Appendix S1: MCMC algorithm used to fit the multivariate probit (MP) occupancy model

Here we describe the likelihood function, prior distribution, and MCMC algorithm (Brooks *et al.*, 2011) used to compute a Markov chain for estimating summaries of the posterior distribution. We use bracket notation (Gelfand & Smith, 1990) to specify probability density functions (pdfs) and probability mass functions (pmfs). For example, $[x, y]$ denotes the joint density of random variables X and Y, $[x|y]$ denotes the conditional density of X given Y, and $[x]$ denotes the unconditional (marginal) density of X.

We developed a MCMC algorithm to generate a Markov chain whose stationary distribution is equivalent to a posterior distribution with the following unnormalized pdf:

$$\begin{aligned}[\boldsymbol{B}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}^{-1}, \boldsymbol{a}, \boldsymbol{W}, \boldsymbol{Z}|\boldsymbol{Y}] \;\; &\propto \;\; [\boldsymbol{B}][\boldsymbol{\alpha}][\boldsymbol{a}][\boldsymbol{\Sigma}^{-1}|\boldsymbol{a}] \prod_{i=1}^{n} [\boldsymbol{w}_i \mid \boldsymbol{B}'\boldsymbol{x}_i, \boldsymbol{\Sigma})] \prod_{j=1}^{m} I(W_{i,j} > 0)^{Z_{i,j}} (1 - I(W_{i,j} > 0))^{1-Z_{i,j}} \\ &\times \left(\prod_{k=1}^{K_i} \left(Z_{i,j}\, \Phi(\boldsymbol{\alpha}_j' \boldsymbol{V}_{i,k})\right)^{Y_{i,k,j}} \left(1 - Z_{i,j}\, \Phi(\boldsymbol{\alpha}_j' \boldsymbol{V}_{i,k})\right)^{1-Y_{i,k,j}} \right)\end{aligned}$$

where the vector $\boldsymbol{a} = (a_1, \ldots, a_m)'$ contains auxiliary parameters that were used to specify the prior distribution of $\boldsymbol{\Sigma}$ hierarchically (Huang & Wand, 2013) as described below.

Our MCMC algorithm uses either Gibbs or Metropolis-Hastings (MH) sampling (Geyer, 2011) depending on the parameter. Each of the following conditional posterior distributions (i.e., full conditionals) was sampled in one iteration of the algorithm.

1. The full conditional distribution of $\boldsymbol{w}_i = (w_{i,1}, \ldots, w_{i,m})'$ is a truncated multivariate normal, that is,

$$\boldsymbol{w}_i \mid \cdot \;\; \sim \;\; \text{TruncNormal}(\boldsymbol{B}'\boldsymbol{x}_i,\, \boldsymbol{\Sigma};\, \boldsymbol{c}_i, \boldsymbol{d}_i)$$

   where $\boldsymbol{c}_i$ and $\boldsymbol{d}_i$ denote the respective lower and upper limits of support for $\boldsymbol{w}_i$ (i.e., $c_{i,j} \leq w_{i,j} \leq d_{i,j}$). The values of $c_{i,j}$ and $d_{i,j}$ depend on the value of $Z_{i,j}$. Specifically, if $Z_{i,j} = 1$, then $c_{i,j} = 0$ and $d_{i,j} = \infty$ (that is, $w_{i,j}$ is lower-truncated at zero). If $Z_{i,j} = 0$, then $c_{i,j} = -\infty$ and $d_{i,j} = 0$ (that is, $w_{i,j}$ is upper-truncated at zero). The limits of support hold jointly for all elements of $\boldsymbol{w}_i$, and it is important to sample the truncated normal distribution such that these joint restrictions are honored.

