Scalable and robust regression models for continuous proportional data

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

July 2025 EAC-ISBA conference
Joint work with

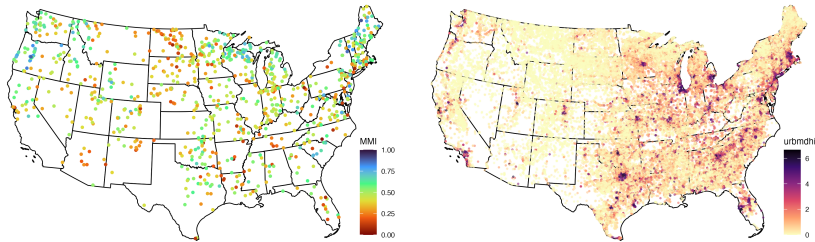
Benjamin Dahl
(Duke U.)

Otso Ovaskainen
(U. Jyväskylä)

David Dunson
(Duke U.)

Benthic macroinvertebrate multimetric index (MMI) analysis

- Regression model for continuous, bounded response $Y \in [0, 1]$



- MMI $\in [0, 1]$ (healthiness of lake), high MMI: diverse lake aquatic insects community
- (Left): MMI of 949 lakes from 2017 NLA survey [U.S. Environmental Protection Agency, 2022]
- (Right): Lake watershed covariate of 50k+ lakes from LakeCat data [Hill et al., 2018]
  - medium/high urban land cover, soil erodibility factor, coniferous forest cover, ...

Beta regression

- **Beta regression** model [Ferrari and Cribari-Neto, 2004, Cribari-Neto and Zeileis, 2010]

$$Y_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \text{Beta}(\mu_i, \phi), \quad \mu_i = E(Y_i \mid \boldsymbol{x}_i) = g^{-1}(\boldsymbol{x}_i^{\text{T}}\boldsymbol{\beta})$$

  - μ, ϕ: Mean & precision parameterization of beta, $E(Y) = \mu$, $\text{var}(Y) = \mu(1-\mu)/(1+\phi)$
- Popularly used for modeling continuous proportional data
  - (Medical imaging) Percentage of tissue area in mammogram [Peplonska et al., 2012]
  - (Econometrics) Central-bank independence index [Berggren et al., 2014]
  - (Political science) Voting rights index [Kubinec, 2023]
  - (Ecology) Percent cover measurements, Diversity indices
    [de Vargas Ribeiro et al., 2022][Korhonen et al., 2024], [Lindholm et al., 2021][Bharti et al., 2023][Rolls et al., 2023]

MMI data analysis with beta regression

- Accounting for spatial dependence is crucial for ecological data [Guélat and Kéry, 2018]
- Fitted beta mixed effects model with spatial random effect $u(s_i)$:

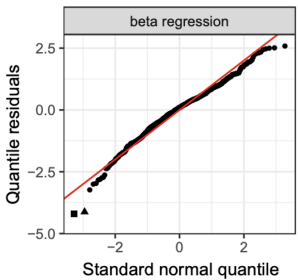
$$Y(s_i) \mid u(s_i) \overset{\text{ind}}{\sim} \text{Beta}(g^{-1}(\boldsymbol{x}(s_i)^\top \boldsymbol{\beta} + u(s_i)), \phi), \quad i = 1, \ldots, n$$
$$u(\cdot) \sim \text{mean zero Gaussian process.}$$

  where $Y(s_i) \in [0, 1]$: MMI at location s_i, $\boldsymbol{x}(s_i) \in \mathbb{R}^p$: covariate at location s_i

- Used nearest neighbor GP (NNGP) prior [Datta et al., 2016] for spatial random effect
- $n = 949$ (removed one lake with 0 MMI), $p = 9$ lake watershed covariates:
  - agkffact (soil erodibility), bfi (base flow index), conif (coniferous forest cover)
  - cbnf (cultivated N fixation), crophay (crop&hay land cover), fert (synthetic N fertilizer use), manure (manure application)
  - urbmdhi (medium/high density urban land cover), pestic97 (1997 pesticide use)

MMI data analysis with beta regression

- Non-Gaussian spatial model; used `Stan` to fit, took 2 hours for 6000 MCMC iterations
- 2 significant covariates (bfi, urbmdhi) based on 95% credible interval
- Assess goodness of fit with quantile residuals [Dunn and Smyth, 1996]
  - Two observations show a lack of fit



