Bayesian Analysis (2019) **14**, Number 4, pp. 1143–1171

Implicit Copulas from Bayesian Regularized Regression Smoothers

Nadja Klein[*] and Michael Stanley Smith[†]

Abstract. We show how to extract the implicit copula of a response vector from a Bayesian regularized regression smoother with Gaussian disturbances. The copula can be used to compare smoothers that employ different shrinkage priors and function bases. We illustrate with three popular choices of shrinkage priors—a pairwise prior, the horseshoe prior and a g prior augmented with a point mass as employed for Bayesian variable selection—and both univariate and multivariate function bases. The implicit copulas are high-dimensional, have flexible dependence structures that are far from that of a Gaussian copula, and are unavailable in closed form. However, we show how they can be evaluated by first constructing a Gaussian copula conditional on the regularization parameters, and then integrating over these. Combined with non-parametric margins the regularized smoothers can be used to model the distribution of non-Gaussian univariate responses conditional on the covariates. Efficient Markov chain Monte Carlo schemes for evaluating the copula are given for this case. Using both simulated and real data, we show how such copula smoothing models can improve the quality of resulting function estimates and predictive distributions.

Keywords: distributional regression, horseshoe prior, penalized splines, radial basis, regression splines.

1 Introduction

A popular way to estimate a smooth unknown function from noisy data is to approximate it with a linear combination of basis functions in a regression with coefficients that are regularized (Ruppert et al., 2003). We refer to such an approximation as a regularized regression smoother. In a Bayesian context, the regularization term arises from adopting a shrinkage prior for the coefficients. When the response is Gaussian, conditional on the signal, it is common to adopt a conditionally Gaussian shrinkage prior. Examples include (but are not limited to) the pairwise priors of penalized splines (Lang and Brezger, 2004), the horseshoe prior (Carvalho and Polson, 2010), and Bayesian variable selection priors (Clyde and George, 2004). In this paper we show how to extract the 'implicit copula' of the distribution of a response vector from such a regularized regression smoother. This captures the dependence structure between the elements of the vector. It can be used to compare the smoothing properties of different combinations of priors and function bases. Moreover, it can also be combined with non-parametric marginal distributions to create new regularized regression smoothers for non-Gaussian data. We

[*] Nadja Klein is an Assistant Professor at the Humboldt University of Berlin. This work was completed while she was an Alexander von Humboldt Feodor-Lynen fellow at the University of Melbourne

[†] Michael Smith is Professor of Management (Econometrics) at the Melbourne Business School, University of Melbourne, mike.smith@mbs.edu

https://doi.org/10.1214/18-BA1138

call these 'copula smoothers', and they have exactly the same dependence structure as that of the original smoother, but are substantially more flexible. As such, they provide an alternative approach to semi-parametric distributional regression (Rigby and Stasinopoulos, 2005; Klein et al., 2015; Wood et al., 2016) for a univariate response.

An implicit copula—also called an 'inversion copula' by Smith and Maneesoonthorn (2018)—is constructed from a random vector $\tilde{\boldsymbol{Z}}$ with distribution function $F_{\tilde{Z}}$, by inverting the usual expression of Sklar's theorem; see Nelson (2006, Section 3.1). For example, the simplest implicit copula is the Gaussian copula, which is obtained when $F_{\tilde{Z}}$ is Gaussian (Song, 2000). In this paper, $\tilde{\boldsymbol{Z}}|\boldsymbol{x}$ is a vector of observations on the response from a regularized regression model, where $\boldsymbol{x}$ are the observed covariate values. If the joint and marginal distribution functions are $F_{\tilde{Z}}(\tilde{\boldsymbol{z}}|\boldsymbol{x})$ and $F_{\tilde{Z}_i}(\tilde{z}_i|\boldsymbol{x})$, respectively, and $\boldsymbol{u} = (u_1, \ldots, u_n)'$, then the resulting implicit copula function for n observations is

$$C_\pi(\boldsymbol{u}|\boldsymbol{x}) = F_{\tilde{Z}}\left(F^{-1}_{\tilde{Z}_1}(u_1|\boldsymbol{x}), \ldots, F^{-1}_{\tilde{Z}_n}(u_n|\boldsymbol{x})|\boldsymbol{x}\right), \qquad (1)$$

which is itself a function of $\boldsymbol{x}$. Throughout the paper we refer to $\tilde{\boldsymbol{Z}}$ as a vector of observations on a 'pseudo-response', because it is not observed directly.

To construct our copula, we first derive the implicit copula of $\tilde{\boldsymbol{Z}}|\boldsymbol{x}$ with the basis coefficients integrated out, but conditional on the regularization parameters. This is a Gaussian copula with a correlation matrix that is a function of $\boldsymbol{x}$ and the regularization parameters. The latter can include parameters that allow the basis to be of varying dimension. We then integrate over the distribution of the regularization parameters, denoted as π, to obtain the desired implicit copula C_π of the regularized regression smoother, which is unavailable in closed form. In a Bayesian context, the integration can be done with respect to either the prior or posterior of the regularization parameters. In either case, we stress here that the resulting implicit copula has a dependence structure that is very different from that of a Gaussian copula – something we illustrate in our empirical work. The implicit copula density can be expressed as an integral that can be computed readily using Bayesian methods – even when the dimension of the copula is high. This approach greatly simplifies computation of the implicit copula compared to direct evaluation of (1) as suggested by Smith and Maneesoonthorn (2018).

Three shrinkage priors for the basis coefficients are considered in detail: an autoregressive prior, a horseshoe prior and a g prior augmented with point mass. These are combined with a number of matching bases, including B-spline, augmented Fourier and regression spline bases for univariate functions, and additive or radial bases for multivariate functions. We show how to compute dependence metrics (such as Spearman's rho or quantile dependence) between the response variable at two different covariate values. Varying these covariate values produces a surface of dependence metric values that characterize the level of smoothing of the regression smoother. Different combinations of shrinkage prior and basis result in large differences between these surfaces. The surfaces provide a tool for comparing the level of smoothing from the different regularized regressions, which is otherwise difficult.

The proposed implicit copula is used to model the dependence between the elements of a vector of response values $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$, conditional on the covariate values $\boldsymbol{x}$. The margin F_Y of Y_i is modeled non-parametrically, while all regression smoothing is through the copula function $C_\pi(\boldsymbol{u}|\boldsymbol{x})$ only, which is why we call it a copula smoother. For this case, efficient Markov chain Monte Carlo (MCMC) schemes to estimate the posteriors are outlined for each choice of shrinkage prior. We show how to estimate the regression function, which is the expectation of Y_i conditional on the covariate values $\boldsymbol{x}$. We also show how to compute the Bayesian predictive density of Y_i conditional on $\boldsymbol{x}$. A simulation study illustrates the effectiveness of the copula smoother for function and predictive density estimation.

The approach can be extended to multiple covariates in two ways. First, we construct the implicit copula of an additive regularized regression smoother for $\tilde{Z}_i$. However, the copula smoother for response Y_i that uses this copula is not additive in the covariates. Therefore, the usual partial residuals (Hastie and Tibshirani, 1990) cannot be computed, and we show how to compute both function estimates and partial residuals on the domain of the pseudo-response instead. To illustrate, the model is applied to the widely studied Boston housing data. The copula smoother captures the non-Gaussian marginal distribution, increasing accuracy of the predictive density. Our second approach, is to construct the implicit copula when using a radial basis for the mean of $\tilde{Z}_i$. We show how to do this in Part A of the Supplementary Material to this paper (Klein and Smith, 2018), and apply it to an example with $n = 11{,}375$ observations, demonstrating the viability of using the copula smoother when the n-dimensional implicit copula is of high dimension.

The rest of this paper is structured as follows. Section 2 outlines the implicit copula; both in general and for the three regularization priors considered in detail. Section 3 employs the proposed copula with arbitrary margins to construct a copula smoother for non-Gaussian data. Section 4 contains the simulation study. Section 5 extends our copula to additive bases, and illustrates using the Boston housing data. Section 6 concludes. The Supplementary Materials contain extensive additional material, including tables and figures referred to in the text with prefix 'S'.

2 Implicit Copula

In this section, we explain our approach for constructing the copula of a regularized regression model for the pseudo-response with a single covariate. It is extended to the case of multiple covariates in Section 5 and Part A of the Supplementary Material.

2.1 The General Idea

Consider the regression model

$$\tilde{Z}_i = \tilde{m}(x_i) + \varepsilon_i,\ \text{for } i = 1, \dots, n \tag{2}$$

for a pseudo-response $\tilde{Z}_i$, where $\tilde{m}$ is an unknown univariate function, x_i is a covariate value, and ε_i is distributed independently $N(0, \sigma^2)$. It is popular to model $\tilde{m}$ as a linear combination of p basis functions $b_1, \dots, b_p$, so that $\tilde{m}(x) = \sum_{j=1}^{p} \beta_j b_j(x)$. In this case, (2) can be rewritten as the linear model

$$\tilde{\boldsymbol{Z}} = B\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3}$$

for $\tilde{\boldsymbol{Z}} = (\tilde{Z}_1, \ldots, \tilde{Z}_n)' \in \mathbb{R}^n$, with $\boldsymbol{\varepsilon} \sim \mathrm{N}(0, \sigma^2 I)$. The $(n \times p)$ design matrix B has ith row $\boldsymbol{b}_i' = (b_1(x_i), \ldots, b_p(x_i))$ evaluated at x_i. There are many bases used in practice, and we consider three common choices here: regression splines, B-splines and an augmented Fourier basis.

In the Bayesian literature, priors are employed on $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ to allow for a data-driven level of shrinkage to provide a smooth, but flexible, estimate of $\tilde{m}$. We follow this approach and employ the prior

$$\boldsymbol{\beta} | \boldsymbol{x}, \sigma^2, \boldsymbol{\theta}, \boldsymbol{\gamma} \sim N(\boldsymbol{0}, \sigma^2 P(\boldsymbol{\theta})^{-1}), \tag{4}$$

where the precision matrix $P(\boldsymbol{\theta})$ is of full rank. The parameters $\boldsymbol{\theta}$ are shrinkage parameters, while $\boldsymbol{\gamma}$ are further parameters that allow for the basis to be of varying dimension (which we discuss later). The matrix P may also be a function of the covariate vector $\boldsymbol{x} = (x_1, \ldots, x_n)'$. In Section 2.2 we consider three different priors of this form. Conditional on $(\boldsymbol{x}, \sigma^2, \boldsymbol{\theta}, \boldsymbol{\gamma})$, $\boldsymbol{\beta}$ can be integrated out to give

$$\tilde{\boldsymbol{Z}} | \boldsymbol{x}, \sigma^2, \boldsymbol{\theta}, \boldsymbol{\gamma} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2 (I - B\Omega B')^{-1}), \tag{5}$$

with $\Omega = (B'B + P(\boldsymbol{\theta}))^{-1}$. Application of the Woodbury formula gives $(I - B\Omega B')^{-1} = I + BP(\boldsymbol{\theta})^{-1}B'$, with ith diagonal element equal to $1 + \boldsymbol{b}_i' P(\boldsymbol{\theta})^{-1} \boldsymbol{b}_i$. Therefore, the ith margin of this distribution is $\tilde{Z}_i | \boldsymbol{x}, \sigma^2, \boldsymbol{\theta}, \boldsymbol{\gamma} \sim \mathrm{N}(0, \sigma^2(1 + \boldsymbol{b}_i' P(\boldsymbol{\theta})^{-1} \boldsymbol{b}_i))$.

The copula of the distribution at (5) is called a Gaussian copula (Song, 2000), and is constructed by standardizing the distribution to have zero mean and unit variances. To do so here, we set $\boldsymbol{Z} = \sigma^{-1} S(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\gamma}) \tilde{\boldsymbol{Z}}$, where $S(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = \mathrm{diag}(s_1, \ldots, s_n)$ is a diagonal scaling matrix with elements $s_i = \left[1 + \boldsymbol{b}_i' P(\boldsymbol{\theta})^{-1} \boldsymbol{b}_i\right]^{-1/2}$. With this standardization, the regression at (2) can be rewritten as

$$Z_i = m(x_i) + \frac{s_i}{\sigma} \varepsilon_i, \tag{6}$$

where $m(x_i) = (s_i/\sigma) \boldsymbol{b}_i' \boldsymbol{\beta}$ and both s_i and $\boldsymbol{b}_i$ are functions of x_i. The conditional distribution of the standardized vector $\boldsymbol{Z} = (Z_1, \ldots, Z_n)'$ is then

$$\boldsymbol{Z} | \boldsymbol{x}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \boldsymbol{\gamma} \sim \mathrm{N}\left(\frac{S(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\gamma})}{\sigma} B\boldsymbol{\beta}, S(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\gamma}) S(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\gamma})' \right). \tag{7}$$

Integrating out $\boldsymbol{\beta}$ as before, gives the unconditional (on $\boldsymbol{\beta}$) distribution of $\boldsymbol{Z}$, which we summarize in the following Theorem.

