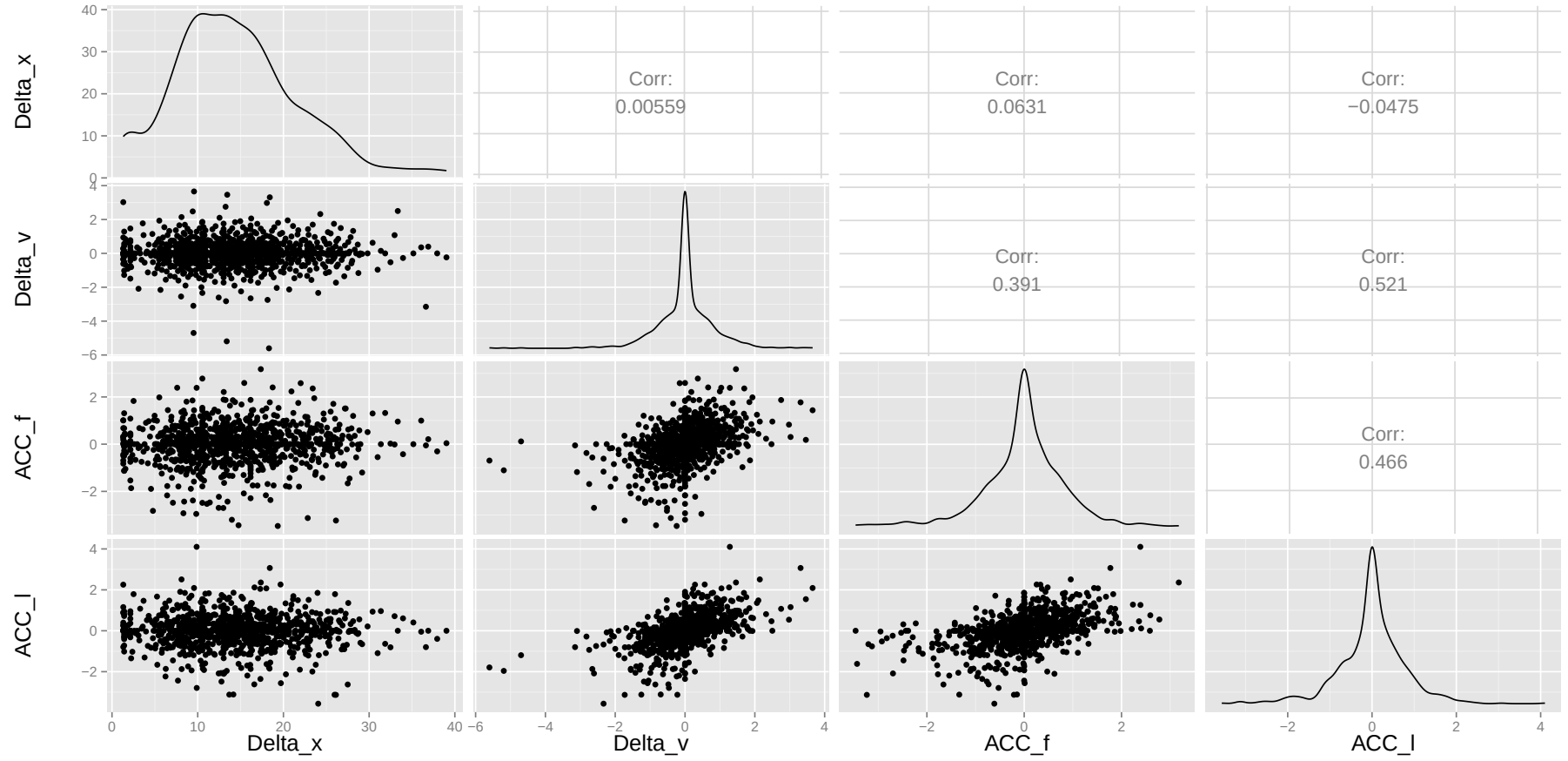
