Bivariate Residual Plots with Simulation Polygons

Rafael A. Moral

23/08/2018

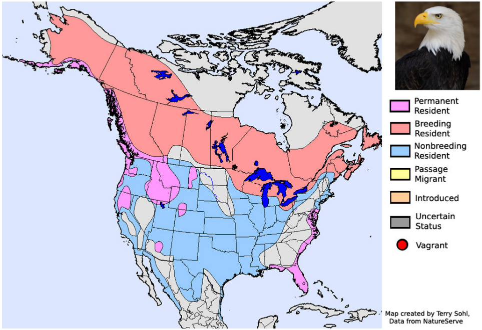
Case study

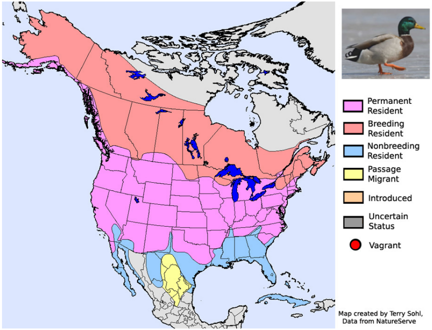
- Bald eagles and mallards

Case study

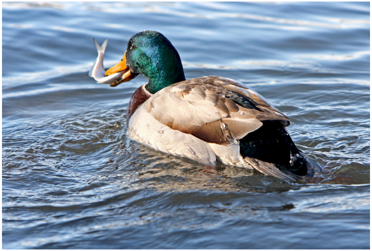
- Bald eagles and mallards

Case study

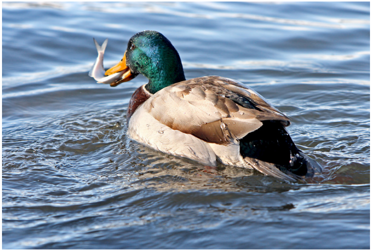




(Ridgely et al., 2003)




alamy stock photo



- How should we assess goodness-of-fit for joint models?

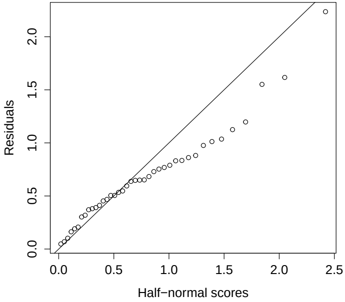
For univariate models

"The ability of the human eye to find patterns in scatters of points is one strong reason for the use of graphical methods." (A.C. Atkinson, 1985)

For univariate models

"The ability of the human eye to find patterns in scatters of points is one strong reason for the use of graphical methods." (A.C. Atkinson, 1985)

- Half-normal plots with simulation envelopes
- Ordered absolute values of a model diagnostic vs. expected order statistics of a half-normal distribution $\Phi^{-1}\left(\frac{i+n-\frac{1}{8}}{2n+\frac{1}{2}}\right)$

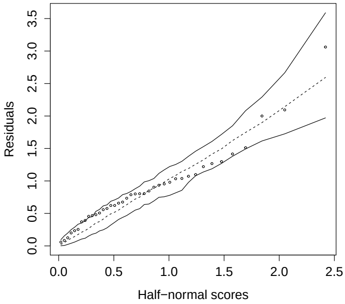


Half-normal plots with simulation envelopes

- Fit model and obtain diagnostics in absolute value and in order
- Simulate 99 response variables using same model matrix, error distribution and fitted parameters
- Refit the model to each simulated sample and obtain the same diagnostics, again, sorted absolute values
- Compute desired percentiles (e.g. 2.5% and 97.5%) to form the envelope

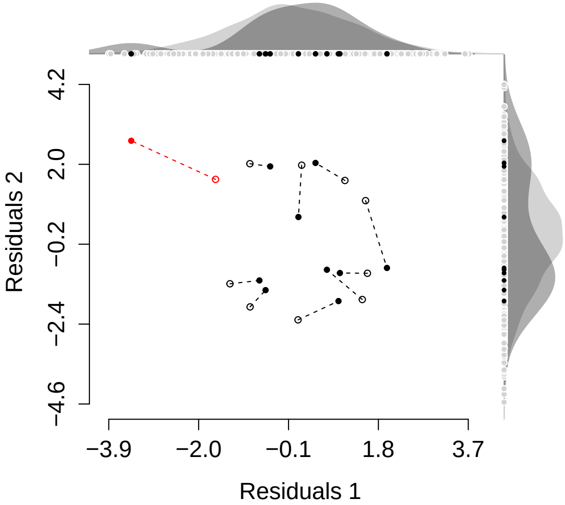
Half-normal plots with simulation envelopes

- Fit model and obtain diagnostics in absolute value and in order
- Simulate 99 response variables using same model matrix, error distribution and fitted parameters
- Refit the model to each simulated sample and obtain the same diagnostics, again, sorted absolute values
- Compute desired percentiles (e.g. 2.5% and 97.5%) to form the envelope

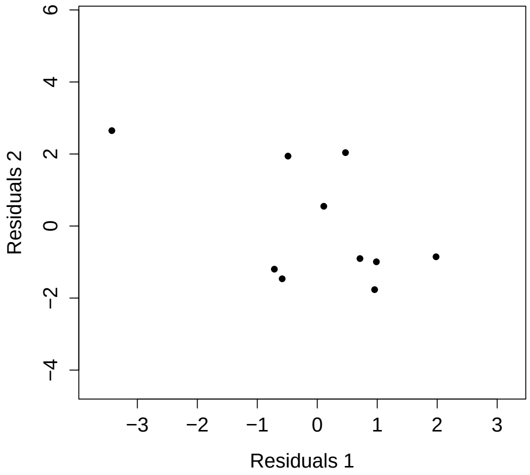


How can we do this for a bivariate model?

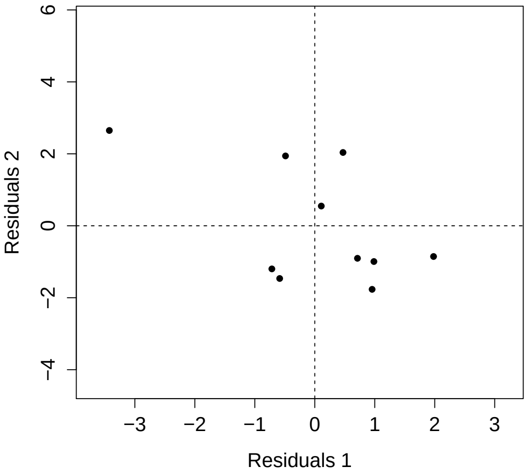
How can we do this for a bivariate model?



