Bayesian Variable Selection - Generalized Linear and Hierarchical Models

Jim Griffin

University College London

Plan

- GLMs: Approximating marginal likelihood, priors, logistic regression, Poisson regression
- Hierarchical models: Ecological example

Generalized Linear Models (GLMs)

Generalized linear models are defined by

- A linear predictor $\boldsymbol{\eta} = \alpha \mathbf{1}_n + \boldsymbol{X}_\gamma$
- A probability distribution for y_i where $\mu_i = \mathrm{E}[y_i] = g^{-1}(\eta_i)$ for a link function g

Challenges

The approach for linear regression can be extended to GLMs but with the following challenges

- How to define a g-prior?
- The marginal likelihood $p(\boldsymbol{y}|\boldsymbol{X}_\gamma)$ is usually not available in closed form. How can we extend the computational methods?

How to define a g-prior?

The g-prior from linear regression can be used directly but this is no longer the conditionally conjugate prior

A g-prior can be developed from the unit information principle. In a GLM

$$\boldsymbol{\beta}_\gamma \sim \mathrm{N}\left(\boldsymbol{\beta}_{\gamma,0}, nJ_\gamma\left(\boldsymbol{\beta}_{\gamma,0}\right)\right)$$

where $\boldsymbol{\beta}_{\gamma,0}$ is the true value and $J_\gamma\left(\boldsymbol{\beta}_{\gamma,0}\right)$ is observed or expected Fisher information.

This suggests the g-prior $\boldsymbol{\beta}_\gamma \sim \mathrm{N}\left(\boldsymbol{\mu}_\gamma, gJ_\gamma\left(\boldsymbol{\mu}_\gamma\right)\right)$ where $\boldsymbol{\mu}_\gamma$ is either $\mathbf{0}_{p_\gamma}$ or the MLE $\hat{\boldsymbol{\beta}}_\gamma$

Extending computational methods

There are three typical approaches

- Use a data augmentation method with latent variables $\boldsymbol{\nu}$ which allows us to calculate $p(\boldsymbol{y}|\boldsymbol{\nu}, \boldsymbol{X}_\gamma)$ analytically
- Use an approximation of $p(\boldsymbol{y}|\boldsymbol{X}_\gamma)$
- Use reversible jump MCMC (Green, 1995)

Example: Probit regression

The responses $y_i = 0, 1$ are binary and $p(y_i) = \Phi\left(\boldsymbol{X}_\gamma \beta_\gamma\right)$

Introduce latent variables $z_1, \ldots, z_n$ where $z_i \sim \mathrm{N}(\boldsymbol{X}_\gamma \beta_\gamma, 1)$ and $y_i = 1 \iff z_i > 0$ (Albert and Chib, 1993) then $p(\boldsymbol{z}|\boldsymbol{X}_\gamma)$ can be analytically calculated.

Gibbs sampler

- Update γ using an MCMC sampler for linear regression replacing $p(\boldsymbol{y}|\boldsymbol{X}_\gamma)$ by $p(\boldsymbol{z}|\boldsymbol{X}_\gamma)$
- Sample $\beta_\gamma|\boldsymbol{z}, \boldsymbol{X}_\gamma$ and $\boldsymbol{z}|\beta_\gamma, \boldsymbol{X}_\gamma, \boldsymbol{y}$

Data Augmentation

Other data augmentation schemes are

- Logistic regression – Pólya-gamma (Polson et al., 2013)
- Poisson regression (Frühwirth-Schnatter et al., 2009)

Approximation

- Use $\text{BIC}(\gamma) = -2\log L(\gamma) + p_\gamma \log n$ where $L(\gamma)$ is the log likelihood of model γ evaluated at its MLE
- Laplace approximation

$$p(\boldsymbol{y}|\boldsymbol{X}_\gamma) \approx p(\boldsymbol{y}|\boldsymbol{X}_\gamma, \hat{\beta}_\gamma)|\Sigma_\gamma|^{1/2}(2\pi)^{p_\gamma/2}$$

where $\hat{\beta}_\gamma$ is the posterior mode and
$\Sigma_\gamma^{-1} = -\,\Delta_{\beta_\gamma}\Delta_{\beta_\gamma} p(\boldsymbol{y}|\boldsymbol{X}_\gamma, \beta_\gamma)\Big|_{\beta_\gamma=\hat{\beta}_\gamma}$

