The Art of Bayesian Model Checking

Paul Conn

NOAA Alaska Fisheries Science Center (MML)

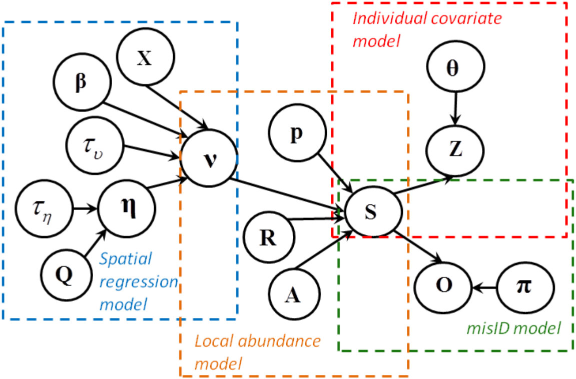
1 Feb 2022

Outline

- Background
- Code: availability and dependencies
- Notation
- Example dataset & estimation models
- Posterior predictive checks
- Discrepancy functions
- Bayesian p-values
- Sampled posterior p-values
- Pivotal discrepancy measures
- Cross validation
- Summary: Bayesian model checking
- GOF for integrated population models (Besbeas and Morgan 2014)

Background

- What this talk is *not*: introduction to Bayesian inference, model convergence diagnostics, model selection
- How do you go about assessing goodness-of-fit in a big hierarchical model?



Code and dependencies

- Presentation and R Markdown code available at www.github.com/pconn/BMC_CAPAM_talk
- Some functions from HierarchicalGOF R package, install available at www.github.com/pconn/HierarchicalGOF

NB: This package accompanied Conn et al. (2018); never intended for production level use!

- Some of these diagnostics are in the DHARMa R package (Hartig 2021)

Notation

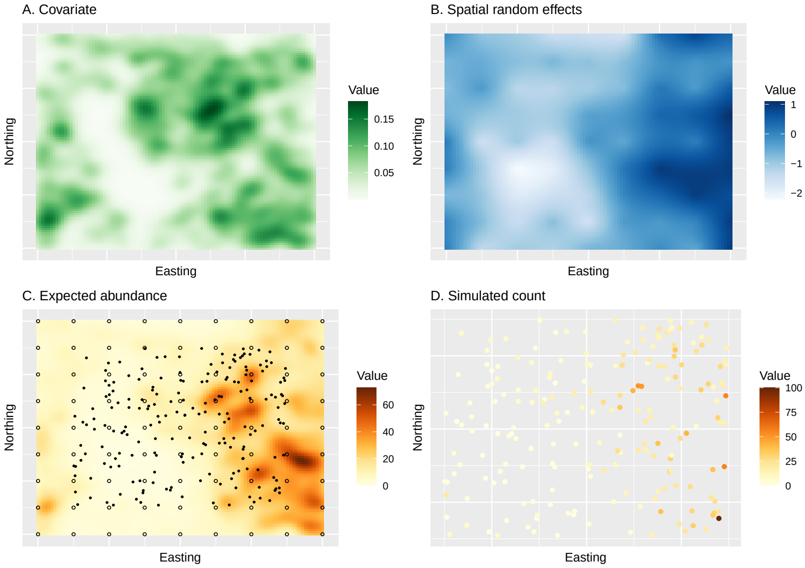
- Bold: vector or matrix
- $[\boldsymbol{\theta}]$: Marginal distribution of $\boldsymbol{\theta}$
- $[\mathbf{y}|\boldsymbol{\theta}]$: Conditional distribution of $\mathbf{y}$ given $\boldsymbol{\theta}$
- $f(y_i|\boldsymbol{\theta})$: Probability mass or density function evaluated at y_i
- $F(y_i|\boldsymbol{\theta}) = \int_{-\infty}^{y_i} f(z|\boldsymbol{\theta})dz$: Cumulative mass or density function evaluated at y_i
- $[\mathbf{y}^{rep}|\mathbf{y}] = \int [\mathbf{y}^{rep}|\boldsymbol{\theta}][\boldsymbol{\theta}|\mathbf{y}]d\boldsymbol{\theta}$: Posterior predictive distribution