Bivariate residuals



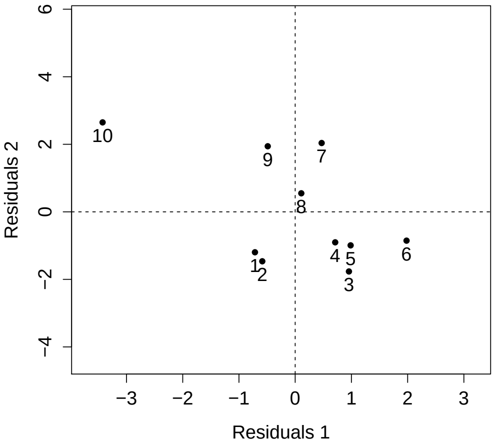
We need to order them



Ordering by angles

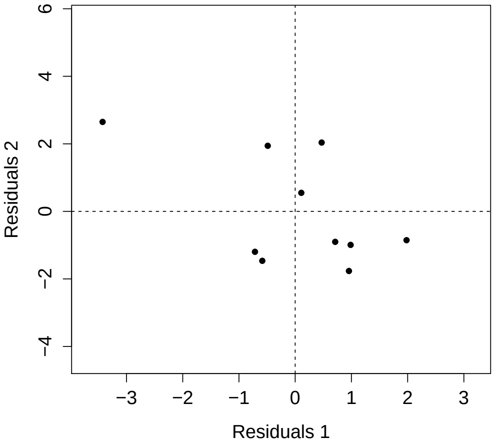
$$\alpha_i = \begin{cases} \tan^{-1}\left(\frac{y_i}{x_i}\right), & x > 0 \\ \tan^{-1}\left(\frac{y_i}{x_i}\right) + \pi, & x < 0 \text{ and } y \geq 0 \\ \tan^{-1}\left(\frac{y_i}{x_i}\right) + \pi, & x < 0 \text{ and } y < 0 \\ \pm\frac{\pi}{2}, & x = 0 \text{ and } y \gtrless 0 \\ \text{undefined}, & x = y = 0 \end{cases}$$