Theorem 1. Let $\tilde{\boldsymbol{Z}}$ follow the linear model at (3), with the prior for $\boldsymbol{\beta}$ as given at (4). Then:

(i) The joint distribution $\boldsymbol{Z} | \boldsymbol{x}, \sigma^2, \boldsymbol{\theta}, \boldsymbol{\gamma} \sim N(\boldsymbol{0}, R(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\gamma}))$ with

$$R(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = S(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\gamma})(I - B\Omega B')^{-1} S(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\gamma})'. \tag{8}$$

(ii) The marginal distributions $Z_i | \boldsymbol{x}, \sigma^2, \boldsymbol{\theta}, \boldsymbol{\gamma} \sim N(0, 1)$ for $i = 1, \ldots, n$.

(iii) The copulas of $\tilde{\boldsymbol{Z}}$ and $\boldsymbol{Z}$, conditional on $(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\gamma})$, are both the same Gaussian copula with copula function $C(\boldsymbol{u}|\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = \Phi_n\left(\Phi_1^{-1}(u_1), \ldots, \Phi_1^{-1}(u_n); \mathbf{0}, R(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\gamma})\right)$, where $\boldsymbol{u} = (u_1, \ldots, u_n)'$, while $\Phi_n(\cdot; \mathbf{0}, R)$ and Φ_1 are the distribution functions of $N_n(\mathbf{0}, R)$ and $N(0,1)$ distributions, respectively.

(iv) The corresponding copula density is

$$c(\boldsymbol{u}|\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = \frac{p(\boldsymbol{z}|\boldsymbol{x}, \sigma^2, \boldsymbol{\theta}, \boldsymbol{\gamma})}{\prod_{i=1}^n p(z_i|\boldsymbol{x}, \sigma^2, \boldsymbol{\theta}, \boldsymbol{\gamma})} = \frac{\phi_n(\boldsymbol{z}; \mathbf{0}, R(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\gamma}))}{\prod_{i=1}^n \phi_1(z_i)}, \tag{9}$$

where $z_i = \Phi_1^{-1}(u_i)$, $\boldsymbol{z} = (z_1, \ldots, z_n)'$ and $\phi_n(\cdot; \mathbf{0}, R)$ and ϕ_1 are the densities of $N_n(\mathbf{0}, R)$ and $N(0,1)$ distributions, respectively.

We make four observations concerning Theorem 1 above. First, σ^2 does not feature in the expression for the copula function or density and is therefore unidentified, so that we simply set it to 1 throughout the rest of the paper. This is because the copula is invariant to the scale of Z_i. Second, if a non-conjugate prior is used for $\boldsymbol{\beta}|\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\gamma}$, then the implicit copula above would not be a Gaussian copula. Third, if an improper prior is employed for $\boldsymbol{\beta}$—such as those popular in the Bayesian spline literature (Lang and Brezger, 2004)—then the distribution $\boldsymbol{Z}|\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\gamma}$ is also improper, and the copula is undefined. Therefore, we only employ strictly proper priors here. Last, while the copula is n-dimensional, the matrix R at (8) is a parsimonious function of $(\boldsymbol{\theta}, \boldsymbol{\gamma})$. In the next subsection we give expressions for R for the three shrinkage priors considered in detail.

While the copula at (9) is Gaussian, mixing over the distribution $\pi(\boldsymbol{\theta}, \boldsymbol{\gamma})$ results in a non-Gaussian copula that cannot in general be expressed in closed form, as summarized in the following corollary.

Corollary 1. *If $\tilde{\boldsymbol{Z}}$ follows the linear model at (3), with the prior for $\boldsymbol{\beta}$ given at (4), and $\pi(\boldsymbol{\theta}, \boldsymbol{\gamma})$ is a proper density, then*

$$c_\pi(\boldsymbol{u}|\boldsymbol{x}) = \int\int c(\boldsymbol{u}|\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\gamma})\pi(\boldsymbol{\theta}, \boldsymbol{\gamma})\mathrm{d}(\boldsymbol{\theta}, \boldsymbol{\gamma})$$

is also a copula density.

The proof of Corollary 1 can be found in Part B of the Supplementary Material. The corresponding copula function is denoted as $C_\pi(\boldsymbol{u}|\boldsymbol{x}) = \int\int C(\boldsymbol{u}|\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\gamma})\pi(\boldsymbol{\theta}, \boldsymbol{\gamma})\mathrm{d}(\boldsymbol{\theta}, \boldsymbol{\gamma})$. In this paper, we consider both the prior $\pi_0(\boldsymbol{\theta}, \boldsymbol{\gamma})$ and the posterior $p(\boldsymbol{\theta}, \boldsymbol{\gamma}|\boldsymbol{y})$ densities for $\pi(\boldsymbol{\theta}, \boldsymbol{\gamma})$. When a regularized smoother is fit to data, it is this mixture copula that captures the dependence structure of the resulting data distribution. Evaluation of (and generation from) c_π and C_π can be undertaken efficiently by Monte Carlo simulation, as we show later. It is C_π that we use in Section 3 to construct the new copula smoothers.

Representation of C_π as a mixture of Gaussian copulas greatly simplifies its computation. It makes computation of the copula much faster, as shown in Section 5 and Part A of the Supplementary Material for two high-dimensional examples. In contrast, C_π is much harder to compute via inversion of the distribution $\tilde{\boldsymbol{Z}}|\boldsymbol{x}$ directly, as in (1).

This is because the marginal distribution function of $\tilde{Z}_i|\boldsymbol{x}$ is

$$F_{\tilde{Z}_i}(\tilde{z}_i|\boldsymbol{x}) = \int \Phi_1\left(\tilde{z}_i; 0, (1 + \boldsymbol{b}_i' P(\boldsymbol{\theta})^{-1}\boldsymbol{b}_i)\right)\pi(\boldsymbol{\theta},\boldsymbol{\gamma})\mathrm{d}(\boldsymbol{\theta},\boldsymbol{\gamma}),$$

where the integral typically requires computation via numerical methods. The direct inversion approach requires evaluation of the quantile functions $\tilde{z}_i = F_{\tilde{Z}_i}^{-1}(u_i|\boldsymbol{x})$, for $i = 1, \ldots, n$, which is prohibitively slow for large sample sizes.

2.2 Three Implicit Copulas

We construct implicit copulas using three popular shrinkage priors for $\boldsymbol{\beta}$. Each prior is of the form at (4), and is usually matched with specific bases. We discuss each in further detail below and summarize them in Table S1 in the Supplementary Material.

P-Spline Copula (PSC) There is an extensive literature on Bayesian P-splines that employ differenced priors, also called random walk priors (Fahrmeir and Lang, 2001). However, these are improper, so that $\boldsymbol{Z}|\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{\gamma}$ with $\boldsymbol{\beta}$ integrated out is also, and the copula at (9) undefined. Therefore, we instead employ a first order stationary autoregression $\beta_i|\beta_{i-1} \sim N(\psi\beta_{i-1}, \tau^2)$, which approximates a first order random walk when $\psi \to 1$. For this prior, $\boldsymbol{\gamma} = \emptyset$, $\boldsymbol{\theta} = \{\psi, \tau\}$, and $P(\boldsymbol{\theta}) = (\tau^2)^{-1}P_0(\psi)$ is a full rank band one matrix. Following Lang and Brezger (2004), we match this prior with a B-spline basis of degree $l = 3$ (i.e. a cubic B-spline) with $m + 2l$ equally-spaced knots, where m is the number of inner knots. In our empirical work, we set m to values between 20 and 30, which is a typical choice, resulting in a dimension of $m + l - 1$ for $\boldsymbol{\beta}$.

For the prior $\pi_0(\boldsymbol{\theta})$ we assume ψ and τ^2 are independent, with $\psi \sim$ Uniform(0.01, 0.99), so that $P_0(\psi)$ is full rank and coefficients are positively correlated. Klein and Kneib (2016) study appropriate priors for τ^2, and we follow these authors and use a Weibull distribution with scale parameter $b_{\tau^2} = 2.5$. From Theorem 1, the correlation matrix

$$R(\boldsymbol{x},\boldsymbol{\theta}) = S(\boldsymbol{x},\boldsymbol{\theta})(I + \tau^2 B P_0(\psi)^{-1}B')S(\boldsymbol{x},\boldsymbol{\theta}),$$

and we label the implicit copula 'PSC'. In Section 2.3 we show that ψ and τ^2 control different aspects of the dependence structure. Last, we note that higher order autoregressive priors for $\boldsymbol{\beta}$ can also be used, similar to the popular higher order random walks (Fahrmeir and Kneib, 2011).

Horseshoe Copula (HSC) The horseshoe prior is attractive due to its robustness, local adaptivity and analytical properties (Carvalho and Polson, 2010). It is a scale mixture, where $\beta_j|\lambda_j \sim \mathrm{N}(0, \lambda_j^2)$, with prior $\pi_0(\lambda_j|\tau) = \text{Half-Cauchy}(0, \tau)$ and $\pi_0(\tau) = \text{Half-Cauchy}(0, 1)$. With this prior $\boldsymbol{\gamma} = \emptyset$, $\boldsymbol{\theta} = \{\boldsymbol{\lambda}, \tau\}$, with $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)'$, while

$$R(\boldsymbol{x},\boldsymbol{\theta}) = S(\boldsymbol{x},\boldsymbol{\theta})(I + B\,\mathrm{diag}(\lambda_1, \ldots, \lambda_p)^2 B')S(\boldsymbol{x},\boldsymbol{\theta}).$$

While we are unaware of any previous usage of the horseshoe prior for smoothing, the localized shrinkage of the prior makes it an attractive choice. Here, we employ the prior with two univariate bases. The first is the same B-spline basis employed for the PSC, while a second is the augmented Fourier basis of $2K$ basis terms $\{\sin(k\pi x), \cos(k\pi x); k = 1,\dots,K\}$, where the covariate is scaled to $[0,1]$ and we typically set $K = 10$ in our empirical work. We label this copula 'HSC'.

Bayesian Variable Selection Copula (BVSC) For this prior, $\boldsymbol{\theta} = \emptyset$, so that we drop reference to it when discussing this implicit copula. Spike-and-slab priors are popular in the Bayesian variable selection literature (Clyde and George, 2004). They allow for bases of varying dimension, with $\boldsymbol{\gamma} = (\gamma_1,\dots,\gamma_p)'$ a vector of binary indicators ($\gamma_i \in \{0,1\}$) denoting whether, or not, each basis term is included or omitted from p candidates. Let $p_\gamma = \sum_{i=1}^{p} \gamma_i$, and at (4) denote $\boldsymbol{\beta}, B$ and P as $\boldsymbol{\beta}_\gamma, B_\gamma$ and P_γ, respectively. We adopt the g prior for the included terms, where $\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma} \sim N(\mathbf{0}, P_\gamma^{-1})$, with $P_\gamma^{-1} = c(B_\gamma' B_\gamma)^{-1}$ and $c = 100$ as in Smith and Kohn (1996). Substituting P_γ into (8), the correlation matrix

$$R(\boldsymbol{x},\boldsymbol{\gamma}) = S(\boldsymbol{x},\boldsymbol{\gamma})(I + cB_\gamma(B_\gamma' B_\gamma)^{-1}B_\gamma')S(\boldsymbol{x},\boldsymbol{\gamma}),$$

$\Omega = \frac{c}{1+c}(B_\gamma' B_\gamma)^{-1}$, and $\boldsymbol{b}_{\gamma,i}$ is the ith row of B_γ. Note that for this prior $s_i = (1 + c\boldsymbol{b}_{\gamma,i}'(B_\gamma' B_\gamma)^{-1}\boldsymbol{b}_{\gamma,i})^{-1/2}$, and is a function of all elements of $\boldsymbol{x}$, not just x_i.

We use the prior mass function $\pi_0(\boldsymbol{\gamma}) =$ Beta$(p - p_\gamma + 1, p_\gamma + 1)$. This has been used extensively in the Bayesian selection literature (e.g. in Smith and Kohn (2002)), and accounts for the multiplicity of the 2^p possible configurations of $\boldsymbol{\gamma}$ (Scott and Berger, 2010). It implies a uniform distribution on $\pi_0(p_\gamma) = 1/(p+1)$ and Bernoulli margins $\Pr(\gamma_i = 1) = 1/2$. We employ this prior with the cubic regression spline basis $\{x, x^2, x^3, (x-k_1)_+^3, \dots, (x-k_K)_+^3\}$, where $\{a\}_+^3 = \min(0, a^3)$ and $k_1,\dots,k_K$ are knots chosen to follow the empirical distribution of the covariate with $K = 25$. We label this implicit copula 'BVSC'.

2.3 Dependence Structure

We use metrics of pairwise dependence from our copulas for two new observations to measure their dependence structure. Possible metrics include quantile dependence and Kendall's tau (Nelson, 2006, Chapter 5), but we illustrate here using Spearman correlation.

