Bayesian Inference for Regression Copulas

Michael Stanley Smith\(^a\) and Nadja Klein\(^b\)

\(^a\)Melbourne Business School, University of Melbourne, Carlton, VIC, Australia; \(^b\)School of Business and Economics, Statistics, Humboldt Universität zu Berlin, Berlin, Germany

**ABSTRACT**

We propose a new semiparametric distributional regression smoother that is based on a copula decomposition of the joint distribution of the vector of response values. The copula is high-dimensional and constructed by inversion of a pseudo regression, where the conditional mean and variance are semiparametric functions of covariates modeled using regularized basis functions. By integrating out the basis coefficients, an implicit copula process on the covariate space is obtained, which we call a “regression copula.” We combine this with a nonparametric margin to define a copula model, where the entire distribution—including the mean and variance—of the response is a smooth semiparametric function of the covariates. The copula is estimated using both Hamiltonian Monte Carlo and variational Bayes; the latter of which is scalable to high dimensions. Using real data examples and a simulation study, we illustrate the efficacy of these estimators and the copula model. In a substantive example, we estimate the distribution of half-hourly electricity spot prices as a function of demand and two time covariates using radial bases and horseshoe regularization. The copula model produces distributional estimates that are locally adaptive with respect to the covariates, and predictions that are more accurate than those from benchmark models. Supplementary materials for this article are available online.

**KEYWORDS**

DISTRIBUTIONAL REGRESSION; HAMILTONIAN MONTE CARLO; IMPlicit COPULA; P-SPLINES; RADIAL BASIS FUNCTIONS; VARIATIONAL BAYES

**ARTICLE HISTORY**

Accepted January 2020

**ARTICLE歷史**

Received February 2019

**1. Introduction**

Non- or semiparametric regression methods typically estimate only the mean of a response variable as an unknown smooth function of covariates. Yet in many applications, other features of the response distributions—such as higher moments and quantiles—also vary with the covariates. For example, to address this Rigby and Stasinopoulos (2005) and Klein et al. (2015) made all the parameters of a response distribution unknown smooth functions of the covariates. However, these authors assumed a specific parametric distribution for the response, conditional on the functions. In this article, we propose a novel class of semiparametric distributional regression models for continuous data that avoids such an assumption. It uses a copula decomposition of the joint distribution of a vector of values from a single response variable. To do so, we employ a new copula with a dependence structure that is an unknown smooth function of the covariate values, and model the marginal distribution of the response variable nonparametrically. The distributional regression is therefore flexible in two ways: nonparametric in a distributional sense with respect to the margin of the response, and semiparametric in a functional sense with respect to the covariates via the copula. It allows the entire distribution of the response to be a smooth unknown function of the covariates.

Copula models (McNeil, Frey, and Embrechts 2005; Nelsen 2006) are popular because the marginal distributions can be modeled arbitrarily and separately from the dependence structure. In this article, the copula has dimension equal to the length of the vector of response values, which can be high (87,648 in one of our examples). Few existing copulas can be used in such a situation, although copulas constructed by the inversion of a parametric distribution (Nelsen 2006, sec. 3.1) can. Such copulas are called either “inversion” or “implicit” copulas, and those constructed by the inversion of Gaussian (Song 2000), t (Demarta and McNeil 2005), and skew t (Smith, Gan, and Kohn 2012) distributions are popular. More flexible implicit copulas can be constructed by inverting the distribution of values of one or more response variables from parametric statistical models. We label these response variables “pseudo-responses” because they are not observed directly. Examples include implicit copulas constructed from factor models (Murray et al. 2013; Oh and Patton 2017), vector autoregressions (Smith and Vahey 2016), nonlinear state space models (Smith and Maneesoonthorn 2018), Gaussian processes (Wauthier and Jordan 2010; Wilson and Ghahramani 2010), and regularized regression (Klein and Smith 2019). These implicit copulas reproduce the dependence structure of the pseudo-response variables, and combining them with arbitrary margins produces a more flexible model that allows for a wide range of data distributions.

In this article, we show how to construct an implicit copula from a heteroscedastic semiparametric regression. Both the mean and variance of the pseudo-response are unknown smooth functions of covariates, each modeled using function bases with regularized coefficients. Because implicit copulas do not retain any information about the marginal (i.e., unconditional on the covariates) location and scale of the pseudo-response, we normalize the pseudo-response to have...
zero mean and unit variance marginally. By integrating out the
basis coefficients of the functions, we derive a copula that is
a smooth function of the covariate values and regularization
parameters only. We call this a “regression copula,” because
when used in a copula model for the vector of response values,
it captures the effect of the covariates. The regularization
parameters become the copula parameters, and these require
estimation.

There are two main challenges when estimating the copula
parameters: (i) the copula function and density are unavailable
in closed form, and (ii) the copula has dimension equal to
the sample size, which may be high. We outline two Bayesian
approaches to overcome these challenges. The first is a Markov
chain Monte Carlo (MCMC) sampler with a Hamiltonian Monte
Carlo (HMC) step (Neal 2011; Hoffman and Gelman 2014)
to evaluate the posterior distribution exactly. The second is a
variational Bayes (VB) estimator (Jordan et al. 1999; Ormerod
and Wand 2010) to compute approximate posterior inference
quickly when the sample size and dimension are high. The
VB estimator is based on a Gaussian approximation with a
sparse factor representation of its covariance matrix (Ong, Nott,
and Wand 2010) to compute approximate posterior inference
and the quantile score (Gneiting and Ranjan 2011), we show
the copula model is more accurate than two benchmark
distributional regression methods.

Finally, we note here that copulas have been used extensively
in multivariate regression frameworks, although our approach
is very different in two ways. First, previous approaches use
a low-dimensional copula to capture the dependence between
multiple response variables with regression margins, which is
often called a “copula regression” (Pitt, Chan, and Kohn 2006),
whereas we use a copula to capture the dependence between
different observations on a single response variable. Second,
most previous methods employ elliptical or vine (Aas et al. 2009)
copulas with closed form densities. In contrast, while our copula
does not have a closed form density, it is tractable and scalable
to higher dimensions, as illustrated in our empirical work.

The article is structured as follows. Section 2.1 shows how to
construct a distributional regression model using a regression
copula and an arbitrary margin. Our regression copula is
outlined in Section 2.2, along with some of its properties in
Section 2.3. Section 3 outlines exact and approximate Bayesian
posterior estimators, along with distributional and functional
prediction. Section 4 discusses the four univariate real data
examples and the comparison with benchmark alternatives
via simulation. Section 5 contains the application to electricity
prices, and Section 6 concludes.

2. Distributional Regression Using Implicit Copulas

In this section, we first introduce the copula model used for
distributional regression. Then we outline our proposed implicit
copula, along with some of its key properties.

2.1. Copula Model

Consider $N$ realizations $Y^{(N)} = (Y_1, \ldots, Y_N)$ of a continuous-
valued response, with corresponding covariate values $\tilde{x}^{(N)} =
(\tilde{x}_1, \ldots, \tilde{x}_N)$. Following Sklar’s theorem, the joint density of
$Y^{(N)} | \tilde{x}^{(N)}$ can always be written as

$$p(y^{(N)} | \tilde{x}^{(N)}) = \ell(F(y_1 | \tilde{x}_1), \ldots, F(y_N | \tilde{x}_N | \tilde{x}^{(N)}))$$

$$\times \prod_{i=1}^N p(y_i | \tilde{x}_i), \quad \text{for } N \geq 2.$$
Here, $c^f$ is the density of an $N$-dimensional copula process, and $F(y|\bar{x}_i)$ is the distribution function of $Y|\bar{x}_i$; both of which are unknown. In this article, we approximate this joint distribution, also conditional on copula parameters $\theta$, with the copula model

$$p(y_{N_i}|\bar{x}_{N_i}, \theta) = c_H \left(F_y(y_1), \ldots, F_y(y_N)|\bar{x}_{N_i}, \theta\right) \prod_{i=1}^{N} p_y(y_i).$$

(1)

The distribution $Y|\bar{x}_i$ is assumed to be invariant with respect to $\bar{x}_i$, and has density $p_y$ and distribution function $F_y$. However, the impact of the copariate values on $Y_{iN}$ is captured by the copula with density $c_H(u_{N_i}|\bar{x}_{N_i}, \theta)$, where $u_{N_i} = (u_1, \ldots, u_N')$ and $u_i = F_y(y_i)$. We call this a “regression copula” because it is a function of $\bar{x}_{N_i}$. It is a copula process on the copariate space with parameters $\theta$ that do not vary with $N$. We use the implicit copula proposed in the subsection below for $c_H$, and a major aim of this article is to show that by doing so, adopting Equation (1) provides a very flexible, but tractable, approach to distributional regression.

