Author/s:
Klein, N;Smith, MS

Title:
Bayesian variable selection for non-Gaussian responses: a marginally calibrated copula approach

Date:
2021-09-01

Citation:
Klein, N. & Smith, M. S. (2021). Bayesian variable selection for non-Gaussian responses: a marginally calibrated copula approach. Biometrics, 77 (3), pp.809-823. https://doi.org/10.1111/biom.13355.

Persistent Link:
https://hdl.handle.net/11343/276227

Bayesian Variable Selection for Non-Gaussian Responses: A Marginally-Calibrated Copula Approach

Nadja Klein

School of Business and Economics, Humboldt-Universität zu Berlin, Berlin, Germany,

email: nadja.klein@hu-berlin.de

and

Michael Stanley Smith

Melbourne Business School, University of Melbourne, 200 Leicester Street, Carlton, VIC 3053, Australia.

email: mike.smith@mbs.edu

SUMMARY: We propose a new highly flexible and tractable Bayesian approach to undertake variable selection in non-Gaussian regression models. It uses a copula decomposition for the joint distribution of observations on the dependent variable. This allows the marginal distribution of the dependent variable to be calibrated accurately using a nonparametric or other estimator. The family of copulas employed are 'implicit copulas' that are constructed from existing hierarchical Bayesian models widely used for variable selection, and we establish some of their properties. Even though the copulas are high-dimensional, they can be estimated efficiently and quickly using Markov chain Monte Carlo (MCMC). A simulation study shows that when the responses are non-Gaussian the approach selects variables more accurately than contemporary benchmarks. A real data example in the Web Appendix illustrates that accounting for even mild deviations from normality can lead to a substantial increase in accuracy. To illustrate the full potential of our approach we extend it to spatial variable selection for fMRI. Using real data, we show our method allows for voxel-specific marginal calibration of the magnetic resonance signal at over 6,000 voxels, leading to an increase in the quality of the activation maps.

KEY WORDS: fMRI; Implicit copula; Mixtures of g-priors; Spatial Bayesian variable selection.

This paper has been submitted for consideration for publication in *Biometrics*

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/biom.13355

1 Introduction

Bayesian approaches to selecting covariates in regression models are well established; see O'Hara and Sillanpää (2009) and Bottolo and Richardson (2010) for overviews. However, most work remains focused on Gaussian regression models, and extensions to the non-Gaussian case are limited. In particular, the importance of 'marginal calibration' in Bayesian variable (i.e. covariate) selection (BVS) is unexplored. Following Gneiting et al. (2007), by marginal calibration we mean the statistical consistency between the unconditional probability distribution of the dependent variable and its observations, with a more formal definition given by Gneiting et al. (2013, Defn. 2.6a). We address this here by proposing a new approach to BVS that is based on a copula decomposition for a vector of observations of length n on the dependent variable. The impact of the covariates on the dependent variable is captured by the copula function only. This separates the task of selecting covariates from that of modeling the marginal distribution of the dependent variable; the latter of which can then be calibrated accurately. For the copula function we propose a new family of 'implicit copulas', which are constructed from existing popular Bayesian hierarchical regression models used for selecting covariates. By an implicit copula we mean the copula that is implicit in a multivariate distribution and obtained by inverting the usual expression of Sklar's theorem as in Sec. 3.1 of Nelsen (2006). The result is a general and tractable approach that extends BVS to have an accurately calibrated margin for the dependent variable.

Low dimensional copulas are often used to capture dependence between multiple variables. Here the copula is used in a different way to capture the dependence between multiple observations on one dependent variable. The specification of this n-dimensional copula is the key ingredient of our method. To do so we consider a Gaussian linear model for n observations on another dependent variable, which we call a 'pseudo-response' because it is not observed directly. Gaussian spike-and-slab priors with selection indicator variables $\boldsymbol{\gamma}$ are employed for

the coefficients. Integrating out these coefficients gives a Gaussian distribution for the pseudo-response vector conditional on the covariates and $\boldsymbol{\gamma}$, and its implicit copula is a Gaussian copula (Song, 2000) with parameter matrix that is a function of the covariate values and $\boldsymbol{\gamma}$. Finally, to obtain our copula family we mix this Gaussian copula over the scaling factor g of the non-zero coefficients with respect to the different hyper-priors suggested by Liang et al. (2008). The resulting implicit copulas are mixtures of Gaussian copulas.

Because of its high dimension, it is difficult to evaluate our copula family directly. However, we show how to construct an MCMC sampler to undertake stochastic search variable selection (George and McCulloch, 1993), where the scaling factor g is sampled using the Hamiltonian Monte Carlo (HMC) method of Hoffman and Gelman (2014). Careful use of matrix identities for the computations makes application of the method to high dimensions practical. A simulation study compares our approach to Gaussian BVS and the method of Rossell and Rubio (2018). It shows that accurate marginal calibration of the distribution the dependent variable—an intrinsic feature of our copula model—can result in more accurate covariate selection and predictive densities.

However, the main application of our approach is to spatial variable selection for functional magnetic resonance imaging (fMRI). In these studies a large vector of binary indicators signifies which voxels in a partition of the brain are active, so that it is large variable selection problem. We follow Smith et al. (2003), Smith and Fahrmeir (2007), Li and Zhang (2010) and Goldsmith et al. (2014) and use an Ising model as a prior to smooth the binary indicators spatially, but employ our copula model to allow for voxel-wise marginal calibration of the magnetic resonance (MR) signal. The neuroimaging literature suggests that accounting for such deviations from normality in the MR signal is important to obtain accurate activation maps (Eklund et al., 2017). Application of our approach to data from a visual experiment with 6192 voxels shows this to be true here, and also produces much more accurate voxel-

wise predictive distributions for the MR signal, as measured by the logarithmic scores. Importantly, the approach is both fast to implement and can be readily generalized, including to other binary random field priors for spatial smoothing of the binary indicators.

A number of other Bayesian approaches consider variable selection for non-Gaussian continuous-valued data. These include conditionally Gaussian models, where the disturbances follow a mixture of normals and/or data transformations of the dependent variable are considered, as in Smith and Kohn (1996) and Gottardo and Raftery (2009). Rossell and Rubio (2018) propose a BVS approach that allows for skewness and heavy tails by employing two-piece Gaussian and Laplace distributions for the errors. Chung and Dunson (2009) consider variable selection for a distributional regression model constructed through a probit stick-breaking process, Kundu and Dunson (2014) consider selection when the errors are modelled non-parametrically. Yu et al. (2013) propose variable selection in Bayesian quantile regression. However, none of these approaches fits our copula framework, nor ensure accurate calibration of the marginal distribution of the response. Sharma and Das (2018) proposed an alternative class of priors for regression coefficients based on copulas, and Kraus and Czado (2017) used a D-vine copula to capture flexibly the dependence between covariates and response in a regression model. However both use copulas in a very different way than suggested here and do not generalize existing BVS schemes as our approach does. Last, Klein and Smith (2019) construct implicit copulas from regularized smoothers, and our copula family extends these to variable selection.

The rest of this paper is structured as follows. Section 2 outlines our approach, including our proposed copula family. Section 3 details how to compute Bayesian inference, including the predictive densities and Bayes factors. Section 4 contains the simulation study, Section 5 extends the approach to spatial variable selection for fMRI data, and Section 6 concludes. A Web Appendix contains extensive additional material, including proofs, copula properties,

details of the estimation algorithms, and the in-depth analysis of an additional regression dataset with $p = 252$ correlated covariates.

2 Variable Selection in Regression using Copulas

2.1 Marginally calibrated variable selection

Consider a vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ of n realisations on a continuous dependent variable, along with an $n \times p$ design matrix $\boldsymbol{X}$ for p regression covariates. Bayesian approaches to variable selection usually proceed by introducing a vector of binary indicator variables $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)'$, such that the jth covariate is included in the regression if $\gamma_j = 1$, and excluded if $\gamma_j = 0$. When the dependent variable is non-Gaussian, the most common approach is to consider non-Gaussian distributions for the disturbance to a linear model; see Kundu and Dunson (2014), Rossell and Rubio (2018) and references therein. Thus, a non-Gaussian distribution is selected for Y_i, conditional on $\boldsymbol{X}$ and $\boldsymbol{\gamma}$. In this paper we suggest an alternative approach based on copulas that allows the marginal distribution of Y_i, unconditional on $\boldsymbol{X}$ and $\boldsymbol{\gamma}$, to be chosen. We model the joint density of $\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\gamma}$ as

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\gamma}) = c_{\text{BVS}}(\boldsymbol{u}|\boldsymbol{X}, \boldsymbol{\gamma}) \prod_{i=1}^{n} p_Y(y_i)\,, \tag{1}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)'$, $\boldsymbol{u} = (u_1, \ldots, u_n)'$, $u_i = F_Y(y_i)$ and the distribution of Y_i is assumed to be *marginally* invariant with respect to $\boldsymbol{X}$ and $\boldsymbol{\gamma}$, with density p_Y and distribution function F_Y. The impact of the covariate values $\boldsymbol{X}$ and model indicators $\boldsymbol{\gamma}$ on $\boldsymbol{Y}$ *jointly* is captured through the copula with density c_{BVS}, which is a function of $\boldsymbol{X}$ and $\boldsymbol{\gamma}$. For this we use the copula proposed in Section 2.2 below.

A major advantage of employing (1) is that it separates the modeling of the marginal distribution F_Y of the data, from the task of selecting the covariates. Therefore, F_Y can be calibrated accurately, and we model it non-parametrically in our work. Variable selection is

based on the posterior distribution of $\boldsymbol{\gamma}$, which is given by

$$p(\boldsymbol{\gamma}|\boldsymbol{X},\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\gamma})p(\boldsymbol{\gamma}) \propto c_{\text{BVS}}(\boldsymbol{u}|\boldsymbol{X},\boldsymbol{\gamma})p(\boldsymbol{\gamma})\,, \tag{2}$$

with model prior $p(\boldsymbol{\gamma})$. A major aim of this paper is to show that adopting (1) with our proposed copula provides a very general, but tractable, approach to undertaking variable selection and model averaging for non-Gaussian regression data.

