Statistical Science
2007, Vol. 22, No. 3, 353–358
DOI: 10.1214/07-STS235D
Main article DOI: 10.1214/07-STS235

Comment: Bayesian Checking of the Second Levels of Hierarchical Models

Valen E. Johnson

This article extends Bayarri and Berger's (1999) proposal for model evaluation using "partial posterior" p values to the evaluation of second-stage model assumptions in hierarchical models. Applications focus on normal-normal hierarchical models, although the final example involves an application to a beta-binomial model in which the distribution of the test statistic is assumed to be approximately normal.

The notion of using partial posterior p values is potentially appealing because it avoids what the authors refer to as "double use" of the data, that is, use of the data for both fitting model parameters and evaluating model fit. In classical terms, this phenomenon is synonymous to masking and is widely known to reduce the power of test statistics for diagnosing model inadequacy. In the present context, masking is avoided by defining the reference distribution of a test statistic t by the partial posterior distribution, defined as

$$\pi(\theta \mid x_{obs}/t_{obs}) \propto \frac{f(x_{obs} \mid \theta)\pi(\theta)}{f(t_{obs} \mid \theta)}. \tag{1}$$

Heuristically, the partial posterior distribution contains information in the data x_{obs} about model parameter θ not reflected in t_{obs}. From this definition, it follows that the partial posterior distribution and (full) posterior distribution are equivalent when t is ancillary, and that the partial posterior distribution and prior distribution coincide when t is sufficient. The latter fact suggests that partial posterior distributions defined with respect to improper prior densities may not be proper when the test statistic is "approximately sufficient" for some subset of parameter values. It also precludes the use of partial posterior model assessment for objective Bayesian models using test statistics that are sufficient, although the authors presumably regard sufficient test statistics as being useful only for assessing the adequacy of (proper) prior distributions. Nonetheless, insight regarding the relative advantages of the proposed methodology as test statistics vary from being "nearly sufficient" to "nearly ancillary" would be useful.

Under regularity assumptions specified in Robins, van der Vaart and Ventura (2000), partial posterior p values also have the important property of being asymptotically uniformly distributed under the null model. Prior-predictive p values and their extensions to p values based on pivotal quantities (described below) share this property—even in finite samples. p values based on posterior predictive and related reference distributions do not, which makes it difficult to interpret these diagnostics for purposes of formal model assessment. Bayarri and Costellanos (B&C) provide convincing examples that illustrate this difficulty and highlight the dangers associated with the naive use of nonuniform p values. However, it should be noted that the extreme p values reported by the authors are perhaps also somewhat suspect given the relatively small sample sizes considered in the examples. That is, even ignoring errors associated with the numerical approximation of the partial posterior density and the resulting distribution of the test statistic, asymptotic uniformity of the partial posterior p values may not have been achieved to the level of accuracy required for the report of partial posterior p values down to the number of significant digits provided. This concern is heightened by the plots in the third column of Figure 1, which suggest that partial posterior p values are anticonservative for moderate sample sizes.

The significant advantage of partial posterior p values—that of reducing masking—does not come without cost, and two potentially difficult tasks must be performed to construct these diagnostics. First, it is necessary to estimate the sampling density of the chosen test statistic as a function of the model parameter θ. In the article, this task is performed only for cases in which the sampling density of the test statistic can be easily approximated by exploiting a translation-invariance property of the normal distribution. Such a strategy is unlikely to work outside of normal family problems or for more sophisticated test statistics

Valen E. Johnson is Professor of Biostatistics, University of Texas MD Anderson Cancer Center, 515 Holcombe Blvd, Unit 447 Houston, Texas 77030-4009, USA (e-mail: ejohnson@mdanderson.org).