What Is a Prior and How to Derive One?

Song Qian

Department of Environmental Sciences
The University of Toledo

September 27, 2018
ICEI 2018, Jena, Germany

Scientists versus Engineers

- Questions asked: “Why?” versus “How?”

Scientists versus Engineers

- Questions asked: “Why?” versus “How?”
- Reasoning mode: induction versus deduction

Scientists versus Engineers

- Questions asked: “Why?” versus “How?”
- Reasoning mode: induction versus deduction
- An uncertain answer versus a definite answer

Scientists versus Engineers

- Questions asked: “Why?” versus “How?”
- Reasoning mode: induction versus deduction
- An uncertain answer versus a definite answer
- Limited versus “unlimited” chances of experimentation

Two Branches of Statistics

- Definition of probability: long-run frequency versus degree of belief
- Neyman and Pearson (1933) – two approaches of testing statistical hypotheses
  - Thomas Bayes – “probabilities *a posteriori* of the possible causes of a given event”
  - Bertrant and Borel – hypothetical deduction
    - “no test of this kind could give reliable result”
    - useful with properly selected *en quelque sorte remarquable* character of data
- Frequentist’s versus Bayesian statistics

Bayesian Statistics

Compared to the long-run frequency view of statistics, Bayesian –

- is coherent: everything under one equation $p(\theta|y) \propto p(\theta)p(y|\theta)$

Bayesian Statistics

Compared to the long-run frequency view of statistics, Bayesian –

- is coherent: everything under one equation $p(\theta|y) \propto p(\theta)p(y|\theta)$
- is consistent with how we think about chance –

Bayesian Statistics

Compared to the long-run frequency view of statistics, Bayesian –

- is coherent: everything under one equation $p(\theta|y) \propto p(\theta)p(y|\theta)$
- is consistent with how we think about chance –

Bayesian Statistics

Compared to the long-run frequency view of statistics, Bayesian –

- is coherent: everything under one equation $p(\theta|y) \propto p(\theta)p(y|\theta)$
- is consistent with how we think about chance –

"Pr(Trump will make Merkel great again)" is meaningful

Bayesian Statistics

Compared to the long-run frequency view of statistics, Bayesian –

- is coherent: everything under one equation $p(\theta|y) \propto p(\theta)p(y|\theta)$
- is consistent with how we think about chance –

"Pr(Trump will make Merkel great again)" is meaningful

Bayesian Statistics

Compared to the long-run frequency view of statistics, Bayesian –

- is coherent: everything under one equation $p(\theta|y) \propto p(\theta)p(y|\theta)$
- is consistent with how we think about chance –

"Pr(Bavaria wins tomorrow) = 0.74" is meaningful

Bayesian Statistics

Compared to the long-run frequency view of statistics, Bayesian –

- is coherent: everything under one equation $p(\theta|y) \propto p(\theta)p(y|\theta)$
- is consistent with how we think about chance –

"Pr(Bavaria wins tomorrow) = 0.74" is meaningful

- results are easy to interpret – no more *p*-value, nor "statistical significance"
- but needs a *prior* distribution.

What is a Prior?

The non-normative definition:

- A prior of an uncertain quantity is the probability distribution that would express one's beliefs about the quantity – Wikipedia

What is a Prior?

The non-normative definition:

- A prior of an uncertain quantity is the probability distribution that would express one's beliefs about the quantity – Wikipedia
  - We do not usually define our uncertainty using probability distribution

What is a Prior?

The non-normative definition:

- A prior of an uncertain quantity is the probability distribution that would express one's beliefs about the quantity – Wikipedia
  - We do not usually define our uncertainty using probability distribution
  - We use probability, but we are bad at getting the probability right (Human judgment relies on heuristics and is often biased. Tversky and Kahneman (1974) *Science* 185:1124-1131).

When Prior Is Known

- The Bayes estimator is the best with respect to Bayes risk

When Prior Is Known

- The Bayes estimator is the best with respect to Bayes risk
- The difficulty used to be computation

When Prior Is Known

- The Bayes estimator is the best with respect to Bayes risk
- The difficulty used to be computation
- MCMC resolved the computation problem

When Prior Is Known

- The Bayes estimator is the best with respect to Bayes risk
- The difficulty used to be computation
- MCMC resolved the computation problem
    - Adrian F.M. Smith (now Sir Adrian Smith) quit statistics because “all the Bayesian problems are solved.”

But How Do We Derive a Prior?

- The personal belief definition is of no help

But How Do We Derive a Prior?

- The personal belief definition is of no help
- There are numerous papers on how to derive a prior that does not contain real information

But How Do We Derive a Prior?