Ordering by angles



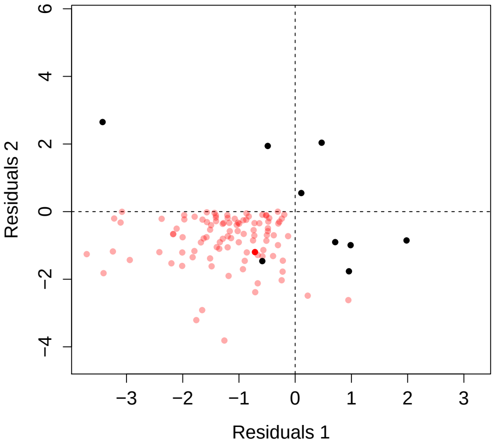
Now we simulate

- Simulate 99 bivariate responses and refit model



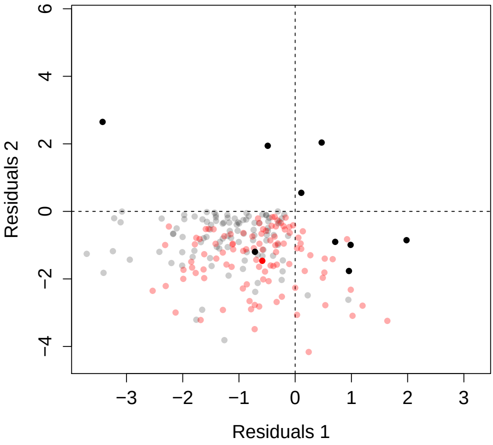
Now we simulate

- Simulate 99 bivariate responses and refit model



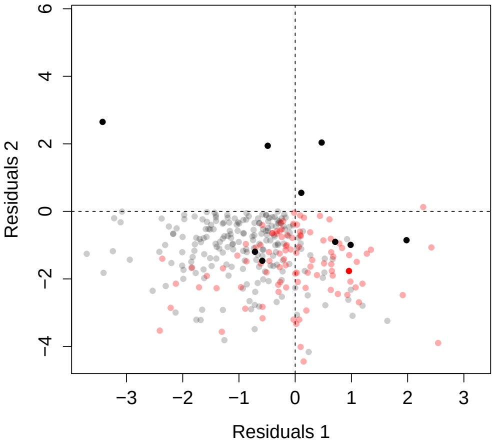
Now we simulate

- Simulate 99 bivariate responses and refit model



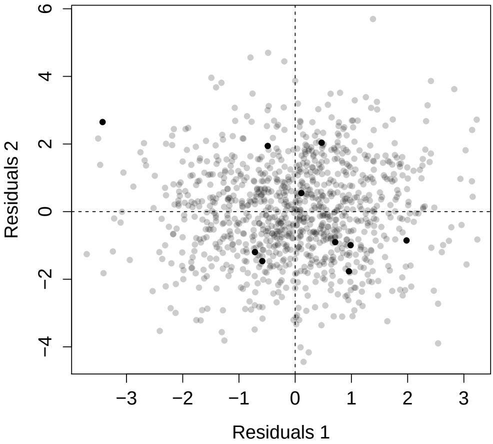
Now we simulate

- Simulate 99 bivariate responses and refit model



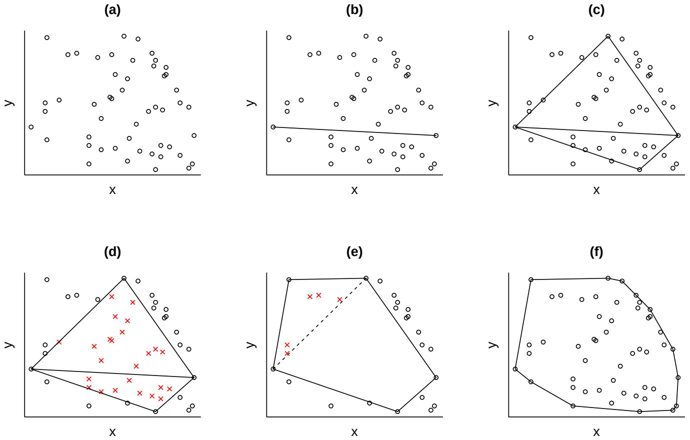
Now we simulate

- Simulate 99 bivariate responses and refit model

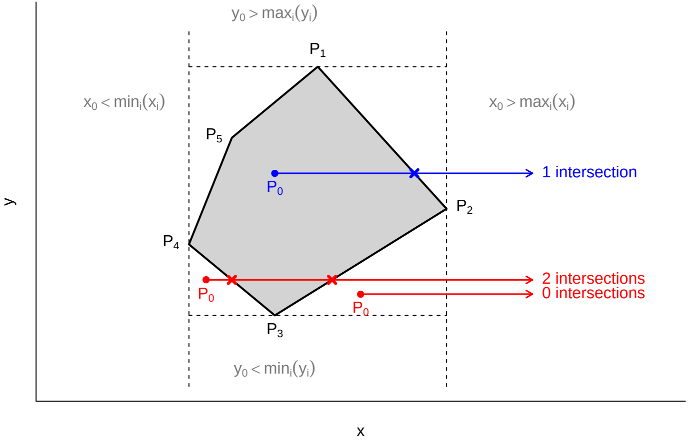


Now we must obtain our "envelope"

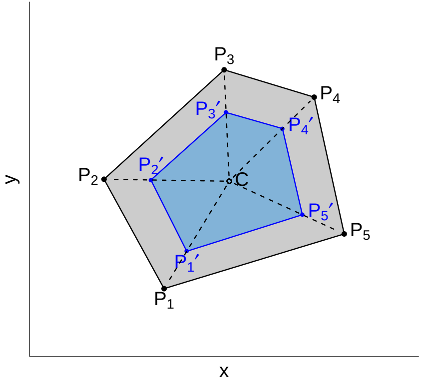
- Convex Hull



Is the point inside?

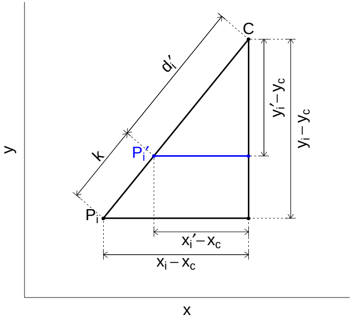


95% of a polygon



$$A_{\mathbf{P}} = \frac{1}{2}\left|\sum_{i=1}^{n-1}(x_i y_{i+1} - x_{i+1} y_i) + (x_n y_1 - x_1 y_n)\right| \qquad A_{\mathbf{P}'} = \gamma A_{\mathbf{P}}, \quad 0 < \gamma < 1$$

95% of a polygon



$$\frac{x_i' - x_C}{x_i - x_C} = \frac{y_i' - y_C}{y_i - y_C} = \frac{d_i'}{d_i}$$

$$x_i' = \frac{d_i'}{d_i}(x_i - x_C) + x_C = \frac{d_i x_i - k\tilde{x}_i}{d_i}$$

$$y_i' = \frac{d_i'}{d_i}(y_i - y_C) + y_C = \frac{d_i y_i - k\tilde{y}_i}{d_i}$$

$\tilde{x}_i = x_i - x_C$ and $\tilde{y}_i = y_i - y_C$

95% of a polygon

$$ak^2 + bk + c = 0$$

$$\begin{aligned}
a &= \sum_{i=1}^{n-1} \frac{\tilde{x}_i \tilde{y}_{i+1} - \tilde{x}_{i+1}\tilde{y}_i}{d_i d_{i+1}} + \frac{\tilde{x}_n \tilde{y}_1 - \tilde{x}_1 \tilde{y}_n}{d_n d_1} \\
b &= \sum_{i=1}^{n-1} \left\{ \frac{d_i(\tilde{x}_{i+1} y_i - x_i \tilde{y}_{i+1}) + d_{i+1}(x_{i+1}\tilde{y}_i - \tilde{x}_i y_{i+1})}{d_i d_{i+1}} \right. \\
&\qquad \left. + \frac{d_n(\tilde{x}_1 y_n - x_n \tilde{y}_1) + d_1(x_1 \tilde{y}_n - \tilde{x}_n y_1)}{d_n d_1} \right\} \\
c &= 2(A \pm \gamma A_{\mathbf{P}})
\end{aligned}$$

$$\hat{k} = \min_i \{k_i \in \mathbb{R}\}$$

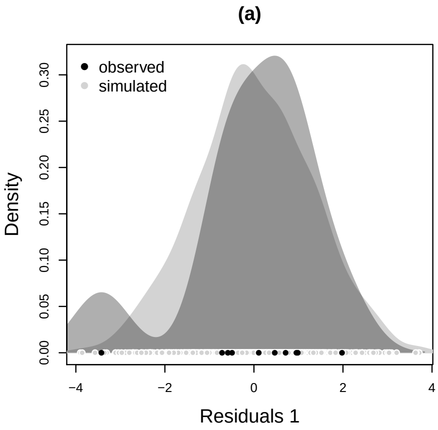
95% of a polygon

- Another possibility is to scale the distances d_i from the centroid to the vertices to $\sqrt{\alpha} \times d_i$ and the resulting coordinates are:

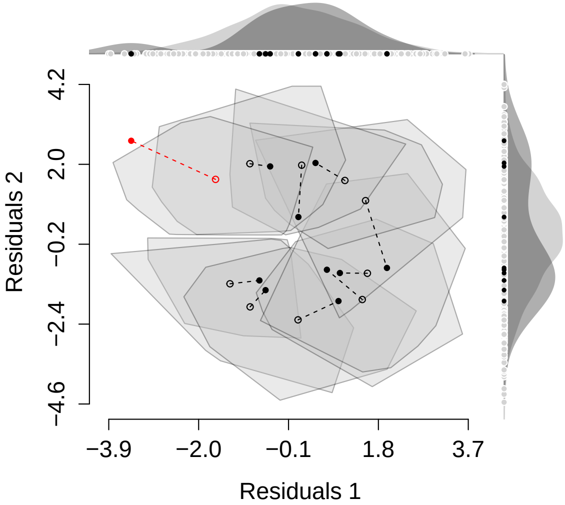
$$
\begin{aligned}
x_i^* &= \sqrt{\alpha}(x_i - x_C) + x_C \\
y_i^* &= \sqrt{\alpha}(y_i - y_C) + y_C
\end{aligned}
$$

for the new polygon $\mathbf{P}^* = \overline{P_1^* \ldots P_v^*}$, with $P_i^* = (x_i^*, y_i^*)$.