Before specifying $c_H$, we stress that even though the distribution $Y|\bar{x}_i$ is assumed invariant with respect to $\bar{x}_i$, $Y_i$ is not marginally independent of $\bar{x}_i$ when also conditioning on the unknown mean and variance functions of the pseudo-response, as shown in Part A.1 of the Web Appendix. Moreover, to see how the response is affected by the covariates in the distributional regression at Equation (1), consider a sample of size $n$ with $y = (y_1, \ldots, y_n)'$, covariate values $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_n)$ and $u = (u_1, \ldots, u_n)'$. Then a new response $y_{n+1}$ with corresponding covariate values $\bar{x}_{n+1}$ has predictive density

$$p(y_{n+1}|\bar{x}_{n+1}, \theta) = \int p(y, y_{n+1}|\bar{x}_{n+1}, \theta) dy$$

$$= \int c_H(u_{n+1}|\bar{x}_{n+1}, \theta) du \cdot p_y(y_{n+1}).$$

(2)

This density is a function of all the copariate values $\bar{x}_{n+1} = [\bar{x}, \bar{x}_{n+1}]$, which includes $\bar{x}_{n+1}$. Moreover, integrating over the posterior of $\theta$ gives the posterior predictive density of $y_{n+1}$ from the regression model as

$$p(y_{n+1}|\bar{x}_{n+1}, \theta) = \int p(y_{n+1}|\bar{x}_{n+1}, \theta) p(\theta|y) d\theta.$$

(3)

Equation (3) forms the basis for our distributional regression predictions as a function of $\bar{x}_{n+1}$, and its first two moments are estimates of the regression mean and variance functions. In Section 3.4, we show how to compute Equations (2) and (3) efficiently for our proposed copula.

### 2.2. Implicit Regression Copula

Key to our approach is the regression copula with density $c_H$, which is derived from a semiparametric heteroscedastic regression model for a pseudo-response. To do so, we first outline the regression and then construct its implicit copula with only the basis coefficients of the mean function integrated out, which is a Gaussian copula. Next, to derive the copula with the basis coefficients of the variance function also integrated out, it is represented as an integral of the Gaussian copula. We show that such a representation is computationally efficient.

#### 2.2.1. Pseudo-Response Regression Model

Consider a regression model for a pseudo-response $\tilde{Z}_i$ with covariates $\tilde{x}_i = [x_i, w_i]$ given by

$$\tilde{Z}_i = \tilde{m}(x_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2_i \epsilon_i^2),$$

$$\sigma^2_i = \exp(g(w_i)), \quad i = 1, \ldots, n,$$

(4)

where the first and second moments are smooth unknown functions $\tilde{m}$ and $g$ of the two copariate vectors. We model these using linear combinations of basis functions $b_1, \ldots, b_{p_1}$ and $v_1, \ldots, v_{p_2}$, such that $\tilde{m}(x) = \sum_{i=1}^{p_1} \beta_i b_i(x)$ and $g(w) = \sum_{j=1}^{p_2} \alpha_j v_j(w)$. Typical choices for the bases include polynomial or B-spline bases for a scalar copariate, and additive or radial bases for multiple covariates. With these approximations, the regression model is usually called semiparametric.

For $n$ pseudo-response values $\tilde{Z} = (\tilde{Z}_1, \ldots, \tilde{Z}_n)'$ the regression at Equation (4) can be written as

$$\tilde{Z} = B\beta + \epsilon, \quad \epsilon = (\epsilon_1, \ldots, \epsilon_n)' \sim N(0, \sigma^2 \Sigma),$$

$$\Sigma = \text{diag}(\sigma^2_1, \ldots, \sigma^2_n),$$

$$\sigma^2_i = \exp(g(w_i)), \quad i = 1, \ldots, n,$$

where $\beta = (\beta_1, \ldots, \beta_n)$, $\alpha = (\alpha_1, \ldots, \alpha_{n-1})$, and the design matrices $B \in \mathbb{R}^{n \times p_1}$ and $V \in \mathbb{R}^{n \times p_2}$ have $i$th rows $b_i = (b_1(x_i), \ldots, b_{p_1}(x_i))'$ and $v_i = (v_1(w_i), \ldots, v_{p_2}(w_i))'$. To produce smooth and efficient function estimates it is usual to regularize the basis coefficients $\beta$ and $\alpha$. In a conjugate Bayesian context, this corresponds to adopting the conditionally Gaussian priors

$$\beta|\theta, \sigma^2 \sim N(0, \sigma^2 P_{\theta, \beta}^{-1}), \quad \alpha|\theta, \sigma^2 \sim N(0, P_{\alpha}^{-1})$$

(6)

with smoothing (or “hyper”) parameters $\theta$ and $\sigma^2$. The forms of the precision matrices $P_{\theta, \beta}, P_{\alpha}$ are typically matched with the choice of bases for $\tilde{m}$ and $g$, for which we give two examples later.

#### 2.2.2. Regression Copula Construction

We extract two copulas from the regression model defined at Equations (4)–(6). They are called “implicit” (McNeil, Frey, and Embrechts 2005, p. 190) or “inversion” (Nelsen 2006, p. 51) copulas because they are constructed by inverting Sklar’s theorem. The copulas are $n$-dimensional with dependence structures that are (smooth) functions of $\bar{x} = [x, w]$, with $x = (x_1', \ldots, x_n')'$ and $w = (w_1', \ldots, w_n')'$. The first regression copula derived is the implicit copula of the distribution $\tilde{Z}|x, w, \sigma^2, \alpha, \theta, \sigma^2_i, \alpha_i$, which we label $C_1$. To construct $C_1$, note that the prior for $\beta$ is conjugate and can be integrated out of the distribution for $\tilde{Z}$ analytically, giving

$$\tilde{Z}|x, w, \sigma^2, \alpha, \theta, \sigma^2_i, \alpha_i \sim N(0, \sigma^2 \Sigma - \Sigma^{-1} B \Omega B' \Sigma^{-1})^{-1} \Sigma + B P_{\theta, \beta}^{-1} B',$$

(7)

where $\Omega = (B' \Sigma^{-1} B + P_{\theta, \beta}^{-1})^{-1}$, and by applying the Woodbury formula

$$\Sigma^{-1} - \Sigma^{-1} B \Omega B' \Sigma^{-1} = \Sigma + B P_{\theta, \beta}^{-1} B'.$$

It is straightforward to show that the copula of a normal distribution is the widely employed Gaussian copula (Song 2000). It is obtained by standardizing the marginal means to zero and the variances to one. The margin in $\tilde{Z}_i$ at Equation (7) is $N(0, \sigma^2_i \exp(g(w_i)) + \sum_{j=1}^{p_2} \alpha_j v_j(w_i))$, so that we normalize $\tilde{Z}$.
by the diagonal matrix \( \sigma^{-1} S(x, w, \alpha, \theta_\beta) = \sigma^{-1} \text{diag}(s_1, \ldots, s_n) \) with 
\[ s_i = \left[ \exp(v' \alpha) + b'_\beta(\theta_\beta)^{-1} b_i \right]^{-1/2}, \] 
to get 
\[ Z = \sigma^{-1} S(x, w, \alpha, \theta_\beta) Z. \] 
With this, the regression at Equation (4) can be rewritten for the standardized pseudo-response as
\[ Z_i = m(x_i, w_i) + \frac{s_i}{\sigma} \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2 \sigma_i^2), \] 
(8)
where \( m(x_i, w_i) = (s_i/\sigma) \tilde{m}(x_i) = (s_i/\sigma) b_i \beta \) is a function of both \( x_i \) and \( w_i \), because \( s_i \) is also.

Denoting \( S \equiv S(x, w, \alpha, \theta_\beta) \) for conciseness, the distribution of the normalized vector with \( \beta \) integrated out is
\[ Z|x, w, \sigma^2, \alpha, \theta_\beta, \theta_\alpha \sim N(0, R), \] 
and \( N(0, 1) \) margins for all elements \( Z_1, \ldots, Z_n \). It is straightforward to show (Song 2000) that the random vectors \( Z \) and \( \tilde{Z} \) (conditional on \( x, w, \alpha, \theta_\beta \)) have the same Gaussian copula function
\[ C_1(u|x, w, \alpha, \theta_\beta) = \Phi \left( \Phi_1^{-1}(u_1), \ldots, \Phi_1^{-1}(u_n); 0, R \right), \] 
where \( u = (u_1, \ldots, u_n) \), and \( \Phi; (0, R) \) and \( \Phi_1 \) are the distribution functions of \( N(0, R) \) and \( N(0, 1) \) distributions, respectively. This is a regression copula because \( R \) is a function of \( \tilde{x} \).

We make a number of observations on \( C_1 \). First, the parameter \( \sigma^2 \) does not feature in the expression for \( R \), and is unidentified in the copula, so that we set \( \sigma^2 = 1 \) throughout the article. Second, if the density for the distribution of \( Z \) at Equation (9) is denoted as \( p_Z \), with marginal densities \( p_{Z_i} \) for \( i = 1, \ldots, n \), then the copula density \( c_1 \) is
\[ c_1(u|x, w, \alpha, \theta_\beta) = \frac{p_Z(z|x, w, \alpha, \theta_\beta)}{\prod_{i=1}^n p_{Z_i}(z_i|x, w, \alpha, \theta_\beta)} = \frac{\phi(z; 0, R)}{\prod_{i=1}^n \phi_1(z_i)} \] 
(10)
where \( z_i = \Phi_1^{-1}(u_i), z = (z_1, \ldots, z_n), \) and \( \phi; (0, 0) \) and \( \phi_1 \) are the densities of \( N(0, R) \) and \( N(0, 1) \) distributions, respectively. Third, if a nonconjugate prior is used for \( \beta \), then \( C_1 \) is not a Gaussian copula (something we do not consider in this article), Last, because \( R \) is a function of \( \alpha \), so is the dependence structure of \( C_1 \). If \( \alpha = 0 \), then \( C_1 \) corresponds to the copula of a homoscedastic regression, as discussed by Klein and Smith (2019).