We make three remarks concerning the appropriateness of the decomposition at (1). First, regression models are usually specified conditional on parameters for the mean, variance and possibly other moments. In contrast, the expressions above are unconditional on such parameters. We show in Part B of the Web Appendix that when also conditioning on additional model parameters introduced below in Section 2.2, the distribution of Y_i is a function of the covariates, as is expected in a regression model. Second, we show in Part A of the Web Appendix that in a Gaussian linear regression model with a zero mean g-prior for the regression coefficients, the margin of Y_i with the coefficients and error variance integrated out, is asymptotically independent of $\boldsymbol{X}$. Third, the predictive density arising from the copula model at (1) is a function of the covariate values and indicator variables $\boldsymbol{\gamma}$. To see this, consider a new realization Y_{n+1}, with a $p \times 1$ vector of covariate values $\boldsymbol{x}_{n+1}$. Let $\boldsymbol{X}^+ = [\boldsymbol{X}'|\boldsymbol{x}_{n+1}]'$ and $\boldsymbol{u}^+ = (\boldsymbol{u}', F_Y(y_{n+1}))'$, then from (1), Y_{n+1} has predictive density

$$p(y_{n+1}|\boldsymbol{X}^+,\boldsymbol{\gamma},\boldsymbol{y}) = \frac{p(y_{n+1},\boldsymbol{y}|\boldsymbol{X}^+,\boldsymbol{\gamma})}{p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\gamma})} = \frac{c_{\text{BVS}}(\boldsymbol{u}^+|\boldsymbol{X}^+,\boldsymbol{\gamma})}{c_{\text{BVS}}(\boldsymbol{u}|\boldsymbol{X},\boldsymbol{\gamma})} p_Y(y_{n+1})\,. \tag{3}$$

This is a function of the observed values of all the covariates and $\boldsymbol{\gamma}$. Moreover, marginalizing over the posterior of $\boldsymbol{\gamma}$ gives the posterior predictive density for $Y_{n+1}|\boldsymbol{X}^+,\boldsymbol{y}$ as

$$p(y_{n+1}|\boldsymbol{X}^+,\boldsymbol{y}) = \sum_{\boldsymbol{\gamma}} p(y_{n+1}|\boldsymbol{X}^+,\boldsymbol{\gamma},\boldsymbol{y})p(\boldsymbol{\gamma}|\boldsymbol{X},\boldsymbol{y})\,. \tag{4}$$

This forms the basis of the predictive distribution of Y_{n+1} from the copula model, and we give a computationally tractable expression for (4) in Section 3.2.2.

2.2 Variable selection copula

Key to our approach is the specification of our proposed copula with density c_{BVS}. To derive this, consider the linear model

$$\widetilde{\boldsymbol{Z}} = \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma + \boldsymbol{\varepsilon}\,, \tag{5}$$

where $\widetilde{\boldsymbol{Z}} = (\tilde{Z}_1, \ldots, \tilde{Z}_n)'$, $\boldsymbol{X}_\gamma$ is an $n \times q_\gamma$ sub-matrix of $\boldsymbol{X}$ that comprises the columns of $\boldsymbol{X}$ where $\gamma_j = 1$ (so $q_\gamma = \sum_{j=1}^{p} \gamma_j$), $\boldsymbol{\beta}_\gamma$ are the corresponding regression coefficients, and $\boldsymbol{\varepsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$. We refer to $\widetilde{\boldsymbol{Z}}$ as a vector of observations on a 'pseudo-response', because it is not observed directly in our model. Following Smith and Kohn (1996); George and McCulloch (1997); Liang et al. (2008) and many others, the conjugate g-prior $\boldsymbol{\beta}_\gamma | \boldsymbol{X}, \sigma^2, \boldsymbol{\gamma}, g \sim \mathrm{N}(\boldsymbol{0}, g\sigma^2(\boldsymbol{X}_\gamma{}'\boldsymbol{X}_\gamma)^{-1})$, $g > 0$, is used for the non-zero coefficients. Its conjugacy and scaling prove attractive features for constructing the variable selection copula.

To construct the variable selection copula, we first extract the implicit copula of the distribution of $\widetilde{\boldsymbol{Z}}$ conditional on $\boldsymbol{X}, \boldsymbol{\gamma}, g, \sigma^2$, but with $\boldsymbol{\beta}_\gamma$ integrated out. This is $\widetilde{\boldsymbol{Z}} | \boldsymbol{X}, \boldsymbol{\gamma}, g, \sigma^2 \sim N(\boldsymbol{0}; \boldsymbol{\Omega})$, where

$$\boldsymbol{\Omega} = \sigma^2 \left(\boldsymbol{I} - \tfrac{g}{1+g} \boldsymbol{X}_\gamma (\boldsymbol{X}_\gamma{}'\boldsymbol{X}_\gamma)^{-1} \boldsymbol{X}_\gamma{}' \right)^{-1} = \sigma^2 (\boldsymbol{I} + g \boldsymbol{X}_\gamma (\boldsymbol{X}_\gamma{}'\boldsymbol{X}_\gamma)^{-1} \boldsymbol{X}_\gamma{}')\,,$$

which follows from integrating out $\boldsymbol{\beta}_\gamma$ as a normal, and applying the Woodbury formula. The implicit copula is the Gaussian copula (Song, 2000), with parameter matrix equal to the correlation $\boldsymbol{R}$ of the distribution, and density $c_{\text{Ga}}(\boldsymbol{u}; \boldsymbol{R}) = |\boldsymbol{R}|^{-1/2} \exp\left(-\frac{1}{2} \boldsymbol{z}'(\boldsymbol{R}^{-1} - \boldsymbol{I})\boldsymbol{z}\right)$, where $\boldsymbol{z} = (\Phi_1^{-1}(u_1), \ldots, \Phi_1^{-1}(u_n))'$ and Φ_1 is the standard normal distribution function. To derive $\boldsymbol{R}$ we standardize $\boldsymbol{\Omega}$ by the variances

$$\mathrm{Var}(\tilde{Z}_i | \boldsymbol{X}, \boldsymbol{\gamma}, g, \sigma^2) = \sigma^2 (1 + g \boldsymbol{x}_{\gamma,i}{}' (\boldsymbol{X}_\gamma{}'\boldsymbol{X}_\gamma)^{-1} \boldsymbol{x}_{\gamma,i}) \equiv \sigma^2 s_i^{-2}\,, \text{ for } i = 1, \ldots, n\,,$$

where $\boldsymbol{x}_{\gamma,i}{}'$ is the ith row of $\boldsymbol{X}_\gamma$. Let $\boldsymbol{S}_\gamma \equiv \boldsymbol{S}(\boldsymbol{X}, \boldsymbol{\gamma}, g) = \mathrm{diag}(s_1, \ldots, s_n)$, then

$$\boldsymbol{R} \equiv \boldsymbol{R}(\boldsymbol{X}, \boldsymbol{\gamma}, g) = \frac{1}{\sigma^2} \boldsymbol{S}_\gamma \boldsymbol{\Omega} \boldsymbol{S}_\gamma = \boldsymbol{S}_\gamma \left(\boldsymbol{I} + g \boldsymbol{X}_\gamma (\boldsymbol{X}_\gamma{}'\boldsymbol{X}_\gamma)^{-1} \boldsymbol{X}_\gamma{}' \right) \boldsymbol{S}_\gamma\,. \tag{6}$$

Note that the location and scale of $\widetilde{\boldsymbol{Z}}$ are unidentified in its copula, and $\boldsymbol{R}$ is not a function

of σ^2. Therefore, without loss of generality, we set $\sigma^2 = 1$ and do not include an intercept in $\boldsymbol{X}$. We stress here that this does not mean the observational data Y_i has zero mean or fixed scale, which is instead captured through F_Y in (1).

[Table 1 about here.]

Finally, we mix over g with respect to its prior $p(g)$ to obtain the variable selection copula as a continuous mixture of Gaussian copulas, as defined below.

DEFINITION 1: Let C_{Ga} and c_{Ga} be the Gaussian copula function and density, respectively. Then if $\widetilde{\boldsymbol{Z}}$ follows the linear model at (5), with the g-prior for $\boldsymbol{\beta}_\gamma$, and $p(g)$ is a proper density for $g > 0$, then we call $C_{\text{BVS}}(\boldsymbol{u}|\boldsymbol{X}, \boldsymbol{\gamma}) = \int C_{\text{Ga}}(\boldsymbol{u}; \boldsymbol{R}(\boldsymbol{X}, \boldsymbol{\gamma}, g))p(g)\mathrm{d}g$ a variable selection copula, with density function

$$c_{\text{BVS}}(\boldsymbol{u}|\boldsymbol{X}, \boldsymbol{\gamma}) = \int c_{\text{Ga}}(\boldsymbol{u}; \boldsymbol{R}(\boldsymbol{X}, \boldsymbol{\gamma}, g))p(g)\mathrm{d}g\,.$$

It is straightforward to show the function $C_{\text{BVS}}(\boldsymbol{u}|\boldsymbol{X}, \boldsymbol{\gamma})$ is a well-defined copula function.

Table 1 depicts the transformations underlying the construction of C_{BVS}. Part C of the Web Appendix gives some properties of C_{BVS}. We consider the three priors discussed by Liang et al. (2008) for g, and a point mass, as listed below:

(a) *Hyper-g prior:* with density $p(g) = \frac{a-2}{2}(1+g)^{-a/2}$, which is proper for $a > 2$. This implies a beta prior on the shrinkage factor $g/(1+g) \sim \text{Beta}(1, 0.5a - 1)$, and we set $a = 4$ leading to a uniform prior on this factor.

(b) *Hyper-g/n prior:* with density $p(g) = \frac{a-2}{2n}(1 + g/n)^{-a/2}$ and $a = 4$.

(c) *Zellner-Siow prior:* with density $p(g) = \frac{\sqrt{n/2}}{\Gamma(1/2)}c^{-3/2}\exp(-n/(2g))$.

(d) *Point mass prior:* We also consider fixing $g = 100$ and $g = n$ for comparison.

While computing the integral over g in Defn. 1 is possible using numerical methods, in general it is difficult to evaluate C_{BVS} or c_{BVS} directly because $\boldsymbol{R}(\boldsymbol{X}, \boldsymbol{\gamma}, g)$ is an n-dimensional matrix. Instead, we generate g as part of an MCMC scheme, as discussed in Section 3. We

use the popular prior $p(\boldsymbol{\gamma}) = B(p - q_\gamma + 1, q_\gamma + 1)$, where B is the beta function, which implies $p(q_\gamma) = 1/(p+1)$.

3 Estimation and Inference

Estimation of (1) requires estimation of both the marginal F_Y and copula parameters $\boldsymbol{\gamma}$. It is common to use two-stage estimators, where F_Y is estimated first, followed by $\boldsymbol{\gamma}$, because they are much faster and often involve only a minor loss of efficiency (Joe, 2005). Grazian and Liseo (2017) and Klein and Smith (2019) integrate out uncertainty for F_Y using a Bayesian non-parametric estimator, but find that this does not improve the accuracy of inference meaningfully, as we also demonstrate in an empirical example in Part F of the Web Appendix. Therefore, we adopt a two-stage estimator, and use the adaptive kernel density estimator (KDE) of Shimazaki and Shinomoto (2010) to estimate F_Y.

3.1 Posterior evaluation

We follow George and McCulloch (1993) and evaluate the posterior of $\boldsymbol{\gamma}$ using MCMC. However, direct computation of the posterior mass at (2) is slow because computing c_{BVS} requires integration over g, so we generate g as part of the MCMC scheme. To implement the sampler we follow Klein and Smith (2019) and express the likelihood conditional on g in closed form by transforming to the (normalized) pseudo-response as follows. Let $\boldsymbol{Z} = (Z_1, \ldots, Z_n)' = \frac{1}{\sigma}\boldsymbol{S}_\gamma \widetilde{\boldsymbol{Z}}$, where $\widetilde{\boldsymbol{Z}}$ is the pseudo-response at (5), then it follows from Section 2.2 that $\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\gamma}, g \sim N(\boldsymbol{0}, \boldsymbol{R}(\boldsymbol{X}, \boldsymbol{\gamma}, g))$. Moreover, Y_i can be expressed in terms of Z_i as $Y_i = F_Y^{-1}(\Phi_1(Z_i))$, so that by a change of variables from $\boldsymbol{Y}$ to $\boldsymbol{Z}$,

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\gamma}, g) = p(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\gamma}, g) \prod_{i=1}^{n} \frac{p_Y(y_i)}{\phi_1(z_i)} = \phi(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{R}) \prod_{i=1}^{n} \frac{p_Y(y_i)}{\phi_1(z_i)}, \tag{7}$$

where the Jacobian of the transformation is $|\frac{d\boldsymbol{z}}{d\boldsymbol{y}}| = \prod_{i=1}^{n} \frac{p_Y(y_i)}{\phi_1(z_i)}$, ϕ_1 is the standard normal density, and $\phi(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{R})$ is the density of a $N(\boldsymbol{0}, \boldsymbol{R})$ distribution.