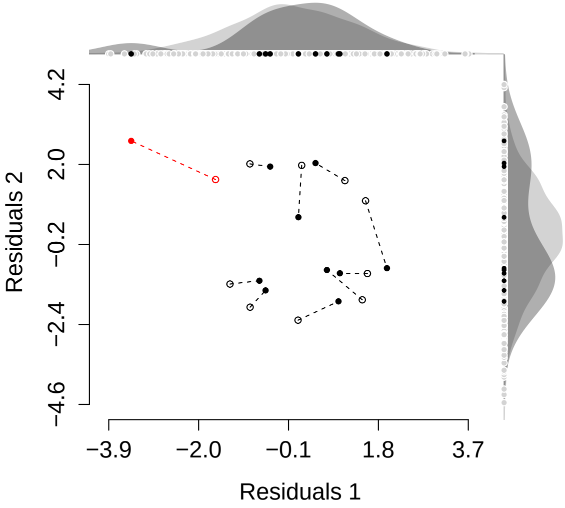
Adding density plots



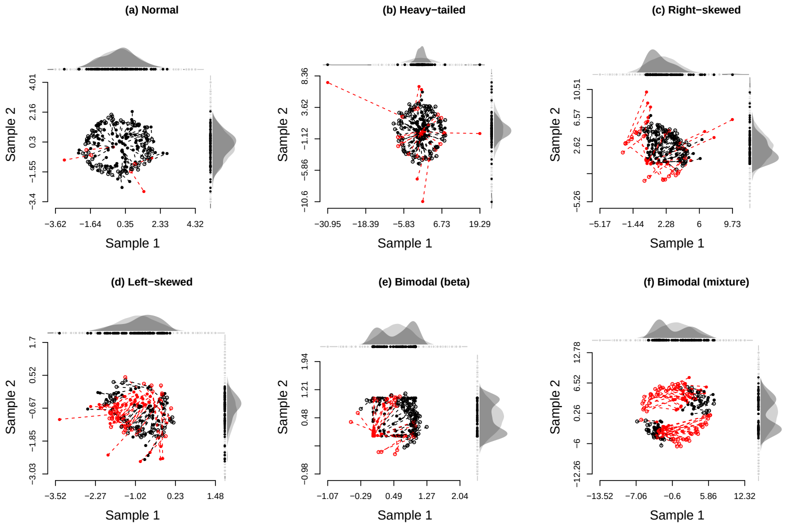
Final display



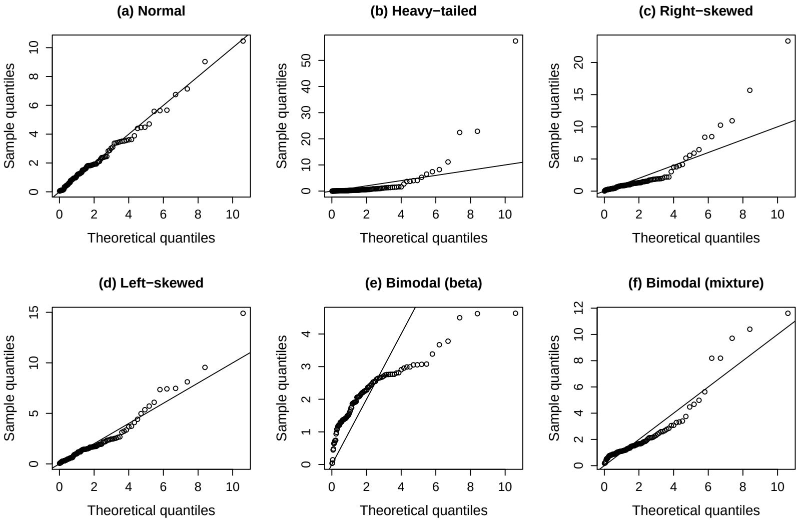
Final display



Expected shapes



Expected shapes



An example using simulated data

$$\mathbf{Y}_i = \begin{bmatrix} Y_{1i} \\ Y_{2i} \end{bmatrix} \sim N_2\left(\begin{bmatrix} \mu_{1i} \\ \mu_{2i} \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}\right),$$

Marginally

$$Y_{ji} \quad \sim \quad N(\mu_{ji}, \sigma_j^2), \quad j = 1, 2,$$

$$\text{Cov}(Y_{1i}, Y_{2i}) = \sigma_{12}$$

Model fitting

$$\mu_{ji} = \beta_{j0} + \beta_{j1} x_i$$

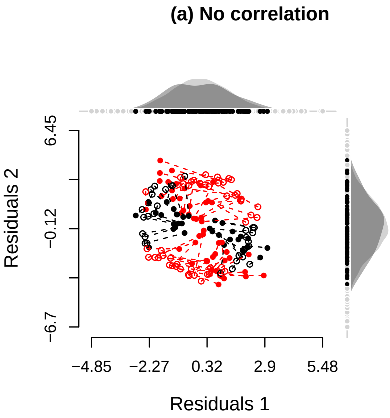
$$L(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \prod_{i=1}^{n} (2\pi)^{-1} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\mathbf{y}_i - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) \right\}$$

Estimates

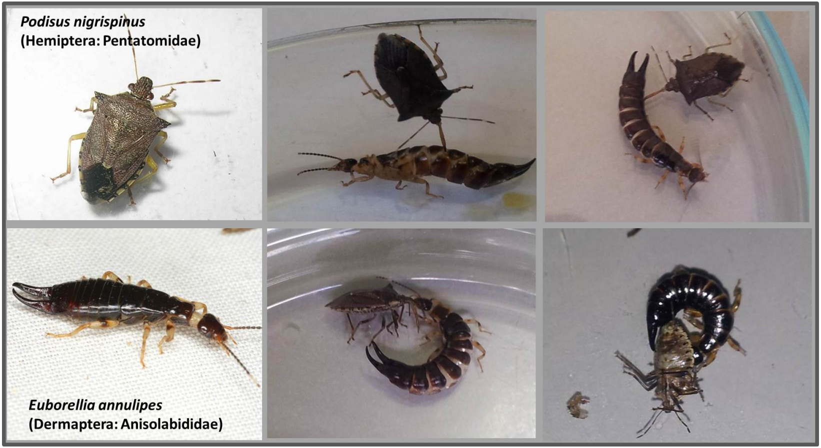
Table 1: Parameter estimates (standard errors) for both models fitted to the simulated correlated bivariate normal data and true values