Example dataset

Simulated spatial count dataset (think CPUE index standardization with spatially autocorrelated random effects) with 200 randomly sampled locations.

$$
\begin{aligned}
y_i &\sim \text{Poisson}(\exp(\mathbf{x}_i'\boldsymbol{\beta} + \eta_i + \epsilon_i)) \\
\boldsymbol{\eta} &\sim \text{Predictive-process-exponential}(\theta, \tau_\eta) \\
\epsilon_i &\sim \text{Normal}(0, 1/\tau_\epsilon)
\end{aligned}
$$

Example dataset



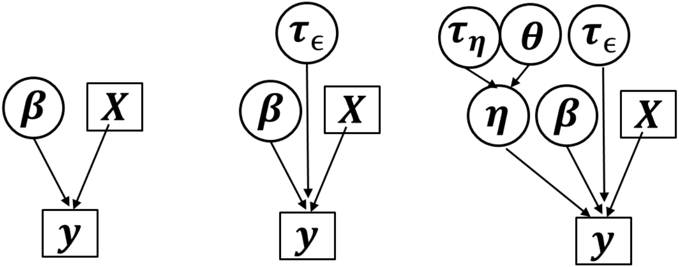
A. Covariate

B. Spatial random effects

C. Expected abundance

D. Simulated count

Example estimation models



Simple Bayesian Poisson GLM (no random effects)

Bayesian Poisson GLMM w/ overdispersion only

Bayesian Poisson GLMM w/ overdispersion and spatially autocorrelated random effects

DAGs for count data

Posterior predictive checks

Posterior predictive distribution:

$[\mathbf{y}^{rep}|\mathbf{y}] = \int [\mathbf{y}^{rep}|\boldsymbol{\theta}][\boldsymbol{\theta}|\mathbf{y}]d\boldsymbol{\theta}$

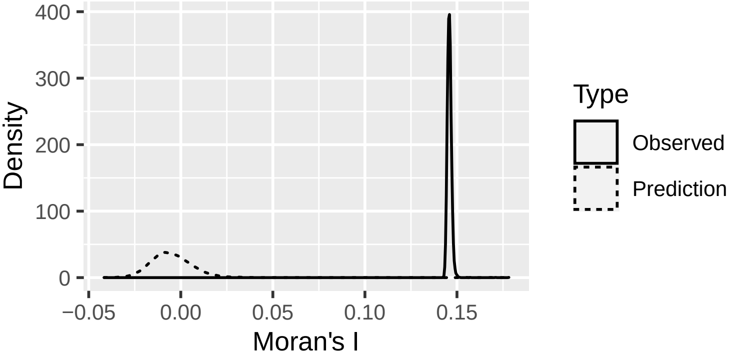
Practically:

1. Sample from the posterior $\boldsymbol{\theta}^{rep} \sim [\boldsymbol{\theta}|\mathbf{y}]$
2. Generate replicated posterior predictive data $\mathbf{y}^{rep}|\boldsymbol{\theta}^{rep}$.
3. Compare real data to simulated data in some fashion. Do they look similar?

Graphical checks

Many possible checks, perhaps used more informally since assessment of fit from graphs is somewhat subjective. See, e.g. `bayesplot` library for `Stan` users.

Example: Generate "null" distribution of Moran-I statistic values for posterior predictions and compare to those for observed data (shown: Bayesian GLM)



Discrepancy functions - $T(\mathbf{y}, \boldsymbol{\theta})$

Are my data similar to those simulated from a model (i.e., posterior predictions)?

-Omnibus: e.g., Chi-square, Freeman-Tukey, Deviance, Likelihood ratio

-Targeted: Quantiles, Proportion of zeros, Moran's I of residuals

-Pivotal: Stay tuned!