Comments

- Since the MLE of a GLM usually converges quickly to a normal then the Laplace approximation will often be good
- The Laplace approximation can be used as an importance distribution to avoid bias and integrated into a pseudo-marginal (Andrieu and Roberts, 2009) or conditional pseudo-marginal sampler (Deligiannidis et al., 2018; Liang et al., 2023)
- Adaptive approximate Laplace approximation (AALA) (Liang et al., 2023) (built on the approximate Laplace approximation of Rossell et al., 2021) using the steps
  1. Calculate β_γ using the new design matrix and an estimate of $\boldsymbol{\eta}$ (from MCMC)
  2. Update β_γ with one step of Weighted Least Squares (WLS) and calculate Laplace estimate

Fine mapping of Systemic Lupus Erythematosus

Problem: identify the SNPs that play a crucial role in predicting Systemic Lupus Erythematosus using a case/control study

The data consists of genotypes from a genome-wide genetic case/control association study involving 4035 cases (SLE patients) and 6959 controls (public repository of European ancestry)

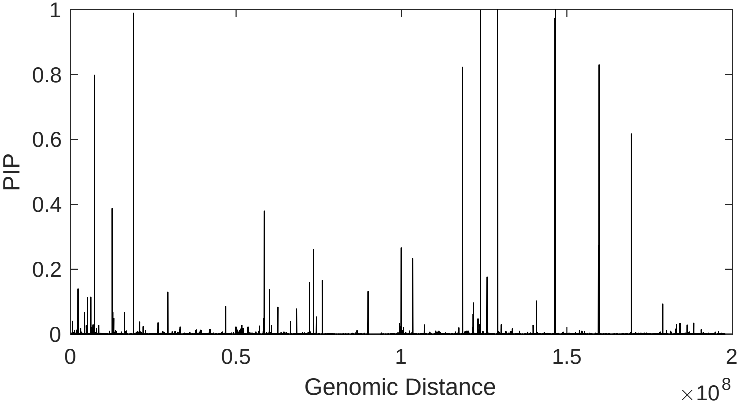
SNPs on 4 Chromosomes: 1 (5771), 3 (42,430), 11 (32,290), 21 (9306)

Liang et al. (2023) use PARNI with an AALA of the marginal likelihood in a correlated pseudo-marginal scheme

Average mean squared errors for inclusion variables

Chromosome	SNPs	ADS-DA	PARNI-CPM	Relative Efficiency
1	5,771	1.84×10^{-5}	7.34×10^{-6}	2.51
3	42,430	2.01×10^{-4}	5.11×10^{-5}	3.94
11	32,290	7.09×10^{-5}	9.89×10^{-6}	7.15
21	9,306	1.18×10^{-5}	1.79×10^{-7}	65.93

Chromosome 3



Hierarchical models

An attraction of MCMC methods for Bayesian variable selection is that they can be easily embedded in larger hierarchical models.

The MCMC scheme for the hierarchical model can be extended to allow for updating of the model indicator γ

This approach allows uncertainty about other model parameters to be propagated to Inference about the included variables

Example: Monitoring of wildlife species

Monitoring wildlife populations is important but challenging:

- Species absence or presence from a particular site cannot be easily verified
- Large scale monitoring is also expensive and slow
- Long-term monitoring of a large number of species can be unsustainable

