# Errata for *Computational Statistics*, 1st Edition, 4rd Printing and Beyond

Geof H. Givens and Jennifer A. Hoeting

March 25, 2014

The final paragraph of the preface indicates if you own a 4rd printing copy.

Here is a list of corrections and other notes. We appreciate comments from our careful readers, including Jim Albert, Shan Ba, Jim Brennan, Shoja'eddin Chenouri, Hugh Chipman, Mark Delorey, Stephanie Fitchett, Doug Gorman, Andrew Hill, Michael Höhle, Quiming Huang, Mori Jamshidian, Yueyang Jiang, Wentao Li, Duncan Murdoch and Jason Song.

**Production Error:**

- A sofware-related problem introduced by the publisher during the production phase has caused page 216 to be unreadable in some early copies of the fourth printing. The correct version of page 216 was placed at page 206, which renders page 206 missing. Error-free versions of both pages are given at the end of this document. **As of 10/30/06, this problem no longer exists.**

**Chapter 1:**

- Page 9, first line below (1.25). Replace "is a convex function" with "is a strictly monotonic function".
- Page 12, last paragraph above Example 1.2. The word *influence* is misspelled.

**Chapter 2:**

- Page 39 first paragraph of section 2.2.2.3 and page 40 first paragraph. Information is gained about the curvature of $\mathbf{g}$ not $\mathbf{g}'$.

**Chapter 3:**

- The AIC values in this chapter are 2 units too low.
- Section 3.2, last sentence of third paragraph. It is slightly clearer to say "If the neighborhood is defined by allowing as many as $k$ changes to the current candidate solution in order to produce the next candidate, then it is a *k-neighborhood*, and the alteration of those features is called a *k-change*."
- Section 3.5.1.1, third sentence. Replace "a individual" with "an individual".
- Section 3.5.2.2, last sentence of third paragraph. Replace "Such an..." with "Such a..."

- Exercise 3.4. The steady state GA should have $G = 1/P$.

**Chapter 4:**

- Page 95, the first equation *below* (4.18), there is a log missing in the last term on the right hand side. In other words, the correct equation is $\mathrm{E}\{\log f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})|\mathbf{x}, \boldsymbol{\theta}^{(t)}\} = \mathrm{E}\{\log f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\theta})|\mathbf{x}, \boldsymbol{\theta}^{(t)}\} - \mathrm{E}\{\log f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})|\mathbf{x}, \boldsymbol{\theta}^{(t)}\}$
- Page 98, fifth line up from the bottom of example 4.4, the end of the line should be $p_{\mathrm{C}}$, not $p_c$.
- Section 4.2.3. We have received the following email from Mori Jamshidian expressing his view of the SEM algorithm.

  > I'm using your text for my computational stats class, and it's been very good, especially in terms of the topics covered. When covering Chapter 4, Section 4.2.3 on EM variance estimation, I noticed that you cover SEM algorithm as one of the main algorithms for EM standard error estimation. In a paper that you have also cited in your book (Jamshidian and Jennrich 2000, JRSS-B) we have noted that SEM does not have a solid theoretical foundation, and have explained why it's prone to all sorts of numerical inaccuracies. Thus, we recommend that the SEM method not be used at all. You mention the method in Jamshidian and Jennrich (2000) as a "more sophisticated numerical differentiation strategy." It turns out that implementation of the methods in Jamshidian and Jennrich (2000) are much simpler than that of SEM, and as we show in our paper they result in highly accurate results. In our view, SEM is a somewhat unsuccessful attempt in using numerical differentiation in the context of EM, as we explain in our paper. Just thought to bring it up, in case you may find this useful for your future editions of the book.
- Page 102, third line of example 4.6 should be $\widehat{p}_{\mathrm{T}} = 0.0132$ not 0.132.
- Page 104, equation (4.49): Omit the $\delta_i$ from the denominator of this expression.
- Page 104, the line above equation (4.51) should begin "for $k = 1, \ldots, C$".
- Page 112, the line above equation (4.78) should begin "Finally, note that $\mathbf{b}^{(t)}, \ldots$".

**Chapter 6:**

- Page 150, example 6.2. Replace $\log \lambda \sim N(4, 0.5^2)$ with $\log \lambda \sim N(\log 4, 0.5^2)$.