Bayesian p-values

Historically, this is the most frequently reported Bayesian model checking procedure.

```
$P \leftarrow 0$
for $i \in 1 : m$ do
    Draw $\boldsymbol{\theta}_i \sim [\boldsymbol{\theta}|\mathbf{y}]$
    Draw $\mathbf{y}_i^{rep} \sim [\mathbf{y}|\boldsymbol{\theta}_i]$
    Calculate $T_i^{rep} = T(\mathbf{y}_i^{rep}, \boldsymbol{\theta}_i)$
    Calculate $T_i^{obs} = T(\mathbf{y}, \boldsymbol{\theta}_i)$
    if $T_i^{obs} < T_i^{rep}$ then
        $P \leftarrow P + 1$
    end if
end for
$P \leftarrow P/m$
```

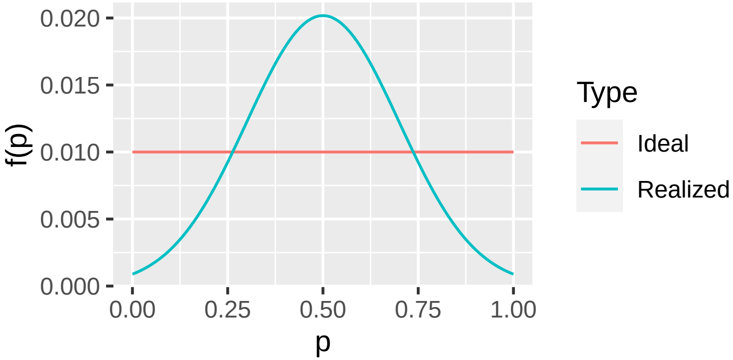
Bayesian p-values for spatial regression example

	F-T	ChiSq	Moran	Zeroes	Tail
GLM	0.00	0.00	0.00	0.50	0.00
GLMM-Simple	0.55	0.55	0.00	0.20	0.90
GLMM-Spatial	0.39	0.42	0.79	0.29	0.61

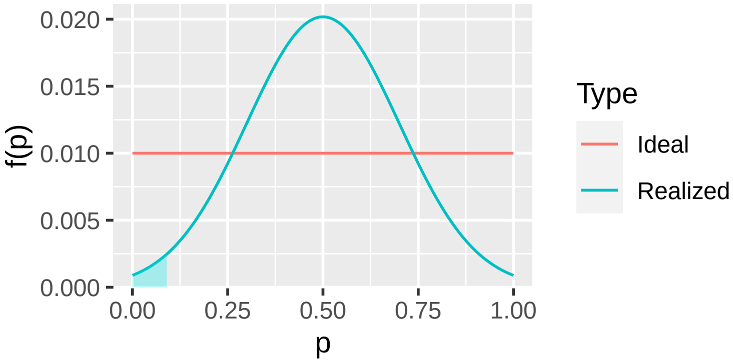
Bayesian P-values: problems with interpretation

Question: If data were repeatedly simulated under the same model that is used for estimation, what distribution of p-values would we hope to get?

Bayesian P-values: problems with interpretation



Bayesian P-values: problems with interpretation



Bayesian p-values are known to be conservative!! An extreme value is indicative of lack-of-fit, but a smallish one (e.g. 0.1) may or may not be problematic (in this case a calibrated p-value is 0.02!!)

Sampled posterior p-values

```
$P \leftarrow 0$
Draw $\boldsymbol{\theta} \sim [\boldsymbol{\theta}|\mathbf{y}]$
for $i \in 1:m$ do
    Draw $\mathbf{y}_i^{rep} \sim [\mathbf{y}|\boldsymbol{\theta}]$
    Calculate $T_i^{rep} = T(\mathbf{y}_i^{rep}, \boldsymbol{\theta})$
    Calculate $T_i^{obs} = T(\mathbf{y}, \boldsymbol{\theta})$
    if $T_i^{obs} < T_i^{rep}$ then
        $P \leftarrow P + 1$
    end if
end for
$P \leftarrow P/m$
```

-Advantage: P-value distribution guaranteed to be uniform

-Disadvantage: Answer depends on random number seed!

Pivotal discrepancy measures (Yuan and Johnson 2012)

Can be used to test lack-of-fit at any stage of a hierarchical model.

Two strategies:

1) Parametric: Use known distributional properties, e.g.,

$Y \sim \mathcal{N}(\mu, \sigma^2) \rightarrow Z = \frac{Y-\mu}{\sigma} \sim N(0,1)$

Here, Z is a pivotal quantity in that it's reference distribution does not depend on μ or σ. Simply keep track of Z and compare to $\mathcal{N}(0,1)$.

