2019-11-01

Bayesian Multivariate Nonlinear State Space Copula Models

Kreuzer, A

Bayesian Multivariate Nonlinear State Space Copula Models

Alexander Kreuzer*‡, Luciana Dalla Valle†, and Claudia Czado‡

Technische Universität München‡and University of Plymouth†

November 4, 2019

Abstract

In this paper we propose a flexible class of multivariate nonlinear non-Gaussian state space models, based on copulas. More precisely, we assume that the observation equation and the state equation are defined by copula families that are not necessarily equal. For each time point, the resulting model can be described by a C-vine copula truncated after the first tree, where the root node is represented by the latent state. Inference is performed within the Bayesian framework, using the Hamiltonian Monte Carlo method, where a further D-vine truncated after the first tree is used as prior distribution to capture the temporal dependence in the latent states. Simulation studies show that the proposed copula-based approach is extremely flexible, since it is able to describe a wide range of dependence structures and, at the same time, allows us to deal with missing data. The application to atmospheric pollutant measurement data shows that our approach is suitable for accurate modeling and prediction of data dynamics in the presence of missing values. Comparison to a Gaussian linear state space model and to Bayesian additive regression trees shows the superior performance of the proposed model with respect to predictive accuracy.

Keywords: Time Series, Bayesian Inference, Hamiltonian Monte Carlo, Vine Copulas

1 Introduction

State space models, also called dynamic models, originated in the field of system theory and were introduced by Kalman (1960) and Kalman and Bucy (1961), with early applications in aerospace-related research (Hutchinson (1984)). Since then, state space models have gained popularity in a number fields and have been applied in different areas, such as economics (Kitagawa and Gersch (1984); Shumway and Stoffer (1982)), medicine (Myers et al (2007); Liu and Guo (2015)) and ecology (Frühwirth-Schnatter (1994)). Durbin and Koopman (2000, 2002, 2012) provide a thorough illustration of state space models for time series analysis.

Linear Gaussian state space models are the most popular models in this class, with several contributions in the literature, including, for example, Ippoliti et al (2012) who applied these approaches to environmental data and Van den Brakel et al (2010), to official statistics. However, the strong assumptions of linear Gaussian state space models prevent their applicability to data showing departures from linearity and normality. In order to overcome these limitations, Chen et al (2012) applied nonlinear state space models to an epidemiological study on measles infection, relaxing the linearity assumptions, yet assuming normality for the model equations. Johns and Shumway (2005) proposed a spatio-temporal model for the analysis of censored dust particle concentrations which overcomes the linearity and normality assumptions, but assumes conditional Gaussian equation errors.

Copula-based approaches have proven to be particularly suitable for modeling data showing departures from multivariate normality. Copulas allow us to model separately the marginals from the dependence structure, and the use of different copula families, particularly Archimedean

*Corresponding author: E-mail: a.kreuzer@tum.de

copulas such as the Clayton and Gumbel, are suitable to accommodate asymmetric tail dependence. The literature of copula applications is vast. For environmental, actuarial and financial applications, see, for example, Genest and Favre (2007), Patton (2006), Jondeau and Rockinger (2006), Cherubini et al (2004), among others. A detailed overview of copulas and their properties is given by Joe (1997) and Nelsen (2007).

A rich class of parametric copula families is available in the bivariate case. However, in higher dimension the applicability of copulas is mostly limited in practice to the multivariate Gaussian or Student t. Vines, constructed using bivariate copulas as building blocks, provide a flexible alternative in the multivariate case. Vine copulas were first introduced by Joe (1996) and organised in a systematic way using graphical model structures by Bedford et al (2002). A thorough introduction to vines is provided by Aas et al (2009) and Czado (2019). Special types of vine copulas are C-vines, where one variable plays the role of the root node in each level, and D-vines, constructed as sequences of bivariate copulas. D-vines were employed by Smith et al (2010) to model the dependence structure of longitudinal data.

Hafner and Manner (2012) and Almeida and Czado (2012) suggest a bivariate state space model, with a bivariate copula in the observation equation and a Gaussian autoregressive process of order one, which describes the time evolution of the copula parameter, in the state equation. Kreuzer et al (2019) propose a univariate nonlinear non-Gaussian state space model, where both the observation and the state equation, are defined in terms of copula specifications. However, the copulas describing the observation and the state equation belong to the same family.

We propose a multivariate nonlinear non-Gaussian state space model, which extends the approach introduced by Kreuzer et al (2019) to multivariate observations, which we assume to be related to an underlying latent variable. This approach allows us to capture cross-sectional as well as temporal dependence in a very flexible way, since the copulas specifying the model can be all different. For each time point, the proposed model can be described as a C-vine truncated at the first tree, with the latent state being the root node. The latent states are treated as parameters, with prior distribution given by a D-vine truncated after the first tree to capture temporal dependence. An advantage of our approach is that missing values are handled in a natural way, since they are treated as latent variables. For model estimation, we cannot rely on the standard Kalman filter approach developed for linear state space models. Therefore, we suggest a Bayesian approach implemented using the Hamiltonian Monte Carlo (HMC) method (Neal et al (2011), Carpenter et al (2017)), where we introduce an indicator variable for the copula families specifying the state space model equations.

We demonstrate the usefulness of our method in a data set containing different air pollutant measurements. Three different pollutants are considered, and for each pollutant, measurements from a high-cost and from a low-cost sensor are utilized. In addition, covariates such as the temperature are available. To model this data we follow a flexible two-step modeling approach, motivated by Sklar's Theorem (Sklar (1959)). First we model the marginal distributions with generalized additive models (Hastie and Tibshirani (1987)) and in the second step we model dependencies with the novel copula state space model. We utilize our model to reconstruct high-cost measurements from low-cost measurements as in De Vito et al (2008) and show that the copula-based state space model, in combination with marginal generalized additive models, does a good job at predicting high-cost measurements. We show that it outperforms a Gaussian state space model and Bayesian additive regression trees with respect to the continuous ranked probability score (Gneiting and Raftery (2007)).

The rest of the paper is organized as follows: Section 2 introduces the novel multivariate copula state space model, Section 3 discusses Bayesian inference for the novel approach, Section 4 is devoted to the air pollutant measurements application and Section 5 concludes.

2 The Model

Copula approaches are very flexible since they can be combined with different marginal distributions. For the air pollution measurements data with additional covariates, as analyzed in Section 4, we propose generalized additive models (GAMs) for the margins in combination with the novel copula state space model to capture dependencies. The GAM explains the effect of the covariates, while the copula-based state space model handles temporal and cross-sectional dependence. In this section, we first introduce the marginal models (Section 2.1), which yield

data on the copula scale. Then, we review the linear Gaussian state space model (Section 2.2) and show an equivalent formulation in terms of Gaussian copulas (Section 2.3). In Section 2.4 we finally introduce the multivariate copula state space model as a generalization of the linear Gaussian state space model. The behavior of this model is illustrated with simulated data in Section 2.5.

2.1 Marginal Models

Consider a multivariate time series $\mathbf{Y}_t = (Y_{t1}, \ldots, Y_{td})'$ corresponding to d-dimensional continuous data, observed at the time points $t = 1, \ldots, T$, that may depend on a q-dimensional covariate vector $\boldsymbol{x}_t = (x_{t1}, \ldots, x_{tq})'$.

In order to allow for more flexibility, we consider Box-Cox transformations (Box and Cox (1964)) of the response variables, i.e. we consider the transformed variables

$$BC(Y_{tj}, \lambda_j) = \begin{cases} \frac{Y_{tj}^{\lambda_j} - 1}{\lambda_j}, \text{ for } \lambda_j \neq 0 \\ \ln(Y_{tj}), \text{ for } \lambda_j = 0 \end{cases} . \tag{1}$$

The relationship between the Box-Cox-transformed variables and the covariates can be expressed in various ways, e.g. using linear or nonlinear regression models. We assume a GAM (Hastie and Tibshirani (1987)) such that

$$BC(Y_{tj}, \lambda_j) = f_j(\boldsymbol{x}_t) + \sigma_j \varepsilon_{tj},$$

where $f_j(\cdot)$ is a smooth function of the covariates, expressing the mean of the GAM, and $\varepsilon_{tj} \sim N(0, 1)$. Let us define the standardized errors of the GAM as

$$Z_{tj} = \frac{BC(Y_{tj}, \lambda_j) - f_j(\boldsymbol{x}_t)}{\sigma_j}. \tag{2}$$

Note that $Z_{tj} \sim N(0, 1)$ holds.

We aim at modeling the errors $\mathbf{Z}_t = (Z_{t1}, \ldots, Z_{td})'$ as a multivariate nonlinear non-Gaussian state space model based on copulas.

2.2 Linear Gaussian State Space Models

State space models relate observations of a response variable to unobserved latent variables or "states". Gaussian linear state space models are defined by a linear observation model and a linear Markovian transition equation (Durbin and Koopman (2000), Durbin and Koopman (2002), Durbin and Koopman (2012)).

Suppose that we model the errors $\mathbf{Z}_t$, with $t = 1, \ldots, T$, extracted from the GAM as explained in Section 2.1, as a linear Gaussian state space model. Here, the variables Z_{tj}, $j = 1, \ldots, d$, are connected to a common continuous state variable W_t. Hence, the model can be formulated as

$$Z_{tj} = \rho_{obs,tj} \, W_t + \sigma_{obs,tj} \, \eta_{obs,tj} \tag{3}$$

$$W_t = \rho_{lat,t} \, W_{t-1} + \sigma_{lat,t} \, \eta_{lat,t}, \tag{4}$$

where $\eta_{obs,tj} \sim N(0, 1)$, $\eta_{lat,t} \sim N(0, 1)$ are independent i.i.d. sequences, $\rho_{obs,tj}$, $\rho_{lat,t}$, $\sigma_{obs,tj}$ and $\sigma_{lat,t}$ are model parameters and $W_0 \sim N(\mu_{lat,0}, \sigma_{lat,0})$, with $\mu_{lat,0}$ and $\sigma_{lat,0}$ generally known. Equation (3) is called observation equation, while Equation (4) is called state equation.