2. We assumed a uniform prior for $\boldsymbol{B}$ (that is, $[\boldsymbol{B}] \propto 1$) so that the full conditional of $\boldsymbol{B}$ is a matrix normal distribution with location matrix $\hat{\boldsymbol{B}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}$, first covariance matrix $\boldsymbol{\Sigma}$, and second covariance matrix $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ (Sinay & Hsu, 2014). Therefore, the full conditonal distribution of $\text{vec}(\boldsymbol{B})$ has a familiar form:

$$\text{vec}(\boldsymbol{B}) \mid \cdot \quad \sim \quad \text{Normal}(\text{vec}(\hat{\boldsymbol{B}}),\, \boldsymbol{\Sigma} \otimes (\boldsymbol{X}'\boldsymbol{X})^{-1})$$

3. We assumed a hierarchical prior distribution for $\boldsymbol{\Sigma}$ (Huang & Wand, 2013) that allowed us to specify marginally noninformative priors for its elements. Specifically, the hyperparameters of this prior can be chosen so that each standard deviation parameter $\sqrt{\sigma_{j,j}}$ has a Half-t prior of arbitrarily high noninformity (Gelman, 2006) and each correlation parameter $r_{j,k} = \sigma_{j,k}/\sqrt{\sigma_{j,j}\,\sigma_{k,k}}$ has a uniform prior on (-1,1). The hierarchical prior distribution for $\boldsymbol{\Sigma}$ is a mixture of Inverse-Wishart and Inverse-Gamma distributions:

$$\begin{aligned} \boldsymbol{\Sigma}^{-1} \mid a_1, \ldots, a_m &\sim \text{Wishart}(\nu + m - 1,\ (2\nu\boldsymbol{A})^{-1}) \\ a_j &\sim \text{Gamma}(1/2, 1/s_j^2) \end{aligned}$$

where $\boldsymbol{A} = \text{diag}(a_1, \ldots, a_m)$. Huang & Wand (2013) showed that the marginal prior density of $r_{j,k}$ is proportional to $(1 - r_{j,k})^{\nu/2-1}$; therefore, we let $\nu = 2$ to specify a marginally uniform prior for each correlation parameter. Huang & Wand (2013) also showed that the marginal prior distribution for $\sqrt{\sigma_{j,j}}$ is a Half-t distribution with ν degrees of freedom and scale parameter s_j; therefore, we specified a noninformative prior for $\sqrt{\sigma_{j,j}}$ by choosing s_j to be arbitrarily high.

The conditional conjugacy of this prior for $\boldsymbol{\Sigma}$ leads to full conditional distributions of familiar form that are relatively easy to sample. Specifically,

$$\begin{aligned} \boldsymbol{\Sigma}^{-1} \mid \cdot &\sim \text{Wishart}(\nu + m - 1 + n,\ (2\nu\boldsymbol{A} + \boldsymbol{E}'\boldsymbol{E})^{-1}) \\ a_j \mid \cdot &\sim \text{Gamma}((\nu + m)/2,\ 1/s_j^2 + \nu(\boldsymbol{\Sigma}^{-1})_{j,j}) \end{aligned}$$

where $\boldsymbol{E} = \boldsymbol{W} - \boldsymbol{X}\boldsymbol{B}$ and $(\boldsymbol{\Sigma}^{-1})_{j,j}$ is the jth diagonal element of $\boldsymbol{\Sigma}^{-1}$.

4. The full conditional distribution of $\boldsymbol{\alpha}_j$ has the following unnormalized pdf:

$$[\boldsymbol{\alpha}_j|\cdot] \;\; \propto \;\; [\boldsymbol{\alpha}_j] \prod_{\substack{i=1: \\ Z_{i,j}=1}}^{n} \prod_{k=1}^{K_i} \left(\boldsymbol{\Phi}(\boldsymbol{\alpha}_j' \boldsymbol{V}_{i,k})\right)^{Y_{i,k,j}} \left(1 - \boldsymbol{\Phi}(\boldsymbol{\alpha}_j' \boldsymbol{V}_{i,k})\right)^{1-Y_{i,k,j}}$$

