Scalable and robust regression models for continuous proportional data

Changwoo Lee
Postdoctoral Associate, Department of Statistical Science, Duke University

December 2025
Joint work with:

Benjamin Dahl
(Duke U.)

Otso Ovaskainen
(U. Jyväskylä)

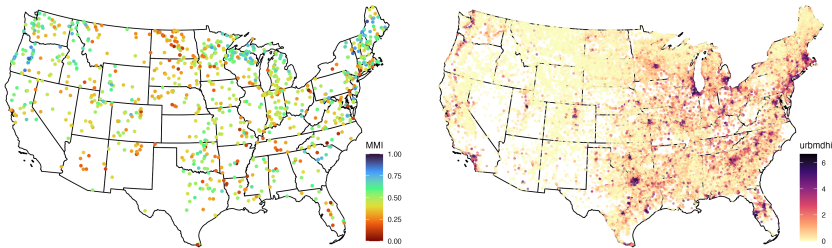
David Dunson
(Duke U.)

Continuous proportional data

- Regression analysis of **continuous proportional data** $Y \in [0, 1]$ (or $[a, b]$ in general)
  - (Economics) Health related quality-of-life measures [Brazier et al., 2002]
  - (Political science) Voting rights index [Kubinec, 2023]
  - (Medical imaging) Percentage of tissue area in mammogram [Peplonska et al., 2012]
  - (Clinical research) Rating scale for Alzheimer's disease [Rosen et al., 1984]
- Continuous proportional data is especially common in **ecology**:
  - Percent cover measurements [Korhonen et al., 2024][van Strien et al., 2024]
  - Species diversity index [Lindholm et al., 2021][Bharti et al., 2023][Rolls et al., 2023][Qiao et al., 2025]
  - [Warton and Hui, 2011]: ≈14% of ecology papers involves non-count based proportions.

Benthic macroinvertebrate multimetric index (MMI)

- MMI $\in [0, 1]$ (healthiness of lake), high MMI: diverse lake aquatic insects community



- (Left) MMI of 950 US lakes from 2017 NLA survey [U.S. Environmental Protection Agency, 2022]
- (Right) Lake watershed covariates (X) of 50,000+ lakes across U.S. [Hill et al., 2018]
  - medium/high urban land cover ("urbanness" around the lake)
- Find association between MMI and covariates & predict MMI at unsampled lakes

MMI data description

- $\mathcal{D}^{\text{orig}}$: $n = 950$ lakes (MMI min 0, median 0.45, max 0.904), $p = 9$ covariates
  - soil erodibility, base flow index, coniferous forest cover, urban land cover, ...
- Some lakes had exceptionally low MMI compared to its surroundings:

Lake name, State	MMI	Mean MMI (10 nearest)	$\mathcal{D}^{\text{orig}}$	$\mathcal{D}^{(-1)}$	$\mathcal{D}^{(-3)}$
Brierpatch Lake, GA	0	0.534	✓		
Jones Pond, NC	0.020	0.460	✓	✓	
Ferguson Lake, AR	0.021	0.414	✓	✓	
⋮	⋮	⋮	✓	✓	✓

- $\mathcal{D}^{(-1)}, \mathcal{D}^{(-3)}$: $n = 949, 947$ (removed 1 lake with 0 MMI and 3 lowest MMI lakes)
- Exceptionally low MMI lakes likely caused by local pollution
  - Ferguson Lake was subject to an administrative order by the Arkansas state government for illegal wastewater discharge from a nearby construction site (AFIN no. 63-03837)

Beta regression

- **Beta regression** for analyzing $Y \in (0,1)$ [Ferrari and Cribari-Neto, 2004]

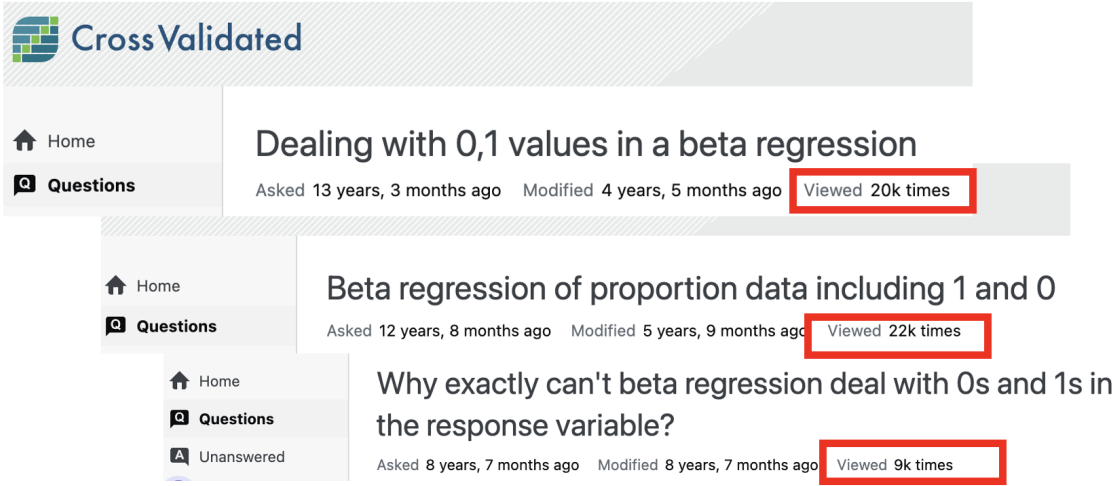
$$Y_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \text{Beta}(\text{mean} = g^{-1}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}), \text{precision} = \phi), \quad i = 1, \ldots, n$$

  with link function $g : (0,1) \to \mathbb{R}$

- Generally preferred over data transformation + linear model approach
  - Modeling $E(Y_i \mid \boldsymbol{x}_i) = g^{-1}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$, not $E\{g(Y_i) \mid \boldsymbol{x}_i\} = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$

- **Key limitations** of beta regression:
  - **Non-robustness**: sensitive to violation of the beta response assumption and outliers.
  - **Poor scalability:** hierarchical settings, models handling complex dependence structure.
  - **Boundary data issues:** cannot handle observations at 0 or 1.

Limitations of beta regression



MMI data analysis with beta regression

- **To show limitations**, we first analyze MMI data $\mathcal{D}^{(-1)}$ and $\mathcal{D}^{(-3)}$ with beta regression
  - Beta regression cannot handle $\mathcal{D}^{\text{orig}}$ due to presence of 0 MMI lake
- $Y(s_i) \in [0, 1]$: MMI at location s_i, $\boldsymbol{x}(s_i) \in \mathbb{R}^p$: covariate at location s_i
- Spatial beta regression with Gaussian process (GP) random effect $u(s_i)$:

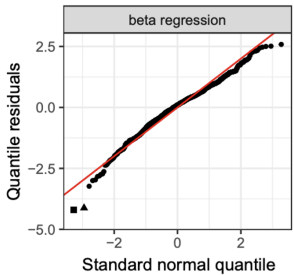
$$Y(s_i) \mid u(s_i) \overset{\text{ind}}{\sim} \text{Beta}(g^{-1}(\boldsymbol{x}(s_i)^\intercal \boldsymbol{\beta} + u(s_i)), \phi), \quad i = 1, \ldots, n$$
$$u(\cdot) \sim \text{mean zero Gaussian process.}$$

  where accounting for spatial dependence is crucial for ecological data [Guélat and Kéry, 2018]

- Used nearest neighbor GP (NNGP) [Datta et al., 2016] for spatial random effect

MMI data analysis with beta regression

- Non-Gaussian spatial model; used `Stan`, took $\approx$ 2 hours for 6000 MCMC iterations
- $\mathcal{D}^{(-1)}$: 2 significant covariates (base flow index, urban land cover) based on 95% CI
- Goodness-of-fit check, two lakes show a lack of fit based on quantile residuals [Dunn and Smyth, 1996]; turns out these two lakes have the lowest MMI



$(n = 949)$	Beta regression	
Variable	Estimate	95% CI
Intercept	-2.363	(-4.160, -0.553)
agkffact	-2.586	(-5.584, 0.330)
bfi	**0.343**	**(0.016, 0.672)**
cbnf	0.165	(-0.081, 0.412)
conif	0.081	(-0.002, 0.164)
crophay	-0.079	(-0.250, 0.091)
fert	-0.073	(-0.310, 0.158)
manure	-0.048	(-0.202, 0.102)
pestic97	-0.014	(-0.106, 0.075)
urbmdhi	**-0.181**	**(-0.288, -0.076)**

MMI data analysis with beta regression

- Re-fit the spatial beta regression model with $\mathcal{D}^{(-3)}$ ($n = 947$), took $\approx$ 2 hours
  - After removing 2 lakes out of 949 lakes, **the conclusion has been changed**
  - $\mathcal{D}^{(-3)}$: 3 significant covariates (soil erodibility, conifer forest cover, urban land cover) selected based on 95% CI. base flow index no longer significant
  - Illustrates non-robustness of beta regression model
  - Even if $n \approx 1000$ and with NNGP prior, computation is still slow, can be prohibitive for larger datasets

($n = 947$)	Beta regression	
Variable	Estimate	95% CI
(Intercept)	-1.829	(-3.621, -0.070)
agkffact	**-3.088**	**(-5.982, -0.206)**
bfi	0.244	(-0.075, 0.568)
cbnf	0.191	(-0.044, 0.430)
conif	**0.096**	**(0.014, 0.175)**
crophay	-0.057	(-0.223, 0.110)
fert	-0.096	(-0.327, 0.135)
manure	-0.001	(-0.148, 0.148)
pestic97	-0.031	(-0.118, 0.057)
urbmdhi	**-0.180**	**(-0.283, -0.076)**