The linear Gaussian state space model can also be expressed using conditional distributions as

$$Z_{tj} \,|\, W_t = w_t \sim N\left(\rho_{obs,tj} \, w_t; \, \sigma^2_{obs,tj}\right)$$
$$W_t \,|\, W_{t-1} = w_{t-1} \sim N\left(\rho_{lat,t} \, w_{t-1}; \, \sigma^2_{lat,t}\right).$$

We assume time stationarity, i.e. $\rho_{obs,tj} = \rho_{obs,j}$, for $j = 1, \ldots, d$, and $\rho_{lat,t} = \rho_{lat}$. Since the model is applied to standardized errors with unit variance we also set $\sigma^2_{obs,tj} = 1 - \rho^2_{obs,j}$

and $\sigma^2_{lat,t} = 1 - \rho^2_{lat}$. In addition, we assume that $\mu_{lat,0} = 0$ and $\sigma_{lat,0} = 1$. These assumptions imply that $Z_{tj} \sim N(0,1)$ unconditionally. Hence, the model expressed through conditional distributions becomes

$$\begin{aligned} Z_{tj} \,|\, W_t = w_t &\sim N\left(\rho_{obs,j}\, w_t;\, 1 - \rho^2_{obs,j}\right) \\ W_t \,|\, W_{t-1} = w_{t-1} &\sim N\left(\rho_{lat}\, w_{t-1};\, 1 - \rho^2_{lat}\right). \end{aligned}$$

Thus, the state space model induces the following bivariate Gaussian distribution

$$\begin{pmatrix} Z_{tj} \\ W_t \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{obs,j} \\ \rho_{obs,j} & 1 \end{pmatrix}\right) \tag{5}$$

$$\begin{pmatrix} W_t \\ W_{t-1} \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{lat} \\ \rho_{lat} & 1 \end{pmatrix}\right). \tag{6}$$

Therefore, we obtain the joint distribution

$$(Z_{11}, \ldots, Z_{d1}, W_1; Z_{12}, \ldots, Z_{d2}, W_2; \ldots, Z_{1T}, \ldots, Z_{dT}, W_T) \sim N_{(d+1)T}(\mathbf{0}, \Sigma)$$

with covariance matrix Σ (see supplementary material). Thus, the joint distribution of Z_{tj} and Z_{t-1j} is given by

$$\begin{pmatrix} Z_{tj} \\ Z_{t-1j} \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho^2_{obs,j}\rho_{lat} \\ \rho^2_{obs,j}\rho_{lat} & 1 \end{pmatrix}\right).$$

2.3 Copula Formulation of a Gaussian State Space Model

The linear Gaussian state space model in equations (5) and (6) can be equivalently expressed in the copula space using Gaussian copulas as follows

$$\begin{aligned} (U_{tj}\,,\, V_t) &\sim \mathbb{C}^{Gauss}_{U_j,V}(\,\cdot\,,\,\cdot;\,\tau_{obs,j}) \\ (V_t\,,\, V_{t-1}) &\sim \mathbb{C}^{Gauss}_{V_2,V_1}(\,\cdot\,,\,\cdot;\,\tau_{lat}), \end{aligned} \tag{7}$$

where

$$U_{tj} = \Phi\left(Z_{tj}\right),\;\; V_t = \Phi\left(W_t\right),\quad j = 1, \ldots, d,\quad t = 1, \ldots, T, \tag{8}$$

with Φ denoting the standard normal cumulative distribution function. The variables U_{tj} and V_t are uniformly distributed as $U_{tj} \sim U\left(0,1\right),\;\; V_t \sim U\left(0,1\right)$, while the variables Z_{tj} and W_t are normally distributed as $Z_{tj} \sim N\left(0,1\right),\;\; W_t \sim N\left(0,1\right)$. The Gaussian copulas in (7) are parametrized by Kendall's τ, such that $\tau_{obs,j} = \frac{2}{\pi}\arcsin(\rho_{obs,j}),\;\; \tau_{lat} = \frac{2}{\pi}\arcsin(\rho_{lat})$.

2.4 Multivariate Nonlinear Non-Gaussian Copula State Space Model

The multivariate nonlinear non-Gaussian copula state space model allows the copula families in (7) to be different from the Gaussian, thus gaining a much greater flexibility to accommodate a wide range of dependence structures.

More precisely, the proposed model can be expressed, in the copula scale, as follows

$$\begin{aligned} (U_{tj}\,,\, V_t) &\sim \mathbb{C}^{m_{obs,j}}_{U_j,V}(\,\cdot\,,\,\cdot;\,\tau_{obs,j}) \\ (V_t\,,\, V_{t-1}) &\sim \mathbb{C}^{m_{lat}}_{V_2,V_1}(\,\cdot\,,\,\cdot;\,\tau_{lat}), \end{aligned} \tag{9}$$

where the copula families $m_{obs,j}$, for $j = 1, \ldots, d$, and m_{lat} are not necessarily equal and belong to a set $\mathcal{M}$ of single parameter copula families, parametrized by $\tau_{obs,j} = g_{m_{obs,j}}(\theta^{m_{obs,j}}_{obs,j})$ and $\tau_{lat} = g_{m_{lat}}(\theta^{m_{lat}}_{lat})$. The functions $g_{m_{obs,j}}, g_{m_{lat}}$ are one-to-one transformation functions and $\theta^{m_{obs,j}}_{obs,j}$ and $\theta^{m_{lat}}_{lat}$ are the parameters of the bivariate copulas $\mathbb{C}^{m_{obs,j}}_{U_j,V}$ and $\mathbb{C}^{m_{lat}}_{V_2,V_1}$, respectively. For example, for the Gumbel copula $g_{Gumbel}(\theta^{Gumbel}_{obs,j}) = 1 - \frac{1}{\theta^{Gumbel}_{obs,j}}$ holds.

The proposed model can also be specified in terms of conditional distribution functions as follows

$$\begin{aligned} (U_{tj} \,|\, V_t = v_t) &\sim \mathbb{C}^{m_{obs,j}}_{U_j|V}(\,\cdot \,|\, v_t;\,\tau_{obs,j}) \\ (V_t \,|\, V_{t-1} = v_{t-1}) &\sim \mathbb{C}^{m_{lat}}_{V_2|V_1}(\,\cdot \,|\, v_{t-1};\,\tau_{lat}), \end{aligned} \tag{10}$$

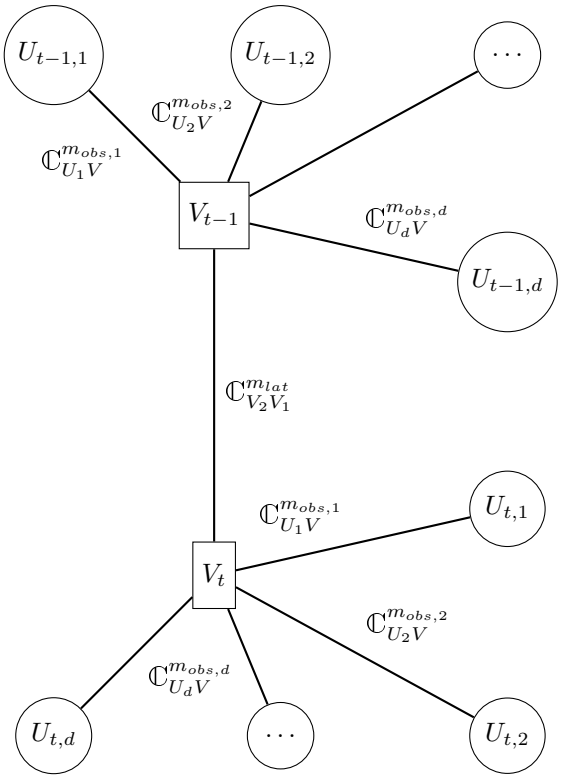

Figure 1: Graphical visualization of the multivariate state space copula model as specified in (9).

Figure 1 shows a graphical representation of the multivariate state space copula model. Each observed variable U_{tj} is linked to the latent state variable V_t via a copula $\mathbb{C}_{U_jV}^{m_{obs,j}}$ and the dependence between the latent states is modeled by the copula $\mathbb{C}_{V_2V_1}^{m_{lat}}$. In the following we denote by $c_{U_j,V}^{m_{obs,j}}$ and $c_{V_2,V_1}^{m_{lat}}$ the density functions of $\mathbb{C}_{U_j,V}^{m_{obs,j}}$ and $\mathbb{C}_{V_2,V_1}^{m_{lat}}$, respectively.

2.5 Illustration of the Copula State Space Model with Simulated Data

We visualize bivariate dependence structures that are obtained from our model with normalized contour plots (see, for example, Czado (2019), Chapter 3). We consider three scenarios which differ in the choice of the family m_{lat} of the latent copula. The parameters are chosen as follows

$$\begin{aligned}
&T = 1000\\
&d = 6\\
&\boldsymbol{m_{obs}} = (\text{Gaussian, Gaussian, Clayton, Clayton, Gumbel, Gumbel})\\
&\boldsymbol{\tau_{obs}} = (0.5, 0.7, 0.5, 0.7, 0.5, 0.7)\\
&\tau_{lat} = 0.7\\
&m_{lat} = \begin{cases} \text{Gaussian, Scenario 1} \\ \text{Clayton, Scenario 2} \\ \text{Gumbel, Scenario 3} \end{cases}
\end{aligned} \tag{11}$$

We consider one symmetric bivariate copula (Gaussian) and two asymmetric bivariate copulas (Gumbel, Clayton). We investigate two types of dependence: cross-sectional and temporal. For the cross-sectional dependence, we consider the pairs $(U_{tj}, U_{tj'})$ with corresponding bivariate copula density

$$c(u_{tj}, u_{tj'}) = \int_0^1 c_{U_jV}^{m_{obs,j}}(u_{tj}, v_t) c_{U_{j'}V}^{m_{obs,j'}}(u_{tj'}, v_t) dv_t. \tag{12}$$

The bivariate marginal density of $(U_{tj}, U_{tj'})$ given in (12) is neither affected by the time t nor by the copula $\mathbb{C}_{V_2V_1}^{m_{lat}}$. So the cross-sectional dependence is not affected by the copula $\mathbb{C}_{V_2V_1}^{m_{lat}}$ and

the corresponding theoretical contour plots are the same for all three scenarios. The empirical normalized contour plots for pairs $(U_{tj}, U_{tj'})$ are shown in Figure 2 for Scenario 1. The contour plots are constructed from 5000 independent simulations of the density in (12) for a fixed $t \in \{1, \dots, T\}$.

We see that if both linking copulas $\mathbb{C}_{U_jV}^{m_{obs,j}}$ and $\mathbb{C}_{U_{j'}V}^{m_{obs,j'}}$ are Gaussian, the contour of $(U_{tj}, U_{tj'})$ looks Gaussian as well (see the panel in the second row and the first column in Figure 2). In this case $\mathbb{C}(u_{tj}, u_{tj'})$ is indeed a Gaussian copula. If we mix a Gaussian and an asymmetric linking copula (see the entries below row 2 in columns 1 and 2 in Figure 2) or if we combine two asymmetric linking copulas (see the lower triangular entries in columns 3, 4 and 5 in Figure 2) we can obtain a variety of different asymmetric contour shapes.