Pivotal discrepancy measures

2) Use a probability integral transform (PIT)

Continuous:

$$F(y_{ij}|\boldsymbol{\theta}) \quad \sim \quad U(0,1)$$

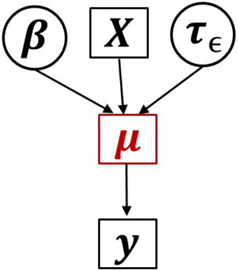
Discrete:

$$\begin{aligned} w_{ij} &\sim U(0,1) \text{ where} \\ w_{ij} &= F(y_{ij}-1|\boldsymbol{\theta}) + u_{ij} f(y_{ij}|\boldsymbol{\theta}) \text{ and} \\ u_{ij} &\sim U(0,1) \end{aligned}$$

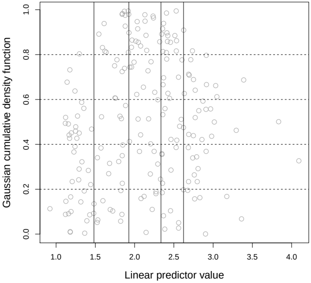
Here w_{ij} is called a "randomized quantile residual" (Dunn and Smyth 1996)

Pivotal discrepancy measures

Application to spatial regression example. PIT test on simple GLMM using a χ^2 test for uniformity,



Bayesian Poisson GLMM
w/ overdispersion only



A χ^2 discrepancy measure using PIT theory using a single sample of $\boldsymbol{\theta}$

The median χ^2 p-value computed in this way (taken across MCMC samples) was 0.37.

Cross validation

Probably the gold standard!! But computationally intensive.

Spatial regression example: K-fold cross validatation with 40 folds of 5 observations each - Simple GLMM model:

-Test for uniformity of empirical CDF: $p = 0.13$

Summary - Bayesian model checking

-Trade off between complexity and performance!

-Posterior predictive p-values, PIT tests are all fast and relatively easy to implement but are conservative. They can tell you when data fit a model terribly, but it is difficult to pinpoint small or moderate lack-of-fit

-Sampled posterior p-values have properly stated p-values but results can differ based on the posterior draw chosen

-Pivotal discrepancy measures allow you to examine fit at different levels of a hierarchical model

-Cross validation tests and calibrated p-values require considerably more investment (running Bayesian analyses interatively)

Bonus!! GOF for integrated population models

"Calibrated simulation" approach (Besbeas and Morgan 2014). Similar in spirit to Bayesian p-values.

1. Fit integrated population model; obtain MLEs, $\hat{\boldsymbol{\theta}}$ and associated variance-covariance matrix $\hat{\boldsymbol{\Sigma}}_{\theta}$
2. For $k = 1, 2, \cdots, n$, simulate data $\mathbf{y}_k \sim [\mathbf{y}|\tilde{\boldsymbol{\theta}}_k]$ where $\tilde{\boldsymbol{\theta}}_k \sim \mathrm{Multivariate\text{-}normal}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}}_{\theta})$
3. Compare $T(\mathbf{y}, \boldsymbol{\theta}_k)$ with $T(\mathbf{y}_k, \boldsymbol{\theta}_k)$ in the same manner as in a Bayesian p-value for each dataset.
4. One can use a simulation study to look at what distribution of p-values one might expect under a "correct" model, and use these for calibration or to select a preferred discrepancy function.

References

Besbeas, Panagiotis, and Byron JT Morgan. 2014. "Goodness-of-Fit of Integrated Population Models Using Calibrated Simulation." *Methods in Ecology and Evolution* 5 (12): 1373–82.

Conn, Paul B, Devin S Johnson, Perry J Williams, Sharon R Melin, and Mevin B Hooten. 2018. "A Guide to Bayesian Model Checking for Ecologists." *Ecological Monographs* 88 (4): 526–42.

Dunn, Peter K, and Gordon K Smyth. 1996. "Randomized Quantile Residuals." *Journal of Computational and Graphical Statistics* 5 (3): 236–44.

Hartig, Florian. 2021. *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*. http://florianhartig.github.io/DHARMa/.

Yuan, Y., and V. E. Johnson. 2012. "Goodness-of-Fit Diagnostics for Bayesian Hierarchical Models." *Biometrics* 68: 156–64.