Parameter	Assuming independence	Estimating covariance	True value
β_{10}	1.50 (0.35)	1.50 (0.35)	2.00
β_{11}	0.43 (0.06)	0.43 (0.06)	0.40
β_{20}	0.78 (0.48)	0.78 (0.48)	0.20
β_{21}	0.18 (0.08)	0.18 (0.08)	0.20
σ_1^2	1.79 (0.28)	1.79 (0.28)	2.00
σ_2^2	3.48 (0.55)	3.49 (0.55)	3.00
σ_{12}	0.00 (–)	−1.61 (0.33)	−1.70
$-2\times$loglik	600.42	557.10	–

Bivariate residual plots with simulation polygons



An example with real data



Bivariate Poisson model

$$\begin{aligned} X_j &\sim \mathsf{P}(\lambda_j), \quad j = 0, 1, 2 \\ Y_1 &= X_0 + X_1 \\ Y_2 &= X_0 + X_2 \end{aligned}$$

$$(Y_1, Y_2) \sim \mathsf{BP}(\lambda_0, \lambda_1, \lambda_2)$$

$$P(Y_1 = y_1, Y_2 = y_2) = e^{-(\lambda_0+\lambda_1+\lambda_2)} \frac{\lambda_1^{y_1} \lambda_2^{y_2}}{y_1! y_2!} \sum_{k=0}^{\min(y_1, y_2)} \binom{y_1}{k} \binom{y_2}{k} k! \left(\frac{\lambda_0}{\lambda_1 \lambda_2} \right)^k$$

$$\begin{aligned} Y_1 &\sim \mathsf{P}(\lambda_0 + \lambda_1) \\ Y_2 &\sim \mathsf{P}(\lambda_0 + \lambda_2) \end{aligned}$$

Model fitting

- Pseudo-likelihood maximization (Gourieroux et al., 1984)
- Newton-Raphson algorithm (Jung & Winkelmann, 1993; Kocherlakota & Kocherlakota, 2001)
- Generalized least squares (Ho & Singer, 2001)
- Bayesian methods (Tsionas, 2001)
- EM algorithm (Karlis & Ntzoufras, 2005)

Model fitting

- Complete-data log-likelihood

$$
\begin{aligned}
l(\lambda_0, \lambda_1, \lambda_2) &= \sum_{i=1}^{n} \log\{P(X_{0i} = x_{0i}, X_{1i} = x_{1i}, X_{2i} = x_{2i})\} \\
&= \sum_{i=1}^{n} \log\{P(X_{0i} = x_{0i})P(X_{1i} = x_{1i})P(X_{2i} = x_{2i})\}
\end{aligned}
$$

- Maximisation is straightforward by fitting three independent Poisson GLMs
- One to variable $x_{1i} = y_{1i} - x_{0i}$, another to variable $x_{2i} = y_{2i} - x_{0i}$, and to variable x_{0i}, replaced by its conditional expectation z_i

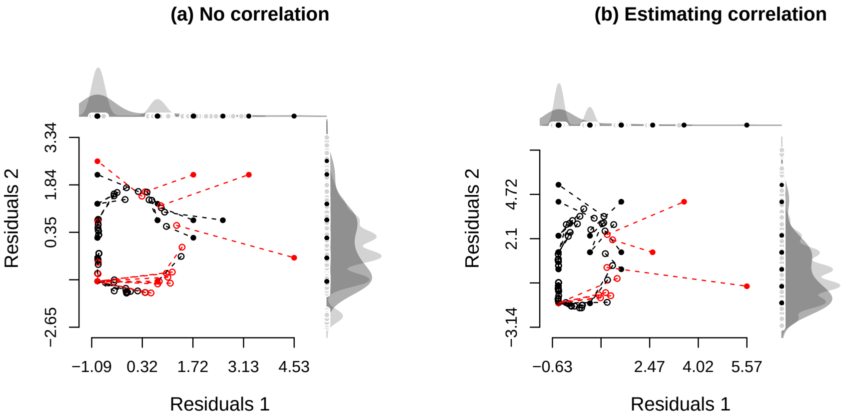
$$
\begin{aligned}
z_i &= \mathsf{E}(X_{0i} | Y_{1i}, Y_{2i}) \\
&= \sum_{x_{0i}=0}^{\infty} x_{0i} P(X_{0i} = x_{0i} | Y_{1i} = y_{1i}, Y_{2i} = y_{2i}) \\
&= \lambda_0 \frac{P(Y_{1i} = y_{1i} - 1, Y_{2i} = y_{2i} - 1)}{P(Y_{1i} = y_{1i}, Y_{2i} = y_{2i})}
\end{aligned}
$$

Estimates

Table 2: Parameter estimates (standard errors) for both models fitted to the attack data

Parameter	Assuming independence	Estimating covariance
$\log \lambda_1$ (stinkbug)	-0.85 (0.20)	-1.43 (0.04)
$\log \lambda_2$ (earwig)	1.00 (0.08)	0.93 (0.00)
$\log \lambda_0$	0.00 $(-)$	-1.66 (0.05)
$-2\times$loglik	332.32	327.83

Bivariate residual plots with simulation polygons



Using PIT diagnostics

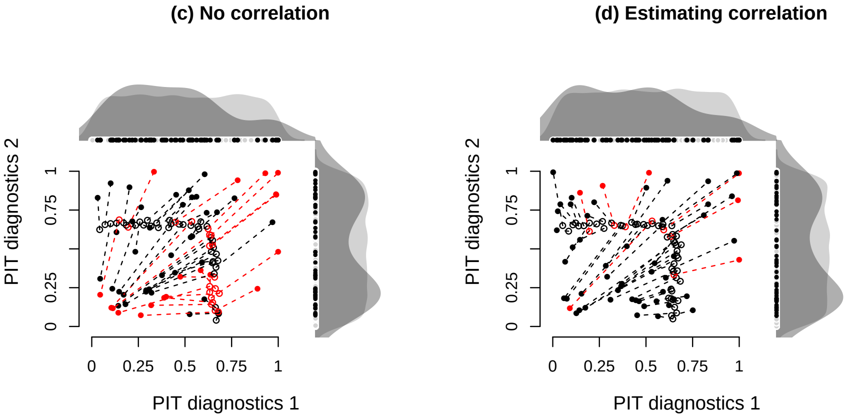
- Randomized PIT diagnostics for discrete models

$$r_{ji}^{rand.pit} = F(y_{ji} - 1; \hat{\boldsymbol{\theta}}_{ji}) + u_i\{F(y_{ji}; \hat{\boldsymbol{\theta}}_{ji}) - F(y_{ji} - 1; \hat{\boldsymbol{\theta}}_{ji})\},$$