For the temporal dependence, we consider the pairs (U_{tj}, U_{t-1j}) with bivariate copula density

$$c(u_{tj}, u_{t-1j}) = \int_{(0,1)^2} c_{U_jV}^{m_{obs,j}}(u_{tj}, v_t) c_{V_2V_1}^{m_{lat}}(v_t, v_{t-1}) c_{U_jV}^{m_{obs,j}}(u_{t-1j}, v_t) dv_t dv_{t-1}. \tag{13}$$

This dependence is affected by three copulas. Figure 3 shows normalized contour plots of the density in (13) obtained from 5000 independent simulations. We can see that if at least one of these copulas is asymmetric we may obtain an asymmetric dependence structure.

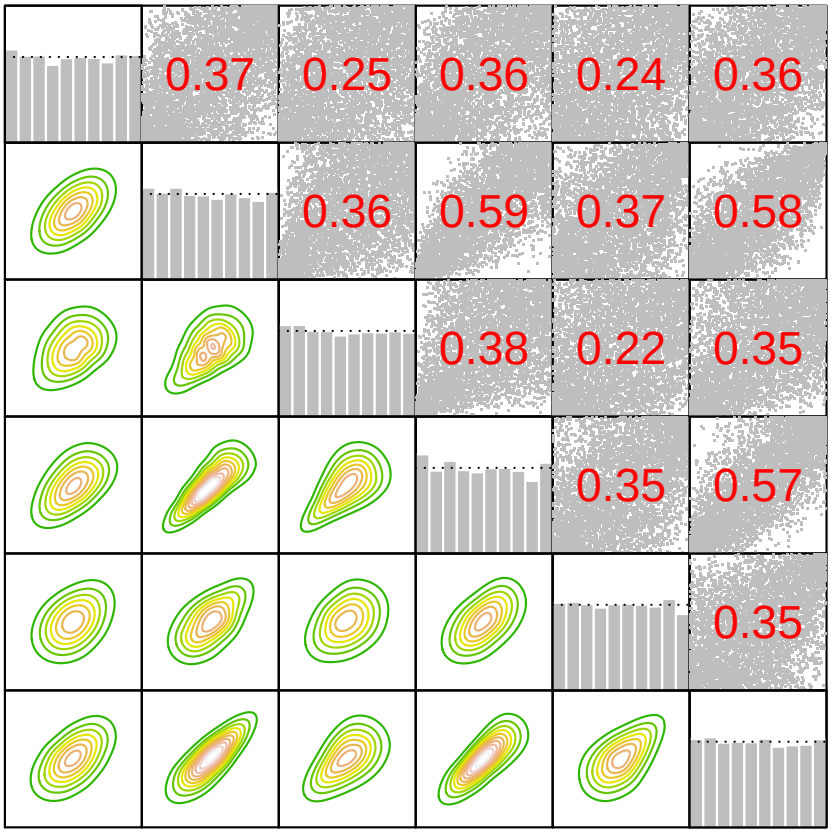

Figure 2: This plot is based on independently simulated data $(u_{tj}^r)_{r=1,\dots,5000, j=1,\dots,6}$ from Scenario 1 for a fixed $t \in \{1, \dots, T\}$. The lower triangular part shows contour plots of all pairs of $(z_{t1}^r, \dots, z_{t6}^r), r = 1, \dots, 5000$, where $z_{tj}^r = \Phi^{-1}(u_{tj}^r)$. The upper triangular part shows corresponding scatter plots and the empirical Kendall's τ for each pair $(u_{tj}, u_{tj'})$. The diagonal shows the histogram of the univariate marginals. More precisely, the plot in the i-th row and j-th column shows the contour plot for the pair (z_{ti}^r, z_{tj}^r) if $i > j$, the scatter plot of (u_{ti}^r, u_{tj}^r) if $i < j$, or the histogram of u_{ti}^r, if $i = j$, with $r = 1, \dots, 5000$.

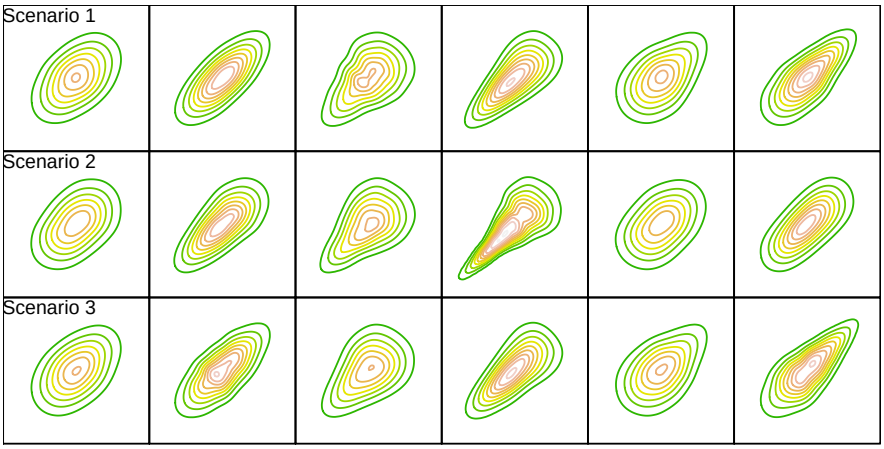

Figure 3: This plot is based on independently simulated data $(u^r_{t-1j}, u^r_{tj})_{r=1,\ldots,5000, j=1,\ldots,6}$ from Scenarios 1–3 for a fixed $t \in \{2,\ldots,T\}$. The data is transformed to the normalized scale as $z^r_{t'j} = \Phi^{-1}(u^r_{t'j})$, $t' = t-1, t$. Contour plots of the pairs $(z^r_{tj}, z^r_{t-1j})_{r=1,\ldots,5000}$ are shown for $j = 1,\ldots,6$. The plot in row m and column j shows the contour plot for $(z^r_{tj}, z^r_{t-1j})_{r=1,\ldots,5000}$, simulated from the parameter specification of Scenario m.

3 Bayesian Inference for the Multivariate Copula State Space Model

For the type of data we are dealing with, missing values are common. We denote the set of time indices of observed/non-missing values for dimension j by $\mathcal{T}_j^{obs}$ and the set of missing values by $\mathcal{T}_j^{miss} = \{1,\ldots,T\} \setminus \mathcal{T}_j^{obs}, j = 1,\ldots,d$. Further, we call $U^{obs} = (u_{tj})_{t\in\mathcal{T}_j^{obs}, j=1,\ldots,d}$ the observed and $U^{miss} = (u_{tj})_{t\in\mathcal{T}_j^{miss}, j=1,\ldots,d}$ the missing values. The missing values can be treated as latent variables. Integrating out the missing values yields the following likelihood for the observed values U^{obs}

$$\begin{aligned}\ell(\boldsymbol{v}, \boldsymbol{\tau_{obs}}, \boldsymbol{m_{obs}} | U^{obs}) &= \int_{(0,1)^{|U_{miss}|}} \prod_{j=1}^{d}\prod_{t=1}^{T} c_{U_jV}^{m_{obs,j}}(u_{tj}, v_t; \tau_{obs,j}) dU^{miss} = \\ &= \prod_{j=1}^{d}\left(\prod_{t\in\mathcal{T}_j^{obs}} c_{U_jV}^{m_{obs,j}}(u_{tj}, v_t; \tau_{obs,j}) \prod_{t\in\mathcal{T}_j^{miss}} \int_{(0,1)} c_{U_jV}^{m_{obs,j}}(u_{tj}, v_t; \tau_{obs,j}) du_{tj}\right) \qquad (14)\\ &= \prod_{j=1}^{d}\prod_{t\in\mathcal{T}_j^{obs}} c_{U_jV}^{m_{obs,j}}(u_{tj}, v_t; \tau_{obs,j}).\end{aligned}$$

Here the latent variable V_t is treated as a parameter v_t, and $\boldsymbol{v} = (v_1,\ldots,v_T)$, $\boldsymbol{\tau_{obs}} = (\tau_{obs,1},\ldots,\tau_{obs,d})$, $\boldsymbol{m_{obs}} = (m_{obs,1},\ldots,m_{obs,d})$. In contrast to a complete case analysis, information from all observed components is utilized in (14). The last equality in (14) uses the fact that in a copula the margins are uniform.

As mentioned above we use a D-vine truncated after the first tree to capture temporal dependence among the latent states, i.e.

$$\pi(\boldsymbol{v}|\tau_{lat}, m_{lat}) = \prod_{t=2}^{T} c_{V_2V_1}^{m_{lat}}(v_t, v_{t-1}; \tau_{lat}) \qquad (15)$$

with Kendall's τ parameter τ_{lat} and copula family indicator $m_{lat} \in \mathcal{M}$. This is a general Markov model of order 1 and collapses to a Gaussian AR(1) process if the Gaussian copula is used.

We restrict $\tau_{obs,1} \in (0,1)$ to be positive to ensure identifiability. This restriction corresponds to restricting the diagonal entries of the factor loading matrix in conventional Gaussian factor

models to be positive (see e.g. Lopes and West (2004)). For the Kendall's τ values of the remaining components we use a vague uniform prior on $(-1, 1)$, reflecting the fact that we do not have prior knowledge about these quantities. The following prior densities are used

$$\tau_{obs,1} \sim Beta(10, 1.5), \;\; \tau_{obs,j} \sim U(-1, 1), \;\; j = 2, \dots, d, \;\; \tau_{lat} \sim U(-1, 1). \tag{16}$$

For the copula family indicators we use discrete uniform priors, i.e.

$$\pi(m_{obs,j}) = \pi(m_{lat}) = \frac{1}{|\mathcal{M}|} \tag{17}$$

for $j = 1, \dots, d$. Further we assume that the Kendall's τ values and the copula family indicators are a priori independent such that the joint prior density is proportional to

$$\pi(\boldsymbol{\tau_{obs}}, \boldsymbol{m_{obs}}, \tau_{lat}, m_{lat}, \boldsymbol{v}) \propto \left(\prod_{t=2}^{T} c_{V_2 V_1}^{m_{lat}}(v_t, v_{t-1}; \tau_{lat}) \right) \pi(\tau_{obs,1}),$$

where $\pi(\tau_{obs,1})$ is the prior density specified in (16). This prior density is a joint density of continuous and discrete parameters. For discrete parameters $\boldsymbol{\delta}^{disc}$ and continuous parameters $\boldsymbol{\delta}^{cont}$ the joint density is defined as

$$f(\boldsymbol{\delta}^{cont}, \boldsymbol{\delta}^{disc}) = f(\boldsymbol{\delta}^{cont} | \boldsymbol{\delta}^{disc}) f(\boldsymbol{\delta}^{disc})$$

where $f(\boldsymbol{\delta}^{cont} | \boldsymbol{\delta}^{disc})$ is a conditional probability density function and $f(\boldsymbol{\delta}^{disc})$ is a joint probability mass function.

