Implicit Copulas from Bayesian Regularized Regression Smoothers

Nadja Klein and Michael Stanley Smith
Humboldt University of Berlin and University of Melbourne

April 5, 2019

Research Seminar WU, Vienna

Partly funded by the Alexander von Humboldt Foundation.

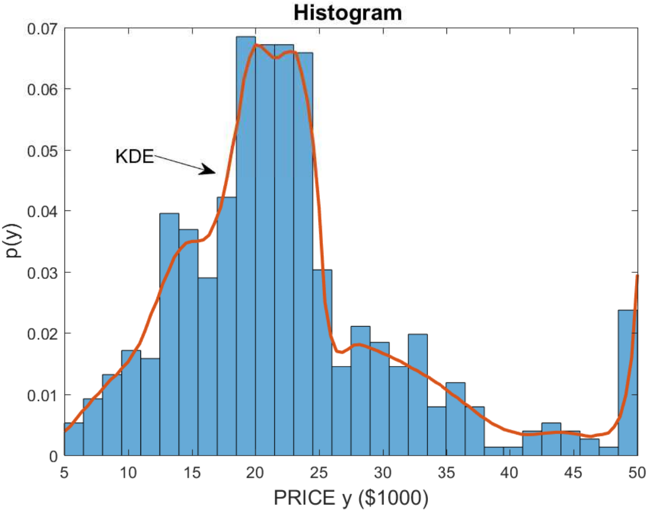
- Typical regression models relate the expectation of the response to covariates.
- 'Statisticians are mean lovers'.
- This exclusive focus on the conditional expectation may however
  - possibly not meaningful and insufficient,
  - often not flexible enough,
  - does not comply to the main goal of the analysis.

Motivating Example: The Boston Housing Data

- Popular nonparametric regression dataset with $n = 506$ and
  - Y: Median house price in a census tract
  - covariates $\boldsymbol{X}$: NOX, RN, DIS, LSTAT, TAX
- Aim: Estimate a nonparametric regression model such that the entire distribution $F(Y|\boldsymbol{X})$ is a function of $\boldsymbol{X}$ and e.g.

$$E(Y|\boldsymbol{X}) = f(\boldsymbol{X}).$$

- However: The marginal distribution of Y is highly non-Gaussian


Histogram
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
p(y)
KDE
5
10
15
20
25
30
35
40
45
50
PRICE y ($1000)

Bayesian Distributional Regression

- Observed data pairs $(\boldsymbol{y}_1, \boldsymbol{x}_1), \ldots, (\boldsymbol{y}_n, \boldsymbol{x}_n)$.
- Model assumption 1: Conditional distribution $F(\boldsymbol{y}_i|\boldsymbol{x}_i)$ given $\boldsymbol{x}_i$, $i = 1, \ldots, n$ is from pre-specified class of K-parametric densities

$$p(\boldsymbol{y}_i|\vartheta_{i1}, \ldots, \vartheta_{iK}).$$

- Model assumption 2: Each parameter ϑ_{ik}, $k = 1, \ldots, K$ is related to a regression predictor $\eta_{ik} = \eta_k(\boldsymbol{x}_i)$:

$$\vartheta_{ik} = h_k(\eta_{ik}) \text{ and } \eta_{ik} = h_k^{-1}(\vartheta_{ik})$$

However, . . .

establishing a good distributional model is difficult in practice because you need to decide

- which parametric distribution assumption to pick,
- which variable goes in which predictor (location, scale, shape of the distribution),

Our General Idea

- Extract the implicit copula of a response vector from a Bayesian regularized smoother
- Construct and compare copulas for:
  - Three popular shrinkage priors (BVS, PS, HS)
  - Differing (matching) bases

Our General Idea

- Extract the implicit copula of a response vector from a Bayesian regularized smoother
- Construct and compare copulas for:
  - Three popular shrinkage priors (BVS, PS, HS)
  - Differing (matching) bases
- Why?