Limitations of beta regression

- **Non-robustness**: sensitive to outliers and violation of beta response assumption.
  - Beta does not belong to a *natural* exponential family, strictly speaking not a GLM
- **Poor scalability**: does not scale well for hierarchical extensions.
  - Mixed models, longitudinal and spatial models: generic methods (e.g. `Stan`) may suffer
- **Boundary data issues**: data with exact 0s and 1s
  - Manipulate ("nudge") the data to lie between open interval $(0, 1)$ [Smithson and Verkuilen, 2006], but results are sensitive to the degree of preprocessing [Kosmidis and Zeileis, 2025]

Main contribution

- **Cobin regression**: continuous binomial (cobin) regression model
  - A proper GLM approach based on **exponential dispersion family** [Jørgensen, 1987]
  - Inherits attractive properties of GLM, including **robustness** of $\hat{\beta}$ to model misspecification
- **Micobin regression**: based on dispersion mixtures of cobin distributions
  - More **flexible** & **robust** family of distribution (cf. t dist as scale mixture of normal)
  - **Can handle exact 0s and 1s**, avoid the need of preprocessing
    - This is different from modeling structural 0/1s with positive prob. mass [Blasco-Moreno et al., 2019]
- We introduce **Kolmogorov-Gamma augmentation** for facilitating computation
  - Converts cobin/micobin likelihood into **conditionally normal likelihood**
  - Seamless integration with **latent Gaussian models**

Review: Generalized linear model (GLM)

- Systematic component and link: $\eta = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}$ (linear predictor), $g(E(Y)) = g\{(B')^{-1}(\theta)\} = \eta$
- Random component: **natural** exponential family [Nelder and Wedderburn, 1972, Morris, 1982] or more generally **exponential dispersion** family [Jørgensen, 1987]

$$f_Y(y;\theta,\phi) = h(y,\lambda)\exp\left[\lambda\left\{y\theta - B(\theta)\right\}\right]$$

  - θ: natural parameter, solely governs the mean of Y: $E(Y) = B'(\theta)$
  - λ: dispersion parameter, controls the variance of Y: $\mathrm{var}(Y) = \lambda^{-1}B''(\theta)$
  - This ensures that for iid data $\{Y_i\}_{i=1}^n$, sufficient statistic for θ is $\sum_{i=1}^n Y_i$
- Beta distribution does not belong to natural exponential family;

$$\mathrm{Beta}(y;\mu,\phi) = B(\mu\phi,\mu-\mu\phi)^{-1}\exp\left[(\mu\phi-1)\log(y) + (\phi-\mu\phi-1)\log(1-y)\right]$$

- For iid data $\{Y_i\}_{i=1}^n$, sufficient statistic for μ is $\sum_{i=1}^n \log(Y_i)$ and $\sum_{i=1}^n \log(1-Y_i)$

Review: Exponential dispersion family

- From one-parameter natural exponential family, exponential dispersion family [Jørgensen, 1987] adds a dispersion parameter λ^{-1} in a systematic way

	exponential tilting by θ $\longrightarrow$	λ-fold convolution and scale by λ^{-1} $\longrightarrow$
$Y \sim N(0,1)$	$Y \sim N(\theta,1)$	$Y \sim N(\theta,\lambda^{-1})$
$Y \sim \text{Exp}(1)$	$Y \sim \text{Exp}(1-\theta)$	$Y \sim \text{Gamma}(\lambda,\lambda(1-\theta))$
$Y \sim \text{InvGamma}(1/2,1/2)$	$Y \sim \text{InvGau}((-2\theta)^{-1/2},1)$	$Y \sim \text{InvGau}((-2\theta)^{-1/2},\lambda)$

Continuous binomial: derivation

- Start from $\text{Unif}(0,1)$ to get exponential dispersion family supported on unit interval:

exponential tilting by θ $\longrightarrow$	λ-fold convolution and scale by λ^{-1} $\longrightarrow$	
$Y \sim \text{Unif}(0,1)$	$Y \sim \text{cobin}(\theta, 1)$	$Y \sim \text{cobin}(\theta, \lambda^{-1})$
	(continuous Bernoulli)	(continuous binomial)

- $\text{cobin}(\theta, 1)$ is known as continuous Bernoulli [Loaiza-Ganem and Cunningham, 2019], with p.d.f.:

$$\text{cobin}(y; \theta, 1) = \theta e^{y\theta}/(e^{\theta}-1) = \exp[y\theta - \log\{(e^{\theta}-1)/\theta\}]$$
$$= \theta \coth(\theta/2) \left(\frac{1}{1+e^{-\theta}}\right)^{y} \left(1 - \frac{1}{1+e^{-\theta}}\right)^{1-y}, \quad y \in [0,1]$$

- We call $\text{cobin}(\theta, \lambda^{-1})$ continuous binomial, in short cobin.

Continuous binomial distribution

Definition 1. (cobin)

We say $Y \sim \text{cobin}(\theta, \lambda^{-1})$, natural param. $\theta \in \mathbb{R}$, dispersion $\lambda^{-1} \in \{1, 1/2, \dots\}$ if

$$p_{\text{cobin}}(y; \theta, \lambda^{-1}) = h(y, \lambda) \exp\left[\lambda\{\theta y - B(\theta)\}\right] = h(y, \lambda) \frac{e^{\lambda\theta y}}{\{(e^\theta - 1)/\theta\}^\lambda}, \quad 0 \le y \le 1$$

with $B(\theta) = \log\{(e^\theta - 1)/\theta\}$ and $h(y, \lambda) = \frac{\lambda}{(\lambda-1)!} \sum_{k=0}^{\lambda} (-1)^k \binom{\lambda}{k} \{\max(\lambda y - k, 0)\}^{\lambda-1}$

- $E(Y) = B'(\theta)$ and $\text{var}(Y) = \lambda^{-1} B''(\theta)$; λ must be an integer
- Supported on $[0, 1]$ if $\lambda = 1$ but supported on open interval $(0, 1)$ if $\lambda \ge 2$
- Relationship between continuous Bernoulli and cobin:

$$Y_1, \dots, Y_\lambda \overset{\text{iid}}{\sim} \text{conti Bernoulli}(\theta) \implies \tfrac{1}{\lambda} \textstyle\sum_{l=1}^{\lambda} Y_l \sim \text{cobin}(\theta, \lambda^{-1})$$

Continuous binomial distribution

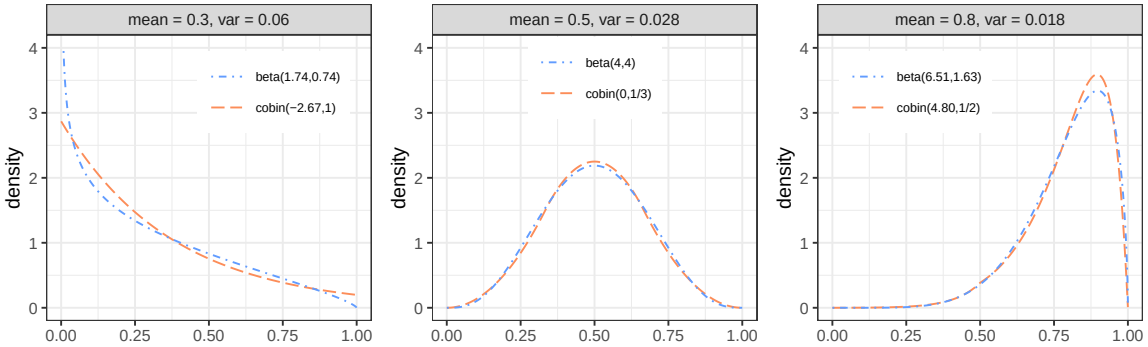


Figure: Comparison between beta and cobin with the same mean and variance

Cobin regression

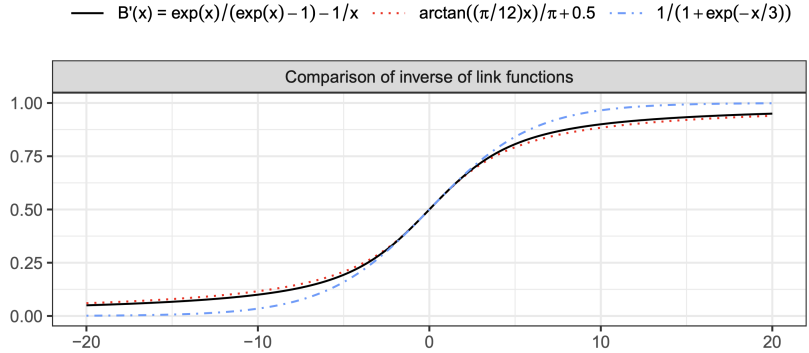
- For $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, cobin regression with link function $g : (0,1) \to \mathbb{R}$ is

$$\begin{aligned} Y_i \mid \theta_i, \lambda &\overset{\text{ind}}{\sim} \text{cobin}(\theta_i, \lambda^{-1}) \\ \theta_i &= (B')^{-1}\{g^{-1}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\}, \quad i = 1, \ldots, n, \end{aligned}$$

  which implies $E(Y_i \mid \boldsymbol{x}_i) = g^{-1}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$.

- Canonical link ("cobit"): $g^{-1}(\eta) = B'(\eta) = e^\eta/(e^\eta - 1) - 1/\eta$ so that $\theta_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$

Canonical link function



Inverse of link functions. (Black) cobin canonical link, (Red dots) Cauchit link, (Blue dashes) logit link.