$(n = 949)$	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 beta regression model with **2 observations removed** ($n = 947$)
- After removing only 2 observations, **the conclusion has been changed**
- 3 significant covariates (agkffact, conif, urbmdhi) based on 95% credible interval, bfi no longer significant
- Illustrates non-robustness of beta regression model
- Even if $n \approx 1000$ and with NNGP prior, computation is still slow (2 hrs for 6000 iterations), 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 violation of beta response assumption.
  - Beta does not belong to a *natural* exponential family, does not belong to GLM
  - μ and ϕ not orthogonal to each other [Ferrari and Cribari-Neto, 2004]
- **Computational challenges**: Bayesian inference in hierarchical settings.
  - Mixed models, longitudinal and spatial models: generic methods (e.g. `Stan`) may suffer
  - Existing scalable methods cannot be easily applied, unless relying on approximation
- **Cannot handle boundary values**: data with exact 0s and 1s
  - Preprocess ("nudge") the data to lie between open interval $(0, 1)$ [Smithson and Verkuilen, 2006]
  - Results are often sensitive to the degree of preprocessing [Kosmidis and Zeileis, 2024]

Contribution

- **Cobin regression**: continuous binomial (cobin) regression model
  - A proper GLM approach based on **exponential dispersion model** [Jørgensen, 1987]
  - Inherits attractive properties of GLM, including **robustness** of $\hat{\beta}$ to model misspecificaton
- **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 Bayesian computation
  - Converts cobin/micobin likelihood into **conditionally normal likelihood**
  - Seamless integration with **latent Gaussian models**

Exponential dispersion model: review

- From 1-param. natural exponential family, exponential dispersion model [Jørgensen, 1987] offers a principled way to add dispersion parameter λ^{-1} that controls variance
- Normal belongs to the exponential dispersion family:

	exponential tilting by θ $\longrightarrow$	λ-fold convolution and scale by λ^{-1} $\longrightarrow$
$Y \sim N(0,1)$	$Y \sim N(\theta,1)$	$Y \sim N(\theta,\lambda^{-1})$

Exponential dispersion model: review

- From 1-param. natural exponential family, exponential dispersion model [Jørgensen, 1987] offers a principled way to add dispersion parameter λ^{-1} that controls variance
- Gamma belongs to the exponential dispersion family:

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

Exponential dispersion model; review

- From 1-param. natural exponential family, exponential dispersion model [Jørgensen, 1987] offers a principled way to add dispersion parameter λ^{-1} that controls variance
- Inverse Gaussian belongs to the exponential dispersion family:

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

Cobin as an exponential dispersion model

- Exponential dispersion family derived from uniform distribution:

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

- $\text{cobin}(\theta, 1)$ corresponds to continuous Bernoulli [Loaiza-Ganem and Cunningham, 2019]
- We call $\text{cobin}(\theta, \lambda^{-1})$ continuous binomial (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 $\lambda = 1$ (having uniform and truncated exponential as special cases) but supported on $(0, 1)$ if $\lambda \ge 2$
- Name motivated from continuous Bernoulli [Loaiza-Ganem and Cunningham, 2019] ($\lambda = 1$ case),

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

Cobin regression model

- Cobin regression with link $g : (0,1) \to \mathbb{R}$ so that $E(Y_i \mid \boldsymbol{x}_i) = g^{-1}(\eta_i) = g^{-1}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$

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

- Canonical link ("cobit"): $g^{-1}(\eta) = B'(\eta) = e^{\eta}/(e^{\eta} - 1) - 1/\eta$ so that $\theta_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$
- Score function (derivative of log-likelihood) is a linear function of $y_i \in [0,1]$

$$\frac{\partial}{\partial \beta_j} \sum_{i=1}^{n} \log p_{\text{cobin}}(y_i; \theta_i, \lambda^{-1}) = \lambda \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{B''(\theta_i)} \frac{\partial \mu_i}{\partial \eta_i}, \quad j = 1, \ldots, p \tag{2}$$

cf. beta score function: $\phi \sum_{i=1}^{n} [\log \frac{y_i}{1-y_i} - \Psi(\mu_i\phi) + \Psi(\phi - \mu_i\phi)] x_{ij} \frac{\partial \mu_i}{\partial \eta_i}$. ($\Psi$: digamma ft).