where $[\boldsymbol{\alpha}_j]$ denotes the density function of a multivariate standard normal prior. The standard normal distribution for scalar α_j implies a uniform prior distribution over the interval $(0, 1)$ for the transformation $\Phi(\alpha_j)$ of α_j. In addition, we centered and scaled continously-valued regressors to have zero mean and unit variance, so it seems sensible to assume the standard normal prior for every element of $\boldsymbol{\alpha}_j$. To sample the full conditional of $\boldsymbol{\alpha}_j$, we used Metropolis-Hastings sampling treating $[\boldsymbol{\alpha}_j|\cdot]$ as the target density. In particular, we used a multivariate normal distribution as a proposal and selected its parameters to approximate the target distribution. The mean and covariance matrix of this proposal were computed by recognizing that $[\boldsymbol{\alpha}_j|\cdot]$ is the product of the prior density and the likelihood function for a binary-regression model of conditional outcome $Y_{i,k,j}$ given $Z_{i,j} = 1$ with success probability $p_{i,k,j} = \Phi(\boldsymbol{\alpha}_j' \boldsymbol{V}_{i,k})$. Therefore, we fit this binary-regression model by the method of maximum likelihood and used the maximum-likelihood estimate $\hat{\boldsymbol{\alpha}}_j$ and its asymptotic covariance matrix as the proposal distribution's mean and covariance matrix, respectively.

5. The full conditional distribution of $Z_{i,j}$ is

$$Z_{i,j}|\cdot \sim \begin{cases} \text{Bernoulli}(1) & \text{if } \sum_{k=1}^{K_i} Y_{i,k,j} \neq 0 \\ \text{Bernoulli}(\theta_{i,j}) & \text{if } \sum_{k=1}^{K_i} Y_{i,k,j} = 0 \end{cases}$$

where

$$\theta_{i,j} = \frac{\Phi(\frac{\mu_w}{\sigma_w}) \prod_{k=1}^{K_i}(1 - p_{i,k,j})}{1 - \Phi(\frac{\mu_w}{\sigma_w}) + \Phi(\frac{\mu_w}{\sigma_w}) \prod_{k=1}^{K_i}(1 - p_{i,k,j})}$$

and

$\mu_w = \mu_{i,j} + \boldsymbol{\sigma}_j' \, \boldsymbol{\Sigma}_{-j,-j} \, (\boldsymbol{w}_{i,-j} - \boldsymbol{\mu}_{i,-j})$

$\sigma_w^2 = \sigma_{j,j} + \boldsymbol{\sigma}_j' \, \boldsymbol{\Sigma}_{-j,-j} \, \boldsymbol{\sigma}_j$

$\boldsymbol{w}_{i,-j}$ = all elements of $\boldsymbol{w}_i = (w_{i,1}, \ldots, w_{i,m})'$ except $w_{i,j}$

$\boldsymbol{\mu}_i = \boldsymbol{B}'\boldsymbol{x}_i$

$\boldsymbol{\mu}_{i,-j}$ = all elements of $\boldsymbol{\mu}_i$ except $\mu_{i,j}$

$\boldsymbol{\Sigma}_{-j,-j}$ = principal submatrix of $\boldsymbol{\Sigma}$ without its jth row and jth column

$\boldsymbol{\sigma}_j$ = all elements in jth column of $\boldsymbol{\Sigma}$ except $\sigma_{j,j}$

References

Brooks, S., Gelman, A., Jones, G.L. & Meng, X.L. (2011) *Handbook of Markov chain Monte Carlo*. Chapman & Hall / CRC, Boca Raton, Florida.

Gelfand, A.E. & Smith, A.F.M. (1990) Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models (Comment on article by Browne and Draper). *Bayesian Analysis*, **1**, 515–534.

Geyer, C.J. (2011) Introduction to Markov chain Monte Carlo. S. Brooks, A. Gelman, G.L. Jones & X.L. Meng, eds., *Handbook of Markov chain Monte Carlo*, pp. 3–48. Chapman & Hall / CRC, Boca Raton, Florida.

Huang, A. & Wand, M.P. (2013) Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis*, **8**, 439–452.

Sinay, M.S. & Hsu, J.S.J. (2014) Bayesian inference of a multivariate regression model. *Journal of Probability and Statistics*, **2014**, 673657.