The second regression copula derived is the implicit copula of \( Z \) with both \( \beta \) and \( \alpha \) integrated out. We label this \( C_H \) (for heteroscedastic regression copula), and it is this copula with density \( c_H \) that is used to model the observed data at Equation (1).

**Theorem 1 (Definition of \( C_H \) and \( c_H \)).** If \( \tilde{Z} \) follows the heteroscedastic regression for the pseudo-response at Equations (4)–(6), \( Z_i = s_i \tilde{Z}_i \) is the normalized response at Equation (8) with \( \sigma = 1 \), \( \tilde{x} = [x, w] \) are the covariate values and \( \theta = [\theta_\beta, \theta_\alpha] \), then the \( n \)-dimensional implicit copula of the distribution \( Z|x, \theta \) has density
\[ c_H(u|x, \theta) = \int c_1(u|x, w, \alpha, \theta_\beta) p(\alpha|\theta_\alpha) d\alpha \] 
\[ = \frac{\int \phi(z; 0, R) p(\alpha|\theta_\alpha) d\alpha}{\prod_{i=1}^n \phi_1(z_i)}, \] 
and copula function
\[ C_H(u|x, \theta) = \int c_H(u|x, w, \alpha, \theta_\beta) p(\alpha|\theta_\alpha) d\alpha \] 
\[ = \int \Phi(z; 0, R) p(\alpha|\theta_\alpha) d\alpha, \] 
where \( u = (u_1, \ldots, u_n) \), the marginal \( u_i = F(z_i|x, \theta) = \Phi_1(z_i) \), so that \( z_i = \Phi^{-1}_1(u_i) \).

**Proof.** See Part A of the Web Appendix.

We make three observations on \( C_H \) defined in Theorem 1. First, integration over \( \alpha \) is required to compute \( C_H \) and \( c_H \). In Section 3, we show how to do this integration exactly using HMC, and approximately using VB methods, when computing posterior inference. Second, the dependence parameters of \( C_H \) are the the smoothing parameters \( \theta = [\theta_\alpha, \theta_\beta] \) of \( \tilde{m}, g \) in the regression for the pseudo-response at Equation (4). Last, it is much simpler to construct the implicit copula of \( Z \), rather than \( \tilde{Z} \). Here is because constructing the latter copula would involve evaluating (and inverting) the \( n \) marginal distribution functions
\[ \tilde{F}_i(z_i|x, \theta) = \int \Phi_1(z_i|\theta) p(\alpha|\theta_\alpha) d\alpha, \quad i = 1, \ldots, n. \] 
Each of these involves computing a \( p_2 \)-dimensional integral using numerical methods. In contrast, the margin of \( Z_i|\tilde{x}, \theta \) is simply a standard normal, which greatly simplifies evaluation of \( C_H \).

**2.3. Properties of \( C_H \)**

Here, we state some properties of the regression copula \( C_H \). First, the independence copula is a limiting case of this copula, as outlined in Theorem 2:

**Theorem 2.** Let \( \Pi(u) = \prod_{i=1}^n u_i \) be the independence copula function (Nelsen 2006, p. 11), and \( \gamma_\beta(\theta_\beta) < \infty \) be the maximum marginal variance of the prior for \( \beta \) at Equation (6), then
\[ \lim_{\gamma_\beta \to 0} C_H(u|x, \theta) = \Pi(u). \] 

**Proof.** See Part A of the Web Appendix.

An implication of Theorem 2 is that the relationship between the response and covariates is weak when the posterior of \( \gamma_\beta \) is close to zero.

Below we give expressions for some common dependence metrics of the bivariate sub-copula \( C_H^{ij} \) of \( C_H \) in elements \( 1 \leq i < j \leq n \). The derivations are given in Part A of the Web Appendix.

(i) For \( q \in (0, 1) \), if \( (U_i, U_j) \sim C_H^{ij} \), the lower and upper quantile dependence are
\[
\lambda_{ij}^L(q | \mathbf{x}, \boldsymbol{\theta}) \equiv \Pr(U_i < q | U_j < q) = \int \lambda_{ij}^L(q | \mathbf{x}, \omega) \rho(\omega | \boldsymbol{\theta}_\omega) d\omega, \\
\lambda_{ij}^U(q | \mathbf{x}, \boldsymbol{\theta}) \equiv \Pr(U_i > q | U_j > q) = \int \lambda_{ij}^U(q | \mathbf{x}, \omega) \rho(\omega | \boldsymbol{\theta}_\omega) d\omega,
\]

where \(\lambda_{ij}^L\) and \(\lambda_{ij}^U\) are the lower and upper pairwise quantile dependences of a bivariate Gaussian copula with correlation parameter \(r_{ij}\) given by the \((i,j)\)th element of \(R\) in Equation (9).

(ii) The lower and upper extremal tail dependence

\[
\lambda_{ij}^L = \lim_{q \to 0} \lambda_{ij}^L(q | \mathbf{x}, \boldsymbol{\theta}) = 0, \quad \text{and} \quad \lambda_{ij}^U = \lim_{q \to 1} \lambda_{ij}^U(q | \mathbf{x}, \boldsymbol{\theta}) = 0.
\]

(iii) Spearman’s rho and Kendall’s tau are

\[
\rho^S_{ij}(\mathbf{x}, \boldsymbol{\theta}) = \frac{6}{\pi} \int \arcsin(r_{ij}/2) \rho(\omega | \boldsymbol{\theta}_\omega) d\omega, \\
\tau^K_{ij}(\mathbf{x}, \boldsymbol{\theta}) = \frac{2}{\pi} \int \arcsin(r_{ij}) \rho(\omega | \boldsymbol{\theta}_\omega) d\omega,
\]

where \(r_{ij}\) as defined above and is a function of \(x, w, \alpha, \beta, \theta\).

These metrics are functions of the copula parameters \(\theta\), and also all \(n\) covariate values \(\mathbf{\bar{x}} = \{x, w\}\), rather than just \(x_i, x_j, w_i, w_j\). (We return to this feature in Section 4, where we show it corresponds to local adaptivity of the distributional estimates from the copula model). The metrics are computed with respect to the posterior of \(\theta\) for the examples in Section 4.

### 3. Estimation

Estimation of the copula model at Equation (1) requires estimation of both the marginal \(F_Y\) and parameters \(\theta\). It is popular to use two stage estimators, where \(F_Y\) is estimated first, followed by \(\theta\), because they are simpler to implement and only involve a minor loss of efficiency (Joe 2005). For \(F_Y\) we use the adaptive kernel density estimator (labeled “KDE”) of Shimazaki and Shinomoto (2010) and a Dirichlet process mixture estimator (Neal 2000) (labeled “DPfit”). For the latter, when estimating \(\theta\) using MCMC, uncertainty with respect to the estimate of \(F_Y\) can also be integrated out by following Grazian and Liseo (2017) and using the draws of \(F_Y\) at each sweep, instead of conditioning on its posterior point estimate. We find in our empirical work that this has only a minor effect on the copula and distributional estimates. Last, in some examples we transform the response variable—for example, by taking its logarithm—before applying the KDE. In this case, the marginal density of the response on the original scale is easily obtained by multiplying the KDE and Jacobian of the transformation in the usual fashion, although we present results on the logarithmic scale for clarity.

#### 3.1. Likelihood

Estimation of \(\theta\) based on Equation (1) with \(N = n\) observations is difficult because \(c_{ij}\) at Theorem 1 is expressed as an integral over \(\alpha\). Nevertheless, the likelihood can be evaluated by expressing it conditional on the coefficients \(\beta\) and \(\alpha\), and then integrating them out using Bayesian methods, which is the approach we employ. The Jacobian of the transformation from \(Z\) to \(Y\) is \(|J_{Z-Y}| = \prod_{i=1}^{n} p_Y(y) / p_Y(z)\), and by a change of variables and Equation (8), the conditional likelihood is

\[
p(y | x, \beta, \alpha, \theta, \omega) = p(z | x, \beta, \alpha, \theta, \omega) / |J_{Z-Y}| = \phi(z; S \Sigma, S \Sigma),
\]

which can be evaluated in \(O(n)\) operations because \(S\) and \(\Sigma\) are diagonal. Below we show how to evaluate the posterior of \(\theta\) exactly by generating \(\alpha\) using a HMC step within a MCMC scheme. However, for large \(n\) and some choices of \(P_\beta, P_\alpha\) exact samplers can be sticky and/or slow, so that we also develop a VB estimator for approximate inference requiring less computation. Both approaches estimate the posterior of the parameters augmented with the basis coefficients, denoted as \(\theta = [\beta, \alpha, \theta_\beta, \theta_\alpha]\) with dimension \(p_\theta\).