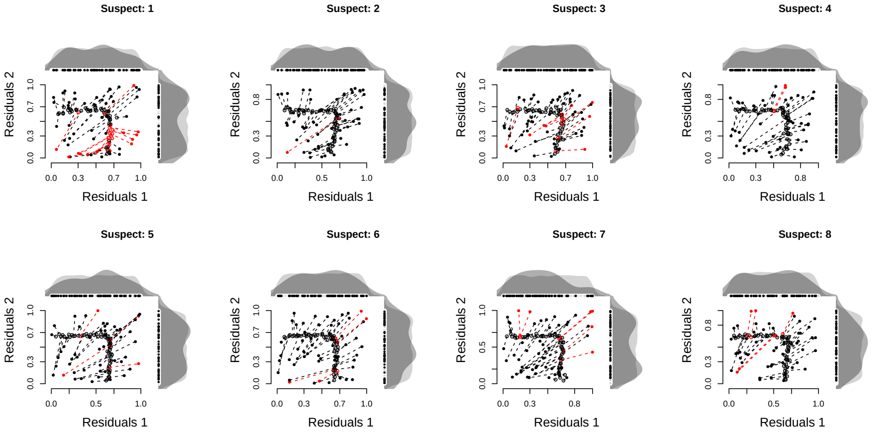
  with $F(-1) = 0$, and u_i is a realization of $U_i \sim$ Uniform$(0, 1)$.
- For the bivariate Poisson model:

$$F(y_{ji}; \hat{\boldsymbol{\theta}}_{ji}) = e^{-\hat{\mu}_j} \sum_{k=0}^{y_{ji}} \frac{\hat{\mu}_j^k}{k!}$$

Bivariate residual plots with simulation polygons



Line-up test


Suspect: 1
Residuals 2
0.0 0.3 0.7 1.0
Residuals 1
0.0 0.3 0.7 1.0
Suspect: 2
Residuals 2
0.0 0.3 0.8
Residuals 1
0.0 0.5 1.0
Suspect: 3
Residuals 2
0.0 0.3 0.7 1.0
Residuals 1
0.0 0.3 0.7 1.0
Suspect: 4
Residuals 2
0.0 0.3 0.8
Residuals 1
0.0 0.3 0.8
Suspect: 5
Residuals 2
0.0 0.3 0.7 1.0
Residuals 1
0.0 0.5 1.0
Suspect: 6
Residuals 2
0.0 0.3 0.7 1.0
Residuals 1
0.0 0.3 0.7 1.0
Suspect: 7
Residuals 2
0.0 0.5 1.0
Residuals 1
0.0 0.3 0.8
Suspect: 8
Residuals 2
0.0 0.3 0.8
Residuals 1
0.0 0.5 1.0

Code availability

- `bivrp` package on CRAN

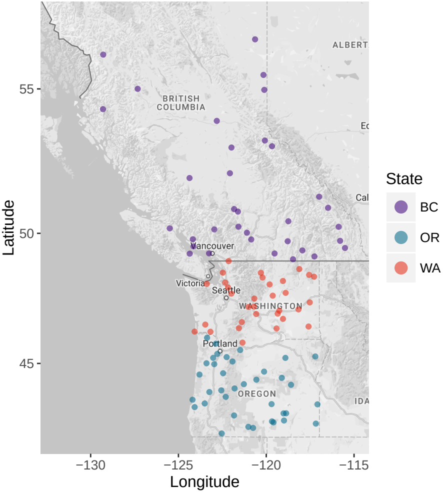
bivrp-package	Bivariate Residual Plots with Simulation Polygons
add.dplots.plot	Internal functions to prepare 'bivrp' objects
add.dplots.prep	Internal functions to prepare 'bivrp' objects
bivrp	Bivariate Residual Plots with Simulation Polygons
chp.perpoint	Internal functions to prepare 'bivrp' objects
get.k	Polygon operations
get.newpolygon	Polygon operations
is.point.inside	Determine if point is inside or outside a simple polygon area
plot.bivrp	Plot Method for bivrp Objects
polygon.area	Polygon operations
sorttheta	Internal functions to prepare 'bivrp' objects

Final considerations

- This is not a formal test!
- Simple tool for assessing goodness-of-fit of bivariate models
- Use of different diagnostics is recommended (e.g. PIT diagnostics)
- Drawbacks include computational burden for complex models and the way outliers may influence convex hulls
- Problematic extension to big data
- Extension to the n-variate setting
- Complementary and (hopefully) helpful approach

But what about the eagles?

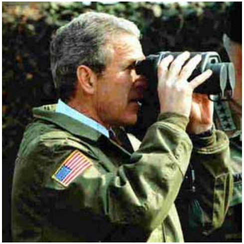
Back to our first motivation example



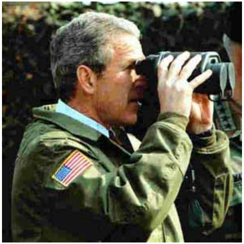
Imperfect detection

Imperfect detection

Imperfect detection



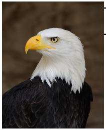
Imperfect detection



Can you spot the cat?

Can you spot the cat?

Case study



Route	Stops 1-10	Stops 11-20	Stops 21-30	Stops 31-40	Stops 41-50
4	0	0	0	1	4
6	0	0	0	1	0
16	0	0	0	1	0
17	0	0	0	0	0
...					
407	0	0	0	1	5
409	0	0	0	0	3
...					

Route	Stops 1-10	Stops 11-20	Stops 21-30	Stops 31-40	Stops 41-50
4	0	0	1	0	4
6	0	0	0	0	0
16	0	0	0	0	0
17	0	2	4	0	0
...					
407	0	0	0	1	0
409	1	0	0	0	7
...					