While all the terms in the right-hand side of (7) are known, the $n \times n$ matrix $\boldsymbol{R}$ cannot be computed directly for large n. To evaluate the posterior, we employ the following sampler.

MCMC Sampler

At each sweep:

Step 1. Randomly partition $\boldsymbol{\gamma}$ into pairs of elements.

Step 2. For each pair (γ_i, γ_j), generate from $p(\gamma_i, \gamma_j | \{\boldsymbol{\gamma} \backslash \gamma_i, \gamma_j\}, \boldsymbol{X}, g, \boldsymbol{y})$.

Step 3. Generate from $p(g|\boldsymbol{X}, \boldsymbol{\gamma}, \boldsymbol{y})$ using Hamiltonian Monte Carlo.

In forming the partition in Step 1, if p is odd-valued one element is simply selected twice, so that pairs of elements (γ_i, γ_j) are always generated in Step 2. Sampling pairs of elements of $\boldsymbol{\gamma}$ in random order helps to improve mixing of the chain. To implement Step 2, from (7) the joint posterior of the indicators is

$$\begin{aligned} p(\boldsymbol{\gamma}|\boldsymbol{X}, g, \boldsymbol{y}) &\propto p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\gamma}, g) p(\boldsymbol{\gamma}) \propto \phi(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{R}(\boldsymbol{X}, \boldsymbol{\gamma}, g)) p(\boldsymbol{\gamma}) \\ &\propto |\boldsymbol{R}(\boldsymbol{X}, \boldsymbol{\gamma}, g)|^{-1/2} \exp\left\{ -\frac{1}{2} \left(\boldsymbol{z}' \boldsymbol{R}(\boldsymbol{X}, \boldsymbol{\gamma}, g)^{-1} \boldsymbol{z} \right) \right\} p(\boldsymbol{\gamma}) \equiv A(\gamma_i, \gamma_j) \,. \end{aligned}$$

Simulating (γ_i, γ_j) involves computing A for the four possible configurations $\mathcal{S} \equiv \{(0,0), (0,1), (1,0), (1,1)\}$, and then setting

$$p((\gamma_i, \gamma_j) | \{\boldsymbol{\gamma} \backslash (\gamma_i, \gamma_j)\}, \boldsymbol{X}, g, \boldsymbol{y}) = \frac{1}{1+h} \,, \quad h = \sum_{(\tilde{\gamma}_i, \tilde{\gamma}_j) \in \{\mathcal{S} \backslash (\gamma_i, \gamma_j)\}} \frac{A(\tilde{\gamma}_i, \tilde{\gamma}_j)}{A(\gamma_i, \gamma_j)} \,, \tag{8}$$

where all ratios are computed on the logarithmic scale. Fast computation of A for the four configurations in $\mathcal{S}$ can be done using a number of matrix identities; see Part D of the Web Appendix, where at no stage is the full $n \times n$ matrix $\boldsymbol{R}$ computed directly.

Generating g at Step 3 uses an HMC step for $\tilde{g} = \log(g)$. We use a variant of the leapfrog integrator of Neal (2011) with the dual averaging approach of Hoffman and Gelman (2014). This requires computation of $\log(p(\tilde{g}|\boldsymbol{X}, \boldsymbol{\gamma}, \boldsymbol{y}))$ up to an additive constant, and its derivative. Part D of the Web Appendix derives analytical expressions for these, and outlines the HMC step in greater detail. We found that the sampler above works well in our applications,

although a referee highlighted that such samplers may mix poorly for large p, in which case adaptive non-Markov samplers as in Griffin et al. (2017) may be preferable.

3.2 Inference

The sampler above produces Monte Carlo draws $\{\boldsymbol{\gamma}^{[k]}, g^{[k]}; k = 1, \ldots, K\}$ from the posterior $p(\boldsymbol{\gamma}, g|\boldsymbol{X}, \boldsymbol{y})$ and is used to compute posterior estimates as detailed below.

3.2.1 Variable selection

Variables can be selected using the marginal posteriors, which are estimated as

$$\Pr(\gamma_i = 1|\boldsymbol{X}, \boldsymbol{y}) \approx \frac{1}{K}\sum_{k=1}^{K} \Pr(\gamma_i = 1|\{\boldsymbol{\gamma}^{[k]} \backslash (\gamma_i, \gamma_j)\}, \boldsymbol{X}, g^{[k]}, \boldsymbol{y})\,.$$

To evaluate the term in this summation, at Step 2 of the sampler for the single pair (γ_i, γ_j) that contains γ_i, the following is computed

$$\Pr(\gamma_i = 1|\{\boldsymbol{\gamma} \backslash (\gamma_i, \gamma_j)\}, \boldsymbol{X}, g, \boldsymbol{y}) = \frac{A(1,0) + A(1,1)}{A(0,0) + A(1,0) + A(0,1) + A(1,1)}\,,$$

where the four values of the bivariate function $A(\gamma_i, \gamma_j)$ are already computed at (8).

3.2.2 Predictive density

In general, direct evaluation of the predictive density of a new observation Y_{n+1} with covariates $\boldsymbol{x}_{n+1}$ at (3) is infeasible because evaluating c_{BVS} is also. However, the posterior predictive density at (4) can still be evaluated as:

$$p(y_{n+1}|\boldsymbol{X}^+, \boldsymbol{y}) = \sum_{\boldsymbol{\gamma}} \int\int p(y_{n+1}|\boldsymbol{X}^+, \boldsymbol{\beta}_\gamma, \boldsymbol{\gamma}, g, \boldsymbol{y}) p(\boldsymbol{\beta}_\gamma, \boldsymbol{\gamma}, g|\boldsymbol{X}, \boldsymbol{y}) \mathrm{d}(\boldsymbol{\beta}_\gamma, g)\,. \tag{9}$$

The predictive density inside the integrals above can obtained by considering a change of variables from Y_{n+1} to $Z_{n+1} = \Phi_1^{-1}(F_Y(Y_{n+1}))$, with Jacobian $\frac{p_Y(y_{n+1})}{\phi_1(z_{n+1})}$, as

$$p(y_{n+1}|\boldsymbol{X}^+, \boldsymbol{\beta}_\gamma, \boldsymbol{\gamma}, g, \boldsymbol{y}) = p(z_{n+1}|\boldsymbol{X}^+, \boldsymbol{\beta}_\gamma, g, \boldsymbol{\gamma}, \boldsymbol{z}) \frac{p_Y(y_{n+1})}{\phi_1(z_{n+1})}\,.$$

From (5), $\tilde{Z}_{n+1}|\boldsymbol{x}_{\gamma,i}, \boldsymbol{\beta}_\gamma, \boldsymbol{\gamma}, g \sim N(\boldsymbol{x}'_{\gamma,n+1}\boldsymbol{\beta}_\gamma, \sigma^2)$ independently when conditioning on $\boldsymbol{\beta}_\gamma$ (whereas the elements of $\widetilde{\boldsymbol{Z}}$ are dependent unconditional on $\boldsymbol{\beta}_\gamma$). Then, because $Z_{n+1} =$

$\frac{s_{n+1}}{\sigma}\tilde{Z}_{n+1}$,

$$p(y_{n+1}|\boldsymbol{X}^+, \boldsymbol{\beta}_\gamma, \boldsymbol{\gamma}, g, \boldsymbol{y}) = \frac{1}{s_{n+1}}\phi_1\left(\frac{z_{n+1} - s_{n+1}\boldsymbol{x}'_{\gamma,n+1}\boldsymbol{\beta}_\gamma}{s_{n+1}}\right)\frac{p_Y(y_{n+1})}{\phi_1(z_{n+1})}, \tag{10}$$

where $s_{n+1} = (1 + g\boldsymbol{x}'_{\gamma,n+1}(\boldsymbol{X}_\gamma{}'\boldsymbol{X}_\gamma)^{-1}\boldsymbol{x}_{\gamma,n+1})^{-1}$, and $\boldsymbol{x}_{\gamma,n+1}$ are the elements of $\boldsymbol{x}_{n+1}$ that correspond to $\boldsymbol{\gamma}$. Notice that σ^2 cancels out in the above because it is unidentified in the copula, and plays no role in the predictions.

An expression for the posterior predictive density is obtained by plugging (10) into (9). The integrals and summation can be evaluated in the usual Bayesian fashion by averaging over Monte Carlo iterates from the posterior $p(\boldsymbol{\beta}_\gamma, \boldsymbol{\gamma}, g|\boldsymbol{X}, \boldsymbol{y})$. However, this requires the additional generation of $\boldsymbol{\beta}_\gamma$ at each sweep of the sampler. A faster approximation that avoids this—and which we have found to be almost as accurate empirically—is to plug in the posterior expectation of $\boldsymbol{\beta}_\gamma$ conditional on $\boldsymbol{\gamma}, g$, given by $\widehat{\boldsymbol{\beta}}_\gamma = \frac{g}{1+g}(\boldsymbol{X}_\gamma{}'\boldsymbol{X}_\gamma)^{-1}\boldsymbol{X}_\gamma{}'\boldsymbol{S}_\gamma^{-1}\boldsymbol{z}$. The main components required for the evaluation of $\widehat{\boldsymbol{\beta}}_\gamma$ are computed previously at each sweep of the sampler. Thus, a fast and accurate predictive density estimator can be constructed using the K Monte Carlo iterates from the sampler as

$$\hat{p}(y_{n+1}|\boldsymbol{X}^+, \boldsymbol{y}) = \frac{\hat{p}_Y(y_{n+1})}{\phi_1(\Phi_1^{-1}(\hat{F}_Y(y_{n+1})))}\left\{\frac{1}{K}\sum_{k=1}^{K}\frac{1}{s_{n+1}^{[k]}}\phi_1\left(\frac{\Phi_1^{-1}(\hat{F}_Y(y_{n+1})) - s_{n+1}^{[k]}\boldsymbol{x}'_{\gamma^{[k]},n+1}\widehat{\boldsymbol{\beta}}_{\gamma^{[k]}}}{s_{n+1}^{[k]}}\right)\right\}, \tag{11}$$

with $s_{n+1}^{[k]} \equiv (1+g^{[k]}\boldsymbol{x}'_{\gamma^{[k]},n+1}(\boldsymbol{X}'_{\gamma^{[k]}}\boldsymbol{X}_{\gamma^{[k]}})^{-1}\boldsymbol{x}_{\gamma^{[k]},n+1})^{-1}$. Last, the mean of (9) is the regression function estimator, as outlined in Part D of the Web Appendix.