- Interpretation of the results in terms of *average slope* (marginal effects)

$$\text{Average slope of } x_j = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial E(y \mid \boldsymbol{x})}{\partial x_j}\Big|_{\boldsymbol{x}=\boldsymbol{x}_i} = \frac{1}{n}\sum_{i=1}^{n} (g^{-1})'(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\beta_j.$$

Cobin regression: robustness properties

- MLE $\hat{\boldsymbol{\beta}}$ obtained by solving score equations (derivative of log-likelihood = 0)

$$\frac{\partial}{\partial \beta_j} \sum_{i=1}^{n} \log p_{\text{cobin}}(y_i; \theta_i, \lambda^{-1}) = \lambda \sum_{i=1}^{n} \frac{(y_i - \mu_i) x_{ij}}{B''(\theta_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0, \quad j = 1, \ldots, p \tag{1}$$

Proposition (consistency of cobin MLE under distributional misspecification)

As long as the mean structure $E(y_i \mid x_i) = g^{-1}(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta})$ is correctly specified, the cobin regression MLE $\hat{\boldsymbol{\beta}}$ is **consistent** [Gourieroux et al., 1984]

- Score function being a linear on y_i is a key for the above proposition to hold
- Since $y_i \in [0, 1]$, **boundary-proximate data has limited impact on log-likelihood landscape**
- cf. beta score function: $\phi \sum_{i=1}^{n} [\log \frac{y_i}{1-y_i} - \frac{\Gamma'(\mu_i \phi)}{\Gamma(\mu_i \phi)} + \frac{\Gamma'(\phi - \mu_i \phi)}{\Gamma(\phi - \mu_i \phi)}] x_{ij} \frac{\partial \mu_i}{\partial \eta_i}$. ($\Psi$: digamma ft).

Micobin: dispersion mixtures of cobin distributions

- Limitations of cobin:
  - λ must be an integer to be a valid distribution $\implies$ **limited flexibility**
  - Unless $\lambda = 1$, cobin is supported on open interval $(0, 1)$, **cannot handle exact 0s and 1s**

Definition 2. (micobin) Dispersion mixture of cobin distribution

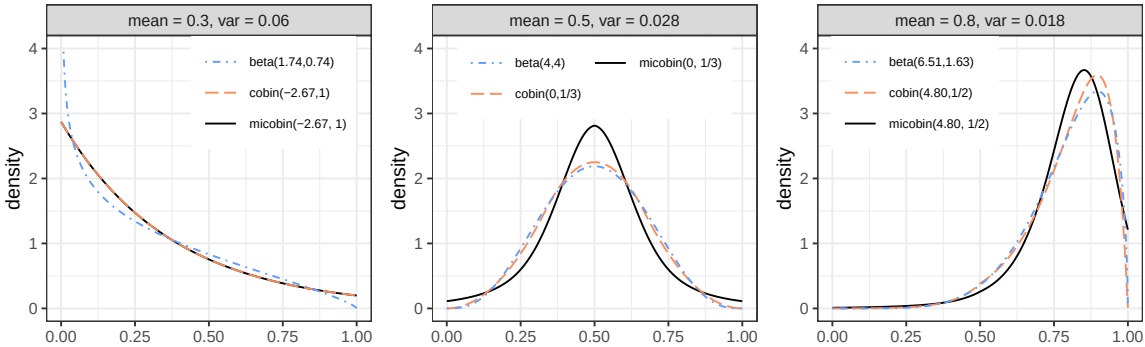
We say $Y \sim \text{micobin}(\theta, \psi)$, natural param. $\theta \in \mathbb{R}$, dispersion $\psi \in (0, 1)$ if

$$Y \mid \lambda \sim \text{cobin}(\theta, \lambda^{-1}), \quad (\lambda - 1) \sim \text{negbin}(2, \psi)$$

- Mean structure preserved $E(Y) = B'(\theta)$
- $(\lambda - 1) \sim \text{negbin}(2, \psi)$ leads to $\text{var}(Y) = \psi B''(\theta)$. (cf. $\text{var}(Y) = \lambda^{-1} B''(\theta)$ for cobin)
- Localizing dispersion ($\lambda \to \lambda_i$), general approach for robustification [Wang and Blei, 2018]

Micobin: dispersion mixtures of cobin distributions

Comparison of beta, cobin, and micobin with the same mean and variance.



Support of micobin is a closed interval $[0, 1]$.

Hierarchical extensions and latent Gaussian models

- GLM assumes independent data: $y_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} f_Y(y; \mu_i, \phi)$, $g(\mu_i) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$, $i = 1, \ldots, n$
- Independence is often unrealistic; various hierarchical extensions:
  - **Grouped data:** random intercept/slope models

$$g(\mu_{ij}) = \boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + u_i, \quad u_i \overset{\text{iid}}{\sim} N(0, \sigma_u^2)$$

  - **Temporal data:** state-space or dynamic models

$$g(\mu_t) = \boldsymbol{x}_t^{\mathrm{T}}\boldsymbol{\beta}_t, \quad \boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t-1} \sim N(G_t\boldsymbol{\beta}_{t-1}, W_t)$$

  - **Spatial data:** spatial generalized linear mixed models

$$g(\mu_i) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + u(s_i), \quad u(\cdot) \sim \text{mean zero Gaussian process}$$

- Share common framework: **latent Gaussian structure** capturing dependence

Hierarchical extension of cobin and micobin regression

- MMI data example: If two lakes at locations s and s' are close to each other, then $E(y(s) \mid \boldsymbol{x}(s)) \approx E(y(s') \mid \boldsymbol{x}(s'))$, adjusting for covariate effect
- **Spatial cobin model**: with canonical link function,

$$\begin{aligned} Y(s_i) \mid u(s_i) &\sim \text{cobin}(\boldsymbol{x}(s_i)^{\mathrm{T}}\boldsymbol{\beta} + u(s_i), \lambda^{-1}) \\ u(\cdot) &\sim \text{mean zero Gaussian process.} \end{aligned}$$

- **Spatial micobin model** :

$$\begin{aligned} Y(s_i) \mid u(s_i) &\sim \text{micobin}(\boldsymbol{x}(s_i)^{\mathrm{T}}\boldsymbol{\beta} + u(s_i), \psi) \\ u(\cdot) &\sim \text{mean zero Gaussian process.} \end{aligned}$$

Hierarchical extensions and latent Gaussian models

- When Y is **non-Gaussian**, latent Gaussian model inference becomes challenging:
  - Optimization (e.g. MLE): Likelihood involves integration over latent Gaussian components $u(s_1), \ldots, u(s_n)$, where analytical results are unavailable unless Y is Gaussian
  - Posterior computation with MCMC: Involve sampling from the full conditional of $(u(s_1), \ldots, u(s_n))$; hard to sample with high-dimensional random effects unless Y is Gaussian
- Common solution: **"Gaussianization"** of the non-Gaussian response model
  - Laplace approximation: Locally *approximates* the likelihood by a Gaussian
    - Core idea behind popular GLMM tools such as `lme4::glmer`.
    - However Laplace approximation can be inaccurate or biased results in certain settings
  - Data augmentation: Introduce aux variables $\{\kappa_i\}$ that lead to conditionally Gaussian structure
    - Enables tractable optimization algorithms (e.g., finding MLE with EM algorithm).
    - Leads to conditionally Gaussian updates for latent parameters in MCMC.

Cobin likelihood

- Consider $Y_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \text{cobin}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \lambda^{-1}), i = 1, \ldots, n$
- Likelihood of $\boldsymbol{\beta}$: (denote $\eta_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$)

$$\mathcal{L}(\boldsymbol{\beta}) \propto \prod_{i=1}^{n} p_{\text{cobin}}(y_i; \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \lambda^{-1}) \propto \prod_{i=1}^{n} \frac{e^{\lambda y_i \eta_i}}{\{(e^{\eta_i} - 1)/\eta_i\}^{\lambda}}$$

- (Bayesian) Under the normal prior $\boldsymbol{\beta} \sim N(0, \Sigma_\beta)$, the posterior is

$$p(\boldsymbol{\beta} \mid \text{data}) \propto \exp\left(-\frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\Sigma_\beta^{-1}\boldsymbol{\beta}\right) \prod_{i=1}^{n} \frac{e^{\lambda y_i \eta_i}}{\{(e^{\eta_i} - 1)/\eta_i\}^{\lambda}}$$

- Likelihood term $\dfrac{e^{\lambda y_i \eta_i}}{\{(e^{\eta_i} - 1)/\eta_i\}^{\lambda}}$ is not a familiar expression in terms of $\eta_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$
- We desire **quadratic log-likelihood function of** η_i (thus quadratic loglikelihood function of $\boldsymbol{\beta}$, "Gaussianization")

Kolmogorov-Gamma augmentation

Theorem 1. (Kolmogorov-Gamma integral identity)

For any $a \in \mathbb{R}$, $b > 0$, and $\eta \in \mathbb{R}$,

$$\frac{(e^{\eta})^a}{\{(e^{\eta}-1)/\eta\}^b} = e^{(a-b/2)\eta} \int_0^{\infty} e^{-\kappa\eta^2/2} p_{\mathsf{KG}}(\kappa; b, 0) \mathrm{d}\kappa, \tag{2}$$

where $p_{\mathsf{KG}}(\kappa; b, 0)$ is the density of a $\mathrm{KG}(b, 0)$ random variable.