- The personal belief definition is of no help
- There are numerous papers on how to derive a prior that does not contain real information
  - Jeffery's prior – $\pi(\mu, \sigma^2) \propto 1/\sigma^2$

But How Do We Derive a Prior?

- The personal belief definition is of no help
- There are numerous papers on how to derive a prior that does not contain real information
  - Jeffery's prior – $\pi(\mu, \sigma^2) \propto 1/\sigma^2$
  - Uninformative (flat) prior – $N(0, 100)$, $gamma(0.001, 0.001)$

But How Do We Derive a Prior?

- The personal belief definition is of no help
- There are numerous papers on how to derive a prior that does not contain real information
  - Jeffery's prior – $\pi(\mu, \sigma^2) \propto 1/\sigma^2$
  - Uninformative (flat) prior – $N(0, 100)$, $gamma(0.001, 0.001)$
  - Reference prior –

But How Do We Derive a Prior?

- The personal belief definition is of no help
- There are numerous papers on how to derive a prior that does not contain real information
  - Jeffery's prior – $\pi(\mu, \sigma^2) \propto 1/\sigma^2$
  - Uninformative (flat) prior – $N(0, 100)$, $gamma(0.001, 0.001)$
  - Reference prior –
- They are either difficult to derive or informative in someways.

But How Do We Derive a Prior?

- The personal belief definition is of no help
- There are numerous papers on how to derive a prior that does not contain real information
  - Jeffery's prior – $\pi(\mu, \sigma^2) \propto 1/\sigma^2$
  - Uninformative (flat) prior – $N(0, 100)$, $gamma(0.001, 0.001)$
  - Reference prior –
- They are either difficult to derive or informative in someways.
  - For example, Gelman (2004)

An Informative Prior is Essencial

- Prior times likelihood is proportional to the posterior
- Without the prior, the posterior is just the likelihood
- Updating is the key

Stein's Paradox

Stein's Paradox

- Efron and Morris (1977) "Stein's Paradox in Statistics" *Scientific American*
  - A class of shrinkage estimators out perform both MLE and Bayes estimator
  - An empirical Bayes interpretation of Stein's paradox
  - The batting average example
  - Qian et al (2015) ES&T 49: 5913-20.

A Normative Definition of a Prior

- Prior distribution of an uncertain parameter for a population is the distribution of the same parameter across similar (exchangeable) populations
- Examples:
  - Prior of a baseball player's batting average – distribution of batting averages of all players in the same league;
  - Prior of the mean phosphorus concentration of Sandusky River – distribution of mean phosphorus concentrations in all similar sized streams in Lake Erie watershed;
  - Expert opinion: a summary of life long observations on the same parameter.

Prior Distribution as an “Among Group” Distribution

- Group can be spatial, temporal, or organizational;
- Deriving prior is a process of assembling and analyzing data from similar “groups”

Prior Distribution as an “Among Group” Distribution

- Group can be spatial, temporal, or organizational;
- Deriving prior is a process of assembling and analyzing data from similar “groups”
- Models based on cross-sectional data are basis for informative prior for a specific site

A Bayesian Hierarchical Modeling Approach

- The hyper-parameter distribution

A Bayesian Hierarchical Modeling Approach

- The hyper-parameter distribution
  - $$\begin{aligned} y_{ij} &\sim N(\mu_j, \sigma_1^2) \\ \mu_j &\sim \underline{N(\mu, \sigma_2^2)} \end{aligned}$$

A Bayesian Hierarchical Modeling Approach

- The hyper-parameter distribution
  - $$\begin{aligned} y_{ij} &\sim N(\mu_j, \sigma_1^2) \\ \mu_j &\sim \underline{N(\mu, \sigma_2^2)} \end{aligned}$$
  - $$\begin{aligned} y_{ij} &\sim N(X\beta_j, \sigma_1^2) \\ \boldsymbol{\beta_j} &\sim \underline{MVN(\mu_\beta, \Sigma)} \end{aligned}$$

Overview

- Qian, et al (2015) ES&T 49: 5913-20
- Monitoring data of TP from streams in the Lake Erie watershed from USGS
- Streams grouped by US EPA's nutrient ecoregion
- One site in NY had few data points (with $>$ 50% censorship)
- Data from rest of the sites used to develop priors