Consider two new covariate values $x_{0,1}, x_{0,2}$, and denote the vector of these two values combined with n existing covariate observations as $\boldsymbol{x}^+ = (x_{0,1}, x_{0,2}, \boldsymbol{x}')'$. If $\boldsymbol{u}^+ = (u_{0,1}, u_{0,2}, \boldsymbol{u}')'$, then from Theorem 1, $C(\boldsymbol{u}^+|\boldsymbol{x}^+,\boldsymbol{\theta},\boldsymbol{\gamma})$ is a Gaussian copula. If a random vector has this copula, then the Spearman correlation between its first two elements $Y_{0,1}$ and $Y_{0,2}$ is

$$\rho^s(x_{0,1}, x_{0,2}|\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{\gamma}) = \frac{6}{\pi}\arcsin(r_{12}(\boldsymbol{x}^+,\boldsymbol{\theta},\boldsymbol{\gamma})),$$

where $r_{12}(\boldsymbol{x}^+,\boldsymbol{\theta},\boldsymbol{\gamma})$ is the first off-diagonal element in the $(n+2)\times(n+2)$ matrix $R(\boldsymbol{x}^+,\boldsymbol{\theta},\boldsymbol{\gamma})$. For the PSC and HSC implicit copulas, it is straightforward to show that r_{12} is a function of only $(x_{0,1}, x_{0,2})$ and not $\boldsymbol{x}$, so that ρ^s is also. However, for the BVSC implicit copula r_{12} depends on all elements of $\boldsymbol{x}^+$ because each element of the diagonal scaling matrix S does so also. It is this feature that makes the smoothing locally adaptive for this copula, as discussed further in Section 2.4. Last, we write ρ^s as a function of $(x_{0,1}, x_{0,2})$ to underline that it is a function of these two values of the regression covariate.

The same dependence metrics for the mixture copula $C_\pi(\boldsymbol{u}^+|\boldsymbol{x}^+)$ at Corollary 1 can be computed via simulation. For example, the Spearman's pairwise correlation between $Y_{0,1}$ and $Y_{0,2}$ from this copula is

$$\begin{aligned}\rho^s_\pi(x_{0,1}, x_{0,2}|\boldsymbol{x}) &= \int \rho^s(x_{0,1}, x_{0,2}|\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\gamma})\pi(\boldsymbol{\theta}, \boldsymbol{\gamma})\mathrm{d}(\boldsymbol{\theta}, \boldsymbol{\gamma}) \\ &\approx \frac{1}{J}\sum_{j=1}^{J} \rho^s(x_{0,1}, x_{0,2}|\boldsymbol{x}, \boldsymbol{\theta}^{[j]}, \boldsymbol{\gamma}^{[j]}),\end{aligned}$$

where $(\boldsymbol{\theta}^{[j]}, \boldsymbol{\gamma}^{[j]})' \sim \pi(\boldsymbol{\theta}, \boldsymbol{\gamma})$ and J is the total number of iterates. Simulating from π is typically straightforward when it is the prior distribution, and can be achieved using the MCMC methods in Section 3 when it is the posterior.

2.4 Empirical Illustration of the Dependence Structure

To illustrate the dependence structure of our proposed copulas, we first consider the PSC with $\boldsymbol{\theta} = \{\psi, \tau^2\}$. Figure 1 shows ρ^s as a function of $(x_{0,1} - x_{0,2})$, where in panel (a) $\psi = 0.5$ and $\tau^2 \in \{0.01, 0.1, 0.5, 1, 10, 100\}$, and in panel (b) $\tau^2 = 1$ and $\psi \in \{0.1, 0.25, 0.5, 0.75, 0.9, 0.95\}$. This reveals that τ^2 determines the overall level of dependence between $Y_{0,1}$ and $Y_{0,2}$, while ψ determines how quickly ρ^s decreases as $|x_{0,1} - x_{0,2}|$ increases. The dependence is symmetric around $(x_{0,1} - x_{0,2}) = 0$.

We next compare the dependence structure of the three (non-Gaussian) implicit copulas $C_\pi(\boldsymbol{u}^+|\boldsymbol{x}^+)$, where the copula parameters are integrated out with respect to

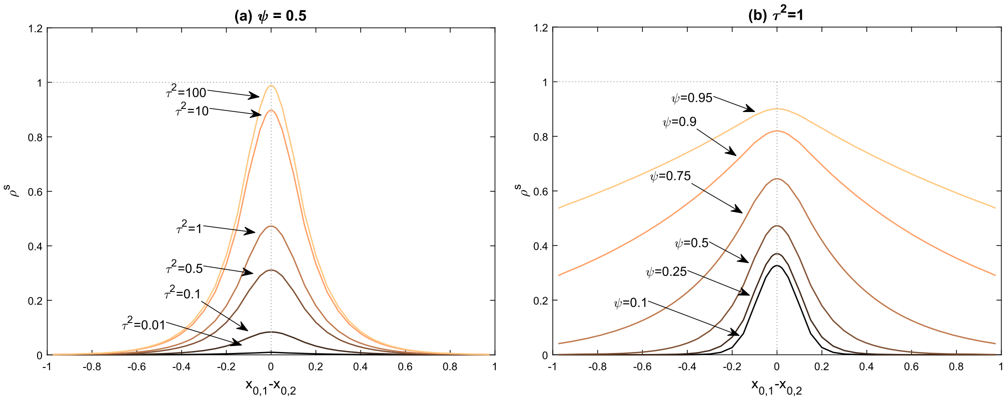

Figure 1: Spearman's rho $\rho^s(x_{0,1}, x_{0,2}|\boldsymbol{x}, \boldsymbol{\theta})$ plotted against $(x_{0,1} - x_{0,2})$ for the PSC with B-spline basis and conditional on $\boldsymbol{\theta}$. In panel (a), $\psi = 0.5$ and $\tau^2 \in \{0, 01, 0.1, 0.5, 1, 10, 100\}$. In panel (b), $\tau^2 = 1$ and $\psi \in \{0.1, 0.25, 0.5, 0.75, 0.9, 0.95\}$.

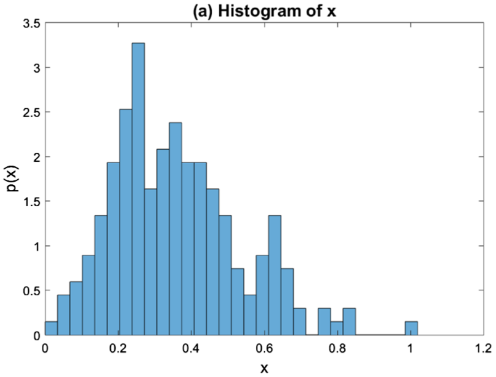

Figure 2: Normalized histogram of the $n = 200$ covariate values $x_1, \ldots, x_n$ from the empirical illustration in Sections 2.4 and 3.5.

the prior $\pi = \pi_0$. Because ρ^s_π is a function of $\boldsymbol{x}$ for the BVSC, $n = 200$ covariate values are generated from a χ^2 distribution and scaled to $[0, 1]$. Figure 2 shows a histogram of these values. We then compute $\rho^s_\pi(x_{0,1}, x_{0,2}|\boldsymbol{x})$ over a bivariate grid for $(x_{0,1}, x_{0,2})$ on the unit square, with $J = 10,000$ iterates simulated from the priors π_0 for each case. Figure 3 plots ρ^s_π as surfaces in the left-hand panels for four copula models: (a) PSC with a B-spline basis, (c) HSC with a B-spline basis, (e) HSC with an augmented Fourier basis, and (g) BVSC with a regression spline basis.

We make five observations. First, in each case ρ^s_π is highest as $|x_{0,1} - x_{0,2}| \to 0$, which is expected for an effective smoother, because response values should be more dependent when their covariate values are closer. Second, even though the function bases are identical in panels (a,c), the level of smoothing is higher with the PSC than HSC. Clearly, the prior for $\boldsymbol{\beta}$ has a strong impact on the dependence structure. Third, even though the prior for $\boldsymbol{\beta}$ is the same in panels (c,e), the bases employed differ, which also has a large effect on the dependence structure. Fourth, 'ripples' in ρ^s_π are observed for the augmented Fourier basis, which is because the basis terms are non-monotonic in $|x_{0,1} - x_{0,2}|$. Fifth, the BVSC is the only case where the n values of $\boldsymbol{x}$ have an impact on ρ^s_π, as seen in panel (g). Smoothing is higher for values of $x_{0,1}$ and $x_{0,2}$ close to 1, and lower for values around 0.3. This is 'local adaptivity' in the level of smoothing to the density of the covariate. We return to Figure 3 in Section 3.5, where we compare the surfaces against those constructed using the posterior of the copula parameters.

3 Copula Smoother for Non-Gaussian Data

The main application of our proposed copula is in conjunction with arbitrary marginal distributions to model non-Gaussian regression data. In this section we outline this model, and Bayesian methods to estimate the copula parameters, regression function and predictive distributions.

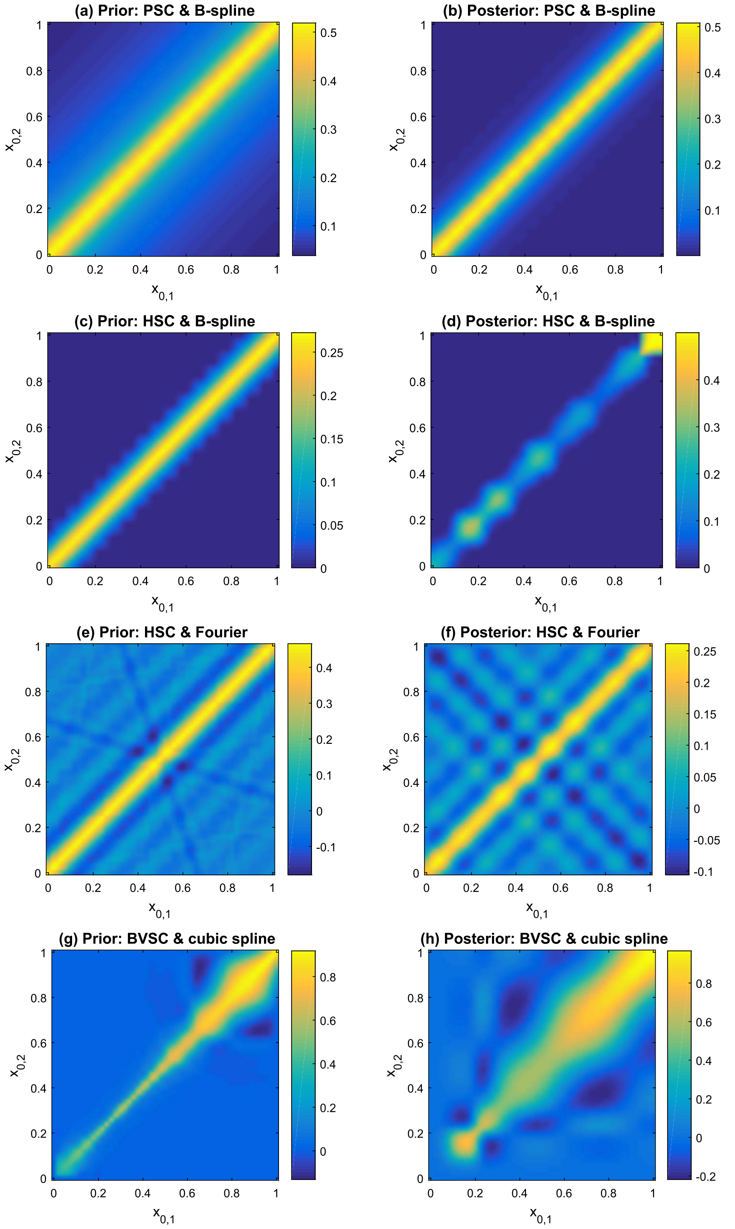

Figure 3: Bivariate surfaces of Spearman's rho $\rho^s_\pi(x_{0,1}, x_{0,2}|\boldsymbol{x})$ values as a function of $(x_{0,1}, x_{0,2}) \in [0,1]^2$. The left column gives results for the copula C_π when $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ is integrated with respect to the prior π_0. The right column gives results for C_π when $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ is integrated with respect to the posterior using the data in Figure 4(b). The panels give results for different shrinkage prior/basis combinations: (a, b) PSC/B-spline; (c, d) HSC/B-spline; (e, f) HSC/augmented Fourier basis; (g, h) BVSC/regression spline.

3.1 Observational Model and Likelihood

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ be n observations on a continuous response, with covariate values $\boldsymbol{x}$. Following Sklar's theorem, the joint density of $\boldsymbol{Y}|\boldsymbol{x}$ can always be written as

$$p(\boldsymbol{y}|\boldsymbol{x}) = c^{\dagger}\left(F(y_1|x_1), \ldots, F(y_n|x_n)|\boldsymbol{x}\right) \prod_{i=1}^{n} p(y_i|x_i),$$

for some copula density $c^{\dagger}(\boldsymbol{u}|\boldsymbol{x})$ and conditional distribution functions $F(y_i|x_i)$ – both of which are unknown. Here, we model the joint distribution of $\boldsymbol{Y}|\boldsymbol{x}$ using a simplified decomposition where: (i) $Y_i|x_i$ has distribution function F_Y and density p_Y that are independent of x_i and do not vary with i, and (ii) the copula $c^{\dagger}$ is modeled using the implicit copula outlined in Section 2. With these assumptions, the density of $\boldsymbol{Y}|\boldsymbol{x}$ is

$$p(\boldsymbol{y}|\boldsymbol{x}) = c_{\pi}\left(F_Y(y_1), \ldots, F_Y(y_n)|\boldsymbol{x}\right) \prod_{i=1}^{n} p_Y(y_i), \qquad (10)$$

where c_{π} is the copula density at Corollary 1 with $\pi = \pi_0$. We call the model at (10) a 'copula smoother', because all regression smoothing is introduced through the copula only, and not the margin F_Y. We show later that this copula model demonstrates excellent regression smoothing properties.