The set of parameters can be summarized as $\mathcal{P} = \{\tau_{lat}, \boldsymbol{\tau_{obs}}, m_{lat}, \boldsymbol{m_{obs}}, \boldsymbol{v}\}$. The posterior density of our model is proportional to

$$f(\mathcal{P} | U^{obs}) \propto \left(\prod_{j=1}^{d} \prod_{t \in \mathcal{T}_j^{obs}} c_{U_j V}^{m_{obs,j}}(u_{tj}, v_t, \tau_{obs,j}) \right) \left(\prod_{t=2}^{T} c_{V_2 V_1}^{m_{lat}}(v_t, v_{t-1}; \tau_{lat}) \right) \pi(\tau_{obs,1}). \tag{18}$$

As in Kreuzer et al (2019), sampling from the posterior in (18) is not straightforward, e.g. Kalman filter recursions cannot be applied. Since the No-U-turn sampler of Hoffman and Gelman (2014) has shown good performance for the univariate copula state space model (Kreuzer et al (2019)), we also use it here. The No-U-Turn sampler is an extension of Hamiltonian Monte Carlo (HMC, Neal et al (2011)) with adaptively selected tuning parameters. To run the sampler we use STAN (Carpenter et al (2017)).

Updating Continuous Parameters

Since HMC cannot deal with discrete variables we integrate over the discrete family indicators which corresponds to summing over them, i.e.

$$\begin{aligned} f(\tau_{lat}, \boldsymbol{\tau_{obs}}, \boldsymbol{v} | U^{obs}) &= \sum_{(m_{lat}, \boldsymbol{m_{obs}}) \in \mathcal{M}^{d+1}} f(\tau_{lat}, \boldsymbol{\tau_{obs}}, m_{lat}, \boldsymbol{m_{obs}}, \boldsymbol{v} | U^{obs}) \\ &\propto \prod_{j=1}^{d} \left(\sum_{m_{obs,j} \in \mathcal{M}} \prod_{t \in \mathcal{T}_j^{obs}} c_{U_j V}^{m_{obs,j}}(u_{tj}, v_t; \tau_{obs,j}) \right) \cdot \\ &\quad \cdot \left(\sum_{m_{lat} \in \mathcal{M}} \prod_{t=2}^{T} c_{V_2 V_1}^{m_{lat}}(v_t, v_{t-1}; \tau_{lat}) \right) \pi(\tau_{obs,1}) \end{aligned} \tag{19}$$

To sample from this density we use STAN's No-U-Turn sampler.

Updating the (Discrete) Copula Family Indicators

In $f(\boldsymbol{m_{obs}}, m_{lat} | \tau_{lat}, \boldsymbol{\tau_{obs}}, \boldsymbol{v}, U^{obs})$, all components of $(\boldsymbol{m_{obs,}}, m_{lat})$ are independent. We have that

$$\begin{aligned} f(m_{obs,j} | \tau_{lat}, \boldsymbol{\tau}_{obs}, \boldsymbol{v}, \boldsymbol{m_{obs,-j}}, m_{lat}, U^{obs}) = \\ \frac{f(\tau_{lat}, \boldsymbol{\tau_{obs}}, m_{lat}, \boldsymbol{m_{obs}}, \boldsymbol{v} | U^{obs})}{\sum_{m'_{obs,j} \in \mathcal{M}} f(\tau_{lat}, \boldsymbol{\tau_{obs}}, m_{lat}, \boldsymbol{m_{obs,-j}}, m'_{obs,j}, \boldsymbol{v} | U^{obs})}, \end{aligned} \tag{20}$$

where $\boldsymbol{m_{obs,-j}}$ is equal to $\boldsymbol{m_{obs}}$ with the j-th component removed. Therefore we obtain

$$f(m_{obs,j}|\tau_{lat}, \boldsymbol{\tau_{obs}}, \boldsymbol{v}, \boldsymbol{m_{obs,-j}}, m_{lat}, U^{obs}) = \frac{\prod_{t\in\mathcal{T}_j^{obs}} c_{U_jV}^{m_{obs,j}}(u_{tj}, v_t; \tau_{obs,j})}{\sum_{m'_{obs,j}\in\mathcal{M}} \prod_{t\in\mathcal{T}_j^{obs}} c_{U_jV}^{m'_{obs,j}}(u_{tj}, v_t; \tau_{obs,j})} \tag{21}$$

Similarly we obtain

$$f(m_{lat}|\tau_{lat}, \boldsymbol{\tau_{obs}}, \boldsymbol{v}, \boldsymbol{m_{obs}}, U^{obs}) = \frac{\prod_{t=2}^{T} c_{V_2V_1}^{m_{lat}}(v_t, v_{t-1}; \tau_{lat})}{\sum_{m'_{lat}\in\mathcal{M}} \prod_{t=2}^{T} c_{V_2V_1}^{m'_{lat}}(v_t, v_{t-1}; \tau_{lat})} \tag{22}$$

Obtaining Updates for the Joint Posterior Density

To obtain R samples from the posterior density given in (18) we first obtain R samples of $\tau_{lat}, \boldsymbol{\tau_{obs}}, \boldsymbol{v}$ from the density given in (19) using STAN. We denote the samples by τ_{lat}^r, $\boldsymbol{\tau_{obs}^r}$, $\boldsymbol{v^r}$, $r = 1, \ldots, R$. Then we sample $m_{obs,j}$ from $f(m_{obs,j}|\tau_{lat}^r, \boldsymbol{\tau_{obs}^r}, \boldsymbol{v^r}, U^{obs})$ (see (21)) to obtain $m_{obs,j}^r$, for $r = 1, \ldots, R$ and $j = 1, \ldots, d$. Further, m_{lat}^r is obtained by sampling from $f(m_{lat}|\tau_{lat}^r, \boldsymbol{\tau_{obs}^r}, \boldsymbol{v^r}, U^{obs})$ (see (22)) , for $r = 1, \ldots, R$.

Predictive Distribution (In-Sample Period)

The predictive density of a new value u_{tj}^{new} for margin j at time $t \in \{1, \ldots, T\}$ is the conditional density of u_{tj}^{new} given U^{obs}, obtained as

$$f(u_{tj}^{new}|U^{obs}) = \int_{domain(\mathcal{P})} f(u_{tj}^{new}, \mathcal{P}|U^{obs})d\mathcal{P} = \int_{domain(\mathcal{P})} f(u_{tj}^{new}|\mathcal{P}, U^{obs}) f(\mathcal{P}|U^{obs})d\mathcal{P}$$

with $f(u_{tj}^{new}|\mathcal{P}, U^{obs}) = c_{U_jV}^{m_{obs,j}}(u_{tj}^{new}, v_t; \tau_{obs,j})$ and $domain(\mathcal{P})$ is the domain of the parameter space $\mathcal{P}$. Note that for the discrete indicator variables the integral is a sum.

To obtain samples from the predictive distribution we sample from the following density

$$f(u_{tj}^{new}, \mathcal{P}|U^{obs}) = f(u_{tj}^{new}|\mathcal{P}, U^{obs}) f(\mathcal{P}|U^{obs}).$$

We proceed as follows:

- We first simulate R samples of $\mathcal{P}$ from $f(\mathcal{P}|U^{obs})$ as described above.
- The $r-th$ sample of u_{tj}^{new}, denoted by $(u_{tj}^{new})^r$, is simulated from $\mathbb{C}_{U_j|V}^{m_{obs,j}^r}(\cdot|v_t^r; \tau_{obs,j}^r)$, for $r = 1, \ldots, R$.

For $t \in \mathcal{T}_j^{miss}$, we can obtain simulated values for the missing values.

Predictive Distribution (Out-of-Sample Period)

To obtain samples from the predictive distribution of a new value u_{tj}^{new} for margin j at time $t \in \{T+1, T+2, \ldots\}$ we consider the following density

$$f(u_{tj}^{new}, \mathcal{P}|U^{obs}) = f(u_{tj}^{new}|\mathcal{P}, U^{obs}) f(\mathcal{P}|U^{obs})$$

with

$$f(u_{tj}^{new}|\mathcal{P}, U^{obs}) = \int_{(0,1)^{t-T}} c_{U_jV}^{m_{obs,j}}(u_{tj}^{new}, v_t; \tau_{obs,j}) \prod_{t'=T+1}^{t} c_{V_2V_1}^{m_{lat}}(v_{t'}, v_{t'-1}; \tau_{lat}) dv_{T+1} \ldots, dv_t.$$

We proceed as follows to obtain samples from this density

- We first simulate R samples of $\mathcal{P}$ from $f(\mathcal{P}|U^{obs})$ as described above.
- For $r = 1, \ldots, R$ and for $t' = T+1, \ldots, t$:
  Sample $v_{t'}$ from $\mathbb{C}_{V_2V_1}^{m_{lat}^r}(\cdot|v_{t'-1}^r; \tau_{lat}^r)$ and denote the sample by $v_{t'}^r$.
- For $r = 1, \ldots, R$: Sample u_{tj}^{new} from $\mathbb{C}_{U_j|V}^{m_{obs,j}^r}(\cdot|v_t^r; \tau_{obs,j}^r)$ and denote the sample by $(u_{tj}^{new})^r$.

Note that the recursive sampling avoids the evaluation of the $t - T$ dimensional integral.

4 Data Analysis

4.1 Data Description

We consider a subset of the data set available at `http://archive.ics.uci.edu/ml/datasets/Air+Quality` (De Vito et al (2008, 2009, 2012)). The data set contains hourly averaged concentration measurements for different atmospheric pollutants obtained at a main road in an Italian city. Here we analyze measurements from June to September 2004, which result in 2928 observations. The measurements for the pollutants were taken from two different sensors, standard (high-cost) sensors and new low-cost (lc) sensors. We refer to a value measured with the standard (high-cost) sensor as a ground truth (gt) value. Ground truth values are available for CO (mg/m^3), NOx (ppb) and NO2 (μg/m^3) and the aim is to predict these values. For each ground truth value we are given a corresponding value obtained from a low-cost sensor, resulting in six different pollution measurements for one time point. The measurements in July for the pollutant CO are visualized in Figure 4. We see that the measurements of the ground truth sensor for CO are missing for several days, i.e. missing observations are present in this data set. The missing values per pollutant range from 4% to 24%, whereas ground truth values have a higher portion of missing values. In addition to the pollution measurements, hourly measurements of the temperature and of relative humidity are also available.