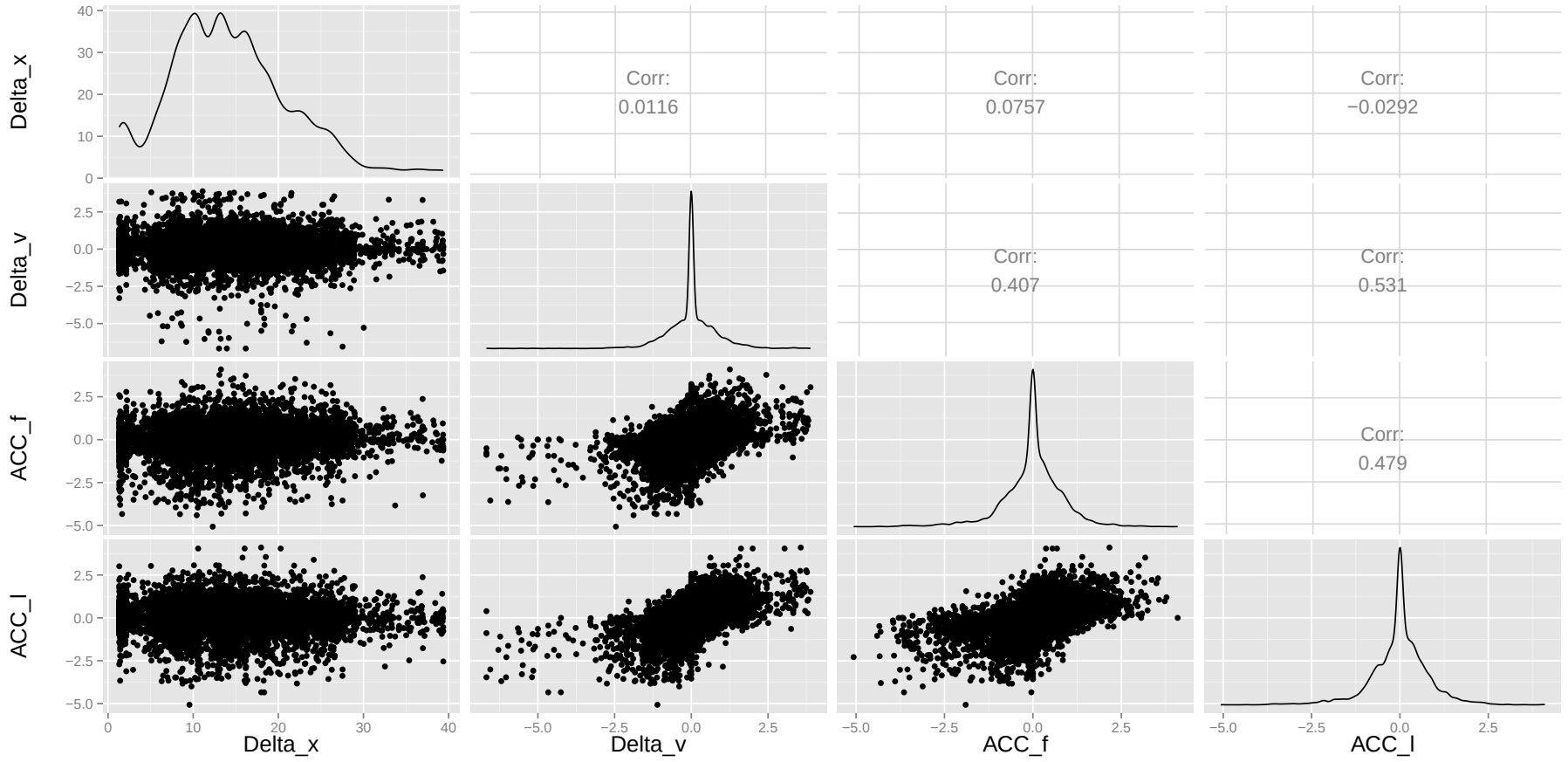