Our General Idea

- Extract the implicit copula of a response vector from a Bayesian regularized smoother
- Construct and compare copulas for:
  - Three popular shrinkage priors (BVS, PS, HS)
  - Differing (matching) bases
- Why?
  - Can be used to compare the shrinkage properties of any Bayesian smoother

Our General Idea

- Extract the implicit copula of a response vector from a Bayesian regularized smoother
- Construct and compare copulas for:
  - Three popular shrinkage priors (BVS, PS, HS)
  - Differing (matching) bases
- Why?
  - Can be used to compare the shrinkage properties of any Bayesian smoother
  - Combined with arbitrary margins, the copula models provide a novel class of semiparametric distributional regression models

Sklar's Theorem

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

$$p(\boldsymbol{y}_{(N)}|\boldsymbol{x}_{(N)}) = c^\dagger(F(y_1|\boldsymbol{x}_1), \ldots, F(y_N|\boldsymbol{x}_N)|\boldsymbol{x}) \prod_{i=1}^{N} p(y_i|\boldsymbol{x}_i)\,, \quad \text{for } N \geq 2$$

Here, $c^\dagger(\boldsymbol{u}_{(N)}|\boldsymbol{x}_{(N)})$ is a N-dimensional copula density and $F(y_i|\boldsymbol{x}_i)$ is the distribution function of $Y_i|\boldsymbol{x}_i$; both of which are unknown

Copula Smoother

- We model the joint density (given any covariates) using the copula decomposition

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

Copula Smoother

- We model the joint density (given any covariates) using the copula decomposition

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

- The distribution $Y_i|\boldsymbol{x}_i$ is assumed to be invariant with respect to $\boldsymbol{x}_i$, and has density p_Y and distribution function F_Y

Copula Smoother

- We model the joint density (given any covariates) using the copula decomposition

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

- The distribution $Y_i|\boldsymbol{x}_i$ is assumed to be invariant with respect to $\boldsymbol{x}_i$, and has density p_Y and distribution function F_Y
- Distributional flexibility comes from choices for p_Y

Copula Smoother

- We model the joint density (given any covariates) using the copula decomposition

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

- The distribution $Y_i|\boldsymbol{x}_i$ is assumed to be invariant with respect to $\boldsymbol{x}_i$, and has density p_Y and distribution function F_Y
- Distributional flexibility comes from choices for p_Y
- However, the impact of the covariate values on $\boldsymbol{Y}_{(N)}$ is captured through the copula with density $c_\pi(\boldsymbol{u}_{(N)}|\boldsymbol{x}_{(N)})$, where $u_i = F_Y(y_i)$

Copula Smoother

- We model the joint density (given any covariates) using the copula decomposition

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

- The distribution $Y_i|\boldsymbol{x}_i$ is assumed to be invariant with respect to $\boldsymbol{x}_i$, and has density p_Y and distribution function F_Y
- Distributional flexibility comes from choices for p_Y
- However, the impact of the covariate values on $\boldsymbol{Y}_{(N)}$ is captured through the copula with density $c_\pi(\boldsymbol{u}_{(N)}|\boldsymbol{x}_{(N)}, \boldsymbol{\theta})$, where $u_i = F_Y(y_i)$
- We call this a copula smoother because the relationship between $\boldsymbol{x}$ and $\boldsymbol{y}$ comes from the copula only

Construction of c_π

- c_π is constructed from a random vector $\tilde{\boldsymbol{Z}}$ with CDF $F_{\tilde{Z}}$ by inversion of Sklar's theorem:

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

- $\tilde{Z}|\boldsymbol{x}$ is called pseudo response as it is not observed directly
- $u_1, \ldots, u_n$ is called the copula data

The Pseudo Response Model

- For $i = 1, \ldots, n$ consider the regression model

$$\tilde{Z}_i = \tilde{m}(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2),$$

The Pseudo Response Model

- For $i = 1, \ldots, n$ consider the regression model

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

  with

$$\tilde{m}(x_i) = \sum_{j=1}^{p} \beta_j B_j(x_i)$$

- and B_j the p basis functions, such as B-spline basis, radial basis, ...

