A New Bayesian Approach to Combining Different Species Data

Leonid Kopylev and Junyong Park

US EPA, Office of Research and Development, National Center for Environmental Assessment – Washington

University of Maryland at Baltimore County, Department of Mathematics and Statistics and ORISE Faculty Fellow at US EPA/ORD/NCEA-Washington

Purpose and Scope

NRC (2014) recommended that IRIS develop the capacity to do Bayesian modeling of chemical hazards. In particular, NRC stated that "...more sophisticated Bayesian approaches have been proposed for combining dose-response estimates for multiple species and multiple chemicals (DuMouchel and Harris 1983; Jones et al. 2009). Those approaches might also be useful to EPA if guidance for selection of appropriate models and priors is developed."

In this research and development effort, EPA evaluated DuMouchel and Harris (1983) approach, developed alternative approach and applied it to the data in Jones et al. 2009

Background

DuMouchel and Harris (1983, JASA) were the first authors that addressed the problem of combining information for multiple species with a non-simplistic approach.

- Proposed a Bayesian approach to interspecies extrapolation
- Special attention to combining dose-response information
- Realized that they need subject matter expertise

DuMouchel and Harris (1983) realized that

- the species do not need to be restricted to humans and animals
- any type of data (including cell potency) is appropriate.
- a lot of toxicological experience is needed to figure out what chemicals dose-response information is combined.

Their ANOVA structure, however, assumes constant relative potency across species, which may not be the case in many examples.

A suggested model: Gaussian graphical model

Example (Jones et al. 2009; Low birth weight)

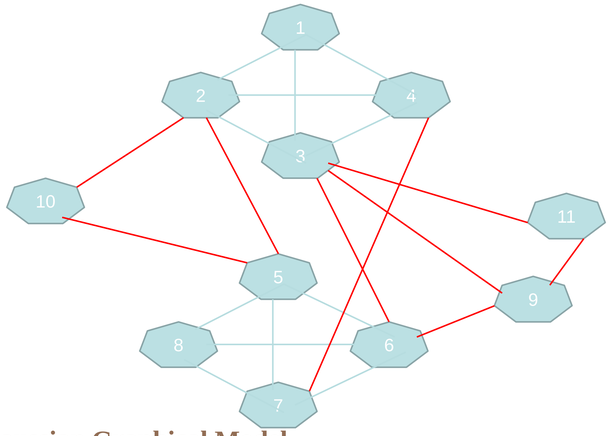
	Total THMs	Chloroform	BDCM	DBCM	Bromoform
Humans	1	2	3	4	12 (missing)
Rat S-D		5	6	7	8
Rat S-F			9		
Rabbit D-B		10			
Rabbit N-Z			11		

- Cell 1 - Cell 11 have the slope of regression model from log(dose) and log(response)
- Empty cells represent no-data

Assumptions

- We assume that species are related for the same chemical and different chemicals are related for the same species.
- We model dependence or relationship among different species and different chemicals through edges in Gaussian graphical model.
- We need to control dependence through prior probabilities for edges based on scientific knowledge rather than subjective choice.

Graphical representation of the Example



Gaussian Graphical Model

- Gaussian graphical model uses the inverse of covariance matrix called a precision matrix.
- Each component in a precision matrix represents the partial correlation between two nodes in the graphical model. Red edges shown for the same chemical data.
- No-edge between two nodes in graphical model is equivalent to partial correlation equal to 0.
- If there is no edge between cells i and j and the correlation is non-zero, then the nonzero correlation is due to all other data.

Formulation of the Bayesian graphical model

y: observed data

$$y \mid \beta, \Sigma \approx N(\beta, \Sigma)$$

Σ: Hyper Inverse Wishart Distribution

$e_{ij} \approx Bernoulli(p_{ij})$: edge between two nodes

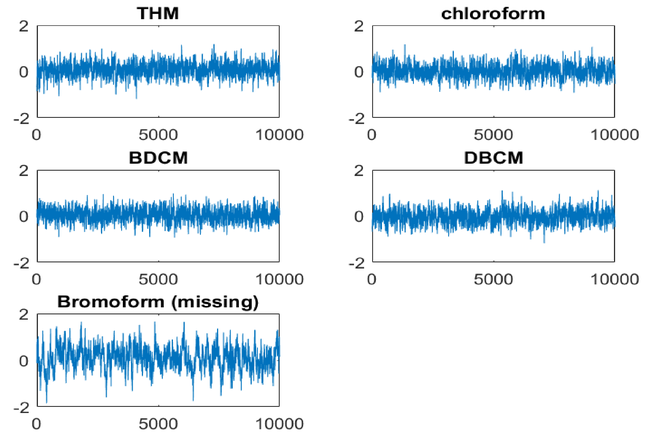
Prior probabilities on edges (representing existence of partial correlations)

- We give high prior probabilities to edges when two nodes have a close relationship.
- Such prior probabilities are still subjective, so they should be determined based on scientific knowledge to minimize subjectivity.

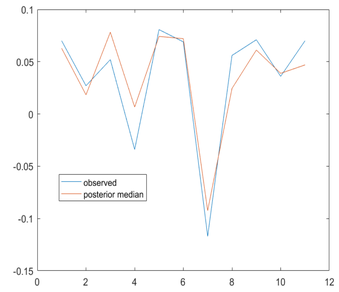
Results

95% Confidence Intervals of Human Data

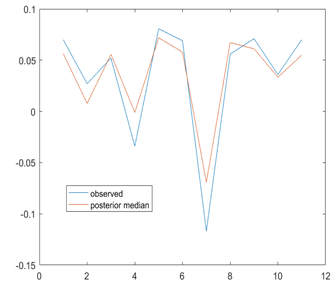
THM	Chloroform	BDCM	DBCM	Bromoform (missing)
0.0855	0.0234	0.0481	-0.0235	0.0878
(-.459,.652)	(-.536,.576)	(-.490,.578)	(-.571,.559)	(-.856,.998)



Validation of the Proposed method (Cross Validation)



Human data 1=THM, 2=Chloroform, 3=BDCM (assumed to be missing), 4=DBCM



Human data 1=THM, 2=Chloroform, 3=BDCM, 4=DBCM(assumed to be missing)

Comments on Results

- Estimate of missing value has more variation.
- When BDCM (or DBCM) is assumed to be missing, the posterior median of predicted values of human BDCM (or DBCM) is close to the observed value of BDCM (or DBCM).
- The patterns of the posterior medians are similar to those of the observed data.

Discussion and Future Directions

- We followed NRC (2014) recommendations on using Bayesian analysis and specifically investigated methodology proposed by DuMouchel and Harris (1983) and Jones et al. (2009).
- We proposed a new Bayesian method and validated recovery of missing human dose-response using Jones et al. (2009) data
- We will use simulation studies to validate our new method and consider its application to additional real data sets.
- We will consider extending the idea of graphical model to the area of combining DNA or RNA sequence data generated from different species.
- We will also consider application of the methodology to more data-poor examples that are more common in IRIS assessment work.

The views expressed in this poster are those of the authors and do not necessarily represent the views or the policies of the U.S. Environmental Protection Agency