- We define Kolmogorov-Gamma $\kappa \sim \mathrm{KG}(b, c)$ as an infinite convolution of gammas:

$$\kappa \overset{\mathrm{d}}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{\epsilon_k}{k^2 + c^2/(4\pi^2)}, \quad \epsilon_k \overset{\mathrm{iid}}{\sim} \mathrm{Gamma}(b, 1), \quad k = 1, 2, \ldots$$

- Conditionally on κ (ignoring integral), log-likelihood becomes **quadratic** in η
- Similar to Pólya-Gamma [Polson et al., 2013], which deals with logistic models $\frac{(e^{\eta})^a}{(e^{\eta}+1)^b}$

Conditionally Gaussian likelihood

- Consider an augmented model $p_{\text{aug}}(y_i, \kappa_i \mid \eta_i)$, denoting $\eta_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$:

$$p_{\text{aug}}(y_i, \kappa_i \mid \eta_i) = h(y_i, \lambda) \exp\left(\lambda(y_i - 0.5)\eta_i - \kappa_i \eta_i^2/2\right) p_{\mathsf{KG}}(\kappa_i; \lambda, 0), \quad i = 1, \ldots, n$$

  by Theorem 1, it recovers cobin regression model upon marginalizing out κ_i, and

$$p(\kappa_i \mid \eta_i, y_i) = p_{\mathsf{KG}}(\kappa_i; \lambda, \eta_i)$$
$$p(y_i \mid \kappa_i, \eta_i) \propto N(\lambda(y_i - 0.5)\kappa_i^{-1}; \eta_i, \kappa_i^{-1}) \text{ in terms of } \eta_i$$

- **Conditional conjugacy** for normal prior models & latent Gaussian models
  - Spatial regression model, e.g. $\eta_i = \boldsymbol{x}(s_i)^{\mathrm{T}}\boldsymbol{\beta} + u(s_i)$, $u(\cdot) \sim$ Gaussian Process.
- Same strategy can be applied to micobin, replacing λ to λ_i
- Straightfoward Gibbs sampler; EM algorithm for MLE $\hat{\boldsymbol{\beta}}$ can also be derived

Kolmogorov-Gamma sampler

- We also develop **fast, exact sampler** for Kolmogorov-Gamma variable with integer b

$$\kappa \sim \mathrm{KG}(b,c) \iff \kappa \overset{\mathrm{d}}{=} \frac{1}{2\pi^2}\sum_{k=1}^{\infty}\frac{\epsilon_k}{k^2 + c^2/(4\pi^2)}, \quad \epsilon_k \overset{\mathrm{iid}}{\sim} \mathrm{Gamma}(b,1), \quad k = 1,2,\ldots$$

- Truncating infinite sum of independent gammas: slow, prone to truncation error
- We have $\mathrm{KG}(\lambda,c) \overset{\mathrm{d}}{=} \sum_{l=1}^{\lambda}\mathrm{KG}(1,c)$ for integer λ
- Exact sampling of $\mathrm{KG}(1,c)$: alternating series method [Devroye, 1986]
- Density of $\mathrm{KG}(1,c)$: $p_{\mathrm{KG}}(x;1,c) = \sum_{n=0}^{\infty}(-1)^n a_n(x;c,t)$ with cutoff $t \in (0.0234, 0.25)$,

$$a_n(x;c,t) = \begin{cases} \{\sinh(c/2)/(c/2)\}\exp(-c^2x/2)a_n^L(x), & 0 < x < t, \\ \{\sinh(c/2)/(c/2)\}\exp(-c^2x/2)a_n^R(x), & t \le x \end{cases} \tag{3}$$

- $a_n^L(x)$, $a_n^R(x)$ are appropriate monotone sequences in n

Kolmogorov-Gamma sampler: Algorithm

1. For $A^L(c,t) = \int_0^t a_0(x;c,t)\mathrm{d}x$ and $A^R(c,t) = \int_t^\infty a_0(x;c,t)\mathrm{d}x$, propose

$$X \sim \begin{cases} \mathrm{GIG}(-1.5, c^2, 1/4)1(0 < X < t) & \text{with prob. } A^L(c,t)/\{A^L(c,t) + A^R(c,t)\} \\ \mathrm{Exp}(c^2/2 + 2\pi^2)1(t \leq X) & \text{with prob. } A^R(c,t)/\{A^L(c,t) + A^R(c,t)\} \end{cases} \tag{4}$$

2. Generate $U \sim \mathrm{Unif}(0, a_0(X;c,t))$
3. Repeat until $U \leq \sum_{n=0}^m (-1)^n a_n(X;c,t)$ (odd m) or $U > \sum_{n=0}^m (-1)^n a_n(X;c,t)$ (even m)
4. Accept X if m is odd, repeat from step 1 again if m is even.

Proposition (KG sampler is fast)

Using the best cutoff point $t^* \approx 0.050239$, the expected number of outer loop & inter loop iterations are bounded above by 1.1456 and 1.1275 for any given c.

Blocked Gibbs sampler for cobin regression

- $Y_i \sim \text{cobin}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \lambda^{-1})$, $i = 1, \ldots, n$ with normal prior $\boldsymbol{\beta} \sim N(\mathbf{0}, \Sigma_\beta)$

---

1. Sample λ from $\text{pr}(\lambda = l \mid \boldsymbol{\beta}) \propto p_\lambda(l) \prod_{i=1}^n p_{\text{cobin}}(y_i; \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}, l^{-1})$, $l = 1, \ldots, L$
2. Sample κ_i from $(\kappa_i \mid \lambda, \boldsymbol{\beta}) \overset{\text{ind}}{\sim} \text{KG}(\lambda, \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$, $i = 1, \ldots, n$
3. Sample $\boldsymbol{\beta}$ from $(\boldsymbol{\beta} \mid \lambda, \boldsymbol{\kappa}) \sim N_p(\boldsymbol{m}_\beta, V_\beta)$, where

$$V_\beta^{-1} = X^{\mathrm{T}}\text{diag}(\kappa_1, \ldots, \kappa_n)X + \Sigma_\beta^{-1}, \quad \boldsymbol{m}_\beta = V_\beta X^{\mathrm{T}}(y_1\lambda - 0.5\lambda, \ldots, y_n\lambda - 0.5\lambda)^{\mathrm{T}}$$

- Some proper prior p_λ for λ and some large upper bound L of λ
- Steps 1,2 jointly updates $(\lambda, \boldsymbol{\kappa})$

Blocked Gibbs sampler for micobin regression

- $Y_i \sim \text{micobin}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \psi)$, $i = 1, \ldots, n$ with priors $\boldsymbol{\beta} \sim N(\mathbf{0}, \Sigma_\beta)$ and $\psi \sim \text{Beta}(a_\psi, b_\psi)$

---

1. Sample λ_i from $\text{pr}(\lambda_i = l \mid \boldsymbol{\beta}, \psi) \propto l(1-\psi)^{l-1} p_{\text{cobin}}(y_i; \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}, l^{-1})$, $l = 1, \ldots, L$, $i = 1, \ldots, n$
2. Sample κ_i from $(\kappa_i \mid \lambda_i, \boldsymbol{\beta}) \overset{\text{ind}}{\sim} \text{KG}(\lambda_i, \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$, $i = 1, \ldots, n$
3. Sample $\boldsymbol{\beta}$ from $(\boldsymbol{\beta} \mid \boldsymbol{\lambda}, \boldsymbol{\kappa}) \sim N_p(\boldsymbol{m}_\beta, V_\beta)$, where
$$V_\beta^{-1} = X^{\mathrm{T}}\text{diag}(\kappa_1, \ldots, \kappa_n)X + \Sigma_\beta^{-1}, \quad \boldsymbol{m}_\beta = V_\beta X^{\mathrm{T}}(y_1\lambda_1 - 0.5\lambda_1, \ldots, y_n\lambda_n - 0.5\lambda_n)^{\mathrm{T}}$$
4. Sample ψ from $(\psi \mid \boldsymbol{\lambda}) \sim \text{Beta}(a_\psi + 2n, b_\psi - n + \sum_{i=1}^n \lambda_i)$

- Steps 1,2 jointly updates $(\lambda, \boldsymbol{\kappa})$, steps 3,4 jointly updates $(\boldsymbol{\beta}, \psi)$

Posterior computation: theory

Theorem 2. (Uniform ergodicity)

The blocked Gibbs samplers for cobin and micobin regressions are uniformly ergodic. That is, there exist constants $M > 0$ and $\rho \in [0, 1)$, independent of initial state, such that $\|P^t(\boldsymbol{\Theta}^{(0)}, \cdot) - \Pi(\cdot)\|_{\text{TV}} \leq M\rho^t$ for all $t \geq 1$.