The Pseudo Response Model

- For $i = 1, \ldots, n$ consider the regression model

$$\tilde{Z}_i = \tilde{m}(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2),$$

with

$$\tilde{m}(x_i) = \sum_{j=1}^{p} \beta_j B_j(x_i)$$

- and B_j the p basis functions, such as B-spline basis, radial basis, ...
- Regularization (smoothness) can be achieved via the prior

$$\boldsymbol{\beta} | \boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma^2 \sim \mathrm{N}(\mathbf{0}, \sigma^2 P_\gamma(\boldsymbol{\theta})^{-1})$$

The Pseudo Response Model

- For $i = 1, \ldots, n$ consider the regression model

$$\tilde{Z}_i = \tilde{m}(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2),$$

with

$$\tilde{m}(x_i) = \sum_{j=1}^{p} \beta_j B_j(x_i)$$

- and B_j the p basis functions, such as B-spline basis, radial basis, ...
- Regularization (smoothness) can be achieved via the prior

$$\boldsymbol{\beta} | \boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma^2 \sim \mathrm{N}(\mathbf{0}, \sigma^2 P_{\gamma}(\boldsymbol{\theta})^{-1})$$

Copula Construction

- Let $S(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = \text{diag}(s_1, \ldots, s_n)$ with

$$\text{Var}(\tilde{Z}_i | \boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = \frac{\sigma^2}{s_i^2}$$

- Set

$$\boldsymbol{Z} = \sigma^{-1} S(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\gamma}) \tilde{\boldsymbol{Z}}$$

- Then, the copula of $\boldsymbol{Z} | \boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\gamma}$ is a Gaussian copula with correlation matrix

$$R(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = S(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\gamma})(I + BP_{\gamma}(\boldsymbol{\theta})^{-1}B')S(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\gamma})$$

- Label the copula function $C(\boldsymbol{u} | \boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\gamma})$

Copula Construction

- If $\pi(\boldsymbol{\theta}, \boldsymbol{\gamma})$ is any proper density, then the implicit copula is

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

- It is easy to show that this is a proper copula
- For the regularization priors, $C_\pi(\boldsymbol{u}|\boldsymbol{x})$ turns out to be far (!) from a Gaussian copula

Three Implicit Copulas

- P-spline copula (PSC)
  - AR(2) prior
  - $\boldsymbol{\theta} = \{\tau^2, \psi_1, \psi_2\}$, $\boldsymbol{\gamma} = \emptyset$
  - Matched with B-spline basis

Three Implicit Copulas

- P-spline copula (PSC)
  - AR(2) prior
  - $\boldsymbol{\theta} = \{\tau^2, \psi_1, \psi_2\}$, $\boldsymbol{\gamma} = \emptyset$
  - Matched with B-spline basis
- Horseshoe copula (HSC)
  - $\beta_j \sim \mathsf{N}(0, \lambda_j^2)$, $\lambda_j \sim \mathsf{C}^+(0, \tau)$, $\tau \sim \mathsf{C}^+(0, 1)$
  - $\boldsymbol{\theta} = \{\lambda_1, \ldots, \lambda_p, \tau\}$, $\boldsymbol{\gamma} = \emptyset$
  - Matched with Fourier basis or radial basis

Three Implicit Copulas

- P-spline copula (PSC)
  - AR(2) prior
  - $\boldsymbol{\theta} = \{\tau^2, \psi_1, \psi_2\}$, $\boldsymbol{\gamma} = \emptyset$
  - Matched with B-spline basis
- Horseshoe copula (HSC)
  - $\beta_j \sim \mathrm{N}(0, \lambda_j^2)$, $\lambda_j \sim \mathrm{C}^+(0, \tau)$, $\tau \sim \mathrm{C}^+(0, 1)$
  - $\boldsymbol{\theta} = \{\lambda_1, \ldots, \lambda_p, \tau\}$, $\boldsymbol{\gamma} = \emptyset$
  - Matched with Fourier basis or radial basis
- Bayesian variable selection copula (BVSC)
  - $\boldsymbol{\beta}_\gamma \sim \mathrm{N}(\mathbf{0}, c(B_\gamma' B_\gamma)^{-1})$, $\pi(\boldsymbol{\gamma}) = \mathrm{Beta}(p - p_\gamma + 1, p_\gamma + 1)$
  - $\boldsymbol{\theta} = \emptyset$, $\boldsymbol{\gamma} = \{\gamma_1, \ldots, \gamma_p\}$
  - Matched with regression splines or radial basis