From Theorem 1, the likelihood conditional on $\boldsymbol{\theta}, \boldsymbol{\gamma}$ (but not $\boldsymbol{\beta}$) is

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\gamma}) \prod_{i=1}^{n} \frac{p_Y(y_i)}{p(z_i|\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\gamma})} = \phi_n(\boldsymbol{z}; \boldsymbol{0}, R(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\gamma})) \prod_{i=1}^{n} \frac{p_Y(y_i)}{\phi_1(z_i)}\,. \qquad (11)$$

For large n, direct computation of the $n \times n$ correlation matrix R is computationally infeasible. However, the likelihood also conditional on $\boldsymbol{\beta}$ is

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) \prod_{i=1}^{n} \frac{p_Y(y_i)}{\phi_1(z_i)} = \phi_n(\boldsymbol{z}; SB\boldsymbol{\beta}, SS') \prod_{i=1}^{n} \frac{p_Y(y_i)}{\phi_1(z_i)}\,,$$

which can be evaluated in $O(n)$ operations because S is diagonal. We exploit this observation to propose MCMC schemes below that avoid direct computation of R.

3.2 Posterior Evaluation

Both the marginal distribution F_Y and the copula parameters $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ require estimation. For F_Y we consider two different non-parametric estimators. The first is an adaptive kernel density estimator (Shimazaki and Shinomoto, 2010), and the second is a Bayesian Dirichlet process mixture of normals estimator.[1] We condition on each density estimate $\hat{F}_Y$ to compute $z_i = \Phi_1^{-1}(\hat{F}_Y(y_i))$, for $i = 1, \ldots, n$, and then use MCMC to evaluate the posterior of the copula parameters given these values.

[1] To estimate the adaptive kernel estimator we use the Matlab routine 'ssvkernel', while for the Bayesian non-parametric estimator we use a modified version of the routine 'DPdensity' in Jara et al. (2011).

In a second approach, we follow Grazian and Liseo (2017) and integrate out the posterior uncertainty for F_Y when estimating the copula parameters using an MCMC scheme. To do so, at each sweep of the MCMC scheme, we re-compute the pseudo data $z_i = \Phi_1^{-1}(F_Y^{[j]}(y_i))$, for $i = 1, \ldots, n$, using the draws $\{F_Y^{[j]}; j = 1, \ldots, J\}$ of the marginal distribution from the Bayesian non-parametric estimator.

For each copula type, we use a different MCMC sampling scheme to estimate their parameters. For the PSC and HSC (where $\boldsymbol{\gamma} = \emptyset$), the proposed sampler evaluates the augmented posterior $p(\boldsymbol{\beta}, \boldsymbol{\theta} | \boldsymbol{x}, \boldsymbol{y})$, while for the BVSC (where $\boldsymbol{\theta} = \emptyset$) the proposed sampler evaluates the posterior $p(\boldsymbol{\gamma} | \boldsymbol{x}, \boldsymbol{y})$. For the PSC and HSC we generate from the conditional posterior $p(\boldsymbol{\beta} | \boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y}) = p(\boldsymbol{\beta} | \boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{z})$, which is Gaussian with mean $\boldsymbol{\mu}_\beta = \Sigma_\beta B' S^{-1} \boldsymbol{z}$ and covariance matrix $\Sigma_\beta = (B'B + P(\boldsymbol{\theta}))^{-1}$. The steps required to generate from the conditional posteriors of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ are outlined separately below for each of the three implicit copulas, and more details are in Part C of the Supplementary Material.

Posterior of the PSC Parameters The conditional posteriors of τ^2 and ψ are not recognizable distributions. A Metropolis-Hastings (MH) step is used to generate $\upsilon = \log(\tau^2)$, where a normal proposal with matching mode and curvature is used. Note that:

$$\begin{aligned} l_\upsilon \equiv \log(p(\upsilon | \boldsymbol{x}, \boldsymbol{\beta}, \psi, \boldsymbol{y})) \propto -\frac{\upsilon}{2} \left(\dim(P_\psi))\right) - 1) - \frac{1}{2\exp(\upsilon)} \boldsymbol{\beta}' P_\psi \boldsymbol{\beta} - \left(\frac{\exp(\upsilon)}{b_{\tau^2}}\right)^{\frac{1}{2}} \\ - \frac{1}{2} \sum_{i=1}^{n} \log(s_i^2) - \frac{1}{2}\left(\boldsymbol{z}'(SS')^{-1}\boldsymbol{z} - 2\boldsymbol{\beta}' B' S^{-1} \boldsymbol{z}\right). \end{aligned}$$

Approximating l_υ by a second order Taylor expansion around the current state $\upsilon^{(c)}$, and taking the exponent, yields the proposal density $\mathrm{N}\left(\mu_\upsilon, \sigma_\upsilon^2\right)$ with $\mu_\upsilon = \sigma_\upsilon^2 \frac{\partial l_\upsilon}{\partial \upsilon} + \upsilon$ and $\sigma_\upsilon^2 = -1 / \frac{\partial^2 l_\upsilon}{\partial \upsilon^2}$. Analytical expressions for the derivatives are given in Supplement C.1. Similarly, we transform ψ to the real line as $\xi = g(\psi) = \log\left((\psi - \epsilon)/(1 - \epsilon - \psi)\right)$, with $\epsilon = 0.01$. The log-posterior is

$$\begin{aligned} l_\xi &\equiv \log(p(\xi | \boldsymbol{x}, \boldsymbol{\beta}, \tau^2, \boldsymbol{y})) \\ &\propto \log\left(\frac{\partial \psi}{\partial \xi}\right) + \log(\pi_0(g^{-1}(\xi))) + \log(p(\boldsymbol{z} | \boldsymbol{x}, \boldsymbol{\beta}, \tau^2, \psi)) + \log(p(\boldsymbol{\beta} | \tau^2, \psi)) \\ &\propto \xi - 2\log(1 + \exp(\xi)) + \log(\det(\Delta(g^{-1}(\xi))) \\ &\quad - \frac{1}{2} \sum_{i=1}^{n} \log(s_i^2) - \frac{1}{2}\left(\boldsymbol{z}'(SS')^{-1}\boldsymbol{z} - 2\boldsymbol{\beta}' B' S^{-1} \boldsymbol{z}\right) - \frac{\boldsymbol{\beta}' P(g^{-1}(\xi)) \boldsymbol{\beta}}{2\tau^2}. \end{aligned}$$

We generate ξ using a MH step in the same fashion as υ, but using the derivatives of l_ξ which are given in Supplement C.1. Because both proposals are based on analytical derivatives, they are fast to compute. In our empirical work, the acceptance rates of υ and ξ were between 60% and 90%. Last, we found joint updates of (τ^2, ψ) had prohibitively low acceptance rates.

Posterior of the HSC Parameters The global scale parameter τ, and each local shrinkage parameter λ_j, are generated separately. MH steps with normal approximations as proposals are used as in the PSC case, where

$$\log(p(\log(\lambda_j^2)|\boldsymbol{x},\boldsymbol{\beta},\boldsymbol{\lambda}_{\setminus j},\tau,\boldsymbol{z})) \propto -\frac{1}{2}\sum_{i=1}^{n}\log(s_i^2) - \frac{1}{2}\boldsymbol{z}'(SS')^{-1}\boldsymbol{z} + \boldsymbol{\beta}'B'S^{-1}\boldsymbol{z}$$
$$- \frac{1}{2}\left[\log(\lambda_j^2) + \frac{\beta_j^2}{\lambda_j^2} + 2\log\left(1+\frac{\lambda_j^2}{\tau^2}\right)\right]$$
$$\log(p(\log(\tau)|\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{z})) \propto -(p-1)\log(\tau) - \log(1+\tau^2) - \sum_{j=1}^{p}\frac{\lambda_j^2}{\tau^2},$$

and $\boldsymbol{\lambda}_{\setminus j} \equiv \{\boldsymbol{\lambda}\setminus\lambda_j\}$. The derivatives of the conditional posteriors of $\log(\lambda_j^2)$ and $\log(\tau)$ are given in the Supplement C.2. Similar to the sampler for the PSC, in our simulations the acceptance rates of these steps were around 70% for $\log(\lambda_j^2)$ and above 90% for $\log(\tau)$.

Posterior of the BVSC Parameters From (11), the posterior

$$p(\boldsymbol{\gamma}|\boldsymbol{x},\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\gamma})\pi_0(\boldsymbol{\gamma}) \propto \phi_n(\boldsymbol{z};\boldsymbol{0},R(\boldsymbol{x},\boldsymbol{\gamma}))\pi_0(\boldsymbol{\gamma})$$
$$\propto \quad |R(\boldsymbol{x},\boldsymbol{\gamma})|^{-1/2}\exp\left\{-\frac{1}{2}\left(\boldsymbol{z}'R(\boldsymbol{x},\boldsymbol{\gamma})^{-1}\boldsymbol{z}\right)\right\}\mathrm{Beta}(p-p_\gamma+1,p_\gamma+1) \equiv A(\gamma_i,\gamma_j).$$

We generate from this posterior using a Gibbs sampler, where $\boldsymbol{\gamma}$ is partitioned into pairs of elements selected at random, and each pair (γ_i,γ_j) is generated conditional on the other elements $\boldsymbol{\gamma}\setminus(\gamma_i,\gamma_j)$. This involves computing $A(\gamma_i,\gamma_j)$ for the four possible configurations $(\gamma_i,\gamma_j) \in \mathcal{S} \equiv \{(0,0),(0,1),(1,0),(1,1)\}$ for that pair of indicator values. This can be undertaken efficiently as outlined in Supplement C.3, where direct computation of R is avoided. We then generate from $p((\gamma_i,\gamma_j)|\boldsymbol{\gamma}\setminus(\gamma_i,\gamma_j),\boldsymbol{x},\boldsymbol{y}) = \frac{A(\gamma_i,\gamma_j)}{\sum_{(\tilde{\gamma}_i,\tilde{\gamma}_j)\in\mathcal{S}}A(\tilde{\gamma}_i,\tilde{\gamma}_j)}$. Unlike for the other two implicit copulas, $\boldsymbol{\beta}$ is not generated as part of the MCMC scheme.

3.3 Function Estimation

For a new observation (Y_0,x_0) on the response and covariate, to estimate the regression function $f(x_0) \equiv \mathbb{E}(Y_0|x_0,\boldsymbol{x})$ we employ the posterior predictive mean

$$\mathbb{E}(Y_0|x_0,\boldsymbol{x},\boldsymbol{y}) = \int \mathbb{E}(Y_0|x_0,\boldsymbol{x},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma})p(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma}|\boldsymbol{x},\boldsymbol{y})\mathrm{d}(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma}).$$

Note that f is different from m in (6), which is the mean function for the pseudo-response. Let $Z_0 = \Phi_1^{-1}(F_Y(Y_0))$, then the expectation in the integrand above is

$$\begin{aligned}\mathbb{E}(Y_0|x_0,\boldsymbol{x},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma}) &= \int F_Y^{-1}(\Phi_1(z_0))p(z_0|x_0,\boldsymbol{x},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma})\mathrm{d}z_0 \\ &= \int F_Y^{-1}(\Phi_1(z_0))\frac{1}{s_0}\phi_1\left((z_0-s_0\boldsymbol{b}_0'\boldsymbol{\beta})/s_0\right)\mathrm{d}z_0, \qquad (12)\end{aligned}$$

where $\boldsymbol{b}_0$ is the vector of basis terms evaluated at the covariate value x_0, and $s_0 = [1 + \boldsymbol{b}_0' P(\boldsymbol{\theta})^{-1} \boldsymbol{b}_0]^{-1/2}$ is the standardizing constant for the new observation. We employ $\hat{F}_Y$ for the marginal distribution function of $Y_0|x_0$, and compute the integral above using standard univariate numerical methods. Finally, the estimator

$$\mathbb{E}(Y_0|x_0, \boldsymbol{x}, \boldsymbol{y}) \approx \frac{1}{J} \sum_{j=1}^{J} \mathbb{E}\left(Y_0|x_0, \boldsymbol{x}, \boldsymbol{\beta}^{[j]}, \boldsymbol{\theta}^{[j]}, \boldsymbol{\gamma}^{[j]}\right) = \hat{f}(x_0) \tag{13}$$

can be computed from the output $\{\boldsymbol{\beta}^{[j]}, \boldsymbol{\theta}^{[j]}, \boldsymbol{\gamma}^{[j]}; j = 1, \ldots, J\}$ of the MCMC scheme. It can also be useful to estimate the conditional mean $m(x_0) \equiv \mathbb{E}(Z_0|x_0, \boldsymbol{x})$ of the pseudo-response at (6). For this we use the posterior predictive mean

$$\begin{aligned}
\mathbb{E}(Z_0|x_0, \boldsymbol{x}, \boldsymbol{y}) &= \int \mathbb{E}(Z_0|x_0, \boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}|\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) \\
&= \int (s_0 \boldsymbol{b}_0' \boldsymbol{\beta})\ p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}|\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) \approx \boldsymbol{b}_0' \left(\frac{1}{J} \sum_{j=1}^{J} s_0^{[j]} \boldsymbol{\beta}^{[j]} \right) = \hat{m}(x_0),
\end{aligned}$$

where $s_0^{[j]} = [1 + \boldsymbol{b}_0' P(\boldsymbol{\theta}^{[j]})^{-1} \boldsymbol{b}_0]^{-1/2}$.