#### 3.2. Exact Estimation Using MCMC

Each scalar element of \(\theta\) (or of a reparameterization) is generated using a normal approximation based on analytical derivatives of the logarithm of its conditional posterior. The coefficients \(\beta\) are generated from a multivariate normal. Details on these two steps are given in Part B.1 of the Web Appendix for the copulas in Section 4.

The most challenging aspect of this sampler is generating from the conditional posterior of \(\alpha\). We found Gaussian or random walk proposals result in poor mixing of the Markov chain, so that a HMC (Neal 2011) step is employed instead. This augments \(\alpha\) by momentum variables, and draws from an extended target distribution that is proportional to the exponential of the Hamiltonian function. Dynamics specify how the Hamiltonian function evolves, and its volume-conserving property results in high acceptance rates of the proposed iterates.

We use the leapfrog integrator (Neal 2011), which employs the logarithm of the target density

\[
l_\alpha \equiv \log(p(\alpha | x, z, \theta \setminus \alpha)) = -\frac{1}{2} \sum_{i=1}^{n} (\log(s_{i1}^2) + \log(s_{i1}^2)) - \frac{1}{2} \left( z'(\Sigma_1 S)^{-1} z - 2 \beta' B' \Sigma^{-1} S^{-1} S \Sigma^{-1} \beta - 1 \right) \alpha', P_\alpha(\theta_\alpha) \alpha
\]

and its gradient

\[
\nabla_\alpha l_\alpha = -P_\alpha(\theta_\alpha) \alpha - \frac{1}{2} V' \left[ \left( \frac{\partial s_{i1}^2}{\partial s_{i1}}, \ldots, \frac{\partial s_{i1}^2}{\partial s_{in}} \right)' + \left( \frac{\partial^2 s_{i1}^2}{\partial s_{i1}^2 s_{i1}}, \ldots, \frac{\partial^2 s_{i1}^2}{\partial s_{i1}^2 s_{in}} \right)' \right] + \frac{1}{2} V' \left( B \beta \right)' \left( \frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_n} \right)' - \left( \frac{\partial \beta_1^2}{\partial \beta_1}, \ldots, \frac{\partial \beta_n^2}{\partial \beta_n} \right)' + V' \left[ z' \left( \frac{\partial \beta_1^2}{\partial \beta_1}, \ldots, \frac{\partial \beta_n^2}{\partial \beta_n} \right)' \left( B \beta \right)' \right].
\]
where $\phi$ is the Hadamard product, $\eta_\theta = V\alpha, \kappa_{1,j} = (\sigma^2_{1,j})^{-1}$, $\kappa_{2,j} = (\sigma^2_{2,j})^{-1}$, a closed form expression for $\frac{\partial^2}{\partial \eta_\theta}$ is given in the Web Appendix and $\frac{\partial^2}{\partial \eta_\theta^2} \sigma^{-2} = 1$. The step size $\epsilon$ and the number of leapfrog steps $L$ at each sweep are set using the dual averaging approach of Hoffman and Gelman (2014) as follows. A trajectory length $t = \epsilon L = 1$ is obtained by preliminary runs of the sampler with small $\epsilon$ (to ensure a small discretization error) and large $L$ (to move far). The dual averaging algorithm uses this trajectory length and adaptively changes $\epsilon, L$ during $M_{\text{adapt}} \leq M$ iterations of the complete sampler with $M$ sweeps, to achieve a desired rate of acceptance $\delta$. In our examples $\delta = 0.75$, while the starting value for $\epsilon$ is given by Algorithm 4 of Hoffman and Gelman (2014). Algorithm 1 gives the HMC step at sweep $m$ of the sampler.

Algorithm 1 Hamiltonian Monte Carlo with dual averaging

Given $\theta^{(m-1)}, \tilde{\theta}_{m-1}, \epsilon_{m-1}, \delta_l, \mu = \log(\epsilon_0), \tilde{H}_{m-1}, M_{\text{adapt}}$:  

1. Set $\epsilon = 0.05, t_0 = 10, \kappa = 0.75$ as in Hoffman and Gelman (2014).
2. Sample $r \sim N_p(0, I)$.
3. Set $\alpha^{(m)} \leftarrow \alpha^{(m-1)}, \tilde{\alpha} \leftarrow \alpha^{(m-1)}, \tilde{r} \leftarrow r, L_m \leftarrow \max(1, \text{round}(t / \epsilon_{m-1})).$
4. For $j = 1, \ldots, L_m$ do $>$ leapfrog integrator steps.
5. Set $\tilde{r} = r + (\epsilon_{m-1}/2)\nabla_a \log|a = \tilde{a}|.$
6. Set $\tilde{\alpha} = \tilde{\alpha} + \epsilon_{m-1}\tilde{r}.$
7. Set $\tilde{r} = \tilde{r} + (\epsilon_{m-1}/2)\nabla_a \log|a = \tilde{a}|.$
8. End for
9. With probability $\tilde{a} = \min(1, e^{(\epsilon_{m-1}-0.5\tilde{r})})$, set $\alpha^{(m)} \leftarrow \tilde{\alpha}.$
10. If $m \leq M_{\text{adapt}}$ then $>$ dual averaging step.
11. Set $\tilde{H}_m = (1-1/(m+t_0))\tilde{H}_{m-1} + (1/(m+t_0))\delta_l \tilde{a}.$
12. Set $\log(\epsilon_m) = \mu - \sqrt{\tilde{r}(m)\tilde{H}_m}(\gamma), \log(\epsilon_m) = m^{-1}\log(\epsilon_m) + (1-m^{-1})(\log(\epsilon_m)).$
13. Else
14. Set $\epsilon_m = \tilde{e}_{M_{\text{adapt}}}.$
15. End if

In our empirical work, a burn-in of 40,000 iterates was employed, after which a Monte Carlo sample of size $J = 50,000$ was collected.

### 3.3. Approximate Estimation Using VB

The VB estimator approximates the augmented posterior $p(\theta|y) \propto p(y|\theta)p(\theta) \equiv h(\theta)$ with a tractable density $q(\theta)$. Here, $p(y|\theta)$ is the conditional likelihood at Equation (11), and $\lambda$ is a vector of “variational parameters” which are calibrated by minimizing the Kullback–Leibler divergence between $q(\theta)$ and $p(\theta|y)$. It is straightforward to show (Ormerod and Wand 2010) that this is equivalent to maximizing the variational lower bound

$$\ell(\lambda) = \int q(\theta) \log \frac{p(y, \theta)}{q(\theta)} \, d\theta = \mathbb{E}_q \left[ \log(h(\theta)) - \log(q(\theta)) \right],$$

with respect to $\lambda$. The expectation in Equation (12) is with respect to the variational approximation (VA) with density $q$, and cannot be computed in closed form. Therefore, a stochastic gradient ascent (SGA) algorithm (Honkela et al. 2010; Salimans and Knowles 2013) is used to maximize $\ell$. This employs an unbiased estimate $\nabla_\lambda \ell(\lambda)$ of the gradient of $\ell$ to compute the update

$$\lambda(t+1) = \lambda(t) + \rho(t) \circ \nabla_\lambda \ell(\lambda(t)),$$

recursively. If $\{\rho(t)\}_{t \in \mathbb{N}}$ is a sequence of vector-valued learning rates that fulfill the Robbins–Monro conditions, then the sequence $\{\lambda(t)\}_{t \in \mathbb{N}}$ converges to a local optimum (Bottou 2010). The learning rates are set adaptively using the ADADELTA method as in Ong, Nott, and Smith (2018).

For the SGA algorithm to be efficient, the estimate $\nabla_\lambda \ell(\lambda)$ should exhibit low variance. To achieve this we use the so-called “reparameterization trick” (Kingma and Welling 2014; Rezende, Mohamed, and Wierstra 2014). This expresses $\theta$ as a function $\theta = a(\xi, \lambda)$ of another random variable $\xi$ that has a density $p_\xi$ that does not depend on $\lambda$. In this case, the lower bound is

$$\ell(\lambda) = \mathbb{E}_{p_\xi} \left[ \log(h(a(\xi, \lambda)) - \log(q(\xi, a(\xi, \lambda))) \right],$$

where $E_{p_\xi}$ is an expectation with respect to $p_\xi$. Note that when differentiating Equation (13) with respect to $\lambda$, information from the posterior density is used, whereas it is not when differentiating Equation (12). Differentiating the expectation in Equation (13) gives

$$\nabla_\lambda \ell(\lambda) = \mathbb{E}_{p_\xi} \left[ \frac{\partial a(\xi, \lambda)}{\partial \lambda} \nabla_\theta \log(h(a(\xi, \lambda)) - \log(q(\xi, a(\xi, \lambda))) \right] - \nabla_\lambda \log(q(\xi, a(\xi, \lambda))),$$

which follows from the “log-derivative trick” ($\mathbb{E}_q(\nabla_\lambda \log(q(\xi))) = 0$). An unbiased estimate of the expectation at Equation (14) can be computed by simulating from $p_\xi$ and efficient implementations typically use just a single iterate of $\xi$.