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]

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

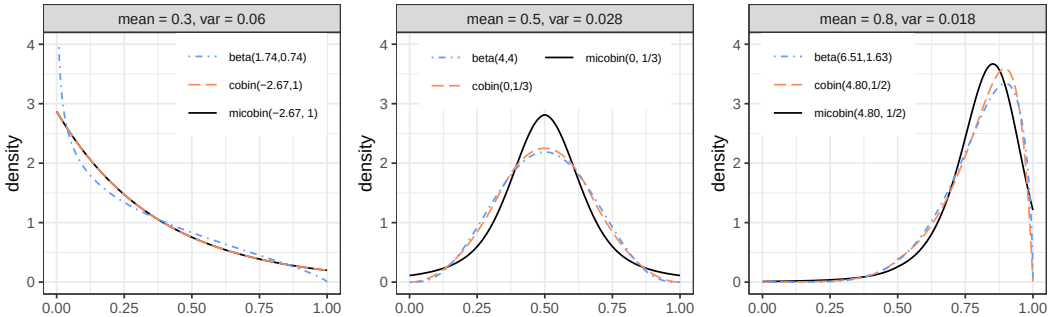
We say $Y \sim \text{micobin}(\theta, \psi)$, natural param. $\theta \in \mathbb{R}$, dispersion $\psi \in (0, 1)$ if

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

- 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)

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 $[0, 1]$ for any θ, ψ.

Micobin regression and extensions

- **Random intercept model** (with canonical link so that θ = linear predictor)

$$Y_{ij} \mid u_i \sim \text{micobin}(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + u_i, \psi)$$
$$u_i \overset{\text{iid}}{\sim} N(0, \sigma_u^2)$$

- **Spatial regression model** with spatially indexed data $(y(s_i), \boldsymbol{x}(s_i))$:

$$Y(s_i) \mid u(s_i) \sim \text{micobin}(\boldsymbol{x}(s_i)^{\mathrm{T}}\boldsymbol{\beta} + u(s_i), \psi)$$
$$u(\cdot) \sim \text{mean zero GP.}$$

- And many other mixed model extensions

Posterior computation

- Consider cobin regression with normal prior $\boldsymbol{\beta} \sim N(0, \Sigma_\beta)$
- Under the canonical link so that $\theta_i = \eta_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$, posterior of $\boldsymbol{\beta}$ is proportional to

$$p(\boldsymbol{\beta} \mid \text{data}) \propto p(\boldsymbol{\beta}) \prod_{i=1}^{n} p_{\text{cobin}}(y_i; \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \lambda^{-1}) \propto \exp(-\frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\Sigma_\beta^{-1}\boldsymbol{\beta}) \prod_{i=1}^{n} \frac{e^{\lambda y_i \eta_i}}{\{(e^{\eta_i}-1)/\eta_i\}^{\lambda}}$$

- Likelihood contribution $\dfrac{e^{\lambda y_i \eta_i}}{\{(e^{\eta_i}-1)/\eta_i\}^{\lambda}}$ is not a familiar expression in terms of $\eta_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$
- We desire a log-likelihood that is quadratic function of η_i (thus quadratic function of $\boldsymbol{\beta}$), which gives conjugacy with normal prior and latent Gaussian models in general

Kolmogorov-Gamma augmentation

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

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

Theorem 1. (Kolmogorov-Gamma integral identity)

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

$$\frac{(e^{\eta})^a}{\{(e^{\eta} - 1)/\eta\}^b} = e^{(a-b/2)\eta} \int_0^{\infty} e^{-\kappa\eta^2/2} p_{\mathsf{KG}}(\kappa; b, 0) \mathrm{d}\kappa, \tag{3}$$

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

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

Conditional conjugacy with latent Gaussian models

- Consider an augmented model with $\eta_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$:

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

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

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

- **Offers 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
- We also develop **fast rejection sampler** for KG variables.

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. (Rapid mixing of Markov chain)

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

- Proof by establishing uniform minorization condition, similar to [Choi and Hobert, 2013]
- Guarantees the existence of CLT for Monte Carlo averages of functions of $\boldsymbol{\beta}$
- Strong result for micobin regression since likelihood is not log-concave

MMI data analysis

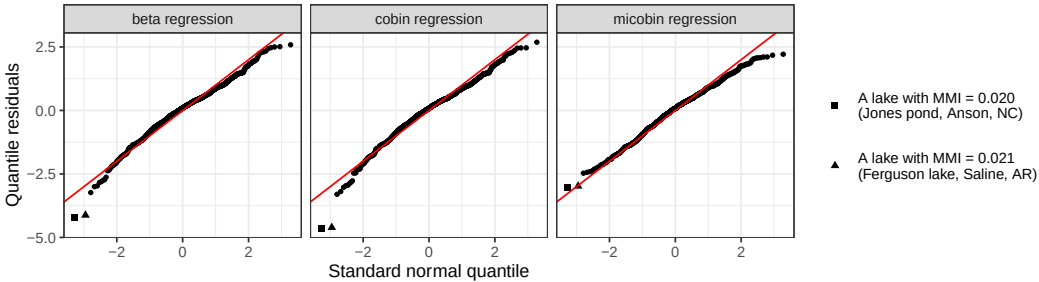
- Return to MMI data analysis, $n = 949$ (removed 1 lake with 0 MMI), $p = 9$
- Fit three different spatial models (beta, cobin, micobin) with cobin canonical link
- Prior $\boldsymbol{\beta} \sim N_p(\mathbf{0}, 100^2 I_p)$, half-Cauchy on random effect standard deviation
- `Stan` for spatial beta;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