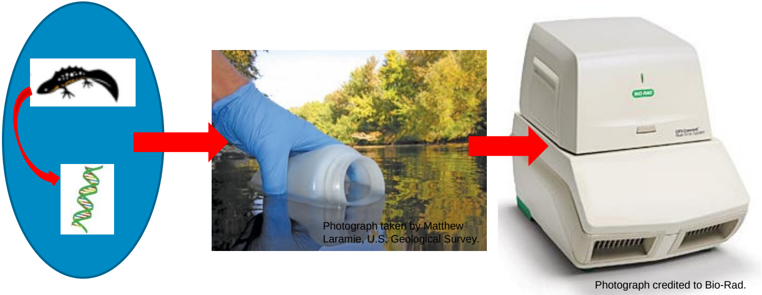
Therefore, ecologists are increasingly replacing human-based sampling methods with technology-based approaches.

eDNA

- Environmental DNA (eDNA) is a survey tool with rapidly expanding applications for assessing presence of a wildlife species at surveyed sites.
- Since the initial proof of concept by Ficetola et al. (2008), the use of eDNA for the assessment of aquatic biodiversity has been rapidly expanding.

eDNA

- eDNA surveys identify recent presence of a species within a waterbody by detecting its shed DNA.



Photograph taken by Matthew Laramie, U.S. Geological Survey.

Photograph credited to Bio-Rad.

eDNA

eDNA surveys are often analysed as presence-absence data.

However, both false positive and false negative errors are possible in the two stages of an eDNA survey:

- the data collection stage (stage 1)
- laboratory analysis stage (stage 2)

We consider single species surveys where quantitative PCR (qPCR) is used to amplify and detect DNA in water samples at stage 2 with species presence at a site often determined using multiple qPCR runs.

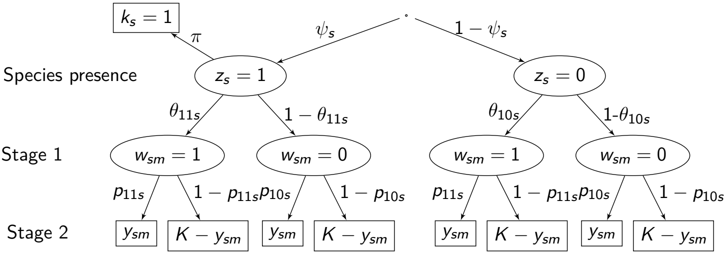
Model with false positives and negatives (Griffin et al., 2020)

- Data is observed at S sites which are each divided into M samples
- There K PCR runs for each sample and $y_{s,m}$ is the number of positives PCR runs
- The model accounts for false positive and false negative errors in both stages of eDNA surveys
- The model is a multi-scale occupancy model, extending the work by Guillera-Arroita et al. (2017)
- In some cases, records that confirm species presence at the site may be available and we show how such records can be incorporated in the model

New model

At site s,

- Observed: $k_s = 1$ if confirmed presence and y_{sm} is number of positive runs from K
- Latent: $z_s = 1/0$ if a species is present/absent and $w_{sm} = 1/0$ if eDNA is present/absent in the m-th sample



Inference

Inference is complicated by likelihood symmetries

Solution	ψ_s	θ_{11}	θ_{10}	p_{11}	p_{10}
1	a	b	c	d	e
2	a	$1-b$	$1-c$	e	d
3	$1-a$	c	b	d	e
4	$1-a$	$1-c$	$1-b$	e	d

These likelihood symmetries make the model only locally identifiable

Bayesian inference

We use Bayesian inference and place a pairs of parameters (*e.g.* θ_{11} and θ_{10} or p_{11} and p_{10})

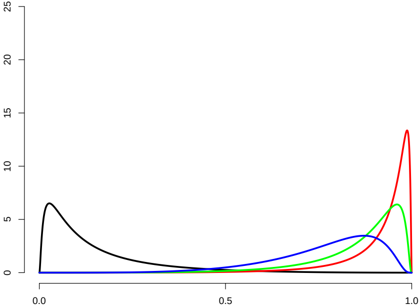
We propose a prior distribution that give a large prior probability that one parameter is larger than another, *e.g.* $\theta_{11} > \theta_{10}$

$$\text{logit}(\theta_{11}) \sim \text{N}\left(\text{logit}(a), \frac{(\text{logit}(a) - \text{logit}(b))^2}{2(\Phi^{-1}(\epsilon))^2}\right)$$

$$\text{logit}(\theta_{10}) \sim \text{N}\left(\text{logit}(b), \frac{(\text{logit}(a) - \text{logit}(b))^2}{2(\Phi^{-1}(\epsilon))^2}\right)$$

Then a and b are the medians of θ_{11} and θ_{10} respectively and $p(\theta_{11} < \theta_{10}) = \epsilon$.