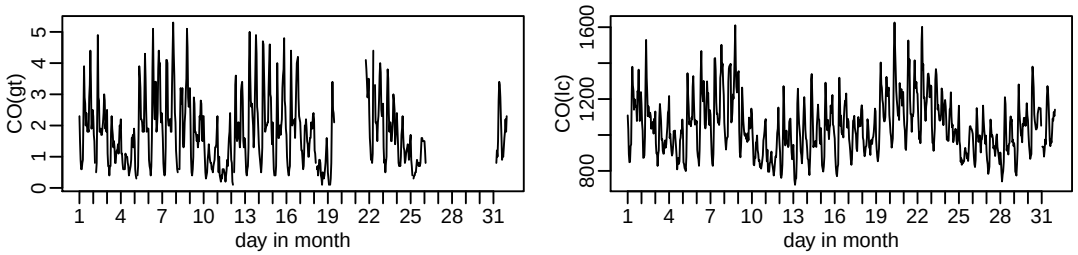

Figure 4: Hourly observed values of one pollutant (CO) from the ground truth (gt) and low-cost (lc) sensors in July 2004. When missing values are present no observations are drawn for the corresponding time points.

In the following, the data containing the pollutant measurements is denoted by $y_{tj}, j = 1, \ldots, 6, t = 1, \ldots, T$, where $T = 2928$ is the length of the time series. As before, $\mathcal{T}_j^{obs}$ is the set of time indices for which observed values are available for the j-th marginal time series. The measurements of relative humidity and temperature are denoted by TEMP_t and RH_t, respectively for $t = 1, \ldots, T$.

4.2 Marginal models

We fit a generalized additive model (GAM) for each pollutant, where temperature, relative humidity, the hour at time t, $\text{H}_t \in \{0, \ldots, 23\}$, and the day at time t, $\text{D}_t \in \{0, \ldots, 6\}$ are used as covariates. We denote the covariates by $\boldsymbol{x_t} = (\text{TEMP}_t, \text{RH}_t, \text{H}_t, \text{D}_t)$. As explained in Section 2.1, we allow for Box-Cox transformations (Box and Cox (1964)) and assume that

$$BC(Y_{tj}, \lambda_j) = f_j(\boldsymbol{x_t}) + \sigma_j \epsilon_{tj} \tag{23}$$

with $\epsilon_{tj} \sim N(0, 1)$ for $t = 1, \ldots, T, j = 1, \ldots, 6$ and $BC(Y_{tj}, \lambda_j)$ as in (1).

For estimating the conditional mean function f_j and σ_j we assume that the errors ϵ_{tj} are independent. Later the dependence among the errors will be modeled with the proposed state space model. We estimate a GAM for different values of λ_j and then choose the model which maximizes the likelihood for given data $y_{tj}, t \in \mathcal{T}_j^{obs}, j = 1, \ldots, 6$. For each GAM we remove the corresponding missing values and rely on the R package `mgcv` of Wood and Wood (2015) for parameter estimation. We obtain estimates $\hat{f}_j$, $\hat{\sigma}_j$ and $\hat{\lambda}_j$ for $j = 1, \ldots, 6$. From Table 1, we see that the estimates for λ_j deviate from 1, which indicates that the Box-Cox transformations are necessary. Figure 5 shows the smooth components of the GAM for four different pollutants. We

see for example a nonlinear effect of the Hour on the pollution measurement. The pollution is high at around 8 am and at around 6 pm, which may correspond to the hours with the highest traffic due to commuting workers.

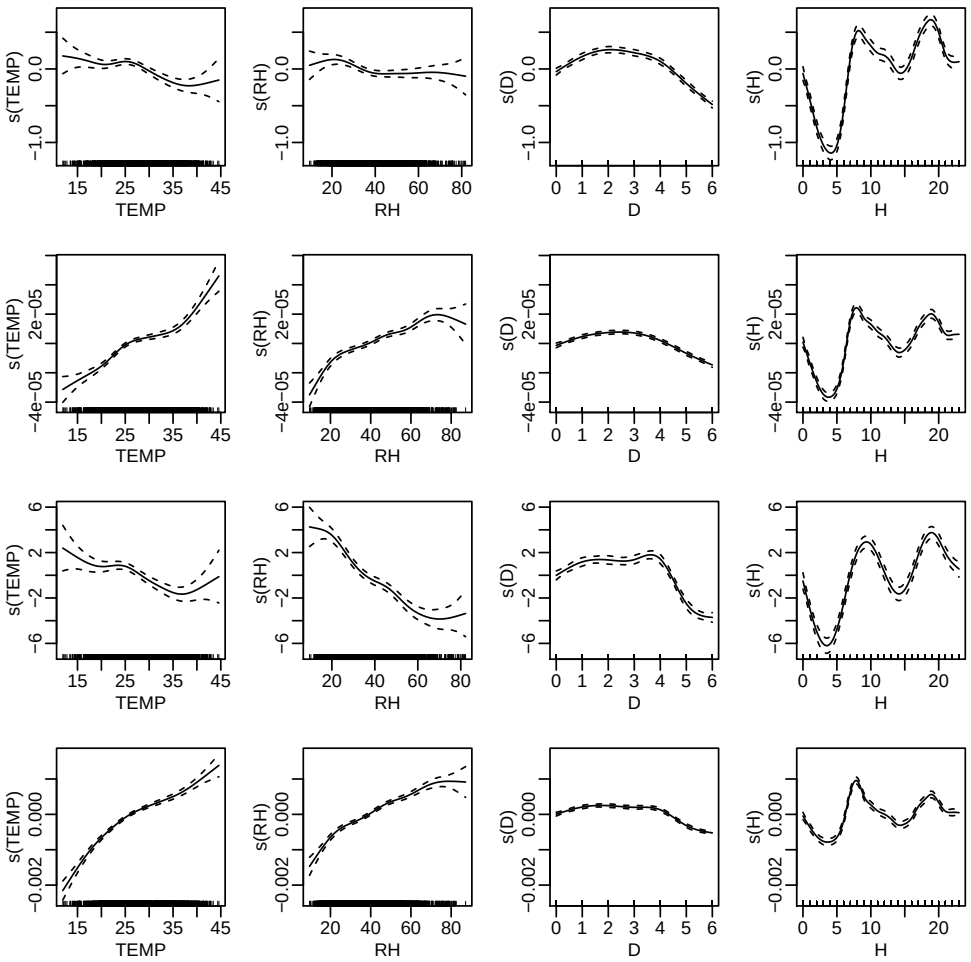

Figure 5: Estimated smooth components of the GAMs for four Box-Cox transformed pollutants: CO(gt), CO(lc), NO2(gt), NO2(lc) (top to bottom row). Each GAM has four covariates, TEMP, RH, D and H. The dashed lines represent a pointwise 95% confidence band.

	CO(gt)	CO(lc)	NOx(gt)	NOx(lc)	NO2(gt)	NO2(lc)
$\hat{\lambda}_j$	0.15	-1.25	0.05	0.05	0.55	-0.70

Table 1: Estimates of $\lambda_1, \ldots, \lambda_6$ for the six GAMs fitted to the six pollution measurements.

4.3 Dependence Model

Recall the standardized errors Z_{tj}, defined in (2), as

$$Z_{tj} = \frac{BC(Y_{tj}, \lambda_j) - f_j(\boldsymbol{x_t})}{\sigma_j}$$

which are $N(0,1)$ distributed. Pseudo observations of Z_{tj} can be obtained from the estimates $\hat{f}_j$, $\hat{\sigma}_j$ and $\hat{\lambda}_j$ as

$$\hat{z}_{tj} = \frac{BC(y_{tj}, \hat{\lambda}_j) - \hat{f}_j(\boldsymbol{x_t})}{\hat{\sigma}_j} \tag{24}$$

for $t = 1, \ldots, T, j = 1, \ldots 6$. To visualize cross-sectional dependencies among the variables Z_{tj} we examine bivariate contour plots for all pairs of $(\hat{z}_{t1}, \ldots, \hat{z}_{t6}), t = 1, \ldots, T$ in Figure 6, ignoring serial dependence. In addition, we examine contour plots of pairs $(\hat{z}_{tj}, \hat{z}_{t-1j}), t = 2, \ldots, T$ for $j = 1, \ldots, 6$ in Figure 7 to visualize temporal dependence. We observe temporal and cross-sectional dependence. Further, the dependence structures seem to be different from a Gaussian one since we observe asymmetries in the contour plots. For example, the contour plot in the bottom left corner of Figure 6 indicates stronger dependence in the upper right corner than in the bottom left corner. Therefore, a linear Gaussian state space model might not be appropriate here, but the proposed copula-based state space model can be a good candidate for this data.