For the BVSC, the vector $\boldsymbol{\beta}$ is not generated as part of the sampler in Section 3.2. Therefore, to compute these function estimators, it is necessary to generate from the Gaussian distribution $\boldsymbol{\beta}_\gamma^{[j]} \sim \boldsymbol{\beta}_\gamma | \boldsymbol{x}, \boldsymbol{\gamma}, \boldsymbol{y}$ at the end of each sweep, and set the remaining elements of $\boldsymbol{\beta}^{[j]}$ to zero. Also, note that for this case s_0 is a function of all covariate values $\{\boldsymbol{x}, x_0\}$, whereas for the HSC and PSC s_0 is a function of x_0 only.

We compute the function estimators $\hat{f}$ and $\hat{m}$ over a grid of values for x_0. Note that at each sweep of the samplers $f^{[j]}(x_0) = \mathbb{E}(Y_0|x_0, \boldsymbol{x}, \boldsymbol{\beta}^{[j]}, \boldsymbol{\theta}^{[j]}, \boldsymbol{\gamma}^{[j]})$ and $m^{[j]}(x_0) = s_0^{[j]} \boldsymbol{b}_0' \boldsymbol{\beta}^{[j]}$ are draws from the posterior distribution of each function at point x_0. Therefore, posterior $(100 - \alpha)\%$ probability intervals can be computed for f and m at point x_0 by ordering these draws and counting off $\alpha/2\%$ of the highest and lowest values.

Evaluation of $\hat{f}(x_0)$ using (13) requires J numerical integrations for each value of x_0. An alternative estimator that is faster to compute, is to plug in the point estimators for the unknown quantities in (12), giving

$$\tilde{f}(x_0) = \int \hat{F}_Y^{-1}(\Phi_1(z_0)) \frac{1}{\hat{s}_0} \phi_1\left((z_0 - \hat{m}(x_0))/\hat{s}_0\right) \mathrm{d}z_0,$$

with $\hat{s}_0 = \frac{1}{J} \sum_{j=1}^{J} s_0^{[j]}$. This involves computing only a single univariate numerical integral. Table S2 summarizes the functional relationships in the copula model, the Bayesian posterior means and their MCMC estimators.

3.4 Predictive Densities

The predictive density $p(y_0|x_0, \boldsymbol{x})$ of a new observation of the response Y_0, given a new covariate value x_0, is estimated using its posterior predictive density

$$p(y_0|x_0, \boldsymbol{x}, \boldsymbol{y}) = p(y_0|x_0, \boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}|\boldsymbol{x}, \boldsymbol{y}).$$

If $z_0 = \Phi_1^{-1}(F_Y(y_0))$, then $\left|\frac{dz_0}{dy_0}\right| = \frac{p_Y(y_0)}{\phi_1(z_0)}$, and by changing variables from y_0 to z_0,

$$\begin{aligned} p(y_0|x_0,\boldsymbol{x},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma}) &= \frac{p_Y(y_0)}{\phi_1(z_0)} p(z_0|x_0,\boldsymbol{x},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma}) \\ &= \frac{p_Y(y_0)}{\phi_1\left(\Phi_1^{-1}\left(F_Y(y_0)\right)\right)} \frac{1}{s_0}\phi_1\left(\frac{\Phi_1^{-1}(F_Y(y_0)) - m(x_0)}{s_0}\right), \end{aligned}$$

which follows from (6). We estimate the predictive regression density using the Monte Carlo iterates and $\hat{p}_Y$, and denote it as:

$$\hat{p}(y_0|x_0) = \frac{\hat{p}_Y(y_0)}{\phi_1(\Phi_1^{-1}(\hat{F}_Y(y_0)))} \left\{ \frac{1}{J}\sum_{j=1}^{J} \frac{1}{s_0^{[j]}}\phi_1\left(\frac{\Phi_1^{-1}(\hat{F}_Y(y_0)) - m^{[j]}(x_0)}{s_0^{[j]}}\right)\right\}. \quad (14)$$

It is also possible to estimate the predictive density $p(z_0|x_0,\boldsymbol{x})$ of the pseudo-response Z_0 given x_0, using the posterior predictive density

$$\begin{aligned} p(z_0|x_0,\boldsymbol{x},\boldsymbol{y}) &= \int p(z_0|x_0,\boldsymbol{x},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma})p(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma}|\boldsymbol{x},\boldsymbol{y})\mathrm{d}(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma}) \\ &= \int \frac{1}{s_0}\phi_1\left(\frac{z_0 - s_0\boldsymbol{b}_0'\boldsymbol{\beta}}{s_0}\right) p(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma}|\boldsymbol{x},\boldsymbol{y})\mathrm{d}(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma}), \end{aligned}$$

which is estimated using the iterates as $\hat{p}(z_0|x_0) = \frac{1}{J}\sum_{j=1}^{J}\frac{1}{s_0^{[j]}}\phi_1((z_0 - m^{[j]}(x_0))/s_0^{[j]})$.

3.5 Empirical Illustration

To illustrate the posterior dependence structure and function estimates from the copula models, we extend the empirical illustration in Section 2.3 by simulating $y_i \sim N(h_3(x_i), 0.5^2)$ using the same covariate values, and function h_3 specified in Section 4. In Figure 4, panel (a) gives a (normalized) histogram of the data $y_1,\ldots,y_n$, along with the kernel (labeled as 'KDE') and Bayesian (labeled as 'DPhat') non-parametric estimates of F_Y, which are very similar. Panel (b) contains a plot of the data and function h_3.

Using the kernel density estimate of F_Y, we generate draws from the posteriors of $(\boldsymbol{\theta},\boldsymbol{\gamma})$ for the same four copula/basis combinations employed previously. We compute the surface of Spearman correlations ρ_π^s, integrating out the copula parameters using these draws. The surfaces are plotted on the right-hand side of Figure 3 to enable comparison with those evaluated previously using draws from the priors π_0. We stress that each point on these surfaces is a pairwise Spearman correlation between $(Y_{0,1}, Y_{0,2})$ arising from $C_\pi(\boldsymbol{u}^+|\boldsymbol{x}^+)$, as discussed in Section 2.3.

The general features of the prior dependence structures discussed in Section 2.3 transfer to the posteriors, although there are some notable differences, and we make four observations. First, the posterior dependence structure of the PSC/B-spline in

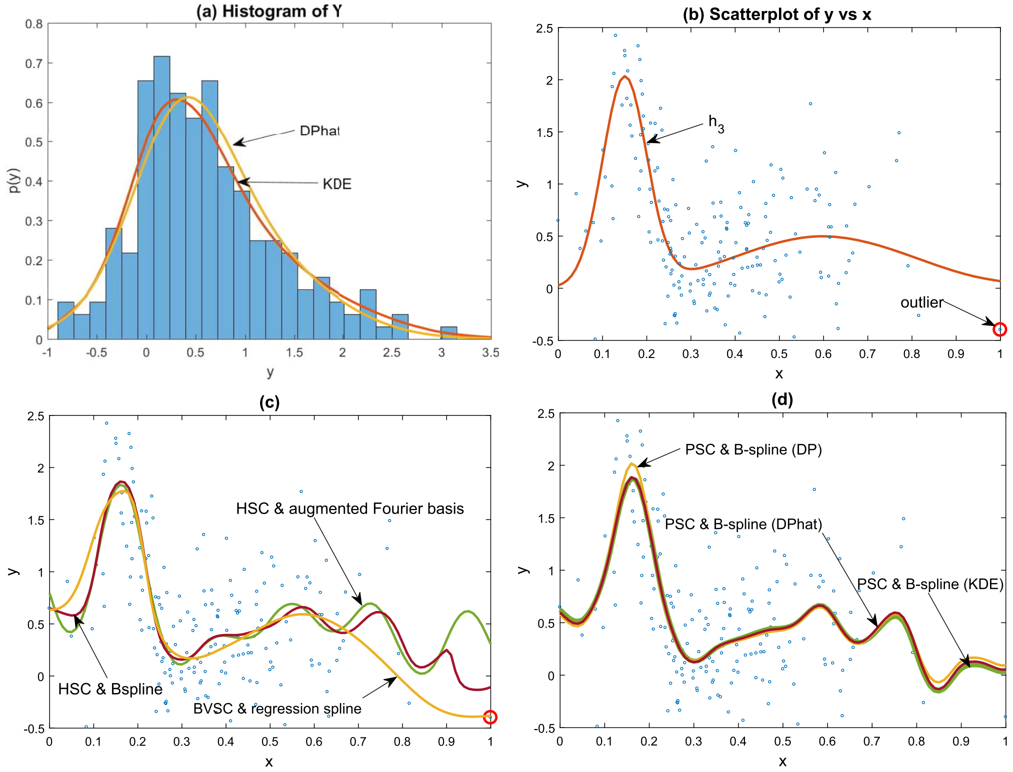

Figure 4: Data summary and function estimates for the empirical illustration in Sections 2.4 and 3.5. Panel (a) plots the (normalized) histogram of $y_1, \ldots, y_n$, plus the kernel and Bayesian density estimates. Panel (b) plots the function h_3 as a red line and a scatterplot of the data. Panel (c) plots function estimates from three copula/basis combinations: (i) BVSC/Regression spline, (ii) HSC/B-spline and (iii) HSC/Fourier basis. Panel (d) plots the function estimates for the PSC/B-spline copula model using three different approaches to marginal estimation: 'KDE' is using the kernel density estimator, 'DPhat' is using the Bayesian density estimator, and 'DP' is for the copula estimator that integrates out marginal density uncertainty.

panel (b) is sharper than its prior in panel (a). Second, the posterior dependence structure of the HSC/B-spline in panel (d) is asymmetric along the line $x_{0,1} = x_{0,2}$, with higher smoothing for covariate values around 0.1 and close to 1. This local variation in smoothing is evident when integrating over the posterior, but not the prior, of the horseshoe copula parameters. Third, when the HSC is combined with the augmented Fourier basis in panel (f), smoothing is non-monotonic in $|x_{0,1} - x_{0,2}|$ because the basis terms are also. Last, the BVSC with a regression spline basis in panel (h) has a posterior level of smoothing that is higher than that of the prior in panel (g). Yet the level of smoothing varies greatly with the value of the covariate, with more smoothing for values greater than 0.5, and less for values around 0.3, reflecting the distribution of the covariate in Figure 2. Last, Figure S1 in the Supplementary Material presents a movie that plots the density of the sub-copula of $C_\pi(\boldsymbol{u}^+|\boldsymbol{x}^+)$ in elements $(u_{0,1}, u_{0,2})$ for a range of values of $(x_{0,1}, x_{0,2})$. It does so for the BVSC with a regression spline basis, further visualizing its dependence structure.

Figure 4(c) compares the posterior function estimates for three of the copula/basis combinations, using the same kernel density estimate of F_Y. The estimator $\hat{f}$ was used for the HSC, and $\tilde{f}$ for the BVSC. All function estimates track the data well, although those from the HSC models under-smooth on the right-hand side. In contrast, the BVSC produces a smoother estimate, which is because Bayesian variable selection is known to be a highly locally adaptive regularization method (Smith and Kohn, 1996). Figure 4(d) compares the impact of different approaches to estimating the marginal F_Y on the function estimates. Three estimates $\hat{f}$ are computed and plotted for the PSC with a B-spline basis. The first (labeled 'KDE') and second (labeled 'DPhat') use the kernel and Bayesian non-parametric estimators for the marginal, respectively. The third (labeled 'DP') employs the draws for F_Y from the Bayesian estimator to integrate out uncertainty in the margin when estimating the copula parameters. All three function estimates are very similar, and are insensitive to the choice of marginal estimator.

Undertaking 1,000 sweeps of the MCMC schemes for estimating the copula parameters took approximately 13, 27 and 3.5 seconds for the HSC, PSC and BVSC, respectively, when implemented in serial using Matlab on a standard desktop.

4 Univariate Simulation

To illustrate the effectiveness of the copula smoother we undertake a simulation study. The PSC and B-spline basis is used with a non-parametric margin F_Y, estimated in the same three ways as in Section 3.5 and labeled as 'PSC/KDE', 'PSC/DPhat' and 'PSC/DP'. The benchmark model is a Bayesian P-spline with the same basis and Gaussian disturbances (labeled as 'PS').