- Total variation distance geometrically decays as the Markov chain progresses
- Proof by establishing uniform minorization condition [Rosenthal, 1995, Choi and Hobert, 2013]
- Guarantees the existence of CLT for Monte Carlo averages of functions of $\boldsymbol{\beta}$

Simulation 1: Consistency of MLE under model misspecification

Proposition (consistency of cobin MLE under potential model misspecification)

As long as the mean structure $E(y_i \mid x_i) = g^{-1}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$ is correctly specified, the cobin regression MLE $\hat{\boldsymbol{\beta}}$ is **consistent** [Gourieroux et al., 1984]

- **Data generation** with 2 links $g(\mu_i) = \beta_0 + \beta_1 x_i$ (logit, cobit) and 4 different distributions (beta, cobin, beta rectangular, 3 mix beta); $n \in \{100, 400, 1600\}$, $(\beta_0^{\text{true}}, \beta_1^{\text{true}}) = (0, 1)$
  - (beta rectangular) $Y_i \sim w_i \text{beta}(\tilde{\mu}_i, \phi) + (1 - w_i)\text{unif}(0, 1)$
    - $\tilde{\mu}_i$ is chosen such that the mean of beta-uniform mixture is μ_i
  - (3 mixture beta) $Y_i \sim 0.25\text{beta}(\mu_i - \epsilon_i, \phi) + 0.5\text{beta}(\mu_i, \phi) + 0.25\text{beta}(\mu_i + \epsilon_i, \phi)$
- **Fit and compare** (1) beta regression MLE, (2) cobin regression MLE
- Correct mean (link) $E(Y_i \mid x_i) = g^{-1}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$, but distribution can be misspecified

Simulation 1: Consistency of MLE under model misspecification

			Beta		Cobin		Beta rectangular		Mixture of beta	
Link	Method	n	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
cobit	beta regression	100	**0.004**	**0.106**	-0.024	0.099	-0.061	0.153	-0.030	0.094
		400	-0.003	**0.054**	-0.030	0.057	-0.069	0.099	-0.037	0.058
		1600	0.002	**0.026**	-0.030	0.038	-0.076	0.084	-0.036	0.042
	cobin regression	100	0.005	0.113	**0.003**	0.099	**0.013**	**0.134**	**0.006**	**0.092**
		400	**-0.002**	0.056	**-0.001**	**0.049**	**0.006**	**0.070**	**-0.001**	**0.046**
		1600	0.002	0.027	**0.000**	**0.023**	**-0.001**	**0.035**	**0.001**	**0.022**
logit	beta regression	100	**0.003**	**0.084**	-0.043	0.080	-0.054	0.117	-0.041	0.074
		400	**0.000**	**0.042**	-0.047	0.059	-0.059	0.080	-0.046	0.055
		1600	0.000	**0.021**	-0.045	0.048	-0.062	0.068	-0.046	0.048
	cobin regression	100	0.015	0.101	**0.005**	**0.066**	**0.020**	**0.116**	**0.005**	**0.067**
		400	0.004	0.051	**0.000**	**0.035**	**0.007**	**0.062**	**0.000**	**0.033**
		1600	0.000	0.026	**0.001**	**0.016**	**0.001**	**0.032**	**0.001**	**0.016**

Simulation 2: Scalability and robustness under spatial models

- Goal: evaluate scalability, estimation&prediction performance in hierarchical settings
- **Data generation**: $Y(s_i) \sim w_i \text{beta}(\tilde{\mu}_i, \phi) + (1 - w_i)\text{unif}(0, 1)$, locations $s_i \sim \text{Unif}([0, 1]^2)$,

$$g_{\text{cobit}}(\mu_i) = \beta_0 + \beta_1 x(s_i) + u(s_i), \quad u(\cdot) \sim \text{mean zero Gaussian process}$$

- $(\beta_0^{\text{true}}, \beta_1^{\text{true}}) = (0, 1)$, $\rho \in \{0.1, 0.2\}$ (spatial dependence, Mátern kernel)
- **Fitted with** (1) spatial beta, (2) spatial cobin, (3) spatial micobin regression models
  - True model is beta-uniform mixture; **all models are wrong** (only mean structure correct)
- `Stan` for spatial beta; Gibbs sampler for spatial cobin/micobin; 5000 MCMC samples.

Simulation 2: Scalability and robustness under spatial models

			Inference ($\hat{\beta}_1$)		Prediction		Sampling ($\boldsymbol{\beta}$)	
ρ	Method	$(n_{\text{train}}, n_{\text{test}})$	Bias	RMSE	negtestLL	MSPE$\times 10^2$	mESS	time (min)
0.1	beta regression	(200, 50)	-0.048	0.118	-0.325	0.427	919.8	44.5
		(400, 100)	-0.052	0.089	-0.354	0.345	978.7	437.7
	cobin regression	(200, 50)	**0.005**	**0.093**	-0.340	0.388	2791.3	2.0
		(400, 100)	**0.005**	**0.067**	-0.372	0.323	3220.9	11.2
	micobin regression	(200, 50)	0.034	0.099	**-0.367**	**0.373**	1908.4	2.4
		(400, 100)	0.037	0.074	**-0.394**	**0.312**	2137.5	11.7
0.2	beta regression	(200, 50)	-0.065	0.120	-0.320	0.329	1187.2	96.3
		(400, 100)	-0.052	0.095	-0.350	0.248	808.0	933.4
	cobin regression	(200, 50)	**0.000**	**0.088**	-0.346	0.306	3366.0	2.2
		(400, 100)	**0.013**	**0.078**	-0.370	0.233	3663.9	12.1
	micobin regression	(200, 50)	0.039	0.092	**-0.373**	**0.293**	2265.3	2.2
		(400, 100)	0.050	0.091	**-0.395**	**0.226**	2575.4	12.7

Simulation 3: Robustness to outliers

- Goal: evaluate the stability of estimation&prediction in the presence of outliers
- **Data generation**: beta or cobin, cobit link $g_{\text{cobit}}(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$, $i = 1, \ldots, 500$, $(\beta_0^{\text{true}}, \beta_1^{\text{true}}, \beta_2^{\text{true}}) = (-6, 1, 0)$
- Consider datasets without and with an outlier: $y^\circ = 0.01$ or $y^\circ = 0.001$, $\boldsymbol{x}^\circ = (6, 6)$
- **Fitted with** (1) beta, (2) cobin, (3) beta rectangular (beta-uniform mixture), (4) micobin regression models
- Evaluate stability of the coefficient estimate (posterior mean) with and without an outlier $\Delta\hat{\beta}_j = |\hat{\beta}_j^\circ - \hat{\beta}_j|$, coverage probability, and stability of linear predictor $\Delta\hat{\eta}_{1:n}$

Simulation 3: Robustness to outliers

Table 3: Outlier robustness simulation results based on 100 replicates in terms of stability of parameter estimates $|\Delta\hat{\beta}_j|$, empirical coverage, and stability of linear predictor $\|\Delta\hat{\eta}_{1:n}\|_2$.

Method	Setting	$\lvert\Delta\hat{\beta}_0\rvert$	$\lvert\Delta\hat{\beta}_1\rvert$	Coverage of $\mathrm{CI}_{.95}(\beta_1)$		$\lvert\Delta\hat{\beta}_2\rvert$	Coverage of $\mathrm{CI}_{.95}(\beta_2)$		$\|\Delta\hat{\eta}_{1:n}\|_2$
				with y°	without y°		with y°	without y°	
beta regression	A1	0.095	0.061	74.0%	94.0%	0.056	69.0%	96.0%	5.948
	A2	0.145	0.091	51.0%		0.083	39.0%		8.852
cobin regression	A1	0.007	0.017	95.0%	97.0%	0.020	95.0%	95.0%	1.768
	A2	0.008	0.016	95.0%		0.020	96.0%		1.762
betarec regression	A1	0.039	0.007	93.0%	92.0%	0.001	95.0%	96.0%	0.987
	A2	0.038	0.006	93.0%		0.001	96.0%		0.964
micobin regression	A1	0.007	0.005	91.0%	90.0%	0.006	93.0%	95.0%	0.601
	A2	0.006	0.005	92.0%		0.006	93.0%		0.562

Setting A1: outlier $y^\circ = 0.01$, Setting A2: outlier $y^\circ = 0.001$

MMI data analysis

- Return to MMI data analysis; original data $n = 950$, $p = 9$
- $\mathcal{D}^{(-1)}$ ($n = 949$, removed 1 lake with 0 MMI) and $\mathcal{D}^{(-3)}$ ($n = 947$)
- Fit three different spatial models (beta, cobin, micobin) with cobin canonical link
- Prior $\boldsymbol{\beta} \sim N_p(\mathbf{0}, 100^2 I_p)$, half-Cauchy on random effect standard deviation
- `Stan` for spatial beta; Gibbs for spatial cobin/micobin; 6000 MCMC iter, 3 chains
  - Leveraging normal conjugacy via KG augmentation, partial collapsing [Van Dyk and Park, 2008] significantly improves mixing, reducing correlation between $\boldsymbol{\beta}$ and $\mathbf{u}$.
  - Took 2 hrs for spatial beta, 5 mins for cobin and micobin per chain
  - **mESS/time difference more than 20x** (multivariate effective sampling size / time)

MMI data analysis results

Results with $\mathcal{D}^{(-1)}$, $n = 949$

($n = 949$)	Beta regression		Cobin regression		Micobin regression	
Variable	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
Intercept	-2.363	(-4.160, -0.553)	-2.106	(-3.859, -0.345)	-1.797	(-3.551, -0.085)
agkffact	-2.586	(-5.584, 0.330)	**-2.888**	**(-5.714, -0.003)**	**-3.457**	**(-6.113, -0.800)**
bfi	**0.343**	**(0.016, 0.672)**	0.293	(-0.022, 0.614)	0.229	(-0.082, 0.548)
cbnf	0.165	(-0.081, 0.412)	0.182	(-0.055, 0.420)	0.191	(-0.035, 0.425)
conif	0.081	(-0.002, 0.164)	**0.093**	**(0.011, 0.176)**	**0.123**	**(0.044, 0.203)**
crophay	-0.079	(-0.250, 0.091)	-0.063	(-0.231, 0.106)	-0.054	(-0.213, 0.105)
fert	-0.073	(-0.310, 0.158)	-0.092	(-0.323, 0.132)	-0.082	(-0.300, 0.138)
manure	-0.048	(-0.202, 0.102)	-0.036	(-0.182, 0.115)	-0.029	(-0.173, 0.118)
pestic97	-0.014	(-0.106, 0.075)	-0.021	(-0.108, 0.067)	-0.025	(-0.108, 0.059)
urbmdhi	**-0.181**	**(-0.288, -0.076)**	**-0.170**	**(-0.273, -0.067)**	**-0.142**	**(-0.243, -0.041)**