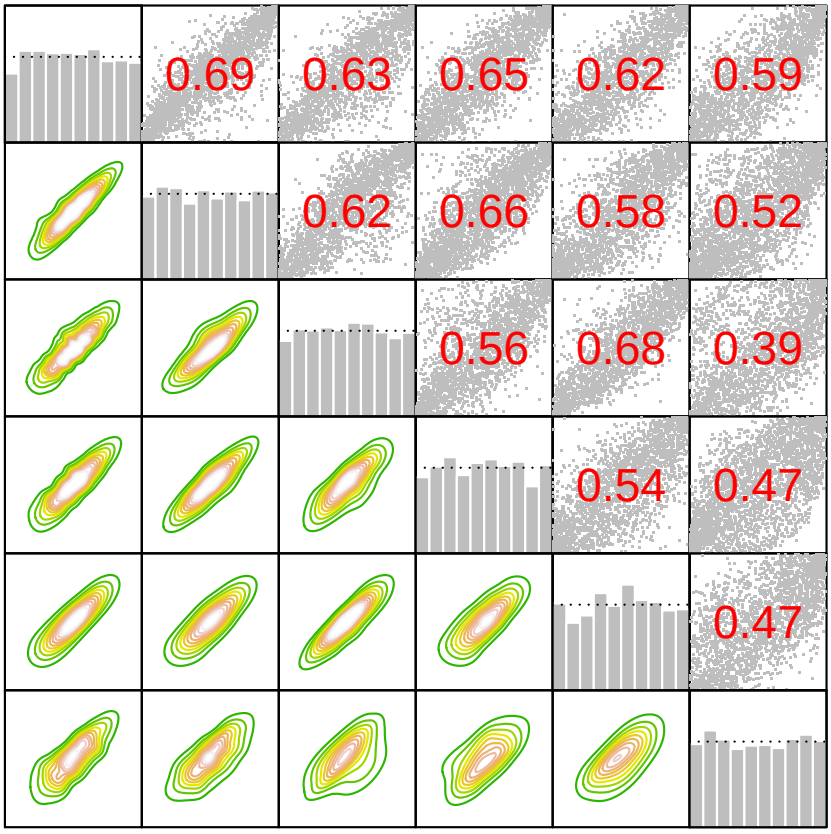

Figure 6: The lower triangular part shows contour plots of all pairs of $(\hat{z}_{t1}, \ldots, \hat{z}_{t6}), t = 1, \ldots, T$ ignoring serial dependence. The upper triangular part shows corresponding scatter plots of all pairs of $(\hat{u}_{t1}, \ldots, \hat{u}_{t6}), t = 1, \ldots, T$ with $\hat{u}_{tj} = \Phi(\hat{z}_{tj})$ and the empirical Kendall's τ for each pair. The diagonal shows the histograms of the univariate marginals. More precisely, the plot in the i-th row and j-th column shows the contour plot for the pair $(\hat{z}_{ti}, \hat{z}_{tj})$ if $i > j$, the scatter plot of $(\hat{u}_{ti}, \hat{u}_{tj})$ if $i < j$, or the histogram of $\hat{u}_{ti}$, if $i = j$. The variables are ordered as follows: 1: CO(gt), 2: CO(lc), 3: NOx(gt), 4: NOx(lc), 5: NO2(gt), 6: NO2(lc).

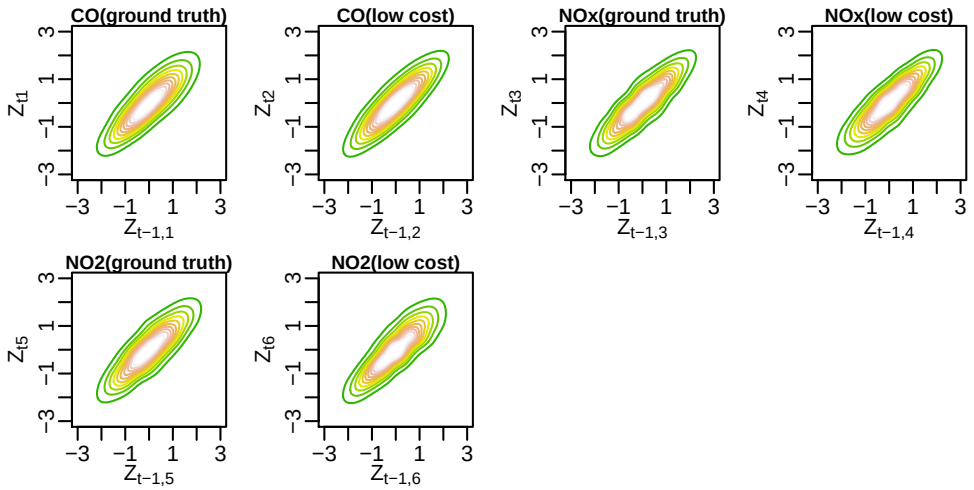

Figure 7: Contour plots of pairs $(\hat{z}_{tj}, \hat{z}_{t-1j})_{t=2,\ldots,T}$ for $j = 1, \ldots, 6$ ignoring serial dependence.

Since our multivariate copula state space model operates on marginally uniform(0,1) distributed data, we obtain uniform(0,1) distributed random variables as follows

$$U_{tj} = \Phi(Z_{tj}) \tag{25}$$

with corresponding pseudo observations

$$\hat{u}_{tj} = \Phi(\hat{z}_{tj}) \tag{26}$$

for $t \in \mathcal{T}_j^{obs}, j = 1, \ldots, 6$. The proposed multivariate copula-based state space model is fitted to the data $\hat{u}_{tj}, t \in \mathcal{T}_j^{obs}, j = 1, \ldots, 6$. Plots of the estimated posterior densities and trace plots are shown in the supplementary material. These plots indicate proper mixing of the Markov Chain. Table 2 shows the selected copula families corresponding to the estimated posterior modes of $m_{obs,j}$ or m_{obs}. We see that four Gaussian, one Student t and two Gumbel copulas were selected. In particular, our model features an asymmetric dependence structure, since the Gumbel copula is included. Simulations of the in-sample period predictive distribution can be obtained as explained in Section 3. Transforming these simulations with the standard normal quantile function, we obtain predictive simulations for the standardized errors, i.e. we obtain draws from the predictive distribution of the error as

$$\epsilon_{tj}^r = \Phi^{-1}(u_{tj}^r), \tag{27}$$

for $r = 1, \ldots, 3000$, where u_{tj}^r is a draw from the in-sample predictive distribution on the copula scale (see Section 3). These simulations are compared to the observed standardized residual of the GAM $(\hat{z}_{tj})$ to assess how well our model fits the data. In particular, we want to asses if a single factor structure is appropriate or if it is too restrictive. According to Figure 8, the model seems to be appropriate. The single factor structure is able to capture the time dynamics of the residuals. The ground truth values for CO are missing from day 26 to day 30. We see that within this period the time dynamic is learned from other series where data is available within this period. While Figure 8 shows plots for two pollutants in July, plots for different pollutants in different months looked similar.

	$\hat{m}_{obs,1}$	$\hat{m}_{obs,2}$	$\hat{m}_{obs,3}$	$\hat{m}_{obs,4}$	$\hat{m}_{obs,5}$	$\hat{m}_{obs,6}$	$\hat{m}_{lat}$
Copula family	Gu	Gu	Ga	S	Ga	Ga	Ga

Table 2: The marginal posterior mode estimates of the copula family indicators $\boldsymbol{m}_{obs}, m_{lat}$. (Ga: Gaussian, S: Student t(df=4), C: Clayton, Gu: Gumbel).

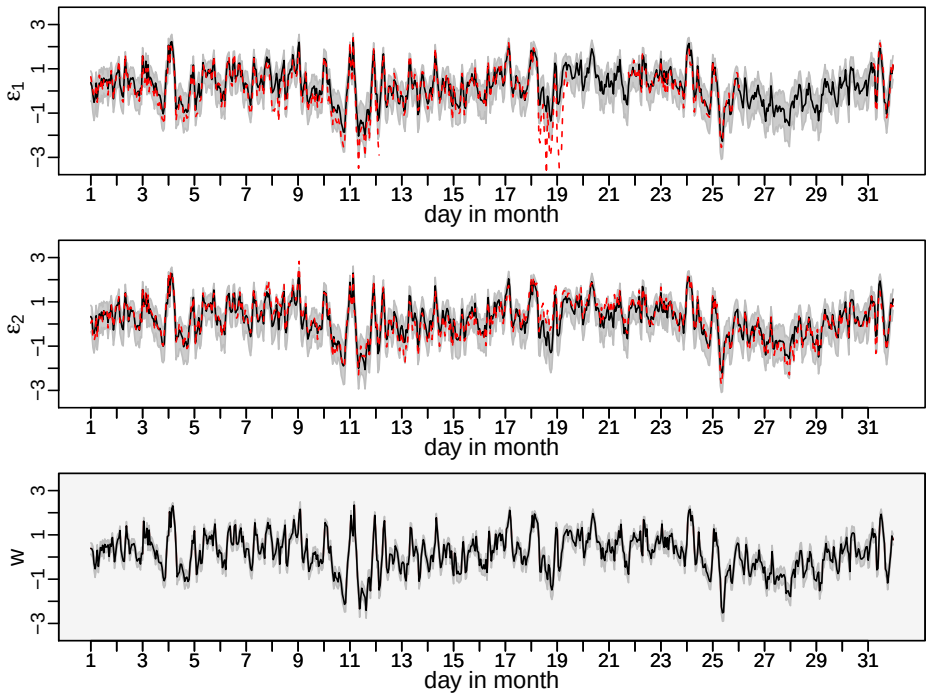

Figure 8: This plot is based on data for July. We show the estimated posterior mode of the predictive distribution of the standardized error ϵ_{tj} plotted against time t for $j = 1, 2$, corresponding to CO(gt) and CO(lc). Draws of the predictive distribution of the error are obtained as $\epsilon^r_{tj} = \Phi^{-1}(u^r_{tj})$, where u^r_{tj} is a sample from the predictive distribution on the copula scale. The observed standardized residual from the GAM is added in red (dashed). In addition, we show the estimated posterior mode of $w_t = \Phi^{-1}(v_t)$ plotted against t in the third row. To all plots we add a 90% credible region constructed from the estimated 5% and 95% posterior quantiles.

4.4 Predictions

We evaluate the proposed model's ability to predict the ground truth values. Therefore we compare the copula state space model to a Gaussian state space model and to Bayesian additive regression trees (Chipman et al (2010)), as a representative for a popular machine learning algorithm. Compared to other machine learning techniques, Bayesian additive regression trees have the advantage that a predictive distribution is obtained instead of a single point estimate. Therefore we can compare models with respect to their forecast distribution, for which we utilize the continuous ranked probability score (Gneiting and Raftery (2007)). The continuous ranked probability score (CRPS) for an observed value $y \in \mathbb{R}$ and a univariate forecast CDF F is defined as

$$CRPS = \int_{\mathbb{R}} (F(z) - 1_{y \leq z})^2 dz. \tag{28}$$

For each of the ground truth values we remove the observations in the last month of the data set and treat them as missing values, which yields the training set. Based on the training set we proceed similarly to what we described above, i.e. we first estimate the GAMs, and then estimate the state space model on the copula scale. Here two state space models are estimated: the copula state space model where the family set $\mathcal{M}$ is chosen as in Section 4.3 and the Gaussian state space model where we restrict the family set to $\mathcal{M} = \{\text{Gaussian}\}$. For each of the two state space models we obtain 2000 simulations from the in-sample predictive distribution u^r_{tj}, $r = 1, \ldots, 2000$, whereas our MCMC approach of Section 3 is run for 3000 iterations and the first 1000 draws are discarded for burn-in. Here t is a timepoint which is among the newly selected missing values for the ground truth value that corresponds to margin j. Based on these simulations we obtain simulations from the predictive distribution of the

Box-Cox transformed response as follows

$$(y_{tj}^{bc})^r = \hat{f}_j(\boldsymbol{x_t}) + \hat{\sigma}_j \Phi^{-1}(u_{tj}^r) \tag{29}$$

for $r = 1, \ldots, 2000$.