4.1 Simulation Design

We consider the three univariate test functions: $h_1(x) = 2x - 1$, $h_2(x) = \sin(10\pi x)$ and

$$h_3(x) = \frac{1}{4}\left[\frac{1}{0.05}\phi_1((x-0.15)/0.05) + \frac{1}{0.2}\phi_1((x-0.6)/0.2)\right].$$

For each function $j = 1, 2, 3$, we generate $n = 100$ observations from three distributions:

$$\begin{aligned}
&\text{Case 1, Normal:} && Y_{1j} = h_j(x) + \varepsilon_1,\ \text{ where } \varepsilon_1 \sim \text{iid}\, \mathrm{N}(0, 0.5^2)\\
&\text{Case 2, Log-normal:} && Y_{2j} = h_j(x) + \varepsilon_2 - \mathbb{E}(\varepsilon_2),\ \text{ where } \varepsilon_2 \sim \text{iid}\, \mathrm{LN}(-2.89, 1.5^2)\\
&\text{Case 3, Implicit Copula:} && Y_{3j} = F_{\text{Gam}}^{-1}\left(\Phi\left(z_j\right); 3, 2\right), z_j = h_j(x) + \varepsilon_3,\\
& && \quad\text{where } \varepsilon_3 \sim \text{iid}\, \mathrm{N}(0, r_j^2).
\end{aligned}$$

Here, F_{Gam} is a Gamma distribution function and LN is the lognormal distribution. The distributions of Y_{lj} are defined conditional on the covariate x, which we generate independently from a uniform on $(0,1)$. Note that the distribution in Case 1 matches that of the Gaussian P-spline, while that in Case 3 matches that of the implicit copula model with a Gamma margin. The distribution in Case 2 matches neither model.

The true regression and noise functions are $f_j(x) \equiv \mathbb{E}(Y_{lj}|x)$ and $v_{lj}(x) \equiv \mathrm{Var}(Y_{lj}|x)$, and in each case the signal-to-noise ratio is $\mathrm{SNR}_{lj} \equiv \mathrm{range}(f_j(x))/(\int v_{lj}(x)\mathrm{d}x)^{1/2} = 4$ over the domain of the covariate $0 \leq x \leq 1$. In Cases 1 and 2 ($l = 1, 2$), it is straightforward to see that $f_j = h_j$, $v_{lj}(x) = \mathrm{Var}(\varepsilon_j)$ is a constant and that $\mathrm{SNR}_{lj} = 4$. However, in Case 3, f_j and v_{lj} are more complex functions of h_j, with

$$f_j(x) = \int y_j p(y_j|x)\mathrm{d}y_j = \int F^{-1}_{\mathrm{Gam}}(\Phi_1(z_j); 3, 2)\frac{1}{r_j}\phi_1\left((z_j - h_j(x))/r_j\right)\mathrm{d}z_j,$$

$$v_{lj}(x) = \mathbb{E}(Y_j^2|x) - f_j(x)^2 = \int [F^{-1}_{\mathrm{Gam}}(\Phi_1(z_j); 3, 2)]^2\frac{1}{r_j}\phi_1\left((z_j - h_j(x))/r_j\right)\mathrm{d}z_j - f_j(x)^2,$$

where the integrals are computed numerically. Setting $\mathrm{SNR}_{3j} = 4$ over $0 \leq x \leq 1$, it is possible to solve the nonlinear optimization problem with respect to r_j to get $r_1 = 0.48$, $r_2 = 0.47$ and $r_3 = 0.58$ for the three functions. For each of the nine combinations of Case l and function h_j we simulated 100 replicates, leading to a total of 900 datasets.

For both the PSC and the PS the same cubic B-spline basis is employed with equally spaced knots and $\dim(\boldsymbol{\beta}) = 32$. As outlined in Section 3.2, the precision matrix of an AR(1) is used for constructing the PSC implicit copula. For the PS the popular first order random walk prior (Lang and Brezger, 2004) is used, although the results are almost identical when the precision matrix of an AR(1) model is employed.

4.2 Measures of Performance

We consider three measures of the quality of the fitted statistical models. The first is a measure of the accuracy of the point estimate of the regression function, and is the root mean square error $\mathrm{RMSE}(f, \hat{f}) = (\frac{1}{n}\sum_{i=1}^n(\hat{f}(x_i) - f(x_i))^2)^{1/2}$ computed over the data points. For the PSC model the regression function estimator is given at (13), whereas for the PS it is $\hat{f}(x_i) = \boldsymbol{b}_i'\mathbb{E}(\boldsymbol{\beta}|\boldsymbol{y})$, which we compute using the BayesX software (Belitz et al., 2015).

The second measure is based on the Kullback-Leibler Divergence (KLD) between the density $p(y|x)$ of the data generating process, and its estimate $\hat{p}(y|x)$, given by

$$\mathrm{KLD}_x(p||\hat{p}) = \int p(y|x)\log\left(\frac{p(y|x)}{\hat{p}(y|x)}\right)\mathrm{d}y.$$

To compute the KLD, note that for Cases 1 and 2 the density $p(y|x)$ is a normal and log-normal distribution, respectively. For Case 3, the density is

$$p(y|x) = \frac{p_{\mathrm{Gam}}(y; 3, 2)}{\phi_1(\Phi_1^{-1}(F_{\mathrm{Gam}}(y; 3, 2)))r_j}\phi_1\left(\frac{\Phi_1^{-1}(F_{\mathrm{Gam}}(y; 3, 2)) - h_j(x)}{r_j}\right),$$

where p_{Gam} is a Gamma density function.

For the PSC, the density estimator is given at (14). For the regular PS, $\hat{p}(y_0|x_0) = (1/\hat{\sigma})\phi_1((y_0 - \hat{f}(x_0))/\hat{\sigma})$, with point estimators $\hat{\sigma}$ and $\hat{f}$. The integral can be computed analytically for the Case 1/PS combination and numerically for the other five combinations of estimator and Case; see Table S3. Finally, we report the mean KLD over an equally-spaced partition $0 = \tilde{x}_1 < \ldots < \tilde{x}_N = 1$ of the covariate, giving $\text{MKLD}(p||\hat{p}) = \frac{1}{N}\sum_{i=1}^{N} \text{KLD}_{\tilde{x}_i}(p||\hat{p})$, where we set $N = 100$. This metric measures the accuracy of $\hat{p}(\cdot|x_0)$.

The third and final measure is of predictive performance. This is the mean logarithmic score computed by ten-fold cross-validation. For a given dataset, we compute this by partitioning the data into ten sub-samples, denoted as $\{(y_{i,k}, x_{i,k}); i = 1, \ldots, n_k\}$ for $k = 1, \ldots, 10$. For sub-sample k, we compute the density estimator using the remaining 9 sub-samples as the training data, and denote these as $\hat{p}_k(y|x)$. The ten-fold mean logarithmic score is then $\text{MLS} = \frac{1}{10}\sum_{k=1}^{10} \frac{1}{n_k}\sum_{i=1}^{n_k} \log \hat{p}_k(y_{i,k}|x_{i,k})$. Here $n = 100$, so that we set $n_k = 10$, giving sub-samples of equal size.

4.3 Results

Figure S2 compares the accuracy of the three copula estimators and the benchmark PS estimator of the regression functions using the RMSE metric. There are nine panels: one for each combination of Cases 1, 2, 3 and test functions h_1, h_2, h_3. The accuracy of the function estimators is similar, even in Case 1 where the PS estimator is the correct model. This is reassuring because the Bayesian P-spline is known to be a highly competitive regression function estimator (Lang and Brezger, 2004; Scheipl et al., 2012). To illustrate, Figure S4 plots the true regression function f_j and the PSC/KDE and PS estimates for a single replicate of data in each case, along with a scatterplot of the data. The function estimates are similar and track the data well. However, the PSC and PS density estimators differ substantially. Figure 5 presents boxplots of the MKLD metric. The PS is slightly more accurate than the three copula estimators in Case 1, which is because the PS matches the data generating process. But in the two non-Gaussian cases—including Case 2 where neither model is correct—the PSC density estimator is substantially more accurate. The same conclusions are drawn from Figure S3, which presents equivalent boxplots for the MLS metric. Thus, using the copula model increases the accuracy of the predictive distributions for the non-Gaussian data substantially here. Last, when comparing the different approaches to estimating F_Y, integrating out uncertainty in its estimate does not increase the accuracy of the function or density estimates, nor the predictive distributions. This is consistent with observations in the broader copula literature, where two stage estimators are widely used (Joe, 2005).

5 Extension to Multiple Covariates

The implicit copula (and the resulting copula smoother) can be extended to account for multiple covariates in two ways. The first is by constructing the implicit copula of an additive model for the pseudo-response, and the second is by employing a radial basis. We explain the first approach below, and the second in the Supplementary Materials.

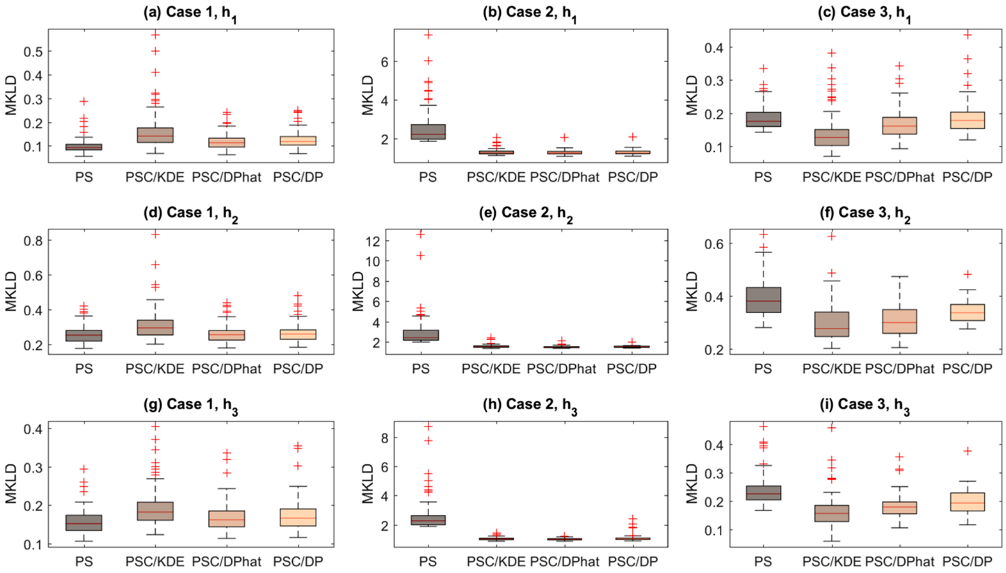

Figure 5: Comparison of density estimate accuracy from the simulation study. Each panel corresponds to a different combination of test function and case. The three columns correspond to Cases 1, 2 and 3, while the three rows correspond to functions h_1, h_2 and h_3. Each boxplot is of the 100 values of the MKLD$(p||\hat{p})$ metric from the simulation replicates. Lower values correspond to increased accuracy. The estimators are the benchmark PS model and the three copulas models PSC/KDE, PSC/DPhat and PSC/DP, as discussed in the text. Analogous boxplots of the two other performance metrics—RMSE and MLS—are given by Figures S2 and S3 in the Supplementary Materials.

5.1 Implicit Copula

Consider replacing (2) with the additive regression

$$\tilde{Z}_i = \sum_{l=1}^{L} \tilde{m}_l(x_{il}) + \varepsilon_i, \text{ for } i = 1, \ldots, n \tag{15}$$

for L smooth functions of covariates $x_1, \ldots, x_L$. As before, each function is modeled as a linear combination of basis functions $\tilde{m}_l(x_l) = \sum_j^{p_l} b_{lj}(x_l)\beta_{lj}$, with corresponding design matrix B_l and coefficient vector $\boldsymbol{\beta}_l = (\beta_{l1}, \ldots, \beta_{lp_l})'$. Then the additive regression can be written as the linear model at (3), but where $B = [B_1|\cdots|B_L]$ is an $(n \times \sum_{l=1}^{L} p_l)$ concatenated design matrix and $\boldsymbol{\beta}' = (\boldsymbol{\beta}'_1, \ldots, \boldsymbol{\beta}'_L)$. Our objective here is to construct the implicit copula of this additive model for the pseudo-response.