- PSIS-LOO: -1084.4 (beta), -1095.5 (cobin), **-1115.4 (micobin)**
- WAIC: -1093.4 (beta), -1103.5 (cobin), **-1119.3 (micobin)**
- Selected variables based on 95% CI are different for beta

MMI data analysis results: robustness

Re-ran the analysis with $\mathcal{D}^{(-3)}$, beta regression result changed

$(n = 947)$	Beta regression		Cobin regression		Micobin regression	
Variable	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
(Intercept)	-1.829	(-3.621, -0.070)	-1.756	(-3.531, 0.006)	-1.741	(-3.520, 0.018)
agkffact	**-3.088**	**(-5.982, -0.206)**	**-3.150**	**(-6.019, -0.258)**	**-3.494**	**(-6.220, -0.822)**
bfi	0.244	(-0.075, 0.568)	0.228	(-0.088, 0.552)	0.219	(-0.097, 0.540)
cbnf	0.191	(-0.044, 0.430)	0.200	(-0.035, 0.437)	0.196	(–0.034, 0.424)
conif	**0.096**	**(0.014, 0.175)**	**0.103**	**(0.021, 0.183)**	**0.125**	**(0.045, 0.204)**
crophay	-0.057	(-0.223, 0.110)	-0.053	(-0.218, 0.114)	-0.050	(-0.210, 0.110)
fert	-0.096	(-0.327, 0.135)	-0.104	(-0.329, 0.122)	-0.089	(-0.316, 0.135)
manure	-0.001	(-0.148, 0.148)	-0.009	(-0.157, 0.138)	-0.022	(-0.167, 0.122)
pestic97	-0.031	(-0.118, 0.057)	-0.030	(-0.119, 0.057)	-0.027	(-0.110, 0.055)
urbmdhi	**-0.180**	**(-0.283, -0.076)**	**-0.169**	**(-0.275, -0.064)**	**-0.143**	**(-0.242, -0.043)**
Change	$\|\|\hat{\boldsymbol{\beta}}^{(949)} - \hat{\boldsymbol{\beta}}^{(947)}\|\|_2 = 0.743$		$\|\|\hat{\boldsymbol{\beta}}^{(949)} - \hat{\boldsymbol{\beta}}^{(947)}\|\|_2 = 0.444$		$\|\|\hat{\boldsymbol{\beta}}^{(949)} - \hat{\boldsymbol{\beta}}^{(947)}\|\|_2 = 0.069$	

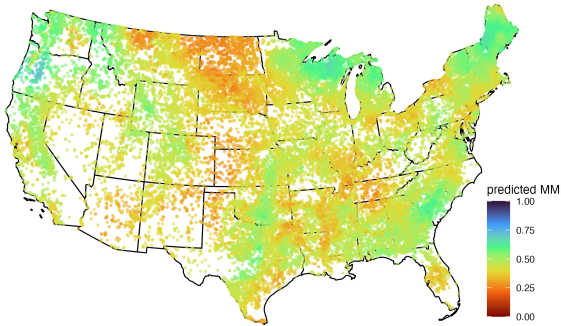
- PSIS-LOO: -1125.2 (beta), **-1134.7 (cobin)**, -1115.4 (micobin)
- WAIC: -1134.7 (beta), **-1142.4 (cobin)**, -1131.1 (micobin)
- Cobin performs best; prediction in the presence of low-MMI lakes is also important (micobin)

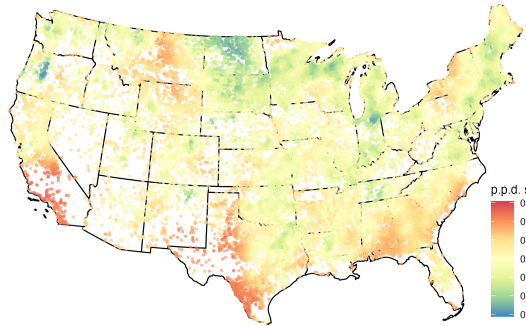
MMI data analysis results: robustness

- Recall that micobin can handle boundary data (0 or 1)
- (Left) Results with $\mathcal{D}^{\text{orig}}$, (Right) Results with $\mathcal{D}^{(-1)}$
- Result almost unchanged under micobin regression

	Micobin regression ($n = 950$)		Micobin regression ($n = 949$)	
Variable	Estimate	95% CI	Estimate	95% CI
(Intercept)	-1.758	(-3.517, -0.04)	-1.797	(-3.551, -0.085)
agkffact	**-3.456**	**(-6.175, -0.794)**	**-3.457**	**(-6.113, -0.800)**
bfi	0.219	(-0.100, 0.537)	0.229	(-0.082, 0.548)
cbnf	0.187	(-0.040, 0.415)	0.191	(-0.035, 0.425)
conif	**0.128**	**(0.048, 0.208)**	**0.123**	**(0.044, 0.203)**
crophay	-0.060	(-0.222, 0.101)	-0.054	(-0.213, 0.105)
fert	-0.071	(-0.296, 0.13)	-0.082	(-0.300, 0.138)
manure	-0.031	(-0.178, 0.115)	-0.029	(-0.173, 0.118)
pestic97	-0.023	(-0.106, 0.059)	-0.025	(-0.108, 0.059)
urbmdhi	**-0.141**	**(-0.243, -0.038)**	**-0.142**	**(-0.243, -0.041)**

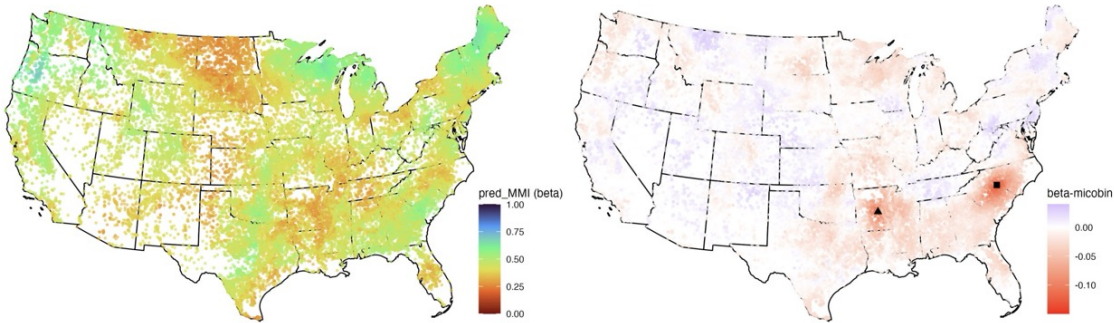
Micobin prediction with $\mathcal{D}^{(-1)}$





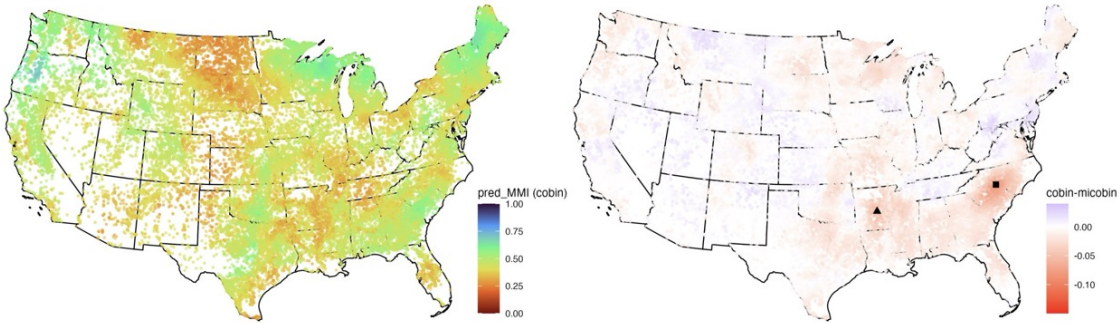
(Left) Predicted MMI from the micobin model. (Right) stdev of predicted MMI

Beta prediction with $\mathcal{D}^{(-1)}$, beta vs micobin comparison



(Left) Predicted MMI from the beta model. (Right) Prediction contrast (beta-micobin)
Square and triangle: two lakes with very low MMI

Cobin prediction with $\mathcal{D}^{(-1)}$, cobin vs micobin comparison



(Left) Predicted MMI from the cobin model. (Right) Prediction contrast (cobin - micobin)
Square and triangle: two lakes with very low MMI

Summary

Overcoming limitations of beta regression models:

- Non-robustness: sensitive to outliers and violation of beta response assumption.
  - **Continuous binomial (cobin) regression** model based on exponential dispersion family, which gives consistent MLE $\hat{\beta}$ even under distributional misspecification.
- Poor scalability: slow for large-scale hierarchical models with latent Gaussian components.
  - **Novel data augmentation scheme** that converts non-Gaussian likelihood into conditionally Gaussian likelihood, facilitating optimization / posterior computation with MCMC
- Boundary data and interpretability: data with exact 0s and 1s must be adjusted, and zero/one-inflated models change interpretation of the model.
  - **Mixture of cobin regression (micobin)** that directly accommodates boundary data without modification, preserving the model structure $g\{E(Y_i \mid x_i)\} = x_i^{\mathrm{T}}\beta$
- R package “cobin” at CRAN: `install.packages("cobin")`

- Lee, C. J., Dahl, B. K., Ovaskainen, O., & Dunson, D. B. (2025). Scalable and robust regression models for continuous proportional data. *arXiv preprint arXiv:2504.15269. (Revision submitted to JASA Theory&Methods)*

Thank you!