Successful application of variational methods requires $q$ to be tractable and a suitable transformation for the reparameterization trick. We follow Ong, Nott, and Smith (2018), who use the Gaussian VA $q(\theta) = \phi(\theta; \mu, \Sigma)$ with a parsimonious factor covariance structure, which meets both conditions. Here, $\Sigma = \Psi \Psi^T + \Delta^2$, where $\Psi$ is a full rank $p_\theta \times K$ matrix with $K \ll p_\theta$, $\Delta = (d_1, \ldots, d_p)$, and $\Delta = \text{diag}(d)$. If $\Psi = (\Psi_{ij})$, then the elements $\Psi_{ij} = 0$ for $j > i$. For uniqueness, it is common to also assume $\Psi_{ij} = 1$, although we do not because the lack of uniqueness does not hinder the optimization, and the unconstrained parameterization is more convenient. To apply the reparameterization trick, set $\theta = \mu + \Psi \xi + \Delta \circ \xi$, where $\xi = (\xi', \xi')', \xi \in \mathcal{R}^K, \Delta \in \mathcal{R}^{p_\theta}$, and $p(\xi)$ is the density of a $N(0, I)$ distribution. In this case, $\lambda = (\mu', \text{vech}(\Sigma'), \delta')$, which has gradient $\nabla_\lambda \ell(\lambda) = (\nabla_\lambda \ell(\lambda'), \nabla_{\text{vech}(\Sigma)} \ell(\lambda'), \nabla_\delta \ell(\lambda'))$, which can be computed analytically and efficiently; see Part B.2 of the Web Appendix. An unbiased estimate $\nabla_\lambda \ell(\lambda)$ is then computed using a sample from $p_\xi$. Algorithm 2 computes the VB estimates.
Algorithm 2 SGA for a Gaussian VA with a factor covariance structure.

Given \( \lambda^{(0)} = [\mu^{(0)}, \psi^{(0)}, d^{(0)}], t = 0:\)

1. \textbf{while} Stopping rule is not satisfied \textbf{do}
2. \hspace{1em} Generate \((\xi^t, \delta^t) \sim N(0, I)\).
3. \hspace{1em} Construct the unbiased estimates \( \nabla_{\mu} \mathcal{L}(\lambda^{(i)}), \)
\( \nabla_{\psi} \mathcal{L}(\lambda^{(i)}) \) \text{ and } \( \nabla_d \mathcal{L}(\lambda^{(i)}) \) using the single sample \((\xi^t, \delta^t)'\).
4. \hspace{1em} Compute the adaptive learning rate vector \( \rho^{(t)} = (\rho_{\mu}^{(t)}, \rho_{\psi}^{(t)}), \rho_{\delta}^{(t)} \) using ADADELTA.
5. \hspace{1em} Set \( \mu^{(t+1)} = \mu^{(t)} + \rho_{\mu}^{(t)} \nabla_{\mu} \mathcal{L}(\lambda^{(i)}), \)
6. \hspace{1em} Set \( \psi^{(t+1)} = \nabla_{\psi} \mathcal{L}(\lambda^{(i)}) \) \text{ and } \( \psi^{(t+1)} = 0 \text{ for } j > i. \)
7. \hspace{1em} Set \( d^{(t+1)} = d^{(t)} + \rho_{d}^{(t)} \nabla_d \mathcal{L}(\lambda^{(i)}). \)
8. \hspace{1em} Set \( \lambda^{(t+1)} \leftarrow (\mu^{(t+1)}, \psi^{(t+1)}, d^{(t+1)}) \) and \( t \leftarrow t + 1. \)
9. \textbf{end while}

In our empirical work, the calibrated value \( \hat{\lambda} \) is set to the average value over the last 10% of steps. A point estimate of the parameters is simply \( \hat{\phi}_{vb} = E_{\phi|Y}(\phi) = \mu. \)

3.4. Distributional and Functional Prediction

For a new observation \( Y_{n+1} \) of the response with covariate values \( x_{n+1} = (x_{n+1}, w_{n+1}) \), the posterior predictive density at Equation (3) is used as a distributional prediction. This can be evaluated by considering a change of variables from \( Y_{n+1} \) to \( Z_{n+1} = \Phi^{-1}_1(F_Y(Y_{n+1})) \) as follows:

\[
p(Y'_{n+1} | x_{n+1}, y) = \int p(y'_{n+1} | x_{n+1}, y) p(\theta | y) d\theta = p_Y(y_{n+1}) \int p(z_{n+1} | x_{n+1}, y) p(\theta | y) d\theta.
\]

From Equation (8) the standardized pseudo-response has a conditional distribution \( Z_{n+1} | x_{n+1}, y, \hat{\theta} \sim N(m(x_{n+1}, w_{n+1}), \sigma_{n+1}^2) \), which is independent of the \((n+1)\) elements of \( x_{n+1} \), except for \( x_{n+1} \). Thus, a Monte Carlo estimator of the posterior predictive density is

\[
\hat{p}(y_{n+1} | x_{n+1}) \equiv \frac{\hat{p}_Y(y_{n+1})}{\phi_1(\Phi^{-1}_1(\hat{F}_Y(y_{n+1})))} \left\{ \frac{1}{j} \sum_{j=1}^J \frac{1}{s_{n+1}^{(j)}} \right\} \phi_1 \left( \Phi^{-1}_1(\hat{F}_Y(y_{n+1})) - m^{(j)}(x_{n+1}, w_{n+1}) \right)\}
\]

(15)

Here, \( m^{(j)}, s^{(j)}_{n+1}, \sigma_{n+1} \) are \( m_{n+1}, \sigma_{n+1} \) computed from draw \( j \) of \( \theta \) from the posterior using the exact sampler in Section 3.2, or a draw from \( q_{vb}(\theta) \) for the VB estimator in Section 3.3. Because Equation (15) is only a function of the element \( x_{n+1} \) of \( x_{n+1} \), we write it henceforth as \( \hat{p}(y_{n+1} | x_{n+1}). \)

A second estimate that is based on a point estimate \( \hat{\theta} \) is

\[
\hat{p}_n(y_{n+1} | x_{n+1}) \equiv \frac{\hat{p}_Y(y_{n+1})}{\phi_1(\Phi^{-1}_1(\hat{F}_Y(y_{n+1})))} \left\{ \frac{1}{\hat{s}_{n+1}\sigma_{n+1}} \right\} \phi_1 \left( \Phi^{-1}_1(\hat{F}_Y(y_{n+1})) - \hat{m}(x_{n+1}, w_{n+1}) \right)\}
\]

with \( \hat{m}, \hat{s}_{n+1}, \sigma_{n+1} \) computed from \( \hat{\theta} \). This second estimate will typically be much faster to evaluate, and can be used with the VB estimate \( \hat{\theta}_{vb} \) or the exact posterior mean.

We denote the regression and variance functions as \( f(x_{n+1}, w_{n+1}) \equiv E(Y_{n+1} | x_{n+1}, w_{n+1}) \) and \( v(x_{n+1}, w_{n+1}) \equiv \text{var}(Y_{n+1} | x_{n+1}, w_{n+1}) \), respectively. We stress that these are different than \( m \) and \( g \) in Equation (4), which are the mean and variance functions for the pseudo-response. Estimates of \( f \) and \( v \) can be computed from the posterior predictive distribution at Equation (3) as follows. Let \( b_{n+1} \) and \( v_{n+1} \) be the vectors of function basis terms evaluated at \( x_{n+1} \) and \( w_{n+1} \), respectively. Then the Bayesian posterior predictive function estimators are

\[
\hat{f}(x_{n+1}, w_{n+1}) \equiv E(Y_{n+1} | \hat{x}_{n+1}, y) = \int E(Y_{n+1} | \hat{x}_{n+1}, \theta) p(\theta | y) d\theta
\]

\[
\hat{v}(x_{n+1}, w_{n+1}) \equiv \text{var}(Y_{n+1} | \hat{x}_{n+1}, y) = \int \text{var}(Y_{n+1} | \hat{x}_{n+1}, \theta) p(\theta | y) d\theta,
\]

where (by a change of variables from \( Y_{n+1} \) to \( Z_{n+1} \)) the terms in the integrands are

\[
E(Y_{n+1} | \hat{x}_{n+1}, \theta) = \int F_Y^{-1}(\phi_1(z_{n+1})) p(z_{n+1} | \hat{x}_{n+1}, \theta) dz_{n+1}
\]

\[
\phi_1 \left( z_{n+1} - s_{n+1} b_{n+1} \right) dz_{n+1}
\]

\[
\phi_0 \left( x_{n+1}, w_{n+1} \right),
\]

\[
\text{var}(Y_{n+1} | \hat{x}_{n+1}, \theta) = \int \left( F_Y^{-1}(\phi_1(z_{n+1})) \right)^2 p(z_{n+1} | \hat{x}_{n+1}, \theta) dz_{n+1}
\]

\[
\phi_1 \left( z_{n+1} - s_{n+1} b_{n+1} \right) dz_{n+1}
\]

\[
\phi_0 \left( x_{n+1}, w_{n+1} \right)^2,
\]

\[
\sigma_{n+1} = \exp(v'_{n+1} \alpha) \text{ and } s_{n+1} = [\exp(v'_{n+1} \alpha) + b_{n+1}' \beta]^{-1} b_{n+1}^{-1/2}.
\]

The integrals with respect to \( z_{n+1} \) above are computed using standard univariate numerical methods. The integrals at Equation (16) can be computed with draws from either the posterior using the exact estimator or the calibrated VA when using the VB estimator. Estimators that are faster to compute can be obtained by simply conditioning on either the
posterior mean or $\hat{\vartheta}_{VB}$, in a similar fashion as with the density estimator.