Joint model formulation (Moral et al., 2018)

- Bivariate N-mixture model

$$
\begin{aligned}
Y_{1_{it}}|N_{1_i} &\sim \text{Binomial}(N_{1_i}, p_{1_{it}}) \\
N_{1_i} &\sim \text{Poisson}(\lambda_{1_i}) \text{ or NB}(\lambda_{1_i}, \phi_1) \\
Y_{2_{it}}|N_{1_i}, N_{2_i} &\sim \text{Binomial}(N_{2_i}, p_{2_{it}}) \\
N_{2_i}|N_{1_i} &\sim \text{Poisson}(\psi_i + \lambda_{2_i} N_{1_i}) \text{ or NB}(\psi_i + \lambda_{2_i} N_{1_i}, \phi_2)
\end{aligned}
$$

Assumptions

- independence among sites
- closed population; no migration
- $\lambda_{2_i} = 0 \Rightarrow$ no correlation between species

Joint N-mixture model

- Likelihood:

$$L(\mathbf{p_1}, \mathbf{p_2}, \boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \boldsymbol{\psi} | \{y_{1_{it}}\}, \{y_{2_{it}}\}) =$$

$$\prod_{i=1}^{R} \left\{ \sum_{n_{1_i}=\max_t\{y_{1_{it}}\}}^{\infty} \left[\prod_{t=1}^{T} \mathsf{Bin}(y_{1_{it}}; n_{1_i}, p_{1_{it}}) \right] f_{N_{1_i}}(n_{1_i}; \boldsymbol{\theta_1}_i) \times \right.$$

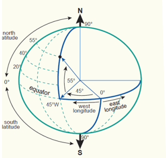
$$\left. \sum_{n_{1_i}=\max_t\{y_{1_{it}}\}}^{\infty} f_{N_{1_i}}(n_{1_i}; \boldsymbol{\theta_1}_i) \sum_{n_{2_i}=\max_t\{y_{2_{it}}\}}^{\infty} \left[\prod_{t=1}^{T} \mathsf{Bin}(y_{2_{it}}; n_{2_i}, p_{2_{it}}) \right] f_{N_{2_i}}(n_{2_i}; \psi_i, \boldsymbol{\theta_2}_i) \right\}$$

Total variation in *observed* eagle abundance

Total variation in *observed* eagle abundance

Habitat suitability

Geographical coordinates



Total variation in *observed* eagle abundance

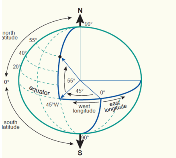
Habitat suitability

Environmental covariates

Affecting eagle

Geographical coordinates

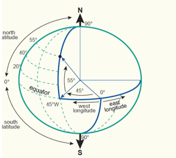
Wind speed
Temperature
...



Total variation in *observed* eagle abundance

Habitat suitability → Geographical coordinates

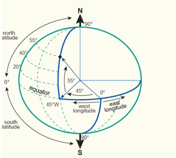
Environmental covariates
- ***Affecting eagle*** → Wind speed, Temperature, ...
- ***Affecting observer*** → Cloud coverage, Expertise, ...



Total variation in *observed* eagle abundance

Habitat suitability

Geographical coordinates



Environmental covariates

Affecting eagle

Wind speed
Temperature
...

Affecting observer

Cloud coverage
Expertise
...

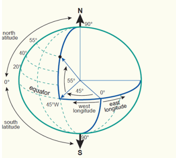
Unexplained variation

Missing covariates
Natural variation
...

Total variation in *observed* eagle abundance

Habitat suitability

Geographical coordinates



Environmental covariates

Affecting eagle

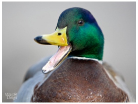
Wind speed
Temperature
...

Affecting observer

Cloud coverage
Expertise
...

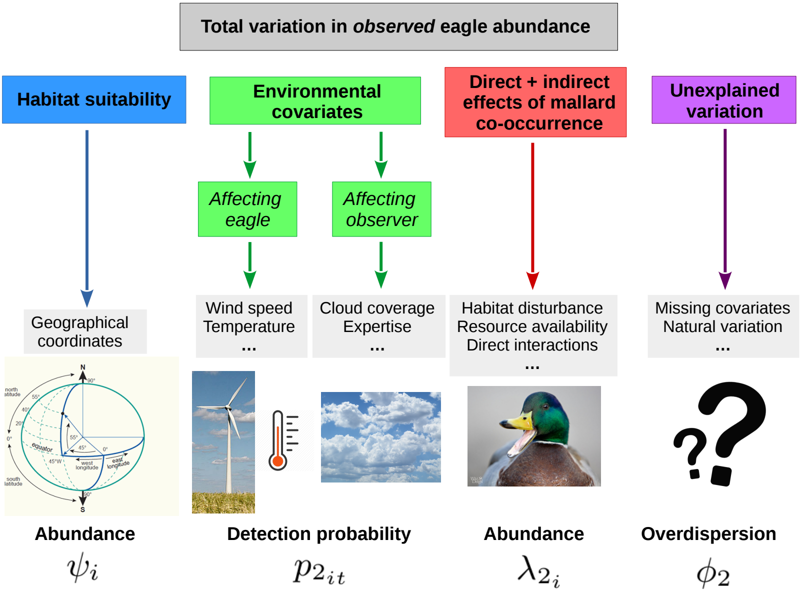
Direct + indirect effects of mallard co-occurrence

Habitat disturbance
Resource availability
Direct interactions
...



Unexplained variation

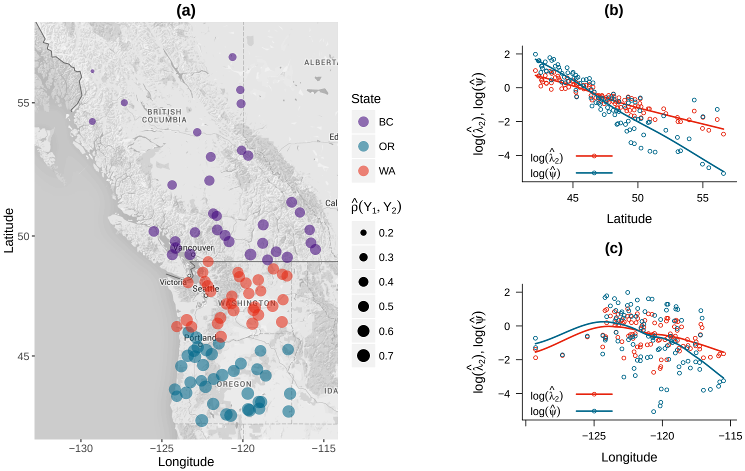
Missing covariates
Natural variation
...