A global intercept parameter is not included in (15) because it is unidentified in its implicit copula. To ensure identifiability of $\boldsymbol{\beta}$, we centre all but one $\tilde{m}_l$ around zero, so that $\mathbf{1}'\tilde{m}_l(\boldsymbol{x}_l) = \mathbf{1}'B_l\boldsymbol{\beta}_l = 0$, for $l = 1, \ldots, L-1$, with $\mathbf{1}$ an n-vector of ones. To regularize each vector $\boldsymbol{\beta}_l$, we assume the same shrinkage prior at (4), but with these constraints, so that

$$p(\boldsymbol{\beta}_l|\boldsymbol{x},\boldsymbol{\theta}_l,\boldsymbol{\gamma}_l) \propto \begin{cases} \phi_{p_l}(\boldsymbol{\beta}_l;\mathbf{0},P(\boldsymbol{\theta})^{-1})I(\mathbf{1}'B_l\boldsymbol{\beta}_l=0) & \text{if } l=1,\dots,L-1 \\ \phi_{p_l}(\boldsymbol{\beta}_l;\mathbf{0},P(\boldsymbol{\theta})^{-1}) & \text{if } l=L \end{cases},$$

where each prior is strictly proper. Setting $P(\boldsymbol{\theta}) = \text{bdiag}(P(\boldsymbol{\theta}_1),\dots,P(\boldsymbol{\theta}_L))$ as a block diagonal matrix and $\boldsymbol{x}_l = (x_{1l},\dots,x_{nl})'$, $\boldsymbol{\beta}$ can be integrated out as a linearly constrained normal, giving

$$\tilde{\boldsymbol{Z}}|\boldsymbol{x}_1,\dots,\boldsymbol{x}_L,\boldsymbol{\theta},\boldsymbol{\gamma} \sim N(\mathbf{0},(I+BP(\boldsymbol{\theta})^{-1}B')),$$

as in Section 2.1. Standardization of $\tilde{\boldsymbol{Z}}$ and formation of the implicit copula then proceeds as in the univariate case, but where $\boldsymbol{b}_i = (\boldsymbol{b}'_{i1},\dots,\boldsymbol{b}'_{iL})'$, $\boldsymbol{b}_{il} = (b_{l1}(x_{il}),\dots, b_{lp_l}(x_{il}))'$,

$$\begin{aligned} s_i &= (1+\boldsymbol{b}'_iP(\boldsymbol{\theta})^{-1}\boldsymbol{b}_i)^{-1/2} = \left(1+\sum_{l=1}^{L}\boldsymbol{b}'_{il}P(\boldsymbol{\theta}_l)^{-1}\boldsymbol{b}_{il}\right)^{-1/2}, \text{ and} \\ \Omega^{-1} &= \text{bdiag}\left(B'_1B_1+P(\boldsymbol{\theta}_1),\dots,B'_LB_L+P(\boldsymbol{\theta}_L)\right), \end{aligned}$$

with 'bdiag' a block diagonal matrix operator. The posterior can be evaluated using the MCMC schemes outlined in the univariate case, with one change. When generating $\boldsymbol{\beta}$, we generate each sub-vector $\boldsymbol{\beta}_l$ conditional on the other elements of $\boldsymbol{\beta}$. For $l=1,\dots,L-1$ this involves generating from a constrained normal using the fast algorithm in Rue and Held (2005, Algorithm 2.6). Further details on how to implement the MCMC scheme for the PSC are given in Supplement D.

5.2 Function Estimation and Partial Residuals

For a new observation $(Y_0,x_{01},\dots,x_{0L})$ on the response and covariates, the regression surface is $f(x_{01},\dots,x_{0L}) \equiv E(Y_0|x_{01},\dots,x_{0L})$. It can be estimated in the same manner as in Section 3.3, but where

$$m(x_{01},\dots,x_{0L}) = s_0\boldsymbol{b}'_0\boldsymbol{\beta} = s_0\sum_{l=1}^{L}\boldsymbol{b}'_{0l}\boldsymbol{\beta}_l = \sum_{l=1}^{L}m_l(x_{0l}),$$

with s_0 as defined above and $m_l(x_{0l}) = s_0\boldsymbol{b}'_{0l}\boldsymbol{\beta}_l$.

Even though the relationship at (15) is additive in the covariates, the regression surface f is not. This means that partial residuals—a popular diagnostic for additive models (Hastie and Tibshirani, 1990)—cannot be easily defined for $\boldsymbol{y}$. However, they can be for the values of the standardized pseudo-response $z_1,\dots,z_n$ as follows.

Definition 1. *For the i-th observation and j-th effect of the additive basis copula, $i=1,\dots,n$ and $j=1,\dots,L$, we define the j-th partial residual $\epsilon_{i,j}$ as*

$$\epsilon_{i,j} = z_i - \sum_{l\neq j}m_l(x_i) = z_i - s_i\sum_{l\neq j}\boldsymbol{b}'_{il}\boldsymbol{\beta}_l,$$

where s_i is defined above.

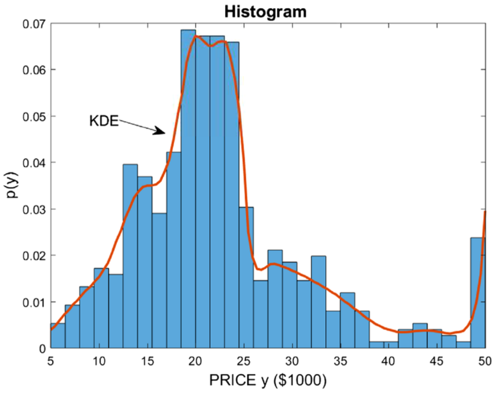

Figure 6: Histogram of PRICE (in $1,000) in the Boston housing dataset. Also shown in red is the adaptive kernel density estimate (KDE).

If the model is correct, then from (6), the partial residual $\epsilon_{i,j}$ is a realization from a $N(m_j(x_i), s_i)$ distribution.

5.3 Example: Boston Housing Data

To illustrate we employ the Boston housing data (Harrison and Rubinfeld, 1978). The data comprise observations on the median value (PRICE) of residential homes in $n = 506$ Boston census tracts. Also recorded are five continuous hedonic variables (NOX, RM, DIS, LSTAT and TAX) defined in Table IV of Harrison and Rubinfeld (1978). The dataset is a common test for flexible regression methods with PRICE as the response. Figure 6 plots the histogram of PRICE, which is far from Gaussian, and regressions with normal errors produce poor estimates of the functional relationships. For example, in their analysis Smith and Kohn (1996) estimate a Box-Cox transform of PRICE and model the errors as a mixture of two normals.

We model PRICE using the PSC smoother with the five continuous variables as covariates. We employ the KDE of PRICE in Figure 6 for the estimate of F_Y. For each covariate, a cubic B-spline basis with equally spaced knots and $\dim(\boldsymbol{\beta}_l) = 22$ is employed. Figure 7 presents summaries of the functional relationships from the fitted copula smoother. The left-hand panels (a, c, e, g, i) plot 'slices' of $\hat{f}$ against each of the five covariates, where in each panel the other four covariates are fixed to their values for the observation with the median PRICE. Also plotted are the equivalent slices of the 95% posterior probability interval for f. For comparison, we estimate an additive P-spline with the same basis (PS) using BayesX software. Panels (a, c, e, g, i) depict the equivalent function estimates from this additive model, and they differ from those of the copula model. The right-hand panels (b, d, f, h, j) show the posterior mean of $m_l(x_{0l}) = s_0 \boldsymbol{b}'_{0l}\boldsymbol{\beta}_l$, along with 95% posterior probability intervals for $m_l(x_{0l})$, for $l = 1, \ldots, 5$. The scatterplots are of the partial residuals $\{\epsilon_{1,l}, \ldots, \epsilon_{n,l}\}$.

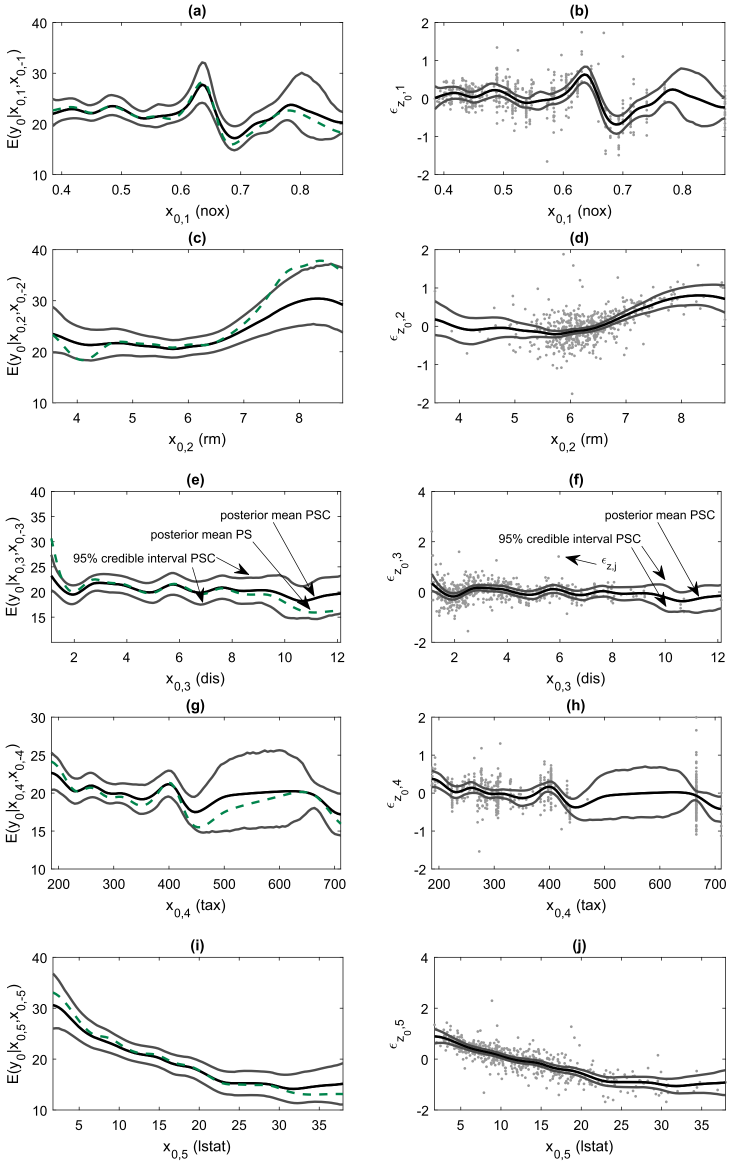

Figure 7: Summary of the copula smoother with an additive PSC fitted to the Boston housing data. The left panels plot slices of the estimated regression surface $\hat{f}$ for each of (a) NOX, (c) RM, (e) DIS, (g) TAX, and (e) LSTAT, fixing the other four covariate values to those of the median priced house. Estimates are given for both the copula smoother (bold line) and additive PS (dashed line) for comparison. The 95% credible intervals are also given for the copula smoother. The right panels plot $\hat{m}_l$ and the partial residuals $\{\epsilon_{1,l},\ldots,\epsilon_{n,l}\}$ for (b) NOX ($l=1$), (d) RM ($l=2$), (f) DIS ($l=3$), (h) TAX ($l=4$), and (j) LSTAT ($l=5$).

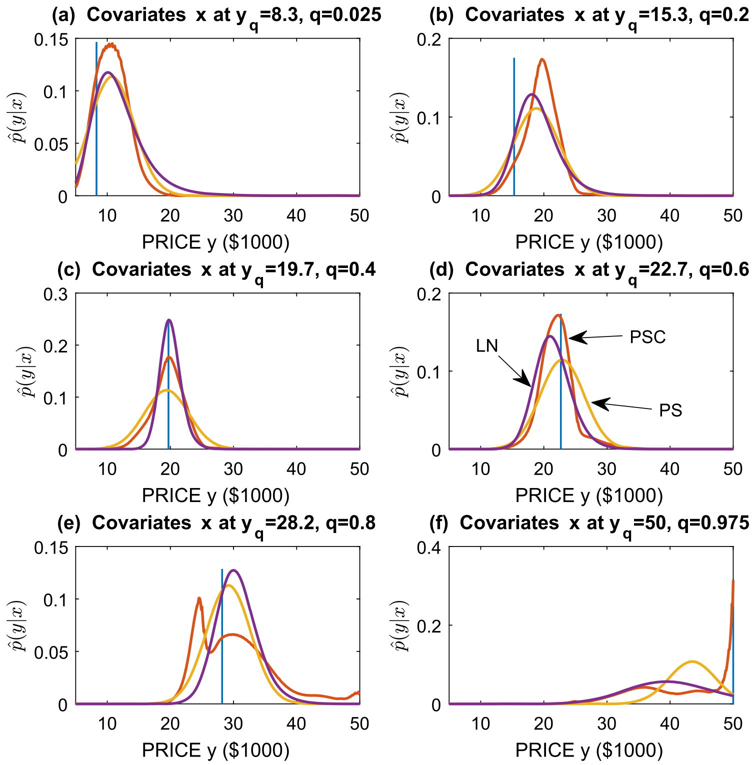

Figure 8: Predictive densities $\hat{p}(y|\boldsymbol{x})$ for six houses in the Boston housing data. Each corresponds to the house at the qth quantile of the observed prices, for (a) $q = 0.025$, (b) $q = 0.2$, (c) $q = 0.4$, (d) $q = 0.6$, (e) $q = 0.8$ and (f) $q = 0.975$. In each panel the predictive density is plotted for the copula smoother (red line), the Gaussian P-spline (yellow line) and the heteroscedastic log-normal model (violet line), while the observed price is marked with a blue vertical line.

To compare the models, we compute the mean logarithmic score for a ten-fold cross validation as in Section 4. For the copula model MLS $= -2.47$, compared to MLS $= -2.86$ for the additive P-spline, indicating that the copula model has more accurate predictive densities. As a second benchmark, we also estimated a distributional regression model (Klein et al., 2015) under a log-normal assumption for PRICE with the same basis for both distributional parameters using the BayesX software. For this estimator MLS $= -2.78$, again indicating lower accuracy than the copula model.

