Bayesian Tools for Synthesis of Ecological Data

Introduction

Tom Hobbs

December 2, 2022

Department of Ecosystem Science and Sustainability

National Science Foundation Award DEB 1753860

"Big data" in ecology

CONCEPTS AND QUESTIONS

156

Big data and the future of ecology

Stephanie E Hampton[1*], Carly A Strasser[2], Joshua J Tewksbury[3], Wendy K Gram[4], Amber E Budden[5], Archer L Batcheller[6], Clifford S Duke[7], and John H Porter[8]

The need for sound ecological science has escalated alongside the rise of the information age and "big data" across all sectors of society. Big data generally refer to massive volumes of data not readily handled by the usual data tools and practices and present unprecedented opportunities for advancing science and informing resource management through data-intensive approaches. The era of big data need not be propelled only by "big science" – the term used to describe large-scale efforts that have had mixed success in the individual-dri
effort – a larg
encourage eco
and societal pr
scientists who
they be ecolog

Front Ecol Environ 201

In the 21st centur
information age
fields of the life sci
logy, and medicine, l

GUEST EDITORIAL GUEST EDITORIAL GUEST EDITORIAL

3

Macrosystems ecology: big data, big ecology

Ecologists are increasingly confronted by questions that, in one way or another, involve analysis or prediction across vast geographic areas or time periods. There is little doubt that many of the problems facing environmental systems have broad-scale components. These problems range from understanding the spatial distributions of invasive species to discerning how the local ecology of forests interacts with regional fire patterns to influence continental fluxes of carbon. Although ecologists have been successful at answering research questions and developing theory at fine scales, they are now rapidly adding new techniques to their toolkit that facilitate the study of broad-scaled regional processes, and interactions with fine-scaled and global processes. This is where "macrosystems ecology" (MSE) fits in.

The papers in this Special Issue were prepared by participants in the US National Science Foundation's MacroSystems Biology program. A common theme throughout most of these articles is a seemingly simple but challenging topic – data! Specifically, it's the data required to study large, complicated, and highly variable objects typical of macrosystems research. The amount of data involved in MSE research is far beyond that which a single research lab can collect and process. What then are the options available to ecologists for conducting data-intensive research if they clearly cannot collect, process, or analyze it all on their own? At least some ecologists will have to develop the concepts and methodology for studying ecological systems at broad scales; revitalize the culture in which they work to be even more collaborative, open, and interdisciplinary than it already is; and embrace the era of "big data".

To date, ecologists have used any of four strategies for acquiring ecological big data: (1) *Collate existing small but information-rich datasets to create spatially, temporally, and thematically extensive datasets.* This strategy is extremely difficult, is unexpectedly expensive, and can result in datasets with geographic or temporal gaps. (2) *Compile data from remote-sensing platforms that are spatially and often temporally extensive.*

Patricia A Soranno
Department of Fisheries and Wildlife, Michigan State University, East Lansing, MI

The *really* big data

Results of Web of Science search using ecol* for publication title

Objective of videos

Learn Bayesian methods for synthesizing existing data and published findings to gain new insight in ecology

Prerequisites

- Basic understanding of Bayesian inference, including derivation of Bayes theorem
- Familiar with common statistical distributions, particularly normal, lognormal, beta, gamma, Poisson, negative binomial.
- Understand directed acyclic graphs and their relationship to the factored joint distribution
- Ability to write expressions for the posterior distribution and the fully factored joint distribution in proper statistical notation
- Grasp essential features of Markov chain Monte Carlo algorithm
- Familiar with basic coding in JAGS, Open Bugs, or Stan