3.2.3 Bayes factor

We derive the Bayes factor for a pair of covariate subsets $\boldsymbol{\gamma}$ and $\widetilde{\boldsymbol{\gamma}}$. For $\boldsymbol{\gamma}$, let $\boldsymbol{U}_\gamma$ be an upper triangular Cholesky factor, such that $\boldsymbol{U}'_\gamma\boldsymbol{U}_\gamma = \boldsymbol{X}'_\gamma\boldsymbol{X}_\gamma$ and $\boldsymbol{M}_\gamma = \boldsymbol{X}_\gamma\boldsymbol{U}_\gamma^{-1}$. Then if $\boldsymbol{M}_\gamma = \{m_{\gamma,ij}\}$, we can express $s_i = (1 + g\sum_{j=1}^{q_\gamma} m_{\gamma,ij}^2)^{-1/2}$ for $i = 1, \ldots, n$, and $\boldsymbol{R}(\boldsymbol{X}, \boldsymbol{\gamma}, g)^{-1} = \boldsymbol{S}_\gamma^{-1}(\boldsymbol{I} - \frac{g}{1+g}\boldsymbol{M}_\gamma\boldsymbol{M}'_\gamma)\boldsymbol{S}_\gamma^{-1}$; with similar expressions for $\widetilde{\boldsymbol{\gamma}}$

PROPOSITION 1: The Bayes factor for model $\boldsymbol{\gamma}$ over model $\widetilde{\boldsymbol{\gamma}}$ is given by

$$\mathrm{BF}(\boldsymbol{\gamma}|\widetilde{\boldsymbol{\gamma}}) = \int_0^\infty \prod_{i=1}^{n} \left[(1+g\sum_{j=1}^{q_\gamma} m_{\gamma,ij}^2)^{\frac{1}{2}} (1+g\sum_{j=1}^{q_{\tilde{\gamma}}} m_{\tilde{\gamma},ij}^2)^{\frac{1}{2}} \right] (1+g)^{-\frac{q_\gamma - q_{\tilde{\gamma}}}{2}}$$
$$\exp\left\{ -\frac{\boldsymbol{z}'\boldsymbol{S}_\gamma^{-1}\boldsymbol{S}_\gamma^{-1}\boldsymbol{z}}{2(1+g)} (1+g\,(1-\tilde{R}_{\gamma,g}^2)) \right\} \exp\left\{ \frac{\boldsymbol{z}'\boldsymbol{S}_{\tilde{\gamma}}^{-1}\boldsymbol{S}_{\tilde{\gamma}}^{-1}\boldsymbol{z}}{2(1+g)} (1+g\,(1-\tilde{R}_{\tilde{\gamma},g}^2)) \right\} p(g)\mathrm{d}g,$$

where we call

$$\tilde{R}_{\gamma,g}^2 = 1 - \frac{\boldsymbol{z}'\boldsymbol{S}_\gamma^{-1}(\boldsymbol{I} - \boldsymbol{M}_\gamma\boldsymbol{M}_\gamma')\boldsymbol{S}_\gamma^{-1}\boldsymbol{z}}{\boldsymbol{z}'\boldsymbol{S}_\gamma^{-1}\boldsymbol{S}_\gamma^{-1}\boldsymbol{z}}$$

the implicit copula coefficient of determination, which (as opposed to the ordinary coefficient of determination) depends on g in the copula model.

Proof of Proposition 1 can be found in Part C of the Web Appendix, and the expression for $\mathrm{BF}(\boldsymbol{\gamma}|\widetilde{\boldsymbol{\gamma}})$ involves an integral which can be evaluated numerically. Part C of the Web Appendix also gives the Bayes factor in the special case where $\boldsymbol{\gamma} = \mathbf{0}$ (i.e. the empty model).

4 Simulation Study

To illustrate the effectiveness of our approach we undertake a simulation study. We employ the copula model at (1) with non-parametric estimates of the margins $\hat{F}_Y$, and label this 'BVSC' throughout the rest of this Section. We consider the four priors for g outlined in Section 2.2 and compare the BVSC to two benchmark methods. The first is that of Liang et al. (2008), which has Gaussian disturbances and the same hyperpriors for g (so that there are also four variants). To evaluate the posterior for this model we use an MCMC sampler similar to that described in Section 3.1, and label this as 'BVS'. The second benchmark is the approach of Rossell and Rubio (2018) as implemented in the R-package `mombf` and labelled 'mombf', where the product MOM (pMOM) non-local prior proposed by Johnson and Rossell (2012) is used. This provides estimates of the posterior model probabilities with normal (N), asymmetric normal (AN), Laplace (L) and asymmetric Laplace (AL) errors, and

we label the benchmark by these distribution types. However, evaluation of the predictive distribution is only available for the normal error model.

4.1 Simulation Design

We generate $n = 200$ observations of $p = 20$ correlated covariates $\boldsymbol{x} = (x_1, \ldots, x_{20})' \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \boldsymbol{D}'\boldsymbol{D}$ and $\boldsymbol{D}$ is an upper triangular Cholesky factor with non-zero elements generated as $\mathrm{N}(0, 0.1^2)$. The resulting $(n \times 20)$ design matrix $\boldsymbol{X}$ is then mean-centered (so that $\mathbf{1}'\boldsymbol{X} = \mathbf{0}$), creating a correlated but numerically stable design. For $j = 1, \ldots, 20$, we set $\beta_j = 0$ with probability 0.75, otherwise we generate β_j from an equally-weighted mixture of the two normals $\mathrm{N}(1, 0.25^2)$ and $\mathrm{N}(-1, 0.25^2)$. Setting $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{20})'$, observations of the dependent variable are generated from the following three distributions:

Case 1, Normal: $Y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}(0, r_1^2)\,,$

Case 2, Log-normal: $Y_i = \exp\left(\boldsymbol{x}_i'\boldsymbol{\beta} + 1.5\varepsilon_i\right), \quad \varepsilon_i \sim \mathrm{N}(0, r_2^2)\,,$

Case 3, Implicit Copula: $Y_i = F_{\mathrm{LN}}^{-1}(\Phi_1(z_i); -2.89, 2),\ z_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}(0, r_3^2)\,,$

for $i = 1, \ldots, n$, and where F_{LN} is the log-normal distribution function. The distribution in case 1 matches that of the Gaussian linear model (which is assumed in the BVS and mombf/N benchmarks), while that in case 3 matches that of the implicit copula model (i.e. BVSC). The distribution in case 2 matches neither model. For each of the three cases we simulated $K = 100$ datasets using the same design matrix $\boldsymbol{X}$, which we refer to as replicates. To make the three cases comparable, we set r_1, r_2 and r_3 to values that give a signal-to-noise ratio (SNR) equal to 8; see the Part E of the Web Appendix for details.

4.2 Results

To compare the approaches we consider two metrics. The first metric measures the accuracy of the predictive densities, and the second measures the correct selection of variables.

4.2.1 Prediction Accuracy

To measure the accuracy of the predictive density of the dependent variable we use the mean logarithmic score computed by ten-fold cross-validation. For a given replicate, we compute this by partitioning the data into ten equally sized sub-samples of sizes n_k, denoted here as $\{(y_{i,k}, \boldsymbol{x}_{i,k}); i = 1, \ldots, n_k\}$ for $k = 1, \ldots, 10$. For each observation in sub-sample k, we compute the predictive density estimator at Eq. (11) using the remaining nine sub-samples as the training data, and denote these densities here as $\hat{p}_k(y_{i,k}|\boldsymbol{x}_{i,k})$. The ten-fold mean logarithmic score is then $\text{MLS} = \frac{1}{10}\sum_{k=1}^{10}\frac{1}{n_k}\sum_{i=1}^{n_k}\log\hat{p}_k(y_{i,k}|\boldsymbol{x}_{i,k})$.

[Figure 1 about here.]

Figure 1 gives boxplots of the MLS of the 100 replicates for the three cases in panels (a–c). In each panel, nine methods are considered: the four variants of both BVSC and BVS, and mombf/N. We make three observations. First, the results for BVS are robust with respect to choice of prior for g, whereas for BVSC fixing $g = 100$ is dominated by the three flexible priors. Second, for the Gaussian data in case 1 BVS is slightly better than BVSC, and mombf/N has the highest variance. Third, BVSC outperforms both BVS and mombf/N substantially in the two non-Gaussian cases 2 and 3, which in case 3 is because the data is generated from a copula model.

4.2.2 Selection Accuracy

The second measure is the precision-recall curve, which is a popular criterion for assessing classification in machine learning. Here, the classification problem is the selection from 20 covariates, using the marginal posteriors $\Pr(\gamma_j = 1|\boldsymbol{y})$, $j = 1, \ldots, 20$, produced by each method. Given a threshold probability value, let TP, FP and FN be true-positive, false-positive and false-negative classification rates, respectively. Then the curve plots Recall=TP/(TP+FN) on the horizontal axis, against Precision=TP/(TP+FP) on the vertical axis, as the threshold

probability value varies from 0 to 1. Simultaneously high values of Recall and Precision (i.e. curves that look like a transposed letter 'L') indicate accurate classification.

[Figure 2 about here.]

Figure 2 plots the average precision-recall curves over the 100 replicates of the simulation. For each case, 12 curves—one for each method and variant—are presented, and we make three observations. First, the asymmetric variants of mombf perform poorly for the non-Gaussian data (cases 2 and 3), despite specifically allowing for this circumstance. Second, the BVSC is much more accurate in cases 2 and 3, where it also outperforms BVS for all variants of priors for g. This robustness to distributional form is a result of the flexibility obtained through the non-parametric calibration of the margins. Third, BVSC is less accurate than BVS and mombf only for the Gaussian data generating process (case 1), although the under-performance is minor when compared to the gains obtained in the two non-Gaussian cases.

5 Extension to Spatial BVS for fMRI

An important application of BVS is to construct activation maps in functional magnetic resonance imaging (fMRI) studies. This involves the extension to spatial data located on a regular lattice of 'voxels' that partition the brain (Smith et al., 2003; Smith and Fahrmeir, 2007; Goldsmith et al., 2014; Lee et al., 2014). We show how to extend our methodology to this case, and demonstrate that the activation maps can be more accurate when taking into account the non-Gaussianity of the magnetic resonance (MR) signal using our implicit copula-based approach to variable selection. This is consistent with recent work that suggests the MR signal is neither conditionally Gaussian nor homoscedastic (Eklund et al., 2017).

5.1 Marginally calibrated variable selection for fMRI

In fMRI studies, a MR signal $\{Y_{i,t}; t = 1, \ldots, T\}$ is observed at each voxel $i \in \{1, \ldots, N\}$. This is matched with a series of scalar observations $\{x_{i,t}; t = 1, \ldots, T\}$ on a transformed stimulus, which is a delayed and continuously modified version of an original stimulus called the 'hemodynamic response' that is derived in a pre-processing step outlined in Smith et al. (2003, p.804). The objective in such analyses is to identify at which voxels the MR signal is related to the transformed stimulus (Bezener et al., 2018).

