Supplement for: A Comparison of the β-substitution Method and a Bayesian Method for Analyzing Left-Censored Data

Tran Huynh[1], Harrison Quick[2], Gurumurthy Ramachandran[1*], Sudipto Banerjee[2], Mark Stenzel[3], Aaron Blair[4], Dale P. Sandler[5], Lawrence S. Engel[5,6], Richard K. Kwok[5], and Patricia A. Stewart[7]

[1] Division of Environmental Health Sciences, University of Minnesota, Minneapolis, MN US
[2] Division of Biostatistics, University of Minnesota, Minneapolis, MN US
[3] Exposure Assessment Applications, LLC, Arlington, VA US
[4] National Cancer Institute, Gaithersburg, MD US
[5] Epidemiology Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC US
[6] Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC US
[7] Stewart Exposure Assessments, LLC, Arlington, VA US
[*] Address correspondence to: Gurumurthy Ramachandran, University of Minnesota, Division of Environmental Health Sciences, School of Public Health, Minneapolis, MN, 420 Delaware St. SE, Minneapolis, MN 55455, USA; *email:* ramac002@umn.edu

This is the supplement for "A Comparison of the β-substitution Method and a Bayesian Method for Analyzing Left-Censored Data" by Huynh et al. Appendix A is a technical appendix detailing the use of PDFs and CDFs when modeling left-censored data. This appendix assumes a knowledge in Bayesian inference, specifically the construction of full-conditional distributions and their use in a Gibbs sampling algorithm for estimating the posterior distribution of various model parameters.

A The Use of PDFs vs. CDFs

In equation (2) of the main manuscript, we use the PDF of a normal distribution to represent the nondetects' contribution to the likelihood, rather than the CDF as is typical in classical maximum likelihood (ML) methods. To illustrate why this was done, we first note the following:

$$\begin{aligned}\int_{-\infty}^{\infty}\phi(x\,|\,\boldsymbol{\theta})I\left\{x\leq \text{LOD}\right\}dx &= \int_{-\infty}^{LOD}\phi(x\,|\,\boldsymbol{\theta})I\left\{x\leq \text{LOD}\right\}dx + \int_{LOD}^{\infty}\phi(x\,|\,\boldsymbol{\theta})I\left\{x\leq \text{LOD}\right\}dx\\ &= \int_{-\infty}^{LOD}\phi(x\,|\,\boldsymbol{\theta})\times(1)dx + \int_{LOD}^{\infty}\phi(x\,|\,\boldsymbol{\theta})\times(0)dx\\ &= \int_{-\infty}^{LOD}\phi(x\,|\,\boldsymbol{\theta})dx = \Phi(\text{LOD}\,|\,\boldsymbol{\theta}) \qquad \text{(A.1)}\end{aligned}$$

where $\Phi(x\,|\,\boldsymbol{\theta})$ and $\phi(x\,|\,\boldsymbol{\theta})$ denote the CDF and the PDF of a normal distribution evaluated at the value x given parameters, $\boldsymbol{\theta}$, respectively. That is, (A.1) shows that if we integrate the hierarchical model in (2) with respect to the nondetects, we will achieve the same likelihood as used in standard ML methods.

The question that remains now is "Why would we want to use the PDFs?" To answer this question, we demonstrate in the following subsections how the hierarchical model would be evaluated using the PDFs and using the CDFs. In these subsections, we will let $\Phi(x)$ and $\phi(x)$ denote the CDF and the PDF of a standard normal distribution evaluated at the value x and let D_i be an indicator variable of the form

$$D_i = \begin{cases} 0, & \text{if } Y_i \leq \text{LOD (i.e., ``nondetected'')} \\ 1, & \text{if } Y_i > \text{LOD (i.e., ``detected'')} \end{cases}. \tag{A.2}$$

We will denote the vector of detected Y_i as $\mathbf{Y}_{det}$, the vector of nondetected Y_i as $\mathbf{Y}_{cen}$, and the vector of indicator variables as $\mathbf{D}$. Furthermore, we will denote the collection of observation indices, i, where $D_i = 1$ as $\{i; D_i = 1\}$; similarly, we denote the set of indices where $D_i = 0$ as $\{i; D_i = 0\}$.

A.1 Using the CDFs

Based on the classical literature, we first define a new data vector, $\mathbf{Y}^* = \{Y_1^*, \ldots, Y_N^*\}'$, with elements

$$Y_i^* = \begin{cases} Y_i, & \text{if } D_i = 1 \\ \text{LOD}, & \text{if } D_i = 0 \end{cases}.$$

Using this, the typical likelihood for a model with left-censored data would be of the form:

$$\begin{aligned} p(\mathbf{Y}^* \mid \mu, \sigma^2, \mathbf{D}) \propto & \prod_{i \in \{i; D_i = 1\}} \left[\left(\sigma^2\right)^{-1/2} \phi\left(\frac{Y_i - \mu}{\sigma}\right) \times I\left\{Y_i > \text{LOD}\right\} \right] \\ & \times \prod_{i \in \{i; D_i = 0\}} \left(\sigma^2\right)^{-1/2} \Phi\left(\frac{\text{LOD} - \mu}{\sigma}\right). \end{aligned} \tag{A.3}$$

Using this and the priors specified in the main manuscript, our hierarchical model would be as follows:

$$\begin{aligned} p\left(\mu, \sigma^2 \mid \mathbf{Y}^*, \mathbf{D}\right) \propto & U\left(\mu \mid a_\mu, b_\mu\right) \times U\left(\ln\sigma \mid a_\sigma, b_\sigma\right) \\ & \times \left(\sigma^2\right)^{-N/2} \prod_{i \in \{i; D_i = 1\}} \phi\left(\frac{Y_i - \mu}{\sigma}\right) \times \prod_{i \in \{i; D_i = 0\}} \Phi\left(\frac{\text{LOD} - \mu}{\sigma}\right). \end{aligned} \tag{A.4}$$