Q & A [1]

[1] This research was partially supported by the National Institutes of Health (grant ID R01ES035625), by the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement No 856506), by the National Science Foundation (NSF IIS-2426762), and by the Office of Naval Research (N00014-24-1-2626).

References I

Bharti, D. K., Pawar, P. Y., Edgecombe, G. D., and Joshi, J. (2023).
Genetic diversity varies with species traits and latitude in predatory soil arthropods (Myriapoda: Chilopoda).
Glob. Ecol. Biogeogr., 32(9):1508–1521.

Blasco-Moreno, A., Pérez-Casany, M., Puig, P., Morante, M., and Castells, E. (2019).
What does a zero mean? Understanding false, random and structural zeros in ecology.
Methods Ecol. Evol., 10(7):949–959.

Brazier, J., Roberts, J., and Deverill, M. (2002).
The estimation of a preference-based measure of health from the SF-36.
J. Health Econ., 21(2):271–292.

Choi, H. M. and Hobert, J. P. (2013).
The Polya-Gamma Gibbs sampler for Bayesian logistic regression is uniformly ergodic.
Electron. J. Stat., 7:2054–2064.

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016).
Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets.
J. Am. Stat. Assoc., 111(514):800–812.

References II

Devroye, L. (1986).
Non-Uniform Random Variate Generation.
Springer New York.

Duan, L., Johndrow, J., and Dunson, D. (2018).
Scaling up data augmentation MCMC via calibration.
J. Mach. Learn. Res., 19(64):64:1–64:34.

Dunn, P. K. and Smyth, G. K. (1996).
Randomized quantile residuals.
J. Comput. Graph. Stat., 5(3):236.

Ferrari, S. and Cribari-Neto, F. (2004).
Beta regression for modelling rates and proportions.
J. Appl. Stat., 31(7):799–815.

Gourieroux, C., Monfort, A., and Trognon, A. (1984).
Pseudo maximum likelihood methods: Theory.
Econometrica, 52(3):681.

References III

Guélat, J. and Kéry, M. (2018).
Effects of spatial autocorrelation and imperfect detection on species distribution models.
Methods Ecol. Evol., 9(6):1614–1625.

Hill, R. A., Weber, M. H., Debbout, R. M., Leibowitz, S. G., and Olsen, A. R. (2018).
The Lake-Catchment (LakeCat) Dataset: characterizing landscape features for lake basins within the conterminous USA.
Freshw. Sci., 37:208–221.

Johndrow, J. E., Smith, A., Pillai, N., and Dunson, D. B. (2019).
MCMC for imbalanced categorical data.
J. Am. Stat. Assoc., 114(527):1394–1403.

Jørgensen, B. (1987).
Exponential dispersion models.
J. R. Statist. Soc. B, 49(2):127–145.

Korhonen, P., Hui, F. K. C., Niku, J., Taskinen, S., and van der Veen, B. (2024).
A comparison of joint species distribution models for percent cover data.
Methods Ecol. Evol., 15(12):2359–2372.

References IV

Kosmidis, I. and Zeileis, A. (2025).
Extended-support beta regression for [0, 1] responses.
J. R. Statist. Soc. C, (In press).

Kubinec, R. (2023).
Ordered beta regression: A parsimonious, well-fitting model for continuous data with lower and upper bounds.
Polit. Anal., 31(4):519–536.

Lindholm, M., Alahuhta, J., Heino, J., and Toivonen, H. (2021).
Temporal beta diversity of lake plants is determined by concomitant changes in environmental factors across decades.
J. Ecol., 109(2):819–832.

Loaiza-Ganem, G. and Cunningham, J. (2019).
The continuous Bernoulli: fixing a pervasive error in variational autoencoders.
Adv. Neural Inf. Process. Syst., 32:13287–13297.

Morris, C. N. (1982).
Natural exponential families with quadratic variance functions.
Ann. Stat., 10(1):65–80.

References V

Nelder, J. A. and Wedderburn, R. W. M. (1972).
Generalized linear models.
J. R. Statist. Soc. A, 135(3):370.

Peplonska, B., Bukowska, A., Sobala, W., Reszka, E., Gromadzinska, J., Wasowicz, W., Lie, J. A., Kjuus, H., and Ursin, G. (2012).
Rotating night shift work and mammographic density.
Cancer Epidemiol. Biomarkers Prev., 21(7):1028–1037.

Polson, N. G., Scott, J. G., and Windle, J. (2013).
Bayesian inference for logistic models using Pólya–Gamma latent variables.
J. Am. Statist. Assoc., 108(504):1339–1349.

Qiao, J., Chu, L., Li, Y., Chu, T., Xie, N., and Yan, Y. (2025).
Unraveling spatial patterns and drivers of fish ecological uniqueness in subtropical streams.
Ecol. Evol., 15(4):e71112.

Rolls, R. J., Wolfenden, B., Heino, J., Butler, G. L., and Thiem, J. D. (2023).
Scale dependency in fish beta diversity–hydrology linkages in lowland rivers.
J. Biogeogr., 50(10):1692–1709.

References VI


Rosen, W. G., Mohs, R. C., and Davis, K. L. (1984).
A new rating scale for Alzheimer's disease.
Am. J. Psychiatry, 141(11):1356–1364.


Rosenthal, J. S. (1995).
Minorization conditions and convergence rates for Markov chain Monte Carlo.
J. Am. Statist. Assoc., 90(430):558.


Smithson, M. and Verkuilen, J. (2006).
A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables.
Psychol. Methods, 11(1):54–71.


U.S. Environmental Protection Agency (2022).
National lakes assessment 2017: Technical support document. EPA 841-R-22-001.


Van Dyk, D. A. and Park, T. (2008).
Partially collapsed Gibbs samplers: Theory and methods.
J. Am. Statist. Assoc., 103(482):790–796.

References VII


van Strien, A. J., Irvine, K. M., and Retel, C. (2024).
Trends in plant cover derived from vegetation plot data using ordinal zero-augmented beta regression.
J. Veg. Sci., 35(4).


Wang, C. and Blei, D. M. (2018).
A general method for robust Bayesian modeling.
Bayesian Anal., 13(4):1163–1191.


Warton, D. I. and Hui, F. K. C. (2011).
The arcsine is asinine: the analysis of proportions in ecology.
Ecology, 92(1):3–10.


Zens, G., Frühwirth-Schnatter, S., and Wagner, H. (2024).
Ultimate Pólya-Gamma samplers–efficient MCMC for possibly imbalanced binary and categorical data.
J. Am. Statist. Assoc, 119(548):2548–2559.

Proof outline of Theorem 1

Theorem 1. (Kolmogorov-Gamma integral identity)

For any $b > 0$, and $\eta \in \mathbb{R}$,

$$\left\{ \frac{\eta/2}{\sinh(\eta/2)} \right\}^b = \int_0^\infty e^{-\kappa\eta^2/2} p_{\mathrm{KG}}(\kappa; b, 0) \mathrm{d}\kappa, \tag{5}$$

where $p_{\mathrm{KG}}(\kappa; b, 0)$ is the density of a $\mathrm{KG}(b, 0)$ random variable.

- Expand LHS with Weierstrass factorization theorem
- Recognize factor terms are Laplace transformation of gamma with decaying scale
- Define Kolmogorov-Gamma as an infinite sum of gamma appropriately

$$\left[\frac{(t/2)^{1/2}}{\sinh\{(t/2)^{1/2}\}} \right]^b = \prod_{k=1}^\infty \left(1 + \frac{t}{2\pi^2 k^2}\right)^{-b} = \prod_{k=1}^\infty E\left\{ \exp\left(-\frac{\epsilon_k t}{2\pi^2 k^2} \right) \right\} = E\{\exp(-\kappa t)\}$$

Future directions on MCMC convergence

Theorem 2. (Uniform ergodicity)

The blocked Gibbs samplers for cobin and micobin regressions are uniformly ergodic. That is, there exist constants $M > 0$ and $\rho \in [0, 1)$, independent of initial state, such that $\|P^t(\boldsymbol{\Theta}^{(0)}, \cdot) - \Pi(\cdot)\|_{\text{TV}} \leq M\rho^t$ for all $t \geq 1$.

- The rate ρ depends on n, p and deteriorates as $n, p \to \infty$
- Recent works[2] [3] establish a tighter bound for similar algorithms that does not deteriorate as $n, p \to \infty$
- Better understanding of when the algorithm works well and when do not

[2] Lee, H., & Zhang, K. (2024). Fast mixing of data augmentation algorithms: Bayesian probit, logit, and lasso regression. arXiv preprint arXiv:2412.07999.

[3] Ascolani, F., & Zanella, G. (2025). Mixing times of data-augmentation Gibbs samplers for high-dimensional probit regression. arXiv preprint arXiv:2505.14343.

Future directions on MCMC convergence

- Performance of Polya-Gamma augmentation deteriorates for imbalanced binary data [Johndrow et al., 2019]
- Similar issue exists for Kolmogorov-Gamma under increasingly boundary-proximate data
- Calibration [Duan et al., 2018] or parameter expansion technique [Zens et al., 2024] could be potentially useful

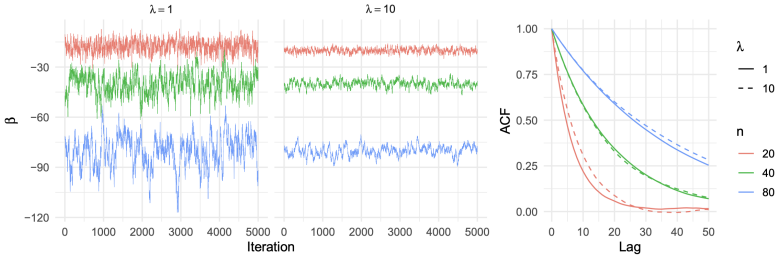


Figure S.10: (Left) Examples of MCMC trace plots (only first 5,000 iterations shown) of β for $p(\beta \mid y_{1:n})$ in (S.21) with Kolmogorov-Gamma augmentation (Right) Autocorrelation function plots with different n and λ settings.