Total variation in *observed* eagle abundance
Habitat suitability
Environmental covariates
Direct + indirect effects of mallard co-occurrence
Unexplained variation
Affecting eagle
Affecting observer
Geographical coordinates
Wind speed
Temperature
...
Cloud coverage
Expertise
...
Habitat disturbance
Resource availability
Direct interactions
...
Missing covariates
Natural variation
...
N
90°
north latitude
55°
40°
20°
0°
equator
45°
45°W
west longitude
east longitude
south latitude
90°
S
Abundance
ψ_i
Detection probability
$p_{2_{it}}$
Abundance
λ_{2_i}
Overdispersion
ϕ_2

Case study: Model selection

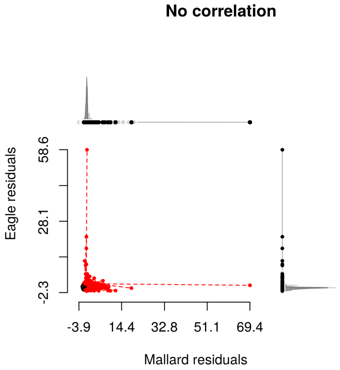
Parameter	Univariate models				Joint models			
	P–P	NB–P	P–NB	NB–NB	P–P	NB–P	P–NB	NB–NB
Detection								
(Mallard – p_{1it})								
Intercept	-4.49	-3.17	-4.49	-3.17	-2.69	-3.14	-4.49	-3.15
Temperature	0.25	0.36	0.25	0.36	0.28	0.36	0.25	0.36
Wind speed	-0.15	-0.28	-0.15	-0.28	-0.11	-0.29	-0.15	-0.28
(Bald eagle – p_{2it})								
Intercept	-4.09	-4.09	-2.88	-2.88	-3.07	-2.87	-2.89	-2.88
Temperature	0.00	0.00	-0.15	-0.15	-0.04	-0.13	-0.19	-0.11
Wind speed	0.14	0.14	0.34	0.34	0.11	0.33	0.34	0.37
Abundance								
(Mallard – λ_{1i})								
Intercept	3.92	2.72	3.92	2.72	2.10	2.63	3.92	2.67
Latitude	0.22	0.06	0.22	0.06	0.37	0.05	0.22	0.05
Longitude	0.32	0.17	0.32	0.17	0.30	0.15	0.32	0.15
Lat×Long	0.04	0.09	0.04	0.09	0.14	0.08	0.04	0.09
(Bald eagle – ψ_i)								
Intercept	3.22	3.22	2.14	2.14	-22.50	-0.76	2.04	-0.48
Latitude	-0.86	-0.86	-0.86	-0.86	-1.86	-1.59	-0.99	-1.65
Longitude	-0.13	-0.13	-0.23	-0.23	12.37	-0.67	-0.42	-0.94
Lat×Long	-0.12	-0.12	0.13	0.13	0.91	-0.16	0.05	-0.01
(Bald eagle – λ_{2i})								
Intercept	—	—	—	—	0.12	-0.53	-28.96	-0.69
Latitude	—	—	—	—	-1.37	-0.83	-6.58	-0.78
Longitude	—	—	—	—	-0.59	-0.34	13.05	-0.27
Lat×Long	—	—	—	—	-0.31	0.04	3.49	0.07
(Dispersion)								
ϕ_1	—	0.51	—	0.51	—	0.39	—	0.46
ϕ_2	—	—	0.46	0.46	—	—	0.48	1.71
$-2 \times$ loglik	3350.28	2986.90	2994.67	2631.29	3196.88	2620.96	2990.69	2617.46

Results

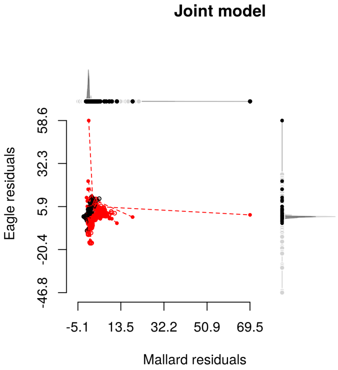


(a)
ALBERTA
BRITISH COLUMBIA
Vancouver
Victoria
Seattle
WASHINGTON
Portland
OREGON
Latitude
Longitude
55
50
45
−130
−125
−120
−115
State
BC
OR
WA
$\hat{\rho}(Y_1, Y_2)$
0.2
0.3
0.4
0.5
0.6
0.7
(b)
$\log(\hat{\lambda}_2)$, $\log(\hat{\psi})$
$\log(\hat{\lambda}_2)$
$\log(\hat{\psi})$
Latitude
2
0
−2
−4
45
50
55
(c)
$\log(\hat{\lambda}_2)$, $\log(\hat{\psi})$
$\log(\hat{\lambda}_2)$
$\log(\hat{\psi})$
Longitude
2
0
−2
−4
−125
−120
−115

Case study: bivariate residual plots



368/515 (71%)



197/515 (38%)

Thank you!

References

- Atkinson, A.C. (1985) *Plots, transformations, and regression: An introduction to graphical methods of diagnostic regression analysis*, Clarendon Press, Oxford.
- Karlis, D. & Ntzoufras, I. (2005) Bivariate poisson and diagonal inflated bivariate poisson regression models in R, *Journal of Statistical Software* 14(10), 1–36.
- Moral, R.A., Hinde, J. & Demétrio, C.G.B. (2017) Half-normal plots and overdispersed models in R: The hnp package, *Journal of Statistical Software* 81(10), 1–23.
- Moral, R.A., Hinde, J. & Demétrio, C.G.B. (2018) bivrp: Bivariate residual plots with simulation polygons. *R package version 1.0-2*.
- Moral, R.A. et al. (2018) Models for Jointly Estimating Abundances of Two Unmarked Site-Associated Species Subject to Imperfect Detection. *JABES* 23, 20–38.
- Pardieck, K.L. et al. (2016) North American breeding bird survey dataset 1966–2015, version 2015.0. U.S. Geological Survey, Patuxent Wildlife Research Center. URL `www.pwrc.usgs.gov/BBS/RawData/`
- Ridgely, R.S., et al. (2003) Digital distribution maps of the birds of the Western hemisphere, version 1.0. NatureServe, Arlington, Virginia, USA.
- Royle, J.A. (2004) N-mixture models for estimating population size from spatially replicated counts. *Biometrics*, 60, 108–115.