Denote voxel activation by the vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_N)'$, such that voxel i is activated by the stimulus if $\gamma_i = 1$, and inactivate if $\gamma_i = 0$. Spatial smoothing is essential to obtaining reliable activation maps, which is achieved by using the mass function of an Ising model as a prior

$$p(\boldsymbol{\gamma}|\theta) \propto \exp\left(\sum_{i=1}^{N} \delta_i \gamma_i + \theta \sum_{i \sim j} \omega_{ij} \mathcal{I}(\gamma_i = \gamma_j)\right).$$

Here, $\mathcal{I}(A) = 1$ if A is true and zero otherwise, and the summation over $i \sim j$ is over all pairwise neighboring elements of $\boldsymbol{\gamma}$ (which are the up to 8 neighbors of each voxel in two dimensions). The weight $\omega_{ij} = 1$ for immediately adjacent voxels i, j, and $\omega_{ij} = 1/\sqrt{2}$ for diagonally adjacent voxels i, j, while the parameter $0 \leqslant \theta \leqslant 0.45$ controls the level of spatial smoothing. The $\delta_1, \ldots, \delta_N$ are coefficients of the external field and determined using a process described in Smith and Fahrmeir (2007, Sec.4.2) that computes each δ_i from the amount of grey matter in voxel i (with more grey matter resulting in a higher value of δ_i). This is important because only grey matter can be activated by the stimulus. Alternative binary random fields may also be used as a prior for $\boldsymbol{\gamma}$ in our framework.

To extend our copula model in Section 2.1 to this case, let $\boldsymbol{Y}_i = (Y_{i,1}, \ldots, Y_{i,T})'$, $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,T})'$, $\boldsymbol{Y} = (\boldsymbol{Y}'_1, \ldots, \boldsymbol{Y}'_N)'$ and $\boldsymbol{x} = (\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_N)'$. Also, let $\boldsymbol{w}_t = (w_{t,1}, \ldots, w_{t,m})'$ be m basis functions (which we specify later) evaluated at time point t used to capture a

localized baseline time trend at each voxel, and $\boldsymbol{W} = [\boldsymbol{w}_1|\ldots|\boldsymbol{w}_T]'$. We assume $Y_{i,t}$ has a voxel-specific marginal distribution function $F_{Y_i}(y_{i,t})$ and density $p_{Y_i}(y_{i,t})$, and adopt the following copula model for the joint density of $\boldsymbol{Y}|\boldsymbol{x}, \boldsymbol{W}, \boldsymbol{\gamma}$:

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{W}, \boldsymbol{\gamma}) = \prod_{i=1}^{N} \left(c_{\text{SBVS}}(\boldsymbol{u}_i|\boldsymbol{x}_i, \boldsymbol{W}, \gamma_i) \prod_{t=1}^{T} p_{Y_i}(y_{i,t}) \right), \tag{12}$$

where $\boldsymbol{u}_i = (u_{i,1}, \ldots, u_{i,T})'$ and $u_{i,t} = F_{Y_i}(y_{i,t})$. At (12), the MR signals $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N$ are independent over voxels when conditioning on $\boldsymbol{\gamma}$, with each vector $\boldsymbol{Y}_i$ following a copula decomposition as at (1). For the T-dimensional copula density c_{SBVS} we employ an implicit copula derived from a pseudo-regression model, as outlined below in Section 5.2.

Spatial dependence between elements of $\boldsymbol{\gamma}$ is introduced using the Ising model prior, giving posterior mass function $p(\boldsymbol{\gamma}|\boldsymbol{x}, \boldsymbol{W}, \boldsymbol{y}) \propto \left(\prod_{i=1}^{N} c_{\text{SBVS}}(\boldsymbol{u}_i|\boldsymbol{x}_i, \boldsymbol{W}, \gamma_i) \right) p(\boldsymbol{\gamma}|\theta)$. As in Section 2, variable selection (i.e. the classification of voxels as active or inactive) using this posterior is separated from the task of marginal calibration of the distributions $F_{Y_1}, \ldots, F_{Y_N}$ of the MR signal. We show in our empirical work that this changes the activation maps and improves the quality of fit (measured using mean logarithmic scores) substantially compared to the Gaussian spatial BVS of Smith et al. (2003) and Smith and Fahrmeir (2007).

5.2 Spatial variable selection copula

The implicit copula c_{SBVS} is of the same form at every voxel, and we derive it here for voxel i. It is obtained from the regression $\tilde{Z}_{i,t} = \boldsymbol{w}_t'\boldsymbol{\alpha} + x_{i,t}\beta_{\gamma_i} + \varepsilon_{i,t}$ for pseudo-response $\tilde{Z}_{i,t}$ at times $t = 1, \ldots, T$. Here, $\varepsilon_{i,t} \sim \text{N}(0, \sigma^2)$ and $\boldsymbol{w}_t'\boldsymbol{\alpha}$ is a time trend with eight low order Fourier terms as basis functions and coefficients $\boldsymbol{\alpha}$, while the transformed stimulus $x_{i,t}$ is a scalar covariate with coefficient β_{γ_i}. Variable selection is performed only on $x_{i,t}$, so that $\beta_{\gamma_i} = 0$ iff $\gamma_i = 0$. Smith et al. (2003) and Smith and Fahrmeir (2007) apply regressions of this form directly to the MR signal at each voxel, but here we instead extract its implicit copula for use at (12).

Setting $\widetilde{\boldsymbol{Z}}_i = (\tilde{Z}_{i,1}, \ldots, \tilde{Z}_{i,T})'$, the regression can be written as the linear model

$$\widetilde{\boldsymbol{Z}}_i = \boldsymbol{W}\boldsymbol{\alpha} + \boldsymbol{x}_i\beta_{\gamma_i} + \boldsymbol{\varepsilon}_i\,, \tag{13}$$

with $\boldsymbol{\varepsilon}_i = (\varepsilon_{i,1}, \ldots, \varepsilon_{i,T})'$, and $\boldsymbol{x}_i, \boldsymbol{W}$ as defined in the previous subsection. Proper priors have to be employed for $\boldsymbol{\alpha}$ and β_{γ_i} to obtain a proper implicit copula. We use the same spike-and-slab prior for β_{γ_i} as previously, so that $\beta_{\gamma_i} = 0|\gamma_i = 0$, and $\beta_{\gamma_i}|\gamma_i = 1, g, \sigma^2 \sim \mathrm{N}\left(0, g\sigma^2(\boldsymbol{x}_i'\boldsymbol{x}_i)^{-1}\right)$. We use a $\mathrm{N}(0, \sigma^2 d\boldsymbol{I})$ prior for $\boldsymbol{\alpha}$, with d set to make the prior uninformative, along with the same four priors for g used previously in Section 2.

The same process is used to construct c_{SBVS} as in Section 2.2, but tailored to account for the time trend in (13). First, the joint distribution $\widetilde{\boldsymbol{Z}}_i|\boldsymbol{x}_i, \boldsymbol{W}, \gamma_i, g, \sigma^2 \sim N(\boldsymbol{0}, \boldsymbol{\Omega})$, with

$$\boldsymbol{\Omega} = \sigma^2 \left(\boldsymbol{I} + d\boldsymbol{W}\boldsymbol{W}' + \gamma_i \left(\frac{g\boldsymbol{x}_i\boldsymbol{x}_i'}{\boldsymbol{x}_i'\boldsymbol{x}_i} \right) \right) ,$$

where β_{γ_i} and $\boldsymbol{\alpha}$ have been integrated out analytically; see Part G.1 of the Web Appendix. Second, the pseudo-response is standardized by the marginal variances of this distribution to give $\boldsymbol{Z}_i = \sigma^{-1}\boldsymbol{S}_{\gamma_i}\widetilde{\boldsymbol{Z}}_i$, with $\boldsymbol{S}_{\gamma_i} \equiv \mathrm{diag}(s_1(\gamma_i), \ldots, s_T(\gamma_i))$ and $s_t(\gamma_i) = \left(1 + d\boldsymbol{w}_t'\boldsymbol{w}_t + \gamma_i \left(\frac{g x_{i,t}^2}{\boldsymbol{x}_i'\boldsymbol{x}_i} \right) \right)^{-\frac{1}{2}}$. Thus, $\boldsymbol{Z}_i|\boldsymbol{x}_i, \boldsymbol{W}, \gamma_i, g, \sigma^2 \sim \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{R}(\boldsymbol{x}_i, \boldsymbol{W}, \gamma_i, g)\right)$, where

$$\boldsymbol{R}(\boldsymbol{x}_i, \boldsymbol{W}, \gamma_i, g) = \boldsymbol{S}_{\gamma_i} \left(\boldsymbol{I} + d\boldsymbol{W}\boldsymbol{W}' + \gamma_i \frac{g\boldsymbol{x}_i\boldsymbol{x}_i'}{\boldsymbol{x}_i'\boldsymbol{x}_i} \right) \boldsymbol{S}_{\gamma_i} , \tag{14}$$

is a correlation matrix, and the implicit copula of both $\widetilde{\boldsymbol{Z}}_i$ and $\boldsymbol{Z}_i$ (conditional on $\boldsymbol{x}_i, \boldsymbol{W}, \gamma_i, g, \sigma^2$) is a Gaussian copula with parameter matrix $\boldsymbol{R}$ defined at (14). Finally, we obtain c_{SBVS} by mixing over the prior for g, which we formalize in the following definition.

DEFINITION 2: Let c_{Ga} be the Gaussian copula density with the correlation matrix $\boldsymbol{R}$ defined at (14). If $p(g)$ is a proper prior density for $g > 0$, then we call

$$c_{\text{SBVS}}(\boldsymbol{u}|\boldsymbol{x}_i, \boldsymbol{W}, \gamma_i) = \int c_{\text{Ga}}(\boldsymbol{u}; \boldsymbol{R}(\boldsymbol{x}_i, \boldsymbol{W}, \gamma_i, g)) p(g) \mathrm{d}g \, .$$

the density function of a spatial variable selection copula.

As in Section 2, σ^2 does not feature in the expression for $\boldsymbol{R}$ nor the copula density. We use this copula at (12), where the transformed stimulus values $\boldsymbol{x}_i$ and activation indicator γ_i vary over voxels, whereas the matrix of linear time trend basis terms $\boldsymbol{W}$ does not.

5.3 Estimation and inference

Parameter estimation and posterior inference is obtained using an adaptation of the sampler in Section 3.1 that is presented in detail in Parts G.2 and G.3 of the Web Appendix. It is a single-site sampler in the elements of $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_N)'$, which is more computationally efficient than multi-site samplers (e.g. Nott and Green (2004)) because key quantities can be pre-computed just once. This is important when undertaking variable selection in fMRI studies due to the large number of voxels N. The parameter g is generated using a Metropolis-Hastings (MH) step with a Gaussian approximation as a proposal. This has an acceptance rate of over 80%, and replaces the HMC step in Section 3. The spatial smoothing parameter θ is generated using an adaptive random walk MH step.