Outline of topics

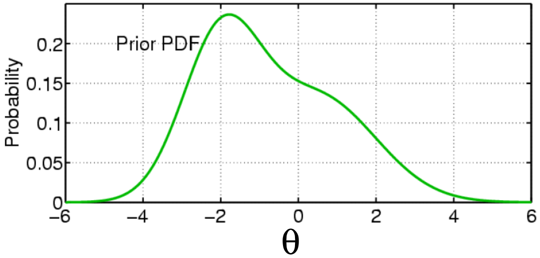
1. This video
   1.1 Review of components of Bayesian inference
   1.2 Why use informed priors?
   1.3 A problem using published means and standard deviations
2. Video 2: Moment matching
3. Video 2: Developing priors from multiple studies
4. Video 4: Hierarchical analysis of data sets from multiple studies

The components of Bayes theorem

$$\overbrace{[\theta|y]}^{\text{Posterior}} = \frac{\overbrace{[y|\theta]}^{\text{likelihood}} \; \overbrace{[\theta]}^{\text{prior}}}{\underbrace{\int_\theta [y|\theta]\,[\theta]\,d\theta}_{\text{marginal distribution of data}}}$$

$$[\theta|y] = \frac{[y|\theta]\,[\theta]}{\int_\theta [y|\theta]\,[\theta]\,d\theta}$$

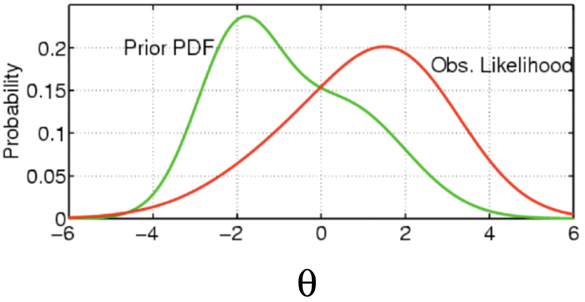
The prior, $[\theta]$, can be informative or vague. [1]



[1] Drawings courtesy of Chris Wikle, University of Missouri

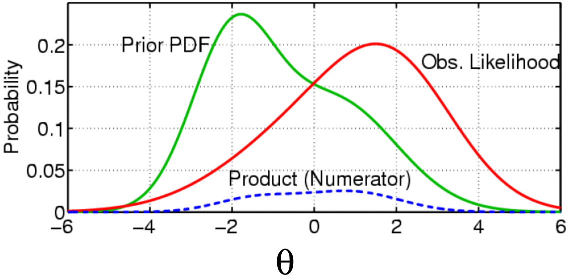
$$[\theta|y] = \frac{[y|\theta][\theta]}{\int_\theta [y|\theta][\theta]\, d\theta}$$

The likelihood (a.k.a. data distribution, $[y|\theta]$)



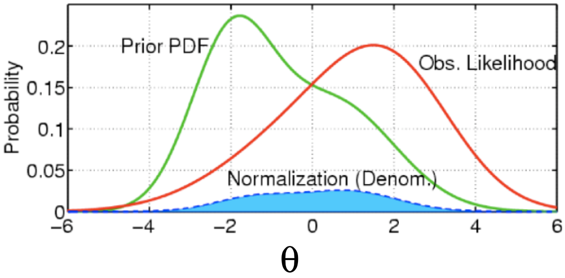
$$[\theta|y] = \frac{[y,\theta]}{[y]} = \frac{[y|\theta][\theta]}{\int_\theta [y|\theta]\,[\theta]\,d\theta}$$

The product of the prior and the likelihood, $[y|\theta][\theta]$, the joint distribution of the parameters and the data, $[y,\theta]$.



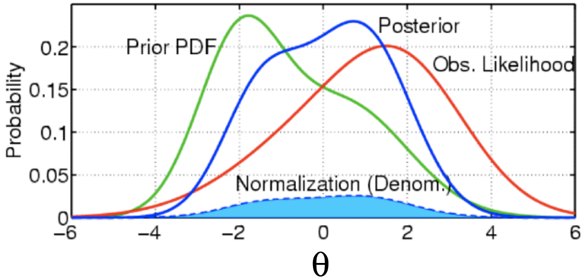
$$[\theta|y] = \frac{[y|\theta][\theta]}{\int_\theta [y|\theta]\,[\theta]\,d\theta}$$

The marginal distribution of the data (the denominator) is the area under the joint distribution.