Since Bayesian additive regression trees rely on the normal distribution, we expect that Box-Cox transformations might also improve the fit for this model. We assume that

$$BC(Y_{tj}, \lambda_j) = g_j(\boldsymbol{x_{tj}^{BART}}) + \sigma_j \epsilon_{tj}, \tag{30}$$

where $\boldsymbol{x_{tj}^{BART}}$ are the covariates, $g_j(\cdot)$ is a sum of regression trees and $\epsilon_{tj} \sim N(0,1)$. In addition to the covariates used for the GAM model, all pollutant measurements except the one corresponding to margin j are included in the covariate vector $\boldsymbol{x_{tj}^{BART}}$. For λ_j we use the same value as for the previously fitted GAM. We have seen that this transformation improves the performance of the Bayesian additive regression trees. McCulloch et al (2018) implement a MCMC sampler in the `R` package `BART` which we use to obtain draws g_j^r, σ_j^r, $r = 1, \ldots, 10000$ from the corresponding posterior distribution. We discard the first 5000 of these draws and then 5000 simulations of the predictive distribution of the response are obtained as

$$(y_{tj}^{bc})^r \sim N(g_j^r(\boldsymbol{x_{tj}^{BART}}), (\sigma_j^r)^2) \tag{31}$$

for $r = 1, \ldots, 5000$.

Based on the simulations $(y_{tj}^{bc})^r, r = 1, \ldots, 5000$ we calculate the empirical CDF and use this to approximate the CRPS (this is implemented in the `R` package `scoringRules` of Jordan et al (2017)) for the different time points and sum them up to obtain the cumulative CRPS. For each of the three methods, we obtain a cumulative CRPS for each the three ground truth indices. In addition, we consider reduced bivariate data sets, where each data set consists of the ground truth value of a pollutant, the corresponding low-cost value and the covariates as in Section 4.2. This yields three reduced data sets, each associated with one of the three pollutants. For each of the reduced data sets we proceed as above, i.e. we first remove ground truth observations in the last month, fit the three different models and calculate the CRPS values.

We refer to the models fitted to the reduced data as bivariate state space models and reduced Bayesian additive regression trees. The models estimated with the full data are referred to as joint models. We want to investigate how the bivariate state space models compare to the six-dimensional ones. The cumulative CRPS values are shown in Table 3. For the pollutant NOx, the state space approach seems not to be the best choice. We have seen (see supplementary material) that for this pollutant, the dependence between the ground truth and the low-cost values varies more over time than for the other pollutants. Relaxing the assumption of a time-constant Kendall's τ, might improve the predictive accuracy for this pollutant. This model extension is subject to future research. Overall, the copula state space model is the best performing model within this comparison, since it outperforms the Gaussian state space model and the Bayesian additive regression trees in two out of three cases.

	CO	NOx	NO2
joint copula state space model	**74.27**	594.50	**569.03**
bivariate copula state space model	84.92	559.22	845.95
joint Gaussian state space model	76.91	594.30	570.55
bivariate Gaussian state space model	87.90	559.18	844.64
joint Bayesian additive regression trees	183.49	**379.31**	1330.90
reduced Bayesian additive regression trees	89.39	520.93	1095.40

Table 3: Cumulative CRPS for the three ground truth values (CO, NOx, NO2) obtained from six different models: joint/bivariate copula state space model, joint/bivariate Gaussian state space model, joint/reduced Bayesian additive regression trees. The best, i.e. the lowest, cumulative CRPS value in marked in bold.

5 Concluding Remarks

We proposed a multivariate nonlinear non-Gaussian copula-based state space model. The model is very flexible: the observation and the state equation are specified with copulas and the

model can be combined with different marginal distributions. We illustrated the model with air pollution measurements data and have shown that the novel copula state space model outperforms a linear Gaussian state space model and Bayesian additive regression trees. As we have seen in Section 4.4, the assumption of a time-constant dependence structure might not always be appropriate. A first extension of the model could allow for dynamic dependence parameters. For this, ideas of the dynamic bivariate copula model of Almeida and Czado (2012) might be used. Another area of future research is the extension to multiple factors.

References

Aas K, Czado C, Frigessi A, Bakken H (2009) Pair-copula constructions of multiple dependence. Insurance: Mathematics and Economics 44(2):182–198

Almeida C, Czado C (2012) Efficient Bayesian inference for stochastic time-varying copula models. Computational Statistics & Data Analysis 56(6):1511–1527

Bedford T, Cooke RM, et al (2002) Vines–a new graphical model for dependent random variables. The Annals of Statistics 30(4):1031–1068

Box GE, Cox DR (1964) An analysis of transformations. Journal of the Royal Statistical Society: Series B (Methodological) 26(2):211–243

Van den Brakel J, Roels J, et al (2010) Intervention analysis with state-space models to estimate discontinuities due to a survey redesign. The Annals of Applied Statistics 4(2):1105–1138

Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P, Riddell A (2017) Stan: A probabilistic programming language. Journal of statistical software 76(1)

Chen S, Fricks J, Ferrari MJ (2012) Tracking measles infection through non-linear state space models. Journal of the Royal Statistical Society: Series C (Applied Statistics) 61(1):117–134

Cherubini U, Luciano E, Vecchiato W (2004) Copula methods in finance. John Wiley & Sons

Chipman HA, George EI, McCulloch RE, et al (2010) BART: Bayesian additive regression trees. The Annals of Applied Statistics 4(1):266–298

Czado C (2019) Analyzing dependent data with vine copulas: a practical guide with R. Springer

De Vito S, Massera E, Piga M, Martinotto L, Di Francia G (2008) On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. Sensors and Actuators B: Chemical 129(2):750–757

De Vito S, Piga M, Martinotto L, Di Francia G (2009) CO, NO2 and NOx urban pollution monitoring with on-field calibrated electronic nose by automatic Bayesian regularization. Sensors and Actuators B: Chemical 143(1):182–191

De Vito S, Fattoruso G, Pardo M, Tortorella F, Di Francia G (2012) Semi-supervised learning techniques in artificial olfaction: A novel approach to classification problems and drift counteraction. IEEE Sensors Journal 12(11):3215–3224

Durbin J, Koopman SJ (2000) Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62(1):3–56

Durbin J, Koopman SJ (2002) A simple and efficient simulation smoother for state space time series analysis. Biometrika 89(3):603–616

Durbin J, Koopman SJ (2012) Time series analysis by state space methods, vol 38. Oxford University Press

Frühwirth-Schnatter S (1994) Data augmentation and dynamic linear models. Journal of Time Series Analysis 15(2):183–202

Genest C, Favre AC (2007) Everything you always wanted to know about copula modeling but were afraid to ask. Journal of Hydrologic Engineering 12(4):347–368

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102(477):359–378

Hafner CM, Manner H (2012) Dynamic stochastic copula models: Estimation, inference and applications. Journal of Applied Econometrics 27(2):269–295

Hastie T, Tibshirani R (1987) Generalized additive models: some applications. Journal of the American Statistical Association 82(398):371–386

Hoffman MD, Gelman A (2014) The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. Journal of Machine Learning Research 15(1):1593–1623

Hutchinson CE (1984) The Kalman filter applied to aerospace and electronic systems. IEEE Transactions on Aerospace and Electronic Systems (4):500–504

Ippoliti L, Valentini P, Gamerman D (2012) Space–time modelling of coupled spatiotemporal environmental variables. Journal of the Royal Statistical Society: Series C (Applied Statistics) 61(2):175–200

Joe H (1996) Families of m-variate distributions with given margins and m (m-1)/2 bivariate dependence parameters. Lecture Notes-Monograph Series pp 120–141

Joe H (1997) Multivariate models and multivariate dependence concepts. CRC Press

Johns CJ, Shumway RH (2005) A non-linear and non-Gaussian state-space model for censored air pollution data. Environmetrics: The official journal of the International Environmetrics Society 16(2):167–180

Jondeau E, Rockinger M (2006) The copula-GARCH model of conditional dependencies: An international stock market application. Journal of International Money and Finance 25(5):827–853

Jordan A, Krüger F, Lerch S (2017) Evaluating probabilistic forecasts with scoringRules. arXiv preprint arXiv:170904743

Kalman RE (1960) A new approach to linear filtering and prediction problems. Journal of Basic Engineering 82(1):35–45

Kalman RE, Bucy RS (1961) New results in linear filtering and prediction theory. Journal of Basic Engineering 83(1):95–108

Kitagawa G, Gersch W (1984) A smoothness priors–state space modeling of time series with trend and seasonality. Journal of the American Statistical Association 79(386):378–389

Kreuzer A, Valle LD, Czado C (2019) A Bayesian Non-linear State Space Copula Model to Predict Air Pollution in Beijing. arXiv preprint arXiv:190308421

Liu Z, Guo W (2015) Modeling diurnal hormone profiles by hierarchical state space models. Statistics in Medicine 34(24):3223–3234

Lopes HF, West M (2004) Bayesian model assessment in factor analysis. Statistica Sinica 14(1):41–68

McCulloch R, Sparapani R, Gramacy R, Spanbauer C, Pratola M (2018) BART: Bayesian additive regression trees. R package version 1

Myers KL, Brockwell AE, Eddy WF (2007) State-space models for optical imaging. Statistics in Medicine 26(21):3862–3874

Neal RM, et al (2011) MCMC using Hamiltonian dynamics. Handbook of Markov Chain Monte Carlo 2:113–162

Nelsen RB (2007) An introduction to copulas. Springer Science & Business Media

Patton AJ (2006) Modelling asymmetric exchange rate dependence. International Economic Review 47(2):527–556

Shumway RH, Stoffer DS (1982) An approach to time series smoothing and forecasting using the EM algorithm. Journal of Time Series Analysis 3(4):253–264

Sklar M (1959) Fonctions de repartition an dimensions et leurs marges. Publ inst statist univ Paris 8:229–231

Smith M, Min A, Almeida C, Czado C (2010) Modeling longitudinal data using a pair-copula decomposition of serial dependence. Journal of the American Statistical Association 105(492):1467–1479