Accurate Monte Carlo mixture estimates of the marginal posteriors $\Pr(\gamma_i = 1|\boldsymbol{x}, \boldsymbol{W}, \boldsymbol{y})$ can be readily computed from the conditional posteriors of the single site sampler. Plots of these are Bayesian estimates of the activation maps. An accompanying output is the map of posterior means of the amplitudes $\beta_{\gamma_1}, \ldots, \beta_{\gamma_N}$, for which an efficient mixture estimate can also be constructed, as outlined in Part G.4 of the Web Appendix.

[Table 2 about here.]

[Figure 3 about here.]

5.4 Empirical results

To illustrate, we construct activation maps for slice 10 of individual B in Smith et al. (2003). This data includes an MR signal observed at $T = 63$ time points and $N = 72 \times 86$ voxels from a simple visual experiment. Activation is therefore largely in the visual cortex. Figure 3 plots the first four sample moments of the MR time series at each voxel, indicating a considerable deviation from normality at many voxels. This is not captured in the Gaussian spatial BVS

model (labelled 'SBVS' here). In contrast, our spatial BVS *copula* model (labelled 'SBVSC' here) does so using non-parametric estimators for $F_{Y_1}, \ldots, F_{Y_N}$.

We estimate the SBVSC parameters for all four priors for g. For comparison, we also estimate the SBVS model of Smith et al. (2003) using the same priors for $g, \theta, \boldsymbol{\gamma}$ as in the copula model. To compare the two sets of results, Table 2 reports the mean (in-sample) logarithmic scores for both the SBVS and SBVSC models, and for each prior of g. To compute these we evaluate the mean in-sample logarithmic scores at each voxel i as $1/T \sum_{t=1}^{T} \log(\hat{p}(y_{i,t}|\boldsymbol{x}))$, and then average the results across active, inactive and all voxels. The predictive densities are computed using a minor adjustment to (11) to account for the time trend terms; see Part G.5 of the Web Appendix. We make three observations. First, in all cases the SBVSC scores are higher than those for SBVS. The relative improvement is 9.5% for voxels classified as active by SBVSC, and 2% across all voxels. Second, for SBVSC setting $g = 100$ degrades the results, although there is no difference between the other three priors for g. Third, the expected number of active voxels is lower for SBVSC, producing a sparser image than SBVS.

[Figure 4 about here.]

[Figure 5 about here.]

Based on these observations, Figure 4 compares activation and amplitude maps for three cases: panels (a,d) SBVSC with hyper-g/n prior; (b,e) SBVS with hyper-g/n prior; and (c,f) SBVS with $g = 100$. The latter case is included because it is the benchmark model suggested by Smith and Fahrmeir (2007). The activation maps in the first row are obtained by defining a voxel as active if and only if $\Pr(\gamma_i = 1|\boldsymbol{x}, \boldsymbol{W}, \boldsymbol{y}) \geqslant 0.8722$. The justification for this threshold value is that $-2\log((1-\Pr(\gamma_i = 1|\boldsymbol{x}, \boldsymbol{W}, \boldsymbol{y}))/\Pr(\gamma_i = 1|\boldsymbol{x}, \boldsymbol{W}, \boldsymbol{y})$ is approximately $\chi^2(1)$ distributed and the threshold corresponds to a p-value of 0.05. The maps for the two SBVS models are close to identical, but differ from those for the SBVSC model, with the latter allowing for sharper edges in the amplitude maps. To further highlight this difference,

Figure 5 plots the difference in the activation probabilities between the copula and Gaussian models. These differ between -0.6 and 0.6, so that allowing for more accurate marginal calibration of the MR signal not only increases the logarithmic scores, but affects activation maps, which are the primary output of fMRI processing.

Both samplers were implemented efficiently in MATLAB. The time to undertake 1000 sweeps was approximately 10 mins (SBVS) and 11 mins (SBVSC) when g is generated, and 2 mins (SBVS) and 2.3 mins (SBVSC) when g is fixed. Thus, marginally-calibrated spatial BVS is only slightly slower than Gaussian spatial BVS, and can be made even faster using a lower level language. Typically, only a few thousand sweeps are needed to obtain highly accurate maps, because the Markov chain converges quickly and mixture estimators are used. While not undertaken here, the speed of the approach allows application to multiple slices using the 3D Markov random field prior for $\boldsymbol{\gamma}$ in Smith and Fahrmeir (2007).

6 Discussion

This paper proposes a new tractable and general approach to undertake BVS for non-Gaussian data. It uses a copula decomposition that allows the marginal distribution of the dependent variable to be calibrated accurately using nonparametric estimation. However, we note that this does not imply the other forms of calibration or dispersion listed by Gneiting et al. (2013). In addition, a referee pointed out that the assumption at (1) that F_Y is independent of $\boldsymbol{X}$ limits the range of distributions for $\boldsymbol{Y}|\boldsymbol{X}$ that can be represented. However, this assumption separates the variable selection problem from that of marginal calibration, which is a major advantage of our proposed approach.

The key ingredient of our methodology is a family of implicit copulas that are constructed from a regression model for a Gaussian pseudo-response with spike-and-slab priors. These mix over the different priors in Liang et al. (2008) for the scaling factor of the g-prior, producing the copula family at Defn. 1 which is a continuous mixture of Gaussian copulas.

We apply our approach to an example with 6192 spatially correlated indicators, and to second example with $p = 252$ correlated covariates in Part F of the Web Appendix. The empirical work demonstrates that mixing over the priors for g (particularly the hyper-g prior) results in much more accurate covariate selection and predictive densities of the dependent variable, compared to fixing g. This is in contrast to Gaussian Bayesian variable selection, where in some examples fixed g (such as $g = 100$ or $g = n$) can provide good results (Smith and Kohn, 1996). Moreover, in the second example we also find that integrating out uncertainty in F_Y as in Grazian and Liseo (2017) does not meaningfully effect covariate selection or prediction. The method is scalable to higher dimensions because estimation by stochastic search over $\boldsymbol{\gamma}$ is fast when exploiting the matrix identities in the Web Appendix. The extension of the method to spatial BVS for fMRI—an important contemporary application (Lee et al., 2014; Bezener et al., 2018)—highlights its tractability.

While the use of copulas to capture dependence between multiple observations on one or more variables is rare, there are some recent examples. In regression these include implicit copulas constructed from Gaussian processes (Wilson and Ghahramani, 2010; Wauthier and Jordan, 2010) or regularized basis functions (Klein and Smith, 2019). In time series analysis, copulas have been used to capture serial dependence in univariate and multivariate data; for example, see Smith (2015) and Shi and Yang (2018). These studies all exploit a copula decomposition to allow for non-Gaussian margins. However, ours is the first study to employ our proposed copula formulation to undertake variable selection.

We finish by making some suggestions for future work. A potential extension is to multivariate regression (i.e. with multiple dependent variables), generalizing the approaches of Brown et al. (1998), Smith and Kohn (2000) and others. To do so requires the specification of an implicit copula analogous to that in Defn. 1. It would also be interesting to explore the relationship between properties of c_{BVS}, such as the dependence metrics in Part C of the Web

Appendix, and covariate selection accuracy. Last, the implicit copula can be extended to other Bayesian models for fMRI data that are based on alternative specifications of (13), such as that in Lee et al. (2014). This would lead to other copula processes on the covariate space with strong potential to further improve voxel classification accuracy.

Acknowledgements

Nadja Klein gratefully acknowledges funding from the Alexander von Humboldt foundation and financial support through the Emmy Noether grant KL 3037/1-1 of the German research foundation (DFG). The authors thank two referees and an associate editor for extensive comments that improved exposition in the manuscript.

Data Availability

The data that support the findings in this paper are available in the Supporting Information.

References

Bezener, M., Hughes, J. and Jones, G. (2018). Bayesian spatiotemporal modeling using hierarchical spatial priors, with applications to functional magnetic resonance imaging (with discussion), *Bayesian Analysis* **13**(4): 1261–1313.

Bottolo, L. and Richardson, S. (2010). Evolutionary stochastic search for Bayesian model exploration, *Bayesian Analysis* **5**(3): 583–618.

Brown, P. J., Vannucci, M. and Fearn, T. (1998). Multivariate Bayesian variable selection and prediction, *Journal of the Royal Statistical Society: Series B* **60**(3): 627–641.

Chung, Y. and Dunson, D. (2009). Nonparametric Bayes conditional distribution modeling with variable selection, *Journal of the American Statistical Association* **104**: 1646–1660.

Eklund, A., Lindquist, M. A. and Villani, M. (2017). A Bayesian heteroscedastic GLM with application to fMRI data with motion spikes, *Neuroimage* **155**: 354–369.

George, E. and McCulloch, R. (1993). Variable selection via Gibbs sampling, *Journal of the American Statistical Association* **88**: 881–889.

George, E. and McCulloch, R. (1997). Approaches to Bayesian variable selection, *Statistica Sinica* **7**: 339–374.

Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness, *Journal of the Royal Statistical Society: Series B* **69**(2): 243–268.

Gneiting, T., Ranjan, R. et al. (2013). Combining predictive distributions, *Electronic Journal of Statistics* **7**: 1747–1782.

Goldsmith, J., Huang, L. and Crainiceanu, C. M. (2014). Smooth scalar-on-image regression via spatial Bayesian variable selection, *Journal of Computational and Graphical Statistics* **23**(1): 46–64.

Gottardo, R. and Raftery, A. (2009). Bayesian robust transformation and variable selection: A unified approach, *Canadian Journal of Statistics* **37**(3): 361–380.

Grazian, C. and Liseo, B. (2017). Approximate Bayesian inference in semiparametric copula models, *Bayesian Analysis* **12**(4): 991–1016.

Griffin, J., Latuszynski, K. and Steel, M. (2017). In search of lost (mixing) time: Adaptive Markov chain Monte Carlo schemes for Bayesian variable selection with very large p, *arXiv preprint, 1708.05678* .

Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo, *Journal of Machine Learning Research* **15**: 1351–1381.

Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models, *Journal of Multivariate Analysis* **94**(2): 401–419.

Johnson, V. E. and Rossell, D. (2012). Bayesian model selection in high-dimensional settings, *Journal of the American Statistical Association* **107**(498): 649–660.

Klein, N. and Smith, M. S. (2019). Implicit copulas from Bayesian regularized regression smoothers, *Bayesian Analysis* **14**(4): 1143–1171.

Kraus, D. and Czado, C. (2017). D-vine copula based quantile regression, *Computational Statistics & Data Analysis* **110**: 1–18.

Kundu, S. and Dunson, D. B. (2014). Bayes variable selection in semiparametric linear models, *Journal of the American Statistical Association* **109**(505): 437–447.

Lee, K.-J., Jones, G. L., Caffo, B. S. and Bassett, S. S. (2014). Spatial Bayesian variable selection models on functional magnetic resonance imaging time-series data, *Bayesian Analysis* **9**(3): 699.

Li, F. and Zhang, N. R. (2010). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics, *Journal of the American Statistical Association* **105**(491): 1202–1214.

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.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics, *in* S. Brooks, A. Gelman, G. Jones and X.-L. Meng (eds), *Handbook of Markov Chain Monte Carlo*, CRC Press, pp. 113–162.

Nelsen, R. (2006). *An Introduction to Copulas*, 2nd edn, Springer.