Distinction with structural zeros

- Micobin has a *positive density* at boundaries but $\mathbb{P}(Y_i = 0) = \mathbb{P}(Y_i = 1) = 0$
- Micobin is appropriate for boundary values arising from, e.g. rounding of data during preprocessing, measurement precision limits.
- Different from structural zeros with positive probability mass (zero/one inflation), which assumes boundary values are qualitatively different from data in $(0, 1)$
- Micobin is inappropriate if data exhibit clear zero/one inflation
- However micobin is directly modeling $E(Y_i \mid \boldsymbol{x}_i)$ not $E(Y_i \mid \boldsymbol{x}_i, Y_i \in (0, 1))$, parsimonious & interpretable

Extension to simplex-valued data

Definition. (Continuous categorical) (Gordon-Rodriguez et al. 2020)

Let $\boldsymbol{y} \in \Delta^{d-1}$. We say $\boldsymbol{y}$ follows continuous categorical with natural parameter $\boldsymbol{\theta} \in \mathbb{R}^{d-1}$ if

$$p(y_1, \ldots, y_{d-1}; \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp(\sum_{j=1}^{d-1} y_j \theta_j) = \frac{\exp(\sum_{j=1}^{d-1} y_j \theta_j)}{(-1)^{d+1} \sum_{j=1}^{d} e^{\theta_j} / \{\prod_{l \neq k} (\theta_l - \theta_k)\}}$$

- Simplex-valued data (compositional data) are also common in ecology & microbiome studies
- It is nontrivial to apply data augmentation to continuous categorical distribution due to complicated $Z(\boldsymbol{\theta})$
- Other formulation, such as normalized cobin distributions, may be potentially useful.

Additional simulation

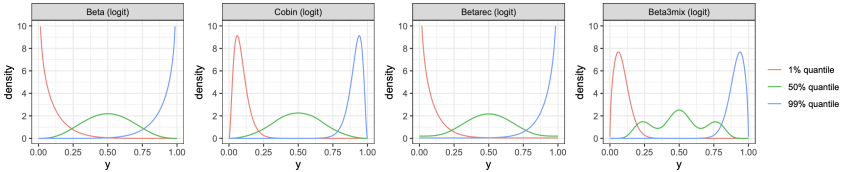
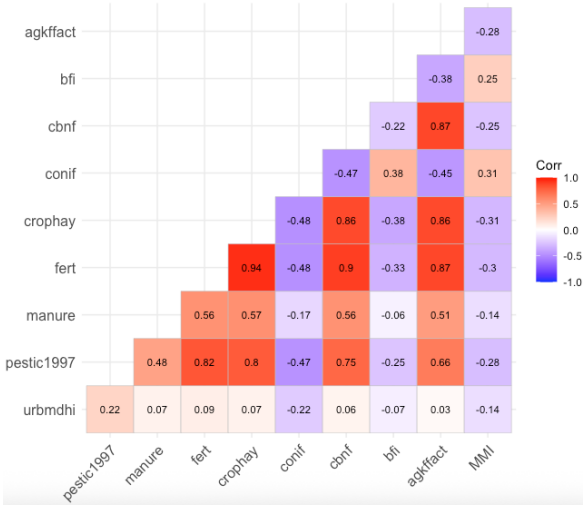


Table S.3: Predictive performance of beta regression, cobin regression, and a data transformation approach under misspecified mean settings. Results are based on 500 simulation replicates.

			Beta data		Cobin data		Betarec data		Beta3mix data	
	Method	n	NTLL	MSPE†	NTLL	MSPE†	NTLL	MSPE†	NTLL	MSPE†
Data from logit link	beta regression, cobit	100	-0.552	0.059	-0.599	0.054	-0.409	0.080	-0.609	0.052
		400	-0.557	0.028	-0.608	0.026	-0.422	0.035	-0.620	0.025
		1600	-0.561	0.020	-0.608	0.018	-0.426	0.023	-0.628	0.018
	cobin regression, cobit	100	-0.485	0.065	-0.612	0.058	-0.397	0.081	-0.618	0.056
		400	-0.488	0.031	-0.622	0.029	-0.408	0.037	-0.632	0.028
		1600	-0.497	0.023	-0.622	0.023	-0.409	0.024	-0.636	0.022
	cobit transform, linear regression with t_3 error	100	-0.233	0.125	-0.488	0.124	-0.186	0.216	-0.491	0.099
		400	-0.236	0.069	-0.502	0.084	-0.191	0.152	-0.503	0.055
		1600	-0.247	0.059	-0.502	0.078	-0.196	0.137	-0.506	0.047

Additional MMI results



Additional MMI results

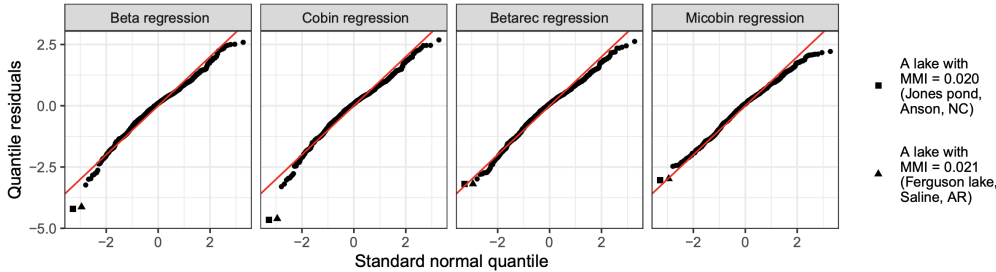


Figure S.7: Comparison of quantile residuals for goodness-of-fit assessment, along with two observations corresponding to the lowest quantiles. The red line corresponds to the $y = x$ line.

Additional MMI results

Table S.7: Estimated fixed-effect coefficients ($\hat{\boldsymbol{\beta}}^{\star}$) from two spatial regression models (beta rectangular, micobin) fitted on $\mathcal{D}^{\star}(n = 950)$ and 95% credible intervals. Bold entries indicate non-intercept coefficients whose 95% credible intervals exclude zero.

	Betarec regression ($\|\|\hat{\boldsymbol{\beta}}^{\star} - \hat{\boldsymbol{\beta}}\|\|_2 = 0.103$)		Micobin regression ($\|\|\hat{\boldsymbol{\beta}}^{\star} - \hat{\boldsymbol{\beta}}\|\|_2 = 0.048$)	
Variable	$\hat{\beta}_j^{\star}$ $(n = 950)$	$\hat{\beta}_j$ $(n = 947)$	$\hat{\beta}_j^{\star}$ $(n = 950)$	$\hat{\beta}_j$ $(n = 947)$
(Intercept)	-1.596 (-3.385, 0.156)	-1.696 (-3.543, 0.092)	-1.758 (-3.517, -0.040)	-1.741 (-3.520, 0.018)
agkffact	**-3.072** (-5.916, -0.288)	**-3.082** (-5.961, -0.200)	**-3.456** (-6.175, -0.794)	**-3.494** (-6.220, -0.822)
bfi	0.202 (-0.113, 0.530)	0.220 (-0.101, 0.551)	0.219 (-0.100, 0.537)	0.219 (-0.097, 0.540)
cbnf	0.200 (-0.033, 0.437)	0.200 (-0.036, 0.437)	0.187 (-0.040, 0.415)	0.197 (-0.034, 0.424)
conif	**0.107** (0.026, 0.186)	**0.103** (0.023, 0.184)	**0.128** (0.048, 0.208)	**0.125** (0.045, 0.204)
crophay	-0.036 (-0.202, 0.132)	-0.042 (-0.209, 0.124)	-0.060 (-0.222, 0.101)	-0.050 (-0.210, 0.110)
fert	-0.127 (-0.357, 0.100)	-0.120 (-0.348, 0.105)	-0.071 (-0.296, 0.130)	-0.089 (-0.316, 0.135)
manure	-0.018 (-0.166, 0.130)	-0.013 (-0.165, 0.137)	-0.031 (-0.178, 0.115)	-0.022 (-0.167, 0.122)
pestic97	-0.034 (-0.121, 0.053)	-0.034 (-0.121, 0.054)	-0.023 (-0.106, 0.059)	-0.027 (-0.110, 0.055)
urbmdhi	**-0.162** (-0.266, -0.058)	**-0.172** (-0.278, -0.067)	**-0.141** (-0.243, -0.038)	**-0.143** (-0.242, -0.043)
mESS($\boldsymbol{\beta}$)/t	3711.9/117 mins	3211.1/96 mins	2995.2/5 mins	3012.6/5 mins
PSIS-LOO	-1101.6($\mathcal{D}^{\star}$)	-1124.8($\mathcal{D}$)	-1107.2($\mathcal{D}^{\star}$)	-1126.9($\mathcal{D}$)
WAIC	-1106.4($\mathcal{D}^{\star}$)	-1131.3($\mathcal{D}$)	-1111.0($\mathcal{D}^{\star}$)	-1131.1($\mathcal{D}$)