We now wish to derive the full conditional distributions for both μ and σ^2 for the purposes of constructing our MCMC algorithm. For μ, we obtain a full conditional distribution of the form:

$$p\left(\mu \mid \mathbf{Y}^*, \mathbf{D}, \sigma^2\right) \propto \prod_{i \in \{D_i = 1\}} \exp\left[-\frac{\left(Y_i - \mu\right)^2}{2\sigma^2}\right] \times \prod_{i \in \{D_i = 0\}} \Phi\left(\frac{\text{LOD} - \mu}{\sigma}\right) \times I\left\{\mu \in \left(a_\mu, b_\mu\right)\right\}.$$

While there are some simplifications that can be made here, this is *not* a known distribution and thus we *cannot* sample directly from this distribution. A similar issue occurs for σ^2:

$$p\left(\sigma^2 \,|\, \mathbf{Y}^*, \mathbf{D}, \mu\right) \propto \left(\sigma^2\right)^{-N/2} \prod_{i \in \{D_i=1\}} \exp\left[-\frac{(Y_i-\mu)^2}{2\sigma^2}\right] \\ \times \prod_{i \in \{D_i=0\}} \Phi\left(\frac{\text{LOD}-\mu}{\sigma}\right) \times I\left\{\sigma \in (a_\sigma, b_\sigma)\right\}.$$

To sample from these distributions in our MCMC algorithm, we may require the use of *Metropolis* steps. Without getting into specifics, it will suffice to say that such an algorithm would not be ideal, particularly if we were to extend our model to include regression parameters, random effects, etc.

A.2 Using the PDFs

We now rewrite our likelihood in terms of the PDFs for the censored observations:

$$p(\mathbf{Y} \,|\, \mu, \sigma^2, \mathbf{D}) = \prod_{i \in \{i; D_i=1\}} \left[\left(\sigma^2\right)^{-1/2} \phi\left(\frac{Y_i-\mu}{\sigma}\right) \times I\left\{Y_i > \text{LOD}\right\}\right] \\ \times \prod_{i \in \{i; D_i=0\}} \left[\left(\sigma^2\right)^{-1/2} \phi\left(\frac{Y_i-\mu}{\sigma}\right) \times I\left\{Y_i \leq \text{LOD}\right\}\right]. \tag{A.5}$$

Note that in this case, rather than *analytically* integrating out the nondetected Y_i, we use a *data augmentation* approach which involves multiply imputing each of the nondetected Y_i; for a recent example of this approach, see eq. (7) of Quick et al. (2014). Using this augmented likelihood and the priors specified in the main manuscript, our updated hierarchical model would be as follows:

$$p\left(\mu, \sigma^2, \mathbf{Y}_{cen} \,|\, \mathbf{Y}_{det}, \mathbf{D}\right) \propto U\left(\mu \,|\, a_\mu, b_\mu\right) \times U\left(\ln\sigma \,|\, a_\sigma, b_\sigma\right) \\ \times \left(\sigma^2\right)^{-N/2} \prod_{i \in \{i; D_i=1\}} \left[\phi\left(\frac{Y_i-\mu}{\sigma}\right) \times I\left\{Y_i > \text{LOD}\right\}\right] \\ \times \prod_{i \in \{i; D_i=0\}} \left[\phi\left(\frac{Y_i-\mu}{\sigma}\right) \times I\left\{Y_i \leq \text{LOD}\right\}\right]. \tag{A.6}$$

As before, we now wish to derive full conditional distributions for μ and σ^2, but here we also require full conditional distributions for each of the nondetected Y_i (the "multiple imputation" component of this model). Our full conditional distributions are as follows:

$$p(\mu \,|\, \mathbf{Y}, \mathbf{D}, \sigma^2) \propto \prod_{i=1}^{N} \exp\left[-\frac{(Y_i-\mu)^2}{2\sigma^2}\right] \times I\left\{\mu \in (a_\mu, b_\mu)\right\} \tag{A.7}$$

$$p(\sigma^2 \,|\, \mathbf{Y}, \mathbf{D}, \mu) \propto \left(\sigma^2\right)^{-N/2} \prod_{i=1}^{N} \exp\left[-\frac{(Y_i-\mu)^2}{2\sigma^2}\right] \times I\left\{\sigma \in (a_\sigma, b_\sigma)\right\} \tag{A.8}$$

$$p(Y_{i,cen} \,|\, \mathbf{Y}_{(i,cen)}, \mu, \sigma^2) \propto \exp\left[-\frac{(Y_{i,cen}-\mu)^2}{2\sigma^2}\right] \times I\left\{Y_{i,cen} \leq \text{LOD}\right\}. \tag{A.9}$$

From these expressions, we can find that (A.7) is the form of an interval truncated Normal distribution, (A.8) is the form of an interval truncated inverse gamma distribution, and (A.9) is the form of a truncated Normal distribution; as such, we can sample directly from each of these full conditional distributions, facilitating a simple MCMC algorithm.

References

Quick, H., Groth, C., Banerjee, S., Carlin, B. P., Stenzel, M. R., Stewart, P. A., Sandler, D. P., Engel, L. S., and Kwok, R. K. (2014). "Exploration of the use of Bayesian modeling of gradients for censored spatiotemporal data from the *Deepwater Horizon* oil spill." *Spatial Statistics*, 9, 166–179.