Figure 6: Marginal distribution, scatter plot and the linear correlation coefficient of the experiment with a sample size of 10000. Plots in the diagonal: marginal distribution of the corresponding parameter. Plots below the diagonal: scatter plots of the two parameters from the corresponding row and column. Plots above the diagonal: linear correlation coefficients of the two parameters from the corresponding row and column.



5 Conclusion

In this paper, we present a general approach for generating samples for dependent parameters. It utilizes the Gaussian copula in the sampling process, which makes it attractive for making samples of parameter from any arbitrary marginal distribution. Furthermore, the Spearman's rank correlation coefficient is employed instead of the traditional linear correlation coefficient, so that the dependence structure of the empirical data can be retained throughout the non-linear transform of the Gaussian copula.

A case study that generates samples for the kinematic parameters of Wiedemann-74 car-following model is included to demonstrate the application of this approach. It has shown that the marginal distributions and correlation coefficients of the generated samples are comparable with that of the empirical data. Specifically, the 1,024 samples, which are generated by employing the Sobol sequence in the sampling process, also present consistent marginal distributions and correlation coefficients as the empirical data. This has demonstrated that the proposed sampling approach is also useful for making samples of computationally expensive models, for which a big number of model runs are not always affordable.

The sampling approach illustrated here represents our first attempt for performing the SA of complex microscopic models with dependent parameters. In the next step, the research will continue with using the correlated samples to derive the global sensitivity indexes for dependent parameters.

Acknowledgement

The authors thank Vincenzo Punzo (University of Napoli Federico II) and Biagio Ciuffo (European Commission Joint Research Center) for discussions on the sensitivity analysis of car-following models and for the access to the car-following data used in this study. The research contained in this paper also benefited from the participation in the Multitude Project, a European Union COST (Cooperation in Science and Technology) Action.

6 References

Ge, Q., B. Ciuffo and M. Menendez (2014a) Combining screening and metamodel-based methods: An efficient sequential approach for the sensitivity analysis of model outputs, *Reliability Engineering & System Safety*, **134**, 334–344.

Ge, Q., B. Ciuffo and M. Menendez (2014b) Comprehensive Approach for the Sensitivity Analysis of High-Dimensional and Computationally Expensive Traffic Simulation Models, *Transportation Research Record: Journal of the Transportation Research Board*, **2422**, 121–130.

Ge, Q., B. Ciuffo and M. Menendez (2014c) An exploratory study of two efficient approaches for the sensitivity analysis of computationally expensive traffic simulation models, *IEEE Transactions on Intelligent Transportation Systems*, **15** (3) 1288–1297.

Ge, Q. and M. Menendez (2014) An efficient sensitivity analysis approach for computationally expensive microscopic traffic simulation models, *International Journal of Transportation*, **2** (2) 49–64.

Helton, J. C., J. D. Johnson, C. J. Sallaberry and C. B. Storlie (2006) Survey of sampling-based methods for uncertainty and sensitivity analysis, *Reliability Engineering & System Safety*, **91** (10) 1175–1209.

Hotelling, H. and M. R. Pabst (1936) Rank correlation and tests of significance involving no assumption of normality, *The Annals of Mathematical Statistics*, **7** (1) 29–43.

Iman, R. L. and W. J. Conover (1982) A distribution-free approach to inducing rank correlation among input variables, **11** (3) 311–334.

Kucherenko, S., S. Tarantola and P. Annoni (2012) Estimation of global sensitivity indices for models with dependent variables, *Computer Physics Communications*, **183** (4) 937–946.

Mara, T. a. and S. Tarantola (2012) Variance-based sensitivity indices for models with dependent inputs, *Reliability Engineering & System Safety*, **107**, 115–121.

MULTITUDE (2015) Multitude: Methods and tools for supporting the use, calibration and validation of traffic simulation models, `http://www.multitude-project.eu`.

Nelsen, R. B. (1999) *An introduction to copulas*, vol. 139, Springer Science & Business Media.

Punzo, V., D. J. Formisano and V. Torrieri (2005) Nonstationary kalman filter for estimation of accurate and consistent car-following data, *Transportation Research Record: Journal of the Transportation Research Board*, **1934** (1) 1–12.

Punzo, V. and F. Simonelli (2005) Analysis and comparison of microscopic traffic flow models with real traffic microscopic data, *Transportation Research Record: Journal of the Transportation Research Board*, **1934** (1) 53–63.

Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana and S. Tarantola (2007) *Global Sensitivity Analysis. The Primer*, John Wiley & Sons, Ltd, Chichester, UK, ISBN 9780470725184.

Sobol, I. M. (1976) Uniformly distributed sequences with an additional uniform property, *USSR Computational Mathematics and Mathematical Physics*, **16** (5) 236–242.

Sobol, I. M. (1993) Sensitivity estimates for nonlinear mathematical models, *Mathematical Modelling and Computational Experiments*, **1** (4) 407 – 414.

Wiedemann, R. (1974) Simulation des straßenverkehrsflusses. heft 8, *Schriftenreihe des Instituts für Verkehrswesen der Universitat Karlsruhe*, **8**.