**Chapter 8:**

- pge 237, equation (8.17), there should be a minus sign directly preceding the summation symbol.
- In the first bullet of exercise 8.5, the matrix should be $22 \times 22$, not $42 \times 42$.

**Chapter 9:**

- Section 9.2.4, last sentence of first paragraph. The bootstrap estimate of the bias is $\sum_{i=1}^{B} (\widehat{\theta}_i^* - \widehat{\theta})/B = \overline{\theta}^* - \widehat{\theta}$.

- Page 271, bottom paragraph. Delete the sentence beginning "For two-sided intervals..." and the clause after the colon in the subsequent sentence. The improvement offered by the nested bootstrap depends on the accuracy of the original interval and the type of interval. In general, nested bootstrapping can reduce the rate of convergence of coverage probabilities by an additional multiple of $n^{-1/2}$ or $n^{-1}$. See the cited references.
- Page 272, Example 9.9, fourth paragraph should begin "Let $t_2 \ldots$".

There are currently no other known errors in this printing.

The following pages are error-free versions of the pages damaged by a production snafu at Wiley.

Now if the within-chain variance for the $j$th chain is $s_j^2 = \frac{1}{L-1}\sum_{t=D}^{D+L-1}\left(x_j^{(t)} - \bar{x}_j\right)^2$, then let

$$W = \frac{1}{J}\sum_{j=1}^{J} s_j^2 \tag{7.19}$$

represent the mean of the $J$ within-chain estimated variances. Finally, let

$$R = \frac{\frac{L-1}{L}W + \frac{1}{L}B}{W}. \tag{7.20}$$

If all the chains are stationary, then both the numerator and the denominator should estimate the marginal variance of $X$. If, however, there are notable differences between the chains, then the numerator will exceed the denominator. As $L \to \infty$, $\sqrt{R} \to 1$. In practice, some authors suggest that $\sqrt{R} < 1.2$ is acceptable [194]. If the chosen burn-in period did not yield an acceptable result, then $D$ should be increased, $L$ should be increased, or preferably both. A conservative choice is to use one-half of the iterations for burn-in. The performance of this diagnostic is improved if the iterates $x_j^{(t)}$ are transformed so that their distribution is approximately normal. Alternatively, a reparameterization of the model could be undertaken and the chain rerun.

There are several potential difficulties with this approach. Selecting suitable starting values in cases of multimodal $f$ may be difficult, and the procedure will not work if all of the chains become stuck in the same subregion or mode. Due to its unidimensionality, the method may also give a misleading impression of convergence for multidimensional target distributions. Enhancements of the Gelman–Rubin statistic are described in [196], and an extension for multidimensional target distributions is given in [65].

Raftery and Lewis [444] propose a very different quantitative strategy for estimating run length and burn-in period. Some researchers advocate no burn-in [202].

### 7.3.2 Practical implementation advice

The discussion above raises the question of what values should be used for the number of chains, the number of iterations for burn-in, and the length of the chain after burn-in. Most authors are reluctant to recommend generic values because appropriate choices are highly dependent on the problem at hand and the rate and efficiency with which the chain explores the region supported by $f$. Similarly, the choices are limited by how much computing time is available. In the last few years, published analyses have used burn-ins from zero to tens of thousands and chain lengths from the thousands to the millions. Diagnostics usually rely on at least three, and typically more, multiple chains. Five to ten years ago, burn-ins and chain lengths were shorter by a factor of 10. As computing power continues to grow, so too will the scope and intensity of MCMC efforts.

In summary, we reiterate our advice from Section 7.3.1.2 here, which in turn echoes [108]. First, create multiple trial runs of the chain from diverse starting values. Next,

where $i = 1, \ldots, I$, $j = 1, \ldots, J_i$, and $k = 1, \ldots, K$. After averaging over $k$ for each $i$ and $j$, we can rewrite the model (7.27) as

$$Y_{ij} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{ij}, \qquad i = 1, \ldots, I, \;\; j = 1, \ldots, J_i, \tag{7.28}$$

where $Y_{ij} = \sum_{k=1}^{K} Y_{ijk}/K$. Assume that $\alpha_i \sim N(0, \sigma_\alpha^2)$, $\beta_{j(i)} \sim N(0, \sigma_\beta^2)$, and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$, where each set of parameters is independent a priori. Assume that $\sigma_\alpha^2$, $\sigma_\beta^2$, and $\sigma_\epsilon^2$ are known. To carry out Bayesian inference for this model, assume an improper flat prior for $\mu$, so $f(\mu) \propto 1$. We consider two forms of the Gibbs sampler for this problem [463]:

(a) Let $n = \sum_i J_i$, $y_{\cdot\cdot} = \sum_{ij} y_{ij}/n$, and $y_{i\cdot} = \sum_j y_{ij}/J_i$ hereafter. Show that at iteration $t$, the conditional distributions necessary to carry out Gibbs sampling for this model are given by

$$\mu^{(t+1)} \big| \left(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)}, \mathbf{y}\right) \sim N\left(y_{\cdot\cdot} - \frac{1}{n}\sum_i J_i \alpha_i^{(t)} - \frac{1}{n}\sum_{j(i)} \beta_{j(i)}^{(t)}, \frac{\sigma_\epsilon^2}{n}\right),$$

$$\alpha_i^{(t+1)} \big| \left(\mu^{(t+1)}, \boldsymbol{\beta}^{(t)}, \mathbf{y}\right) \sim N\left(\frac{J_i V_1}{\sigma_\epsilon^2}\left(y_{i\cdot} - \mu^{(t+1)} - \frac{1}{J_i}\sum_j \beta_{j(i)}^{(t)}\right), V_1\right),$$

$$\beta_{j(i)}^{(t+1)} \big| \left(\mu^{(t+1)}, \boldsymbol{\alpha}^{(t+1)}, \mathbf{y}\right) \sim N\left(\frac{V_2}{\sigma_\epsilon^2}\left(y_{ij} - \mu^{(t+1)} - \alpha_i^{(t+1)}\right), V_2\right),$$

where $V_1 = \left(\frac{J_i}{\sigma_\epsilon^2} + \frac{1}{\sigma_\alpha^2}\right)^{-1}$ and $V_2 = \left(\frac{1}{\sigma_\epsilon^2} + \frac{1}{\sigma_\beta^2}\right)^{-1}$.

(b) The convergence rate for a Gibbs sampler can sometimes be improved via reparameterization. One approach to reparameterization is called hierarchical centering. For this model, hierarchical centering can be described as follows. Let $Y_{ij}$ follow (7.28), but now let $\eta_{ij} = \mu + \alpha_i + \beta_{j(i)}$ and $\epsilon_{ij} \sim N\left(0, \sigma_\epsilon^2\right)$. Then let $\gamma_i = \mu + \alpha_i$ with $\eta_{ij}|\gamma_i \sim N\left(\gamma_i, \sigma_\beta^2\right)$ and $\gamma_i|\mu \sim N\left(\mu, \sigma_\alpha^2\right)$. As above, assume $\sigma_\alpha^2$, $\sigma_\beta^2$, and $\sigma_\epsilon^2$ are known, and assume a flat prior for $\mu$. Show that the conditional distributions necessary to carry out Gibbs sampling for this model are given by

$$\mu^{(t+1)} \big| \left(\boldsymbol{\gamma}^{(t)}, \boldsymbol{\eta}^{(t)}, \mathbf{y}\right) \sim N\left(\frac{1}{I}\sum_i \gamma_i^{(t)}, \frac{1}{I}\sigma_\alpha^2\right),$$

$$\gamma_i^{(t+1)} \big| \left(\mu^{(t+1)}, \boldsymbol{\eta}^{(t)}, \mathbf{y}\right) \sim N\left(V_3\left(\frac{1}{\sigma_\beta^2}\sum_j \eta_{ij}^{(t)} + \frac{\mu^{(t+1)}}{\sigma_\alpha^2}\right), V_3\right),$$

$$\eta_{ij}^{(t+1)} \big| \left(\mu^{(t+1)}, \boldsymbol{\gamma}^{(t+1)}, \mathbf{y}\right) \sim N\left(V_2\left(\frac{y_{ij}}{\sigma_\epsilon^2} + \frac{\gamma_i^{(t+1)}}{\sigma_\beta^2}\right), V_2\right)$$

where $V_3 = \left(\frac{J_i}{\sigma_\beta^2} + \frac{1}{\sigma_\alpha^2}\right)^{-1}$.