MMI data analysis results: association

	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)**

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

MMI data analysis results: goodness of fit



- Quantile residual plot [Dunn and Smyth, 1996]: $\Phi^{-1}(F(y_i \mid \hat{\mu}_i, \hat{\phi}))$ against normal quantiles
- Two influential observations with the lowest MMI values of 0.02 and 0.021
- Re-run the analysis, removing those two lakes

MMI data analysis results: robustness

($n = 947$)	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	$\|\hat{\boldsymbol{\beta}}^{(949)} - \hat{\boldsymbol{\beta}}^{(947)}\|_2 = 0.743$		$\|\hat{\boldsymbol{\beta}}^{(949)} - \hat{\boldsymbol{\beta}}^{(947)}\|_2 = 0.444$		$\|\hat{\boldsymbol{\beta}}^{(949)} - \hat{\boldsymbol{\beta}}^{(947)}\|_2 = 0.069$	

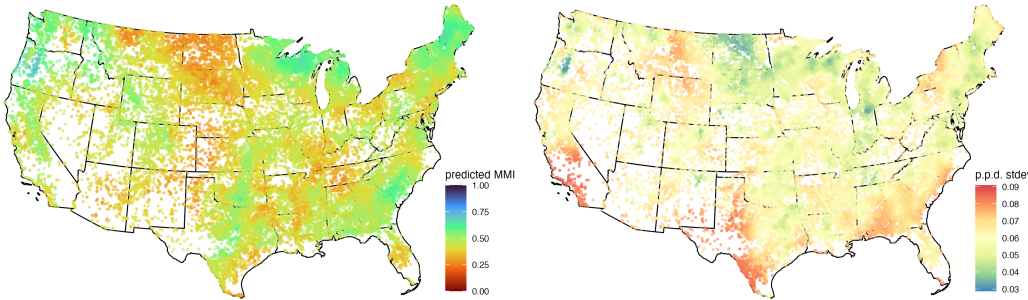
- Beta regression results changed, cobin/micobin results remained same

MMI data analysis results: robustness

- Recall that micobin can handle boundary data
- Re-run micobin model with $n = 950$, a lake with MMI = 0, result mostly unchanged

	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)**

MMI data analysis results: prediction



- Left: predicted MMI, Right: stdev of predicted MMI using spatial micobin regression

- Cobin and micobin regression models with robustness property
- Conditionally normal likelihood with Kolmogorov-Gamma augmentation
- Offers significant computational benefits with latent Gaussian (esp. spatial) models

- 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.
- Reproducing code: `https://github.com/changwoo-lee/cobin-reproduce`
- R package ”cobin”: `https://github.com/changwoo-lee/cobin`

Thank you!

References I

Berggren, N., Daunfeldt, S.-O., and Hellström, J. (2014).
Social trust and central-bank independence.
Eur. J. Polit. Econ., 34:425–439.

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.

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.

Cribari-Neto, F. and Zeileis, A. (2010).
Beta Regression in R.
J. Stat. Softw., 34(2):1–24.

References II

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.

de Vargas Ribeiro, F., Pessarrodona, A., Tucket, C., Mulders, Y., Pereira, R. C., and Wernberg, T. (2022).
Shield wall: Kelps are the last stand against corals in tropicalized reefs.
Funct. Ecol., 36(10):2445–2455.

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

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

Feller, W. (1948).
On the Kolmogorov-Smirnov limit theorems for empirical distributions.
Ann. Math. Stat., 19(2):177–189.

References III

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.

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.

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

References IV

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.

Kosmidis, I. and Zeileis, A. (2024).
Extended-support beta regression for $[0, 1]$ responses.
arXiv preprint arXiv:2409.07233.

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.

References V

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.

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.

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.
https://www.epa.gov/national-aquatic-resource-surveys/nla.