What we are seeking: The posterior distribution, $[\theta|y]$.

$$[\theta|y] = \frac{[y|\theta][\theta]}{\int_\theta [y|\theta][\theta]\,d\theta}$$



Note that we are dividing each point on the dashed line by the area under the dashed line to obtain a probability density function reflecting our prior and current knowledge about θ.

What are informative priors?

Informative priors are distributions that are not diffuse relative to the posterior. These distributions may be based on

- statistics reported in the literature
- posterior distributions from previous studies
- meta-analyses
- "plausible" assumptions

Why use informed priors?

- They enhance insight by combing information from multiple sources
- They speed convergence of MCMC.
- They reduce problems of indetifyability
- They may allow estimation of quantities that would be impossible to estimate with vague priors.

The concept of support

The support of the random variable z is includes all values of z such that

$$[z] > 0.$$

Do not confuse this definition of support with the concept of support from maximum likelihood.

Support of parameter dictates distribution for prior

Example data	Likelihood	Parameter	Support of parameter	Distribution for prior
Counts of occupied sites	binomial or Bernoulli	Probability occupancy	$0 \to 1$	beta
Species richness	Poisson or negative binomial	mean	continuous $0 \to +\infty$	gamma
Above ground net primary production	gamma or lognormal	mean	continuous $0 \to +\infty$	gamma or lognormal
Water balance	normal	mean	continuous $-\infty \to +\infty$	normal
Regression coefficients	normal	mean	continuous $-\infty \to +\infty$	normal

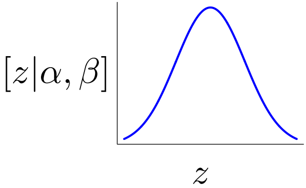
What do we usually get from the literature?

				Quantile	
Parameter	Mean	Median	SD	0.025	0.975
β	1.9	1.9	0.16	1.5	2.2
f_n	0.77	0.77	0.045	0.68	0.85
f_p	0.54	0.54	0.059	0.43	0.66
f_c	0.18	0.17	0.088	0.045	0.38
m	0.47	0.47	0.041	0.39	0.55
p_1	0.95	0.96	0.044	0.84	1.00
p_2	0.89	0.88	0.023	0.84	0.93
p_3	0.93	0.93	0.039	0.84	0.99
ψ	0.031	0.027	0.021	0.004	0.082
σ_p	0.21	0.21	0.029	0.16	0.27
v	0.099	0.081	0.075	0.0068	0.29

Notes: Definitions are: β, the continuous time rate of frequency-dependent transmission (yr^{-1}); f_n, number of juveniles recruited per susceptible adult female; f_p, number of juveniles recruited per recovered adult female; f_c, number of juveniles recruited per infected and infectious adult female; m, sex ratio of juveniles surviving to yearlings; ψ, probability of recrudescence; p_1, juvenile survival probability; p_2, adult and yearling female survival probability; p_3, yearling and adult male survival probability; σ_p, process standard deviation; v, probability of vertical transmission.

The problem

All distributions have parameters:



α and β are parameters of the distribution of the random variable z.

Types of parameters

Parameter name	Function
intensity, centrality, location	sets position on x axis
shape	controls dispersion and skew
scale, dispersion parameter	shrinks or expands width
rate	scale^{-1}

The problem: How do we find parameters using tabulated means and standard deviations?

The normal and the Poisson are the only distributions for which the parameters of the distribution are the mean and the variance. The parameters of all other distributions are *functions* of the mean and the variance.

$$
\begin{aligned}
\alpha &= f_1(\mu, \sigma^2) \\
\beta &= f_2(\mu, \sigma^2)
\end{aligned}
$$

How do we find parameters given published means and variances?

Take home

Prior distributions are powerful tools for synthesizing results from ecological studies. Using them reliably requires choosing finding parameters for distributions based on published means and standard errors.