The Hierarchical Model

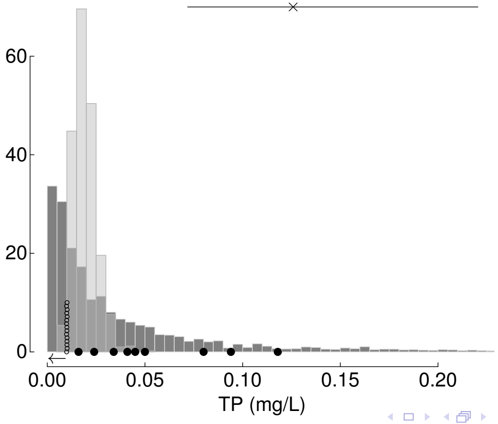
- TP concentrations were modeled using a multilevel model

$$
\begin{array}{rcl}
y_{ijkl} & = & \beta_0 + \beta_{1i} + \beta_{2j} + \beta_{3k} + \epsilon_{ijkl} \\
\beta_{1i} & \sim & N(0, \sigma^2_{\beta_1}) \\
\beta_{2ji} & \sim & N(0, \sigma^2_{\beta_2}) \\
\beta_{3k} & \sim & N(0, \sigma^2_{\beta_3})
\end{array}
$$

- *ijkl*: *l*th observation is in *i*th site, *j*th year, and *k*th season.
- Estimated parameters are used to develop priors
  - Assuming σ^2's follow an inverse-gamma distribution

Bayesian Updating

TP distribution for the NY site – updated using informative priors derived using data from sites in the same ecoregion



The Chla - Nutrient Relationship

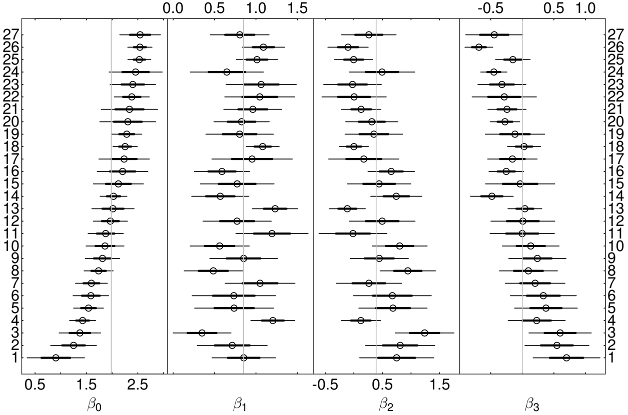
- Qian, et al. (2018) (submitted to ES&T)
- Establishing nutrient criteria using data from multiple lakes
  - Why nutrient criteria should be lake-specific
  - The role of cross-lake data – derive a common prior

$$\begin{aligned}\log(chla_{ij}) &= \beta_{0j} + \beta_{1j}\log(TP_{ij}) + \beta_{2j}\log(TN_{ij}) + \\ &\quad \beta_{3j}\log(TP_{ij})\log(TN_{ij}) + \varepsilon_{ij}\end{aligned}$$

$$\begin{pmatrix}\beta_{0j}\\ \beta_{1j}\\ \beta_{2j}\\ \beta_{3j}\end{pmatrix} \sim MVN\left[\begin{pmatrix}\mu_{\beta_0}\\ \mu_{\beta_1}\\ \mu_{\beta_2}\\ \mu_{\beta_3}\end{pmatrix}, \Sigma\right]$$

- The hyper-parameter distribution (RHS) is the prior for individual lates.

Spatial Variation



Bayesian Updating

- Updating is not limited to similar data collection methods
- Updating can be sequential – we can gradually refine the model
- Maintain a robust sampling program for model updating over time –
  - improving the model
  - detecting temporal changes

Summary

- Bayesian method is always better, provided we have the right prior

Summary

- Bayesian method is always better, provided we have the right prior
- James-Stein estimator and empirical Bayes – prior distribution is “among-group” distribution

Summary

- Bayesian method is always better, provided we have the right prior
- James-Stein estimator and empirical Bayes – prior distribution is “among-group” distribution
- The Bayesian hierarchical modeling

Summary

- Bayesian method is always better, provided we have the right prior
- James-Stein estimator and empirical Bayes – prior distribution is “among-group” distribution
- The Bayesian hierarchical modeling
- Models based on cross-sectional data should be considered as prior models

Summary

- Bayesian method is always better, provided we have the right prior
- James-Stein estimator and empirical Bayes – prior distribution is "among-group" distribution
- The Bayesian hierarchical modeling
- Models based on cross-sectional data should be considered as prior models
- Inference about individual "group" should be based on the posterior derived using group-specific data

A Beginning, Rather Than the End

- Developing a prior – the first step of a modeling work
- Group-specific inference – posterior
- A practice dictated by the Simpson's paradox

Acknowledgment

- Craig Stow, Laura Steinberg, Mike Messner, Bob Miltner