Last, other distributional summaries—for example, quantiles, higher order moments, or Gini coefficients—can be computed similarly.

4. P-Spline Copulas

In this section, we construct regression copulas for a single covariate using cubic B-spline bases for $m_g$. The knots are equally spaced over the range of the covariate, selected so that $\dim(\beta) = 22$, and $\dim(\alpha) = 12$. The matrices $P_\beta, P_\alpha$ are the precisions of stationary AR(2) models, each parameterized in terms of the disturbance variance $\tau_j^2$ and two partial autocorrelations $-1 < \psi_{j,1}, \psi_{j,2} < 1$, for $j = \alpha, \beta$; see Barndorff-Nielsen and Schou (1973). Thus, $P_\beta, P_\alpha$ are of full rank, and $\theta_\beta = \{\tau^2_\beta, \psi_{\beta,1}, \psi_{\beta,2}\}$, $\theta_\alpha = \{\tau^2_\alpha, \psi_{\alpha,1}, \psi_{\alpha,2}\}$. This combination of basis and type of prior for the basis coefficients is widely called a “P-spline,” although random walk priors are more popular. However, the precision matrix of a random walk is of reduced rank, in which case the distribution of $Z$ with $\beta, \alpha$ integrated out is improper, and it does not have a proper copula density, so that such a prior cannot be used. An alternative is to employ the proper prior suggested by Chib and Jeliazkov (2006), although we do not do so here.

4.1. Real Data Examples

We illustrate our approach using the four real datasets listed in Table 1. Each has one covariate (although we consider multiple

| Dataset | n  | Covariate            | Response                                      | Source                  |
|---------|----|----------------------|-----------------------------------------------|-------------------------|
| Geyser  | 299| Waiting time (min)   | Eruption time (min)                           | Venables and Ripley (2002) |
| Rents   | 3082| Apartment area (m²) | Residential rent (EUR/m²)                    | Fahrmeir et al. (2013)  |
| Amazon  | 31,925| Website visit duration (min) | log(sales) (log(USD)) | Panagiotelis, Smith, and Danaher (2014) |
| Incomes | 40,981| Worker age (years) | log(income) (EUR)                            | Klein et al. (2015)     |

NOTE: The columns give (in order) the dataset name, number of observations, covariate, response variable, and published source of the data.

![Figure 1](image.png)

Figure 1. Marginal distributions of the response variables in the four test datasets. Normalized histograms of the response ($Y$) of the four datasets in Section 4, along with kernel (KDE) and Bayesian (DPhat) nonparametric density estimates of $p_Y$. The datasets are (a) Geyser, (b) Rents, (c) Amazon, and (d) Incomes.
covariates in the next section), and we set $x_l = w_l$ throughout. Figure 1 plots histograms of the four response variables, along with KDE and Dirichlet process mixture (DPhat) nonparametric density estimates. These are very similar, and we employ the KDE for $F_Y$, except where mentioned otherwise. Note that the response in the Geyser dataset in Figure 1(a) is bimodal, with which most existing distributional regression methods would struggle. In contrast, it is straightforward to account for this feature in our copula approach through the use of a bimodal copula function. The second employs transformations of the response because the copula data are unaffected.

We fit two variants of the copula model. The first employs the copula function $C_{Ht}$, and is labelled “HPSC” for “heteroscedastic P-spline copula.” The second employs $C_1$ with the constraint $\alpha = 0$ and is labelled “PSC” for (homoscedastic) “P-spline copula,” and is one of the regression copulas proposed by Klein and Smith (2019). Table 2 lists key quantities of the two copulas. Three benchmark models are also considered: the first is labeled PS and is the P-spline smoother with Gaussian errors of Lang and Brezger (2004), the second is labeled “HPS” and is the heteroscedastic P-spline smoother of Klein et al. (2015), while the third is labeled “MLT” and is the “most likely transformation” model of Hothorn, M"ost, and B"uhlmann (2017). For the latter we use Bernstein polynomials as suggested by the authors, and the method is known to be flexible and robust.

4.1.1. Exact Versus Approximate Estimation

We first compare the VB approximate and the HMC exact posterior estimators for the HPSC copula model. The VB estimator was fit using $K = 0, 1, 2, 3, 4, 5, 10, 15$ and 20 factors, and for 1, 5, 10, and 15 thousand steps. Figure 2(b, d, f, h) plots the mean lower bound value over the last 10% of steps (LB) against $K$ for each of the four step sizes and each of the datasets. Increasing $K$ up to 5 improves the accuracy of the VA, but further increases have little impact. Figure 2(a, c, e, g) plots $L(\lambda)$ against the step number for a VA with $K = 20$ factors, and in each case the SGA algorithm converges rapidly. Figures A and B in the Web Appendix plot the mean and standard deviation of the coefficients $(\beta, \alpha)$ from the VA, against their exact posterior means and standard deviations. These show that the variational estimates of the posterior means are highly accurate, but—as is usual in Gaussian VB inference—the posterior standard deviations are underestimated. Computation times are reported in Table A in the Web Appendix and show that the VB estimator is much faster than the exact method and practical to implement, even for a copula of dimension $n = 40,981$.

4.1.2. Predictive Accuracy

To compare the accuracy of the five models (PSC, HPSC, PS, HPS, and MLT) we compute the predictive logarithmic score by 10-fold cross-validation. For a given dataset, we partition the data into 10 (approximately) equally sized subsamples, denoted as $\{(y_{i,k}, x_{i,k}, w_{i,k}); i = 1, \ldots, n_k\}$ for $k = 1, \ldots, 10$. For subsample $k$, we compute the density estimator using the remaining 9

Table 2. Key quantities of the two P-spline regression copulas PSC and HPSC.

| Quantity | PSC | HPSC |
|----------|-----|------|
| $\theta$  | $\theta = (\beta, \theta_\beta)$ | $\theta = (\beta, \alpha, \theta_\beta, \theta_\alpha)$ |
| $s_i$    | $s_i = (1 + b_p(\theta_\beta^{-1}b_\beta))^{-1/2}$ | $s_i = (\exp(v(\alpha)) + b_p(\theta_\beta^{-1}b_\beta))^{-1/2}$ |
| $S$      | $S(x_i, \theta_\beta)$ | $S(x_i, w_i, \alpha, \theta_\beta)$ |
| $R$      | $R(x_i, \theta_\beta) = S(i + BP_\beta(\theta_\beta^{-1}b_\beta))$ | $R(x_i, \alpha, \theta_\beta) = S(\exp(V(\alpha)) + BP_\beta(\theta_\beta^{-1}b_\beta))$ |

NOTE: Reported from the top to bottom rows are: augmented parameters, normalizing factor, normalizing matrix, and parameter matrix of the Gaussian copula $C_1$. 

Figure 2. Summaries of the variational lower bound $L(\lambda)$ in the SGA. The datasets are (a)–(b) Geyser, (c)–(d) Rents, (e)–(f) Amazon, and (g)–(h) Incomes. Panels (b), (d), (f), (h) plot the average lower bound (LB) over the last 10% of steps, against the number of factors $K$ in the Gaussian factor variational approximation. Panels (a), (c), (e), (g) plot the variational lower bound against step number for the approximation with $K = 20$ factors.
subsamples as the training data, and denote these as \( \hat{p}_k(y|x, w) \). For our copula model we use the density estimator \( \hat{p}_{PE} \) in Section 3.4. The 10-fold mean logarithmic score is then \( \bar{L}_{CV} = \frac{1}{10} \sum_{k=1}^{10} \frac{1}{n_k} \sum_{i=1}^{n_k} \log \hat{p}_k(y_{ik} | x_{ik}, w_{ik}) \).

Table 3 reports the \( \bar{L}_{CV} \) values, where the posterior of the copula is computed either exactly using MCMC or HMC, or approximately using VB, with scores given for both cases; we make four observations. First, in all examples both copula models—which account fully for the non-Gaussian distribution of the responses—outperform the two benchmark PS and HPS models. Second, the performance of the copula models estimated using VB is very similar to that of the copula models estimated by exact methods. Third, in every case the HPSC outperforms the PSC, showing that the added flexibility of the heteroscedastic copula translates into improved distributional predictions—something that we demonstrate further below. Last, in all examples HPSC is competitive with the benchmark MLT model, which also allows the entire predictive distribution to vary with the covariates.