Wood S, Wood MS (2015) Package 'mgcv'. R package version 1:29

Supplementary material

1 Additional Material for Section 2.2

The covariance matrix Σ of the joint distribution

$$(Z_{11}, \ldots, Z_{d1}, W_1; Z_{12}, \ldots, Z_{d2}, W_2; \ldots, Z_{1T}, \ldots, Z_{dT}, W_T) \sim N_{(d+1)T}(\mathbf{0}, \Sigma)$$

takes the form

$$\Sigma = \begin{pmatrix} A & \rho_{lat}(A+B) & \rho_{lat}^2(A+B) & \ldots & \rho_{lat}^{T-1}(A+B) \\ \rho_{lat}(A+B) & A & \rho_{lat}(A+B) & \ldots & \rho_{lat}^{T-2}(A+B) \\ \rho_{lat}^2(A+B) & \rho_{lat}(A+B) & A & \ldots & \rho_{lat}^{j-2}(A+B) \\ \rho_{lat}^3(A+B) & \rho_{lat}^2(A+B) & \rho_{lat}(A+B) & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho_{lat}^{T-1}(A+B) & \rho_{lat}^{T-2}(A+B) & \rho_{lat}^{T-3}(A+B) & \ldots & A \end{pmatrix}$$

where the matrices A and B take the following forms

$$A = \begin{pmatrix} 1 & \rho_{obs,1}\rho_{obs,2} & \rho_{obs,1}\rho_{obs,3} & \cdots & \rho_{obs,1}\rho_{obs,d} & \rho_{obs,1} \\ \rho_{obs,1}\rho_{obs,2} & 1 & \rho_{obs,2}\rho_{obs,3} & \cdots & \rho_{obs,2}\rho_{obs,d} & \rho_{obs,2} \\ \rho_{obs,1}\rho_{obs,3} & \rho_{obs,2}\rho_{obs,3} & 1 & \cdots & \rho_{obs,3}\rho_{obs,d} & \rho_{obs,3} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_{obs,1}\rho_{obs,d} & \rho_{obs,2}\rho_{obs,d} & \rho_{obs,3}\rho_{obs,d} & \cdots & 1 & \rho_{obs,d} \\ \rho_{obs,1} & \rho_{obs,2} & \rho_{obs,3} & \cdots & \rho_{obs,d} & 1 \end{pmatrix}$$

$$B = \begin{pmatrix} \rho_{obs,1}^2 - 1 & 0 & 0 & \ldots & 0 & 0 \\ 0 & \rho_{obs,2}^2 - 1 & 0 & \ldots & 0 & 0 \\ 0 & 0 & \rho_{obs,3}^2 - 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & \rho_{obs,d}^2 - 1 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 \end{pmatrix}.$$

2 Additional Material for Section 4.3

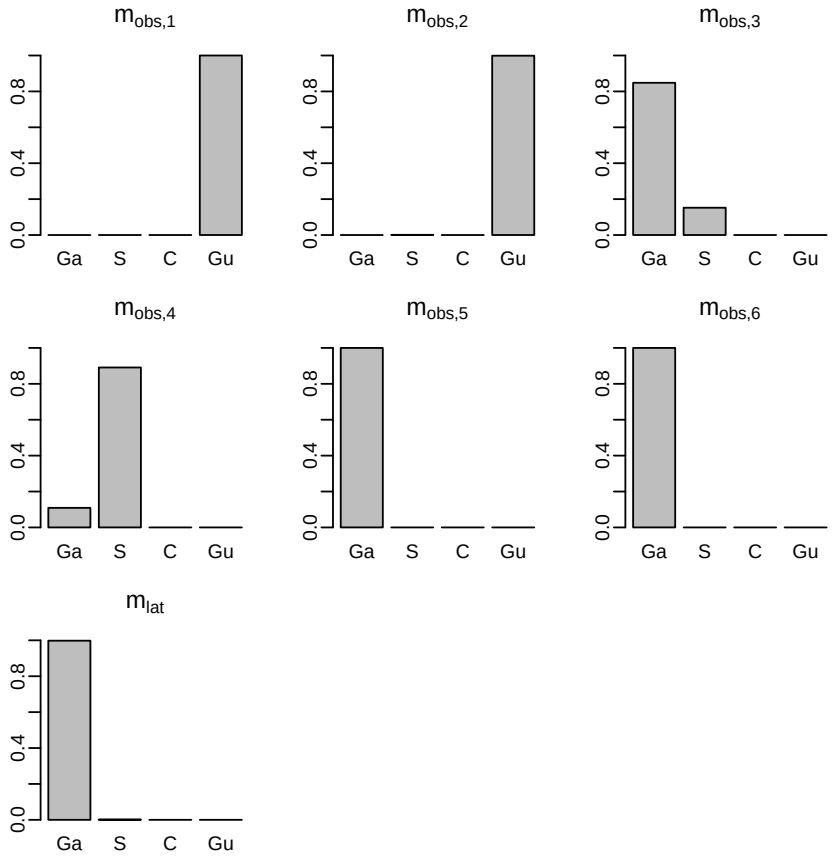

Figure 9: Estimated posterior distribution of the copula family indicators $m_{obs1}, \ldots, m_{obs,6}, m_{lat}$ obtained from 2000 iterations after a burn-in of 1000.

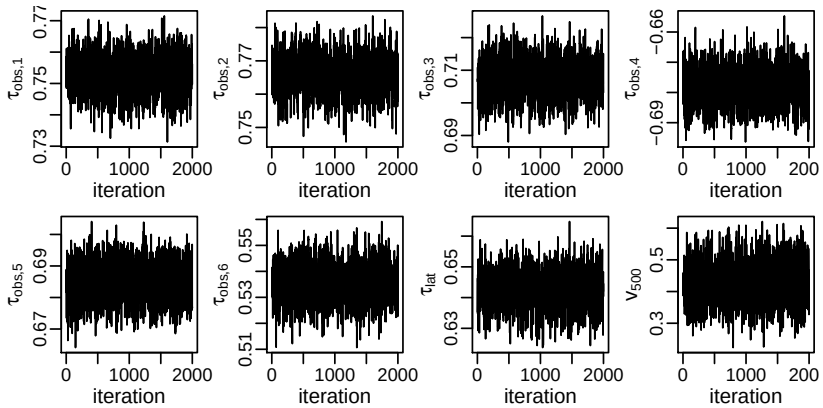

Figure 10: Trace plots of 2000 draws after a burn-in of 1000 for selected parameters of the copula state space model. The variables are ordered as follows: 1: CO(gt), 2: CO(lc), 3: NOx(gt), 4: NOx(lc), 5: NO2(gt), 6: NO2(lc).

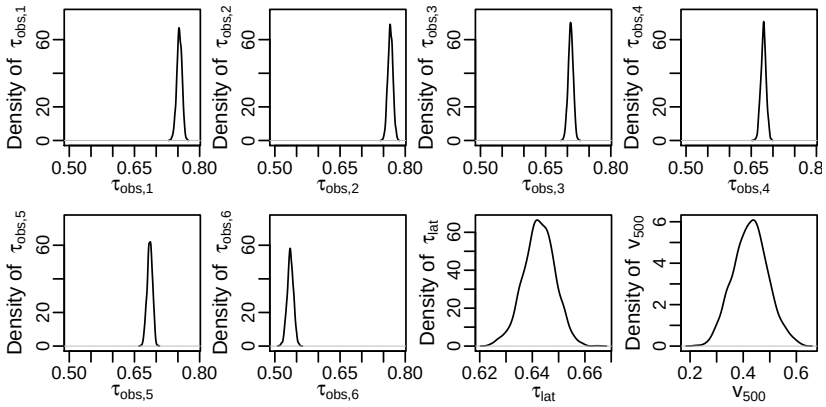

Figure 11: Estimated posterior density for selected parameters of the copula state space model. The posterior density is estimated as the kernel density estimate based on 2000 draws after a burn-in of 1000. For better comparability we multiplied the draws of $\tau_{obs,4}$ by -1. The variables are ordered as follows: 1: CO(gt), 2: CO(lc), 3: NOx(gt), 4: NOx(lc), 5: NO2(gt), 6: NO2(lc).

3 Additional Material for Section 4.4

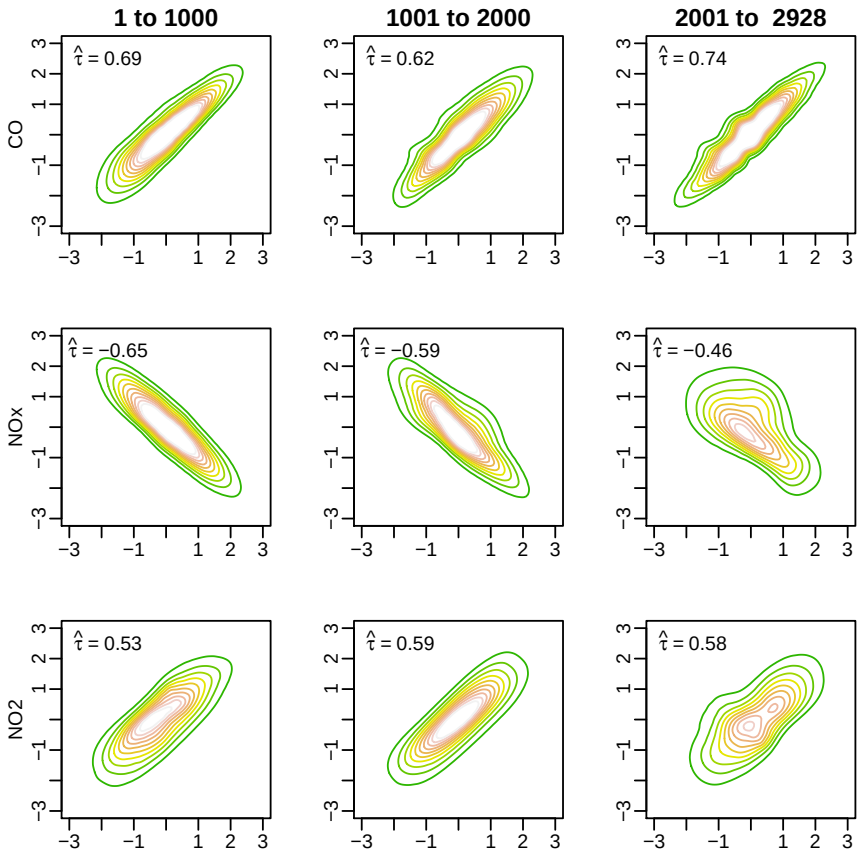

Figure 12: Contour plots for pairs $(\hat{z}_{tj}, \hat{z}_{tj'})_{t\in P_i}$, $i = 1, 2, 3$ where j corresponds to a ground truth and j' to the corresponding low-cost value within a time period P_i ($P_1 : 1, \ldots, 1000$, $P_2 : 1001, \ldots, 2000$, $P_3 : 2001, \ldots, 2928$). For example the top row shows contour plots for the (CO(gt), CO(lc)) pair for the three different time periods. In the top left corner we added the corresponding empirical Kendall's τ, based on the data $(\hat{z}_{tj}, \hat{z}_{tj'})_{t\in P_i}$.