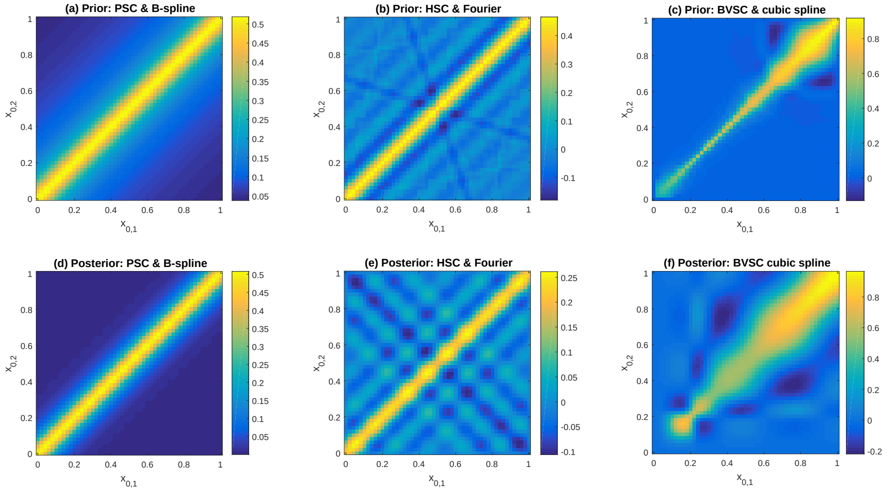
Dependence Structure

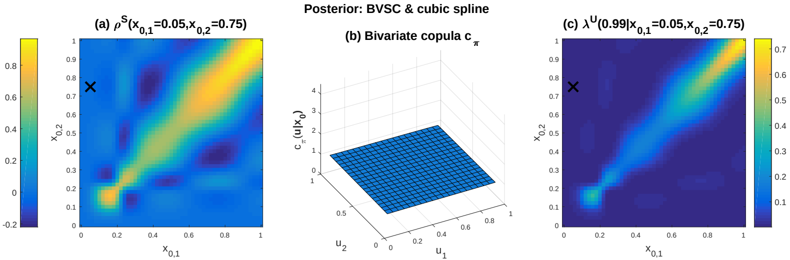
- For a univariate function $m(x)$ consider two new response values $Y_{0,1}, Y_{0,2}$ with covariate values $x_{0,1}, x_{0,2}$
- Compute the Spearman correlation

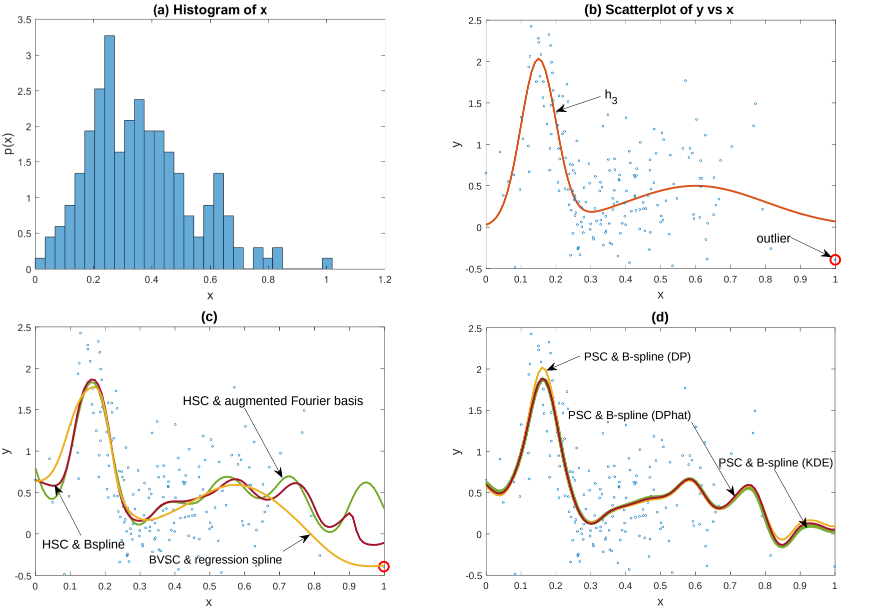
$$\rho_\pi^S(Y_{0,1}, Y_{0,2}|\boldsymbol{x}) \equiv \rho_\pi^S(Y_{0,1}, Y_{0,2}|\boldsymbol{x}, x_{0,1}, x_{0,2})$$

  and plot this as over a grid of $x_{0,1}, x_{0,2}$.
- We do this for $\pi(\boldsymbol{\theta}, \boldsymbol{\gamma})$ equal to the prior and the posterior