Nott, D. J. and Green, P. J. (2004). Bayesian variable selection and the Swendsen-Wang algorithm, *Journal of Computational and Graphical Statistics* **13**(1): 141–157.

O'Hara, R. and Sillanpää, M. (2009). A review of Bayesian variable selection methods: What, How, and Which, *Bayesian Analysis* **4**: 85–118.

Rossell, D. and Rubio, F. J. (2018). Tractable Bayesian variable selection: beyond normality, *Journal of the American Statistical Association* **113**(524): 1742–1758.

Sharma, R. and Das, S. (2018). Regularization and variable selection with copula prior,

arXiv:1709.05514v2 .

Shi, P. and Yang, L. (2018). Pair copula constructions for insurance experience rating, *Journal of the American Statistical Association* **113**(521): 122–133.

Shimazaki, H. and Shinomoto, S. (2010). Kernel bandwidth optimization in spike rate estimation, *Journal of Computational Neuroscience* **29**(1-2): 171–182.

Smith, M., Pütz, B., Auer, D. and Fahrmeir, L. (2003). Assessing brain activity through spatial Bayesian variable selection, *NeuroImage* **20**(2): 802–815.

Smith, M. S. (2015). Copula modelling of dependence in multivariate time series, *International Journal of Forecasting* **31**: 815–833.

Smith, M. S. and Fahrmeir, L. (2007). Spatial Bayesian variable selection with application to functional magnetic resonance imaging, *Journal of the American Statistical Association* **102**(478): 417–431.

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

Smith, M. S. and Kohn, R. (2000). Nonparametric seemingly unrelated regression, *Journal of Econometrics* **98**: 257–281.

Song, P. (2000). Multivariate dispersion models generated from Gaussian copula, *Scandinavian Journal of Statistics* **27**(2): 305–320.

Wauthier, F. L. and Jordan, M. I. (2010). Heavy-tailed process priors for selective shrinkage, *Advances in Neural Information Processing Systems*, pp. 2406–2414.

Wilson, A. G. and Ghahramani, Z. (2010). Copula processes, *Advances in Neural Information Processing Systems*, pp. 2460–2468.

Yu, K., Chen, C., Reed, C. and Dunson, D. (2013). Bayesian variable selection in quantile regression, *Statistics and its interface* **6**: 261–274.

Supporting Information

Online Appendices, Tables, and Figures referenced in Sections 1 through 5 are available with this paper at the Biometrics website on Wiley Online Library. MATLAB code to reproduce the results for the fMRI data in Section 5 is also available at the same website.

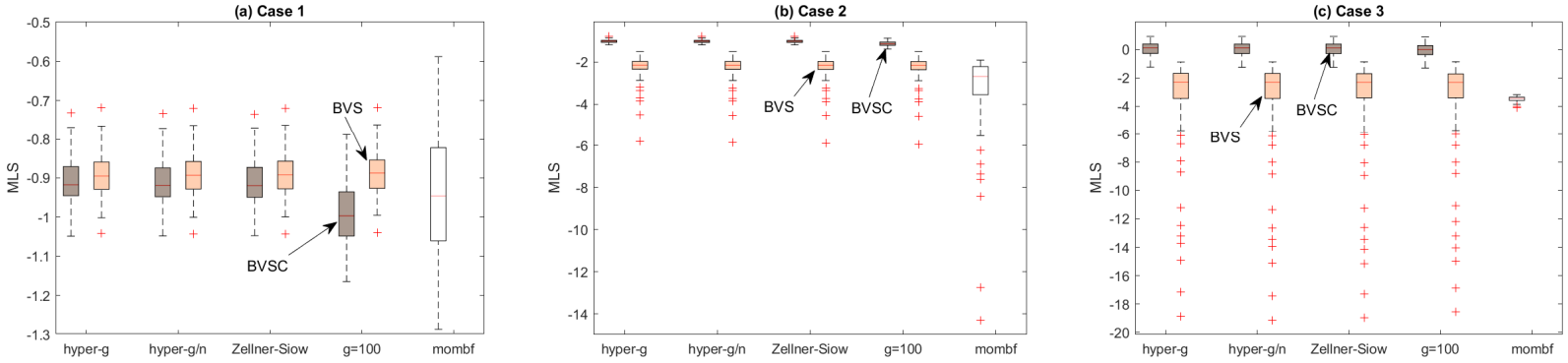

Figure 1: Comparison of the predictive MLS from the simulation study. The three panels provide results for the three cases. Each boxplot is of the 100 values of the MLS from the 100 simulation replicates, where higher values correspond to increased accuracy. In each panel, the first eight boxplots correspond to combinations of the methods BVSC and BVS with the four priors for g. The last boxplot (white) corresponds to the mombf/N method. BVSC dominates all other methods in the non-Gaussian cases 2 and 3. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

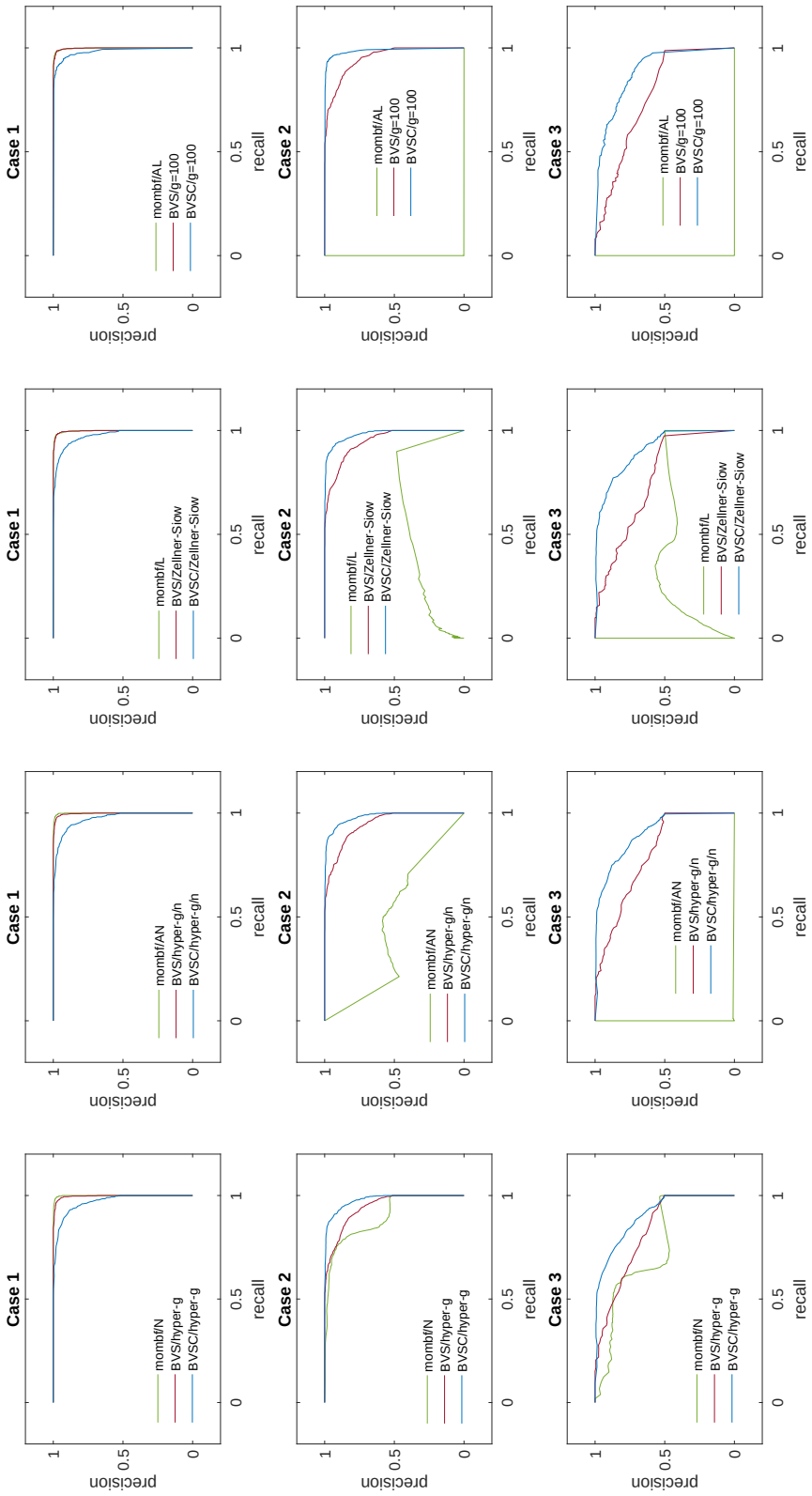

Figure 2: Comparison of the average recall-precision curves from the simulation study for a grid of thresholds in $(0, 1)$. The first row (i.e. panels a,b,c,d) show results for case 1, the second row (i.e. panels e,f,g,h) for case 2, and the third row (i.e. panels i,j,k,l) for case 3. Each panel contains curves for the BVSC (blue) and the BVS (red) methods for one specific prior for g, along with curves from mombf (green) with one distributional assumption from N (normal), AN (asymmetric normal), L (Laplace) and AL (asymmetric Laplace) in each column. Curves with simultaneously higher recall and precision are more accurate classifiers. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

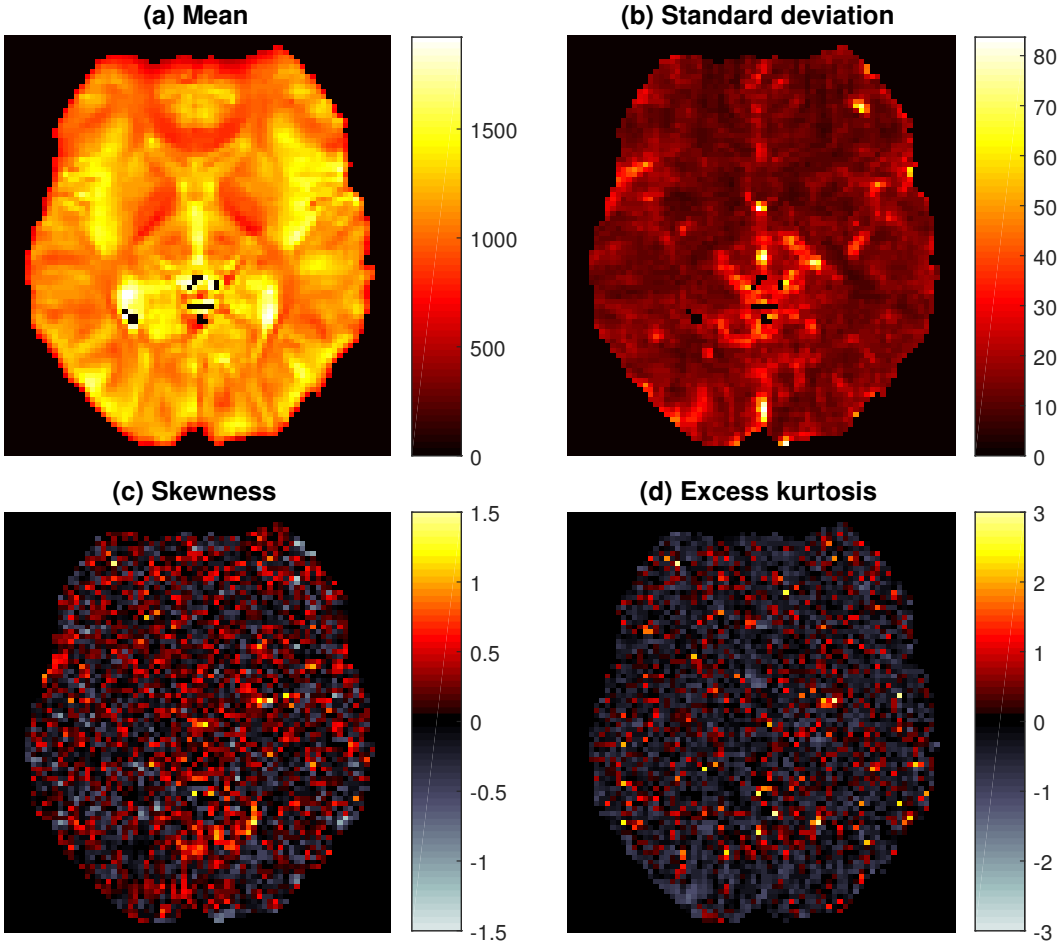

Figure 3: Sample moments of the MR signal at each voxel in the fMRI example. The four panels plot the (a) mean, (b) standard deviation, (c) Pearson skew and (d) excess kurtosis values for each voxel. Computation of the global Moran's I spatial correlation coefficients indicates that there is strong spatial correlation in all four sample moments. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