### 4.1.3. Mean and Variance Function Estimates

To compare the distributional regression estimates, Figure 3 plots the posteriors of \( f, \nu \) computed as in Section 3.4 for the Rents dataset. Posterior mean and 95% intervals are given for \( f \) in the left-hand panels, and for \( \nu \) in the right-hand panels. Panels (a,b) compare the posteriors from the HPSC model computed exactly using HMC, and approximately using VB, and they are very similar, further illustrating the high accuracy of the VB estimator. Panels (c,d) compare the posteriors from the HPSC model using the three different approaches to estimating \( F_y \). These are the kernel estimator (KDE), the Dirichlet process mixture (DPHat), and integrating out \( F_y \) using its draws (DP) as in Grazian and Liseo (2017). The posteriors are similar, and the approach used to estimate the margin has little effect on the distributional regression estimates for this example. Panels (e,f) compare the function estimates from the HPSC model against those of the benchmark MLT model, and they differ substantially—particularly for the variance function \( \nu \).

Finally, panels (g,h) compare the function estimates from the two different regression copulas HPSC and PSC. While the estimates of \( f \) are similar, those for \( \nu \) are very different. This is because the regression model for the pseudo-response of the PSC is homoscedastic with respect to the covariate, whereas that for the HPSC is heteroscedastic. Figure C in the Web Appendix plots the posterior estimates of \( f, \nu \) for the other three datasets. Similar results to the Rents dataset are found, where estimates of \( f \) from the two regression copula models are similar, but those of \( \nu \) differ substantially.

### 4.1.4. Dependence Metrics and Prediction

The improved fit of the HPSC over PSC is because the dependence structure of \( C_H \) is a much more flexible function of the covariates than \( C_1 \) is with \( \alpha = 0 \). To illustrate this we construct pairwise dependence metrics as follows. Set \( x^+ = w^+ = (x^+, x_{n+1}, x_{n+2})^T \), and \( \tilde{x}^+ = (x^+, x^+) \), then compute Spearman’s \( \rho \) for the bivariate sub-copula \( C_{H}^{n+1,n+2} \) with \( \theta \) integrated out with respect to its posterior; that is

\[
\hat{\rho}^S(x_{n+1}, x_{n+2}) \equiv \int \rho_{n+1,n+2}^S(\tilde{x}^+, \theta) p(\theta | y) d\theta \, ,
\]

where \( \rho_{n+1,n+2}^S \) is given in Section 2.3 part (iii). The integration is computed using draws from the posterior. For the PSC, the coefficients \( \alpha = 0 \), and integration is only with respect to \( \theta_{\tilde{b}} \). The metric \( \hat{\rho}^S \) is evaluated on a bivariate grid for \( (x_{n+1}, x_{n+2}) \) over the range of the covariate, and its values plotted as a surface. The process can be replicated for the other dependence metrics.

Figure 4 plots the surfaces of \( \hat{\rho}^S \) for the Rents dataset and both regression copulas. For both copulas, \( \hat{\rho}^S(x_{n+1}, x_{n+2}) \) declines as \( |x_{n+1} - x_{n+2}| \) increases, which is to be expected from any effective regression smoothing method. However, correlation is locally varying for the HPSC only. For example, correlation is higher for values of the covariate (area) around 20 and 80. Equivalent surfaces for the other three datasets, along with the upper quantile dependence and Kendall’s tau, are given in Part C of the Web Appendix, and the same features can be seen. This local variation in the dependence structure of the HPSC ensures the level of smoothing in the regression and variance functions in Figure 3 (and Figure C of the Web Appendix) are “locally adaptive” with respect to the covariate.

In fact, the entire distributional regression fit is locally adaptive to the value of the covariate. To illustrate this, we compute predictive densities for the Incomes dataset from both copula models. Figure 5 plots these for four values of the covariate (age), along with those from the benchmark HPS and MLT models. Because age is measured discretely, we also provide histograms of the salaries of all individuals of these ages. First, because the HPSC model is conditionally Gaussian, the predictive distributions are also, and are inconsistent with the histograms. Second, even though the two copula models share the same marginal \( F_y \), their predictive densities differ. Those from the HPSC copula model are more consistent with the histograms, which accords with the increased accuracy measured by the scores in Table 3.
Figure 3. Comparison of different posterior estimates of the regression function $f$ (left-hand side) and variance function $v$ (right-hand side) for the Rents data. The posterior means of $f$ and $v$ are given as solid lines, and 95% posterior intervals by dashed lines. Scatterplots of the data are included on the left-hand panels. Panels (a)–(b) compare function estimates from the HPSC model computed using HMC and VB. Panels (c)–(d) compare function estimates computed using HMC from the HPSC model using three different marginal estimators discussed in the text: KDE, DPhat, and DP. Panels (e)–(f) compare function estimates from the HPSC model (computed using HMC) with those from the benchmark MLT model. Panels (g)–(h) show estimates of the regression function $f$ (left-hand side) and variance function $v$ (right-hand side) from the HPSC model, compared to those from the PSC model.
Figure 4. Spearman’s rho from the regression copulas for the Rents dataset. Each panel plots estimates of Spearman’s rho $\hat{\rho}_S(x_{n+1}, x_{n+2})$ as bivariate functions of $(x_{n+1}, x_{n+2})$ over the range of the covariate (area). Panel (a) gives values for the HPSC, and panel (b) for the PSC. The localized variation in $\hat{\rho}_S$ in panel (a) corresponds to local adaptivity in the distributional regression from the copula model. Analogous plots for the other three datasets, and other dependence metrics (Kendall’s tau, upper and lower quantile dependence) are given in the Web Appendix.

Figure 5. Predictive densities for the Incomes dataset for four different values of age. The different ages are (a) 22 years old, (b) 24 years old, (c) 45 years old, and (d) 58 years old. Densities are from the PSC (red), HPSC (black), HPS (green), and MLT (blue) regression models. Also plotted are histograms of the subsamples of individuals with these four ages in the Incomes dataset.

The MLT densities are similar to those of the PSC copula model and are also dominated by those from the HPSC copula model.

4.2. Simulation Study

We undertake a simulation study to illustrate the efficacy of our copula-based approach to semiparametric distributional regression. Constructing a simulation design is challenging because all aspects of the distribution are unknown functions of the covariates. Therefore, we base our designs on the five distributional regression methods fitted to the four real datasets in the previous subsection, giving 20 data generating processes (DGPs). From each DGP we simulate 100 datasets (called
“replicates” here), and then refit all five methods to every replicate. Accuracy of a method for each fitted replicate is assessed by using it to predict the densities of the observations in an additional 101st replicate. Thus, we are assessing the accuracy of out-of-sample density forecasting. Full details on the simulation study are given in Part D of the Web Appendix.

Figure 6 gives boxplots of the mean logarithmic score (IS) and mean continuous ranked probability score (CRPS) (Gneiting and Raftery 2007) of predictions for the DGPs based on the Incomes dataset (the other 15 DGPs are given in Part D of the Web Appendix). The shaded boxplots are for the method that matches the DGP in each panel, and these are (unsurprisingly) either the best, or equal best, at recapitulating the DGP. Our focus is therefore on the next best performers, and we make three observations. First, when the DGP is either copula model, the HPSC either equals or out-performs the PSC, highlighting its superiority as a regression copula. Second, when the DGP is the HPS, then the regression copula HPSC is best at recapitulating this DGP. Last, the HPSC either equals or out-performs the MLT benchmark method for all DGPs and metrics, except for the PS DGP. Results for the other 15 DGPs are similar.

5. Radial Basis Copula for Electricity Prices

The relationship between intra-day electricity spot price and demand is used by participants in wholesale markets to formulate optimal bidding strategies (Kirschen and Strbac 2004, pp. 53–72). However, its estimation using regression methods is difficult because prices have a very heavy right tail, and all aspects of their distribution vary extensively with demand, day and time of day (Bunn et al. 2016). To account for this, we construct a regression copula from trivariate radial bases for \( \tilde{m} \), \( g \), combined with horseshoe priors for regularization. We apply it to high-frequency Australian electricity price and demand data, and compare our approach to other distributional regression methods.

5.1. Electricity Data and Regression Copula Model

The Australian NEM is a wholesale market where generators, distributors and third party participants bid for the sale and purchase of electricity one day ahead of transmission; see Ignatieva and Trück (2016) and Smith and Shively (2018) for current descriptions of the market. We consider half-hourly market-wide price \( P_i \) and total market demand \( D_i \) from January 1, 2014 to December 31, 2018, so that \( n = 87,648 \). Total market demand is the sum of demand across the five regions in the NEM, while the market-wide price is the demand-weighted average price across the five regions, constructed from data available at www.aemo.com.au. The three covariates are demand, time of day (TOD\(_i\)) and day number (Day\(_j\)), and set \( x_i = w_i = (D_i, \text{TOD}_i, \text{Day}_j)' \), with each covariate scaled to the unit interval. For \( \tilde{m} \) and \( g \) we employ thin plate spline radial basis functions (RBFs) of the form \( b_j(x) = \delta(x - k_j)^2 \log(\delta(x - k_j)) \) for knot \( k_j \). The distance function \( \delta(x_1, x_2, x_3) = ||(x_1, \sin(\pi x_2), x_3)|| \) is the Euclidean distance with a sine transformation on the second element to ensure the basis is periodic on \([0, 1]\) for TOD\(_i\). The knots are set equal to a random sample (stratified by time of day) of 240 and 96 covariate values for \( \tilde{m} \) and \( g \), respectively. We follow Klein and Smith (2019) and use the horseshoe prior for the regularization at Equation (6), and provide details in Part E of the Web Appendix.