Dependence Structure




Posterior: BVSC & cubic spline
(a) ρ^S(x_{0,1}=0.05,x_{0,2}=0.75)
(b) Bivariate copula c_π
(c) λ^U(0.99|x_{0,1}=0.05,x_{0,2}=0.75)
x_{0,1}
x_{0,2}
c_π(u|x_0)
u_1
u_2


(a) Histogram of x
p(x)
x
(b) Scatterplot of y vs x
h_3
outlier
y
x
(c)
HSC & augmented Fourier basis
HSC & Bspline
BVSC & regression spline
y
x
(d)
PSC & B-spline (DP)
PSC & B-spline (DPhat)
PSC & B-spline (KDE)
y
x

Posterior Estimation

- c_π cannot be expressed in closed form
- The conditional likelihood

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

  is also computationally infeasible for large n because R is $(n \times n)$ and full
- Instead, we use the augmented likelihood

$$p(\boldsymbol{y}|\boldsymbol{x}, \beta, \boldsymbol{\theta}, \gamma)$$

  and MCMC
- Note that in contrast to the Bayesian linear model the posterior of $\boldsymbol{\theta}$ is often not available in closed form

Predictive Densities

- Predict the density of a new observation $Y_0|x_0$ using the posterior predictive density

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

Predictive Densities

- Predict the density of a new observation $Y_0|x_0$ using the posterior predictive density

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

Predictive Densities

- Predict the density of a new observation $Y_0|x_0$ using the posterior predictive density

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

- where

$$p(y_0|x_0, \boldsymbol{x}, \beta, \boldsymbol{\theta}, \gamma) = p(z_0|x_0, \boldsymbol{x}, \beta, \boldsymbol{\theta}, \gamma) \frac{p_Y(y_0)}{\phi_1(z_0)},$$

with $z_0 = \Phi^{-1}(F_Y(y_0))$

Predictive Densities

- Predict the density of a new observation $Y_0|x_0$ using the posterior predictive density

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

- where

$$p(y_0|x_0, \boldsymbol{x}, \beta, \boldsymbol{\theta}, \gamma) = p(z_0|x_0, \boldsymbol{x}, \beta, \boldsymbol{\theta}, \gamma) \frac{p_Y(y_0)}{\phi_1(z_0)},$$

  with $z_0 = \Phi^{-1}(F_Y(y_0))$
- Easy to compute MC estimates of density, and its moments (or other summaries) accurately

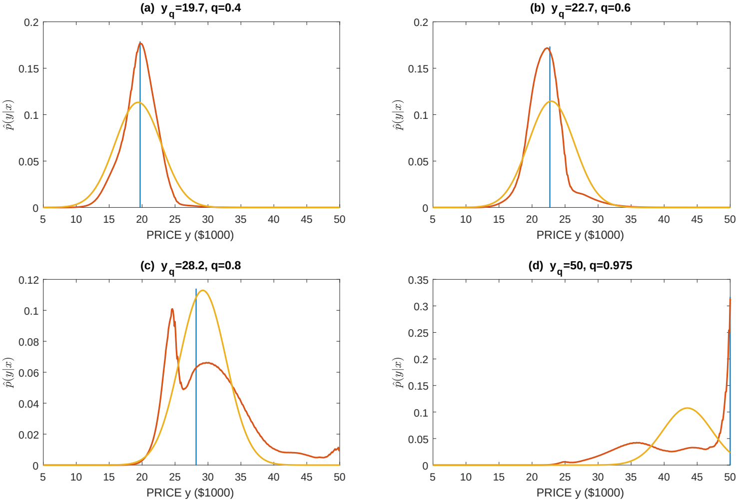
Motivating Example: Predicting House Prices

- Pseudo response model:

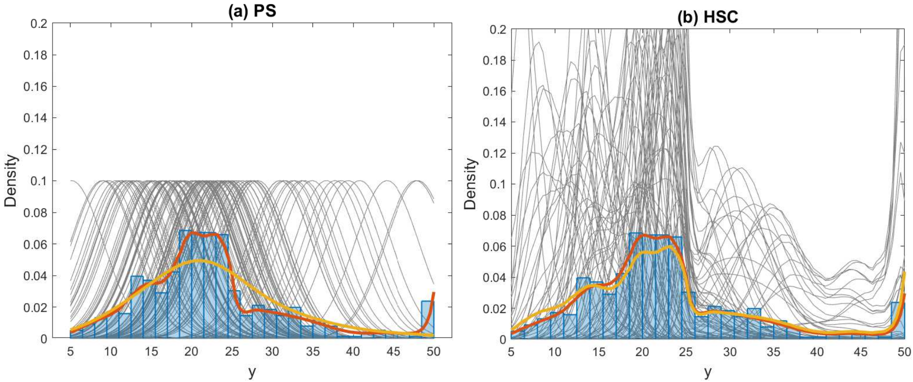
$$\tilde{Z}_i = \sum_{k=1}^{5} f(x_{ik}) + \varepsilon_i$$

- Major aim: Predictive densities of four house prices
- These are at 0.4,0.6,0.8,0.975 quantiles of the data distribution
- Comparison with a regular P-spline regression model (with Gaussian disturbances)
- Log-scores clearly favour the copula model
- We also compared results to other distributions (log-normal and Gamma) but results stayed similar.

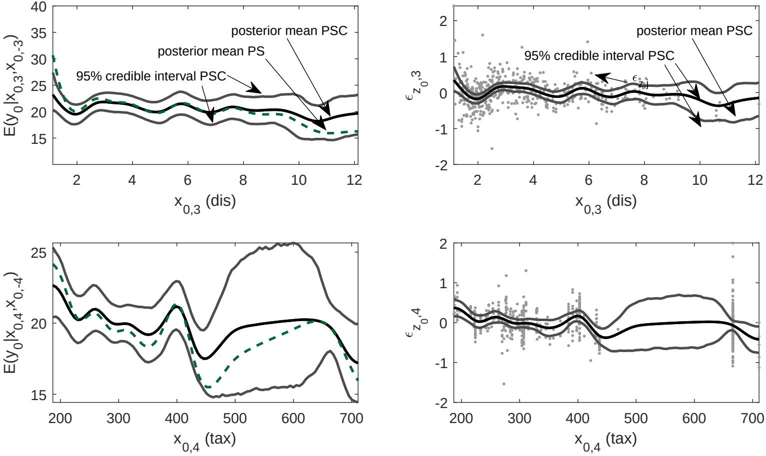
Predicting House Prices I



Predicting House Prices II



Predicted Expectations and Pseudo Residuals



Discussion

- Framework for comparison of Bayesian regularized regression smoothers

Discussion

- Framework for comparison of Bayesian regularized regression smoothers
- Implicit copula has a dependence structure very different from that of a Gaussian copula

Discussion

- Framework for comparison of Bayesian regularized regression smoothers
- Implicit copula has a dependence structure very different from that of a Gaussian copula
- Also, very different from the implicit copula of a Gaussian process prior!

Discussion

- Framework for comparison of Bayesian regularized regression smoothers
- Implicit copula has a dependence structure very different from that of a Gaussian copula
- Also, very different from the implicit copula of a Gaussian process prior!
- New distributional regression method

Discussion

- Framework for comparison of Bayesian regularized regression smoothers
- Implicit copula has a dependence structure very different from that of a Gaussian copula
- Also, very different from the implicit copula of a Gaussian process prior!
- New distributional regression method
- All dependence between Y and x is captured through a flexible implicit copula

Discussion

- Framework for comparison of Bayesian regularized regression smoothers
- Implicit copula has a dependence structure very different from that of a Gaussian copula
- Also, very different from the implicit copula of a Gaussian process prior!
- New distributional regression method
- All dependence between Y and x is captured through a flexible implicit copula
- Improves predictive accuracy

Discussion

- Framework for comparison of Bayesian regularized regression smoothers
- Implicit copula has a dependence structure very different from that of a Gaussian copula
- Also, very different from the implicit copula of a Gaussian process prior!
- New distributional regression method
- All dependence between Y and x is captured through a flexible implicit copula
- Improves predictive accuracy
- Applicable to multiple covariates and large n (e.g. 40,000 in other work)