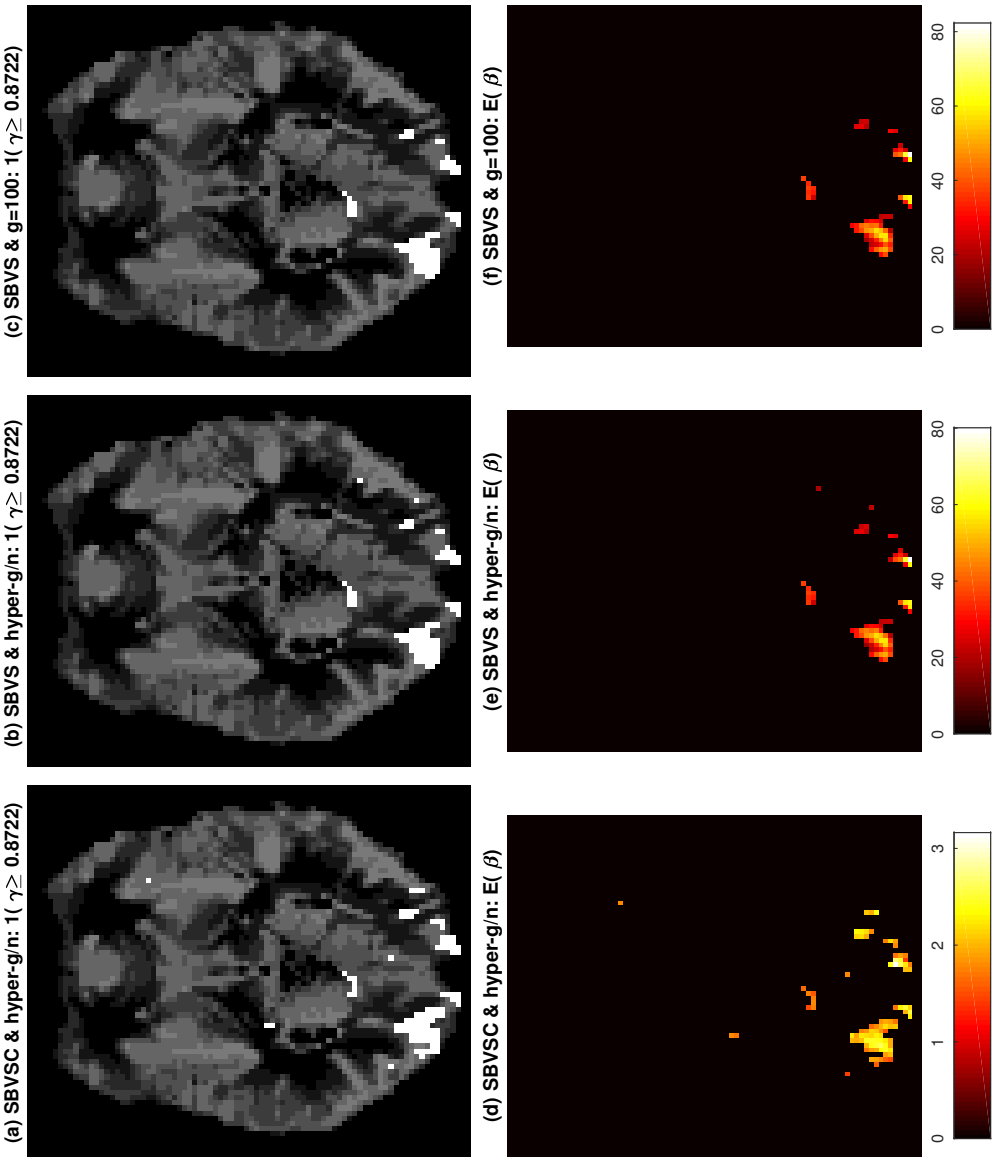

Figure 4: Posterior activation and amplitude maps for the fMRI data from three different estimators. The first row shows the activation maps (where white voxels are those classified as active), and the second shows the mean activation amplitudes. The three different estimators are: (a,d) SBVSC with hyper-g/n prior; (b,e) SBVS with hyper-g/n prior; and, (c,f) SBVS with $g = 100$. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

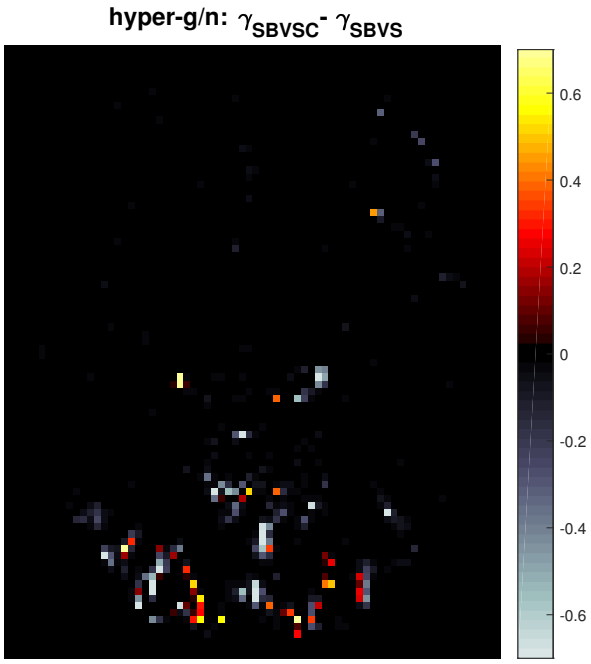

Figure 5: Difference between the activation probabilities of the copula (SBVSC) and Gaussian (SBVS) spatial Bayesian variable selection models for the fMRI data. The difference is SBVSC minus SBVS, and both models employ the same Ising prior for $\boldsymbol{\gamma}$ and hyper-g/n prior for g. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

Table 1: Summary of the one-to-one transformations between the dependent variable Y_i, copula variable U_i specified in Section 2, and the standardized pseudo-response $Z_i = \frac{s_i}{\sigma}\tilde{Z}_i$ specified in Section 3.1. Also given are the joint densities of $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$, $\boldsymbol{U} = (U_1, \ldots, U_n)'$ and $\boldsymbol{Z} = (Z_1, \ldots, Z_n)'$ conditioning on $\boldsymbol{X}, \boldsymbol{\gamma}$ and with/without g. Above, $\boldsymbol{y} = (y_1, \ldots, y_n)'$, $\boldsymbol{u} = (u_1, \ldots, u_n)'$, $\boldsymbol{z} = (z_1, \ldots, z_n)'$ and the $n \times n$ correlation matrix $\boldsymbol{R} \equiv \boldsymbol{R}(\boldsymbol{X}, \boldsymbol{\gamma}, g)$ is specified at (6).

	Observed Data	Copula Data	Standardized Pseudo-data
Variable	Y_i	$U_i = F_Y(Y_i)$	$Z_i = \Phi_1^{-1}(U_i)$
Domain	$\mathbb{R}$	$[0, 1]$	$\mathbb{R}$
Marginal distribution	F_Y	Uniform	Standard Normal
Joint density conditional on $\boldsymbol{X}, \boldsymbol{\gamma}, g$	$p(\boldsymbol{y}\|\boldsymbol{X}, \boldsymbol{\gamma}, g) = \phi(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{R}) \times \prod_{i=1}^n \frac{p_Y(y_i)}{\phi_1(z_i)}$	$p(\boldsymbol{u}\|\boldsymbol{X}, \boldsymbol{\gamma}, g) = c_{\text{Ga}}(\boldsymbol{u}\|\boldsymbol{X}, \boldsymbol{\gamma}, g)$	$p(\boldsymbol{z}\|\boldsymbol{X}, \boldsymbol{\gamma}, g) = \phi(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{R})$
Joint density conditional on $\boldsymbol{X}, \boldsymbol{\gamma}$	$p(\boldsymbol{y}\|\boldsymbol{X}, \boldsymbol{\gamma}) = c_{\text{BVS}}(\boldsymbol{u}\|\boldsymbol{X}, \boldsymbol{\gamma}) \times \prod_{i=1}^n p_Y(y_i)$	$p(\boldsymbol{u}\|\boldsymbol{X}, \boldsymbol{\gamma}) = c_{\text{BVS}}(\boldsymbol{u}\|\boldsymbol{X}, \boldsymbol{\gamma})$	$p(\boldsymbol{z}\|\boldsymbol{X}, \boldsymbol{\gamma}) = \int \phi(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{R}) p(g) \mathrm{d}g$

Table 2: Results from applying both the copula (SBVSC) and Gaussian (SBVS) spatial Bayesian variable selection methods to the fMRI data. The four different priors are used for g, giving a total of eight methods. The top rows report the mean logarithmic scores (MLS) of the MR signal (multiplied by 100 for presentation), broken down voxels classified as active, inactive and overall. Higher MLS values indicate greater accuracy. The bottom rows report the posterior mean and standard deviation of the number of active voxels $q_\gamma = \sum_{j=1}^{N} \gamma_j$.

Prior	hyper-g		hyper-g/n		Zellner-Siow		$g = 100$	
Model	SBVSC	SBVS	SBVSC	SBVS	SBVSC	SBVS	SBVSC	SBVS
Active	-411.65	-450.72	-411.65	-450.72	-411.65	-450.72	-418.37	-451.08
Inactive	-408.12	-416.48	-408.12	-416.48	-408.12	-416.48	-408.32	-416.53
Overall	-408.18	-416.92	-408.18	-416.92	-408.17	-416.92	-408.42	-416.96
$\mathbb{E}(q_\gamma \vert \boldsymbol{y})$	110.8	162.0	110.8	162.0	110.8	162.0	72.1	119.5
$\mathrm{Std}(q_\gamma \vert \boldsymbol{y})$	3.1	13.9	3.1	13.9	3.1	13.9	2.3	8.6