To highlight why the copula model predictions are more accurate, Figure 8 plots the predictive densities $\hat{p}(y_0|\boldsymbol{x}_0)$ from the three fitted models for six representative observations. These are the observations at quantiles $q = 0.025, 0.2, 0.4, 0.6, 0.8, 0.975$ of the PRICE distribution. The predictive distributions from the copula model are generally tighter (i.e. more 'sharp'), and feature a high degree of asymmetry throughout. The predictive density in panel (f) has a spike at PRICE=$50,000, which is caused by a few high-valued observations that are unexplained by the covariates. Earlier analysis (Smith and Kohn, 1996) treats these as outliers, but in the copula model they are captured by the estimated marginal $\hat{F}_Y$ in Figure 6. In contrast, these observations are not well modelled using either the P-spline or distributional regression in panel (f).

6 Discussion

The paper presents a general approach to construct the implicit copula of regularized regression smoothers with Gaussian disturbances. Three diverse shrinkage priors are considered in detail, although the approach can also be employed with other conjugate priors. A Gaussian copula is first constructed by integrating out $\boldsymbol{\beta}$, but conditioning on the regularization parameters $(\boldsymbol{\theta}, \boldsymbol{\gamma})$. The implicit copula is then formed by mixing over their prior or posterior distributions. This conditioning trick greatly simplifies the computation of the implicit copula, which is much harder to compute via inversion of the distribution $\tilde{\boldsymbol{Z}}|\boldsymbol{x}$ directly. We stress here that the implicit copula is not a Gaussian copula, and can have a very different dependence structure as illustrated in Figure 3.

The implicit copulas of elliptical (Fang et al., 2002) and skew-elliptical (Demarta and McNeil, 2005; Smith et al., 2012) distributions are employed widely. More recently, interest has grown in computing the implicit copulas of pseudo-response values of more complex statistical models. Examples include implicit copulas of Gaussian vector autoregressions (Smith and Vahey, 2016), factor models (Murray et al., 2013; Oh and Patton, 2017) and state space models (Smith and Maneesoonthorn, 2018). However, as far as we are aware, ours is the first paper to consider constructing the implicit copula of the regularized regression smoothing models of the type considered here. Acar et al. (2011) and Craiu and Sabeti (2012) consider copulas with dependence parameters that are functions of one or more covariates. However, these are low-dimensional copulas capturing the dependence between two or more response variables, and are very different from those considered here. In contrast, our implicit copulas capture the dependence between multiple values of a single response variable as a function of the covariates, with the dependence structure inherited from the regularized regression smoother.

In the machine learning literature Gaussian process-based regression smoothers—such as support or relevance vector machines (Tipping, 2001)—are a popular alternative to regularized smoothers of the type considered here. While a number of authors extend Gaussian processes by constructing their implicit copulas (Wilson and Ghahramani, 2010; Wauthier and Jordan, 2010), we are unaware of any work constructing the implicit copula of vector machines. Moreover, these copulas are Gaussian copulas, whereas the implicit copulas constructed here are not. Gaussian processes have also been used as building blocks along with conditional copulas to model non-Gaussian regression or time series data (Wauthier and Jordan, 2010; Levi and Craiu, 2016). However, these approaches employ low-dimensional closed form parametric copulas. In contrast, the implicit copulas proposed here are high-dimensional and unavailable in closed form, and are very different.

We finish by mentioning promising directions for extension of our proposed approach. First, the implicit copulas for other popular conjugate priors for regularization (Liang et al., 2008; Scheipl et al., 2012) may be constructed. Second, regression smoothers with elliptical error distributions beyond the Gaussian can be considered. When combined with conjugate priors, application of the conditioning trick will result in the implicit copula being a mixture over the corresponding elliptical copula. Third, while we use the copula smoother to model non-Gaussian continuous data, the copula can also be employed for modeling discrete-valued or mixed data. For these cases, new ways to evaluate the posterior distribution of the regularization parameters are required.

Supplementary Material

Supplementary Material for "Implicit Copulas from Bayesian Regularized Regression Smoothers" (DOI: 10.1214/18-BA1138SUPPA; .pdf). This contains extensive additional material organized into five Parts A–F. It includes implementation details, proofs, additional examples, and tables and figures referred to throughout the text.

MATLAB code for "Implicit Copulas from Bayesian Regularized Regression Smoothers" (DOI: 10.1214/18-BA1138SUPPB; .zip). This contains MATLAB files to implement the Bayesian regularized regression smoothers outlined in the paper.

References

Acar, E. F., Craiu, R. V., and Yao, F. (2011). "Dependence calibration in conditional copulas: A nonparametric approach." *Biometrics*, 67: 445–453. MR2829013. doi: https://doi.org/10.1111/j.1541-0420.2010.01472.x. 1167

Belitz, C., Brezger, A., Klein, N., Kneib, T., Lang, S., and Umlauf, N. (2015). "BayesX—Software for Bayesian inference in structured additive regression models. Version 3.0.2." Available from http://www.bayesx.org. 1160

Carvalho, C. M. and Polson, G., Nicholas (2010). "The horseshoe estimator for sparse signals." *Biometrica*, 97: 465–480. MR2650751. doi: https://doi.org/10.1093/biomet/asq017. 1143, 1148

Clyde, M. and George, E. I. (2004). "Model uncertainty." *Statistical Science*, 19(1): 81–94. MR2082148. doi: https://doi.org/10.1214/088342304000000035. 1143, 1149

Craiu, V. R. and Sabeti, A. (2012). "In mixed company: Bayesian inference for bivariate conditional copula models with discrete and continuous outcomes." *Journal of Multivariate Analysis*, 110: 106–120. MR2927512. doi: https://doi.org/10.1016/j.jmva.2012.03.010. 1167

Demarta, S. and McNeil, A. J. (2005). "The t copula and related copulas." *International Statistical Review*, 73(1): 111–129. 1167

Fahrmeir, L. and Kneib, T. (2011). *Bayesian Smoothing and Regression for Longitudinal, Spatial and Event History Data*. New York: Oxford University Press. MR2850683. doi: https://doi.org/10.1093/acprof:oso/9780199533022.001.0001. 1148

Fahrmeir, L. and Lang, S. (2001). "Bayesian inference for generalized additive mixed models based on Markov random field priors." *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 50: 201–220. MR1833273. doi: https://doi.org/10.1111/1467-9876.00229. 1148

Fang, H.-B., Fang, K.-T., and Kotz, S. (2002). "The meta-elliptical distributions with given marginals." *Journal of Multivariate Analysis*, 82: 1–16. MR1918612. doi: https://doi.org/10.1006/jmva.2001.2017. 1167

Grazian, C. and Liseo, B. (2017). "Approximate Bayesian inference in semiparametric copula models." *Bayesian Analysis*, 12(4): 991–1016. MR3724976. doi: https://doi.org/10.1214/17-BA1080. 1154

Harrison, D. and Rubinfeld, D. L. (1978). "Hedonic prices and the demand for clean air." *Journal of Environmental Economics and Management*, 5: 81–102. 1164

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. New York/Boca Raton: Chapman & Hall/CRC. MR1082147. 1145, 1163

Jara, A., Hanson, T., Quintana, F., Müller, P., and Rosner, G. (2011). "DPpackage: Bayesian semi- and nonparametric modeling in R." *Journal of Statistical Software*, 40(5): 1–30. MR3309338. doi: https://doi.org/10.1007/978-3-319-18968-0. 1153

Joe, H. (2005). "Asymptotic efficiency of the two-stage estimation method for copula-based models." *Journal of Multivariate Analysis*, 94(2): 401–419. MR2167922. doi: https://doi.org/10.1016/j.jmva.2004.06.003. 1161

Klein, N. and Kneib, T. (2016). "Scale-dependent priors for variance parameters in structured additive distributional regression." *Bayesian Analysis*, 11(4): 1071–1106. MR3545474. doi: https://doi.org/10.1214/15-BA983. 1148

Klein, N., Kneib, T., Lang, S., and Sohn, A. (2015). "Bayesian structured additive distributional regression with an application to regional income inequality in Germany." *The Annals of Applied Statistics*, 9: 1024–1052. MR3371346. doi: https://doi.org/10.1214/15-AOAS823. 1144, 1166

Klein, N. and Smith, M. S. (2018). "Supplementary Material containing MATLAB code for "Implicit Copulas from Bayesian Regularized Regression Smoothers"." *Bayesian Analysis*. doi: https://doi.org/10.1214/18-BA1138SUPPA, https://doi.org/10.1214/18-BA1138SUPPB 1145

Lang, S. and Brezger, A. (2004). "Bayesian P-splines." *Journal of Computational and Graphical Statistics*, 13: 183–212. MR2044877. doi: https://doi.org/10.1198/1061860043010. 1143, 1147, 1148, 1160, 1161

Levi, E. and Craiu, R. V. (2016). "Gaussian Process Single Index Models for Conditional

Copulas." *arXiv preprint arXiv:1603.03028*. MR3765819. doi: https://doi.org/10.1016/j.csda.2018.01.013. 1167

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). "Mixtures of g priors for Bayesian variable selection." *Journal of the American Statistical Association*, 103(481): 410–423. MR2420243. doi: https://doi.org/10.1198/016214507000001337. 1168

Murray, J. S., Dunson, D. B., Carin, L., and Lucas, J. E. (2013). "Bayesian Gaussian copula factor models for mixed data." *Journal of the American Statistical Association*, 108(502): 656–665. MR3174649. doi: https://doi.org/10.1080/01621459.2012.762328. 1167

Nelson, R. (2006). *An Introduction to Copulas*. Springer, 2nd edition. MR2197664. 1144, 1149

Oh, D. H. and Patton, A. J. (2017). "Modeling dependence in high dimensions with factor copulas." *Journal of Business & Economic Statistics*, 35(1): 139–154. MR3591542. doi: https://doi.org/10.1080/07350015.2015.1062384. 1167

Rigby, R. A. and Stasinopoulos, D. M. (2005). "Generalized additive models for location, scale and shape (with discussion)." *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 54: 507–554. MR2137253. doi: https://doi.org/10.1111/j.1467-9876.2005.00510.x. 1144

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields*. New York/Boca Raton: Chapman & Hall/CRC. MR2130347. doi: https://doi.org/10.1201/9780203492024. 1163

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press. MR1998720. doi: https://doi.org/10.1017/CBO9780511755453. 1143

Scheipl, F., Fahrmeir, L., and Kneib, T. (2012). "Spike-and-slab priors for function selection in structured additive regression models." *Journal of the American Statistical Association*, 107(500): 1518–1532. MR3036413. doi: https://doi.org/10.1080/01621459.2012.737742. 1161, 1168

Scott, J. G. and Berger, J. O. (2010). "Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem." *The Annals of Statistics*, 38(5): 2587–2619. MR2722450. doi: https://doi.org/10.1214/10-AOS792. 1149

Shimazaki, H. and Shinomoto, S. (2010). "Kernel bandwidth optimization in spike rate estimation." *Journal of Computational Neuroscience*, 29(1-2): 171–182. MR2721339. doi: https://doi.org/10.1007/s10827-009-0180-4. 1153

Smith, M. S., Gan, Q., and Kohn, R. (2012). "Modeling dependence using skew t copulas: Bayesian inference and applications." *Journal of Applied Econometrics*, 27(3): 500–522. MR2905037. doi: https://doi.org/10.1002/jae.1215. 1167

Smith, M. S. and Kohn, R. (1996). "Nonparametric regression using Bayesian variable selection." *Journal of Econometrics*, 75(2): 317–343. 1149, 1159, 1164, 1167

Smith, M. S. and Kohn, R. (2002). "Parsimonious covariance matrix estimation for longitudinal data." *Journal of the American Statistical Association*, 97: 1141–1153. MR1951266. doi: https://doi.org/10.1198/016214502388618942. 1149

Smith, M. S. and Maneesoonthorn, W. (2018). "Inversion copulas from nonlinear state space models with application to inflation forecasting." *International Journal of Forecasting*, 34: 389–407. 1144, 1167

Smith, M. S. and Vahey, S. (2016). "Asymmetric forecast densities for U.S. macroeconomic variables from a Gaussian copula model of cross-sectional and serial dependence." *Journal of Business and Economic Statistics*, 34(3): 416–434. MR3523785. doi: https://doi.org/10.1080/07350015.2015.1044533. 1167

Song, P. (2000). "Multivariate dispersion models generated from Gaussian copula." *Scandinavian Journal of Statistics*, 27(2): 305–320. MR1777506. doi: https://doi.org/10.1111/1467-9469.00191. 1144, 1146

Tipping, M. E. (2001). "Sparse Bayesian learning and the relevance vector machine." *Journal of Machine Learning Research*, 1(Jun): 211–244. MR1875838. doi: https://doi.org/10.1162/15324430152748236. 1167

Wauthier, F. L. and Jordan, M. I. (2010). "Heavy-Tailed Process Priors for Selective Shrinkage." In *Advances in Neural Information Processing Systems*, 2406–2414. 1167

Wilson, A. G. and Ghahramani, Z. (2010). "Copula Processes." In *Advances in Neural Information Processing Systems*, 2460–2468. 1167

Wood, S. N., Pya, N., and Säfken, B. (2016). "Smoothing parameter and model selection for general smooth models." *Journal of the American Statistical Association*, 111: 1548–1563. MR3601714. doi: https://doi.org/10.1080/01621459.2016.1180986. 1144

Acknowledgments

We thank a referee, the Associate Editor and an Area Editor for their careful reading of our paper and helpful comments.