References VI


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

Structural zeros vs Random zeros

RESEARCH ARTICLE

Methods in Ecology and Evolution | BRITISH ECOLOGICAL SOCIETY

What does a zero mean? Understanding false, random and structural zeros in ecology

Anabel Blasco-Moreno[1,3] | Marta Pérez-Casany[2] | Pedro Puig[3] | Maria Morante[4] | Eva Castells[4,5]

Type of zeros	Source	Generator process	Over-dispersion	Zero inflation
False zeros	Design errors	Poor experimental design	–	–
	Observer errors	Lack of experience	–	–
True zeros	Random	Sampling variability	No	No
			Yes	No
	Structural	Outside the count process	No	Yes
			Yes	Yes

- Micobin regression handles **random zeros** arising from sampling variability, not structural zeros

Sampling Kolmogorov-Gamma random variable

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

- 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 $t \in (0.0234, 0.25)$,

$$a_n(x;c,t) = \begin{cases} \{\sinh(c/2)/(c/2)\}\exp(-c^2x/2)a_n^L(x), & 0 < x < t, \\ \{\sinh(c/2)/(c/2)\}\exp(-c^2x/2)a_n^R(x), & t \le x \end{cases} \tag{4}$$

- $a_n^L(x)$, $a_n^R(x)$ derived from dual density representation of Kolmogorov r.v. [Feller, 1948]

Sampling Kolmogorov-Gamma $(1, c)$ random variable

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

$$X \sim \begin{cases} \mathrm{GIG}(-1.5, c^2, 1/4)1(0 < X < t) & \text{with prob. } A^L(c,t)/\{A^L(c,t) + A^R(c,t)\} \\ \mathrm{Exp}(c^2/2 + 2\pi^2)1(t \le X) & \text{with prob. } A^R(c,t)/\{A^L(c,t) + A^R(c,t)\} \end{cases} \tag{5}$$

2. Generate $U \sim \mathrm{Unif}(0, a_0(X;c,t))$
3. Repeat until $U \le \sum_{n=0}^m (-1)^n a_n(X;c,t)$ (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.

Simulation 1: consistency of point estimate

Proposition (consistency of cobin MLE under potential misspecification)

If the mean structure $E(y_i \mid x_i) = g^{-1}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$ is correctly specified, the solution $\hat{\boldsymbol{\beta}}$ of the cobin regression likelihood equations (2) is **consistent** [Gourieroux et al., 1984]

- Data generation with cobit/logit link functions and 4 different distributions
  - $Y_i \sim \text{beta}(\mu_i, \phi)$, $g(\mu_i) = \beta_0 + \beta_1 x_i$
  - $Y_i \sim \text{cobin}((B')^{-1}(\mu_i), \lambda^{-1})$, $g(\mu_i) = \beta_0 + \beta_1 x_i$
  - $Y_i \sim \text{beta rectangular}(\mu_i, \alpha, \phi) = w_i \text{beta}(\tilde{\mu}_i, \phi) + (1 - w_i)\text{unif}(0, 1)$, $g(\mu_i) = \beta_0 + \beta_1 x_i$
  - $Y_i \sim 0.25\text{beta}(\mu_i - \epsilon_i, \phi) + 0.5\text{beta}(\mu_i, \phi) + 0.25\text{beta}(\mu_i + \epsilon_i, \phi)$, $g(\mu_i) = \beta_0 + \beta_1 x_i$
- Correct link & mean structure $g(E(Y_i \mid x_i)) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$, but distribution can be misspecified
- $n \in \{100, 400, 1600\}$, $(\beta_0^{\text{true}}, \beta_1^{\text{true}}) = (0, 1)$, compare betareg $\hat{\boldsymbol{\beta}}$ and cobinreg $\hat{\boldsymbol{\beta}}$

Simulation 1 results

			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**

- Cobin regression $\hat{\beta}$ is consistent even under the misspecified distribution

Simulation 2: spatial regression

- Resembling spatially indexed MMI data $Y(s_i)$
- Data generation: $Y(s_i) \sim \text{beta rectangular}(\mu_i, \alpha, \phi)$, locations $s_i \in [0,1]^2$ uniformly, and

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

- $(\beta_0^{\text{true}}, \beta_1^{\text{true}}) = (0,1)$, $\rho \in \{0.1, 0.2\}$ (spatial dependence)
- Fit data with (1) spatial beta, (2) spatial cobin, (3) spatial micobin regression models
  - **All models are misspecified**
- `Stan` for spatial beta; Gibbs sampler for spatial cobin/micobin; 5000 MCMC samples.
- Compare (1) Inference of β, (2) predictive performance, (3) sampling performance

Simulation 2 results

			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

Monte Carlo standard errors are all less than 0.015 for negtestLL, 0.013 for MSPE, 127.2 for mESS.

- Cobin gives lowest bias/RMSE of $\hat{\beta}_1$, micobin achieves best predictive performance
- Multivariate ESS per time: cobin and micobin better than 40x or more