Prior distribution for θ_{11} and θ_{10}



$b = 0.1$, $\epsilon = 0.025$
θ_{10} – Black
θ_{11} – $a = 0.8$ (Blue), $a = 0.9$ (Green), $a = 0.95$ (Red).

Bayesian inference - covariates

We extends this prior to regression by choosing a prior distribution with maintains the ordering averaging over all sites.

This allows us to:

- estimate the probability of species presence at a site without requiring additional sources of information
- perform effective variable selection

We use a Gibbs sampler updating the latent variables and using Pólya-Gamma (Polson et al., 2013) for efficient computation in each logistic regression

Great crested newt data

- Samples were collected as part of a national distribution modelling assessment for great crested newts, commissioned by Natural England.
- Surveyors were also asked to collect information on additional pond-specific environmental covariates, which we consider as potential predictors for species presence as well as the probabilities of error at the two stages.
- We have $M = 1$ samples at $S = 189$ sites. There were $K = 12$ qPCR runs.

Results

Parameter	Posterior mean	95% posterior credible interval
ψ	0.14	(0.04, 0.42)
θ_{11}	0.73	(0.45, 0.79)
θ_{10}	0.15	(0.05, 0.27)
p_{11}	0.81	(0.71, 0.90)
p_{10}	0.05	(0.03, 0.07)

- We did not identify any covariates that are linked to the probability of species presence, ψ, or to the probabilities of a stage 1 error, as they all have PIP well below 50%.
- On the other hand, four covariates with PIP $>$ 50% have been identified for p_{11} (maximum pond depth, PIP: 1.00, and pond length, PIP: 0.63, presence of macrophytes, PIP: 0.71 and pond density, PIP: 0.91) and one for p_{10} (fish presence, PIP: 0.97).
- Maximum pond depth and presence of macrophytes have a positive effect on stage 2 true positive probability, while pond length and pond density have a negative effect. Finally, the presence of fish decreases the probability of a stage 2 false positive result.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data, *JASA* **88**: 669–679.

Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations, *Annals of Statistics* **37**: 697–725.

Deligiannidis, G., Doucet, A. and Pitt, M. K. (2018). The correlated pseudo marginal method, *JRSSB* **80**: 839–870.

Frühwirth-Schnatter, S., Frühwirth, R., Held, L. and Rue, H. (2009). Improved auxiliary mixture sampling for hierarchical models of non-Gaussian data, *Statistics and Computing* **19**: 479.

Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination, *Biometrika* **82**: 711–732.

Griffin, J. E., Matechou, E., Buxton, A. S., Bormpoudakis, D. and Griffiths, R. A. (2020). Modelling environmental DNA data; Bayesian variable selection accounting for false positive and false negative errors, *JRSS C* **69**: 377–392.

Guillera-Arroita, G., Lahoz-Monfort, J. J., Rooyen, A. R., Weeks, A. R. and Tingley, R. (2017). Dealing with false-positive and false-negative errors about species occurrence at multiple levels, *Methods in Ecology and Evolution* **8**: 1081–1091.

Liang, X., Livingstone, S. and Griffin, J. (2023). Adaptive MCMC for Bayesian variable selection in generalised linear models and survival models, *Entropy* **25**.

Polson, N. G., Scott, J. G. and Windle, J. (2013). Bayesian inference for logistic models using pølya-gamma latent variables, *JASA* **108**: 1339–1349.

Rossell, D., Abril, O. and Bhattacharya, A. (2021). Approximate Laplace approximations for scalable model selection, *JRSS B* **83**: 853–879.