Due to the extreme skew in electricity prices we set \( Y = \log(P_i + 101) \), where we add 101 before taking the logarithm because the minimum observed price in our data is \(-99.82\) (prices can be negative in the NEM). Figure 7(a) plots a histogram of the response and KDE \( \hat{F}_y \), showing that even on the logarithmic scale the distribution of prices is positively skewed and heavy-tailed. Panel (b) gives a quantile-quantile plot...
Figure 7. Marginal distribution summaries of the logarithm of half-hourly electricity prices. Summaries are for \( Y = \log(\text{Price} + 101) \). Panel (a) plots a histogram and the KDE \( \hat{F}_Y \) over the range \( 4 < Y < 6.5 \), although the tails of the marginal distribution extend further. Panel (b) provides a quantile-quantile plot of the residuals from the marginal fit \( \hat{F}_Y \). Panel (c) provides boxplots of \( Y \) broken down by time of day, revealing the diurnal variation in the distribution.

highlighting the accuracy of the KDE estimator for \( F_Y \). Panel (c) contains boxplots of the response broken down by the time of day, and reveals the strong diurnal variation in the entire distribution of prices.

5.2. Empirical Results

We estimate our regression copula (labeled “HRBFC”) using VB with \( K = 20 \) factors for the approximation. A plot of \( L(\lambda) \) against step number (see Web Appendix) indicates reliable convergence of the SGA algorithm. There is a strong (nonadditive) effect of the three covariates on price. For example, Figure 8 plots a "slice" of the trivariate mean \( f \) and variance \( v \) functions against demand \( D \), for May 12, 2018 at 19:00, which is the time of day with the highest mean price. They are the variational posteriors, computed as at Equation (16), and show the positive relationship between the first two moments of price and demand.

To illustrate the impact of demand on the entire distribution, Figure 10 plots the predictive densities of \( Y \) on May 12, 2018 at (a) 06:00, (b) 12:00, (c) 18:00, and (d) 24:00. In each panel, densities are constructed at four levels of demand that correspond to the 0.25, 0.5, 0.95, and 0.99 percentiles of demand at each time of day. Increases in demand accentuate the upper tail, consistent with the nonlinear impact of demand shocks on price spikes documented previously (Higgs and Worthington 2008; Smith and Shively 2018). Further plots of predictive densities over the four years (see Part F of the Web Appendix) show the upper tail is increasingly sensitive to demand, matching the increasing frequency of price spikes during the period.

Last, we compare our regression copula to two benchmarks models. The first is the approach of Rigby and Stasinopoulos (2005) (labeled "GAMLSS") where we tried several distributions and found the ST2 to give the best fit. We found convergence problems when specifying all parameters as additive splines of the three covariates, and were restricted to only allow the mean and variance to do so. The second benchmark is a heteroscedastic regression model with additive P-spline terms for the three covariates (labeled “HPS”). To measure the accuracy of the distributional forecasts for the three models, Table 4 reports the cross-validated mean score metric \( \text{IS}_{CV} \) defined in Section 4.1.2, plus a 10-fold cross-validated mean CRPS metric (CRPSCV). The radial basis regression copula model clearly dominates the GAMLSS and HPS benchmarks. Figure 9 plots the (cross-validated) mean quantile score \( \overline{Q}_S_{CV}(\alpha) = \frac{1}{10} \sum_{k=1}^{10} \frac{1}{n_k} \sum_{i=1}^{n_k} \overline{Q}_S(y_{ik}; x_{ik}, w_{ik}, \alpha) \) for each method and \( \alpha \in (0,1) \), where \( \overline{Q}_S(y; x, w) \equiv -Q_S(\hat{F}_k^{-1}(\alpha), y) \) is defined in Gneiting and Ranjan (2011). All scores are orientated so that higher values indicate greater accuracy. The figure reveals the...
greater accuracy of the HRBFC model at all quantiles, except for the extreme tails where the three models are similar. Predictive density forecasts from GAMLSS and HPS are provided in Part F of the Web Appendix, and show they (unlike those from the HRBFC model) are only weakly effected by increases in demand, which is inconsistent with previous analyses (Higgs and Worthington 2008; Ignatieva and Trück 2016; Smith and Shively 2018).

6. Discussion

This article proposes modeling the entire distribution of a vector of regression response values, conditional on covariates, using a copula decomposition. To do so, a new copula $C_H$ is constructed from a heteroscedastic semiparametric regression for a pseudo-response. When combined with nonparametric or other margins, the resulting regression model is flexible in both the distributional shape and the functional relationship between the covariates and response. Our approach is very general, scalable and numerically stable. We show in our empirical work that it improves predictive accuracy for non-Gaussian data, relative to a number of leading benchmark regression approaches.

A number of authors construct the $n$-dimensional implicit Gaussian copulas of Gaussian processes (Wauthier and Jordan 2010; Wilson and Ghahramani 2010). However, these are very different copulas than those constructed here. Klein and Smith (2019) propose constructing the copula of Bayesian regularized smoothers, which is equivalent to our implicit copula when $\alpha = 0$. This article extends their work by allowing for heteroscedasticity in the pseudo-response, which yields a copula with a much richer dependence structure as shown in Figure 4. This makes the distributional regression locally adaptive, as can be seen in the mean and variance function estimates in Figure 3, and increases predictive accuracy. However, our proposed copula is more difficult to estimate, and the MCMC schemes discussed by Klein and Smith (2019)—who do not consider alternatives—are infeasible. To address this, we develop efficient exact estimation with a HMC step for generating $\alpha$, and approximate estimation using VB. The empirical work demonstrates the efficacy these methods using five diverse real datasets. In every case, our fitted copula model is more accurate than both the simpler regression copula $C_1$ with $\alpha = 0$ and the benchmark models. Moreover, estimation and prediction is fast, allowing the application of the distributional regression methodology to large datasets.

Supplementary Materials

The supplementary materials include an extensive Web Appendix which contains derivations, proofs, further simulation results, tables and figures referred to in the text. Furthermore, Matlab code is provided for reproducing important results in the article.
Figure 10. Predictive distributions of the logarithm of electricity prices from the regression copula model (HRBF). The four panels provide predictions for May 12, 2018 at (a) 06:00, (b) 12:00, (c) 18:00, and (d) 24:00. In each panel, the predictive densities are constructed at four levels of demand corresponding to the 0.25, 0.5, 0.95, and 0.99 percentiles of demand at each time of day.

Acknowledgments

The authors thank the editor, associate editor, and three referees whose comments improved the article.

Funding

Nadja Klein gratefully acknowledges funding from the Alexander von Humboldt Foundation and the German research foundation (DFG) through the Emmy Noether grant KL 3037/1-1.

References

Aas, K., Czado, C., Frigessi, A., and Bakken, H. (2009), “Pair-Copula Constructions of Multiple Dependence,” *Insurance: Mathematics and Economics*, 44, 182–198. [713]

Barndorff-Nielsen, O., and Schou, G. (1973), “On the Parametrization of Autoregressive Models by Partial Autocorrelations,” *Journal of Multivariate Analysis*, 3, 408–419. [719]

Bottou, L. (2010), “Large-Scale Machine Learning With Stochastic Gradient Descent,” in *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT2010)*, eds. Y. Lechevallier and G. Saporta, Springer, pp. 177–187. [717]

Bunn, D., Andresen, A., Chen, D., and Westgaard, S. (2016), “Analysis and Forecasting of Electricity Price Risks With Quantile Factor Models,” *The Energy Journal*, 37, 101–122. [724]

Chib, S., and Jeliazkov, I. (2006), “Inference in Semiparametric Dynamic Models for Binary Longitudinal Data,” *Journal of the American Statistical Association*, 101, 685–700. [719]

Demarta, S., and McNeil, A. J. (2005), “The t Copula and Related Copulas,” *International Statistical Review*, 73, 111–129. [712]

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013), *Regression—Models, Methods and Applications*, Berlin: Springer. [719]

Gianfreda, A., and Bunn, D. (2018), “A Stochastic Latent Moment Model for Electricity Price Formation,” *Operations Research*, 66, 1189–1203. [713]

Gneiting, T., and Raftery, A. E. (2007), “Strictly Proper Scoring Rules, Prediction, and Estimation,” *Journal of the American Statistical Association*, 102, 359–378. [724]

Gneiting, T., and Ranjan, R. (2011), “Comparing Density Forecasts Using Threshold and Quantile-Weighted Scoring Rules,” *Journal of Business & Economic Statistics*, 29, 411–422. [713,725]

Grazian, C., and Liseo, B. (2017), “Approximate Bayesian Inference in Semiparametric Copula Models,” *Bayesian Analysis*, 12, 991–1016. [716,721]

Higgs, H., and Worthington, A. (2008), “Stochastic Price Modeling of High Volatility, Mean-Reverting, Spike-Prone Commodities: The Australian Wholesale Spot Electricity Market,” *Energy Economics*, 30, 3172–3185. [725,726]
