University of Texas, MD Anderson Cancer Center
UT MD Anderson Cancer Center Department of Biostatistics
Working Paper Series

Year 2004 *Paper* 1

Combining Bayesian Chi-Squared Statistics

Valen Johnson*

*vejohnson@mdanderson.org

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

http://biostats.bepress.com/mdandersonbiostat/paper1

Copyright ©2004 by the author.

Combining Bayesian Chi-Squared Statistics

Valen Johnson

Abstract

In an earlier paper, I described how chi-squared goodness-of-fit statistics could be defined for a broad class of Bayesian statistical models. Pivotal to the definition of these chi-squared statistics was the use of a sampled parameter value from the posterior distribution to define either a random set of bin counts or a random vector of cell probabilities. In this article, I demonstrate how chi-squared statistics based on the same posterior distribution and data vector can be combined to obtain a single chi-squared statistic for assessing model fit. A relation between the resulting statistic and the classical chi-squared statistic based on the maximum likelihood estimate is also discussed.

COMBINING BAYESIAN χ^2 STATISTICS

VALEN E. JOHNSON

Abstract. In an earlier paper, I described how χ^2 goodness-of-fit statistics could be defined for a broad class of Bayesian statistical models. Pivotal to the definition of these χ^2 statistics was the use of a sampled parameter value from the posterior distribution to define either a random set of bin counts or a random vector of cell probabilities. In this article, I demonstrate how χ^2 statistics based on the same posterior distribution and data vector can be combined to obtain a single χ^2 statistic for assessing model fit. A relation between the resulting statistic and the classical χ^2 statistic based on the maximum likelihood estimate is also discussed.

1. Background

In Johnson (2004), I defined a χ^2 goodness-of-fit statistic for assessing the adequacy for a broad class of Bayesian statistical models. A shortcoming of this method was that it was capable of producing an arbitrarily large number of correlated χ^2 statistics with no obvious mechanism for combining their values to obtain a single goodness-of-fit diagnostic. In this article, I show how χ^2 values based on distinct posterior samples of the parameter vector can be combined to obtain a single, summary goodness-of-fit diagnostic. The diagnostic has an asymptotic χ^2 distribution and is easy to compute within standard Markov chain Monte Carlo (MCMC) algorithms. Along the way, I derive an estimate of a covariance matrix that determines the distribution of Pearson's χ^2 statistic when based on the maximum likelihood estimate for continuous data (cf., Chernoff and Lehmann, 1954).

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Following the notation used in Johnson (2004), let $\mathbf{y} = \{y_i\}$, $i = 1, \ldots, n$ denote scalar-valued, continuous, identically distributed, conditionally independent observations drawn from probability density function $f(\cdot \,|\, \boldsymbol{\theta})$ defined with respect to Lebesgue measure and indexed by an s-dimensional parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbf{R}^s$. Denote by $F(\cdot \,|\, \boldsymbol{\theta})$ the cumulative distribution corresponding to $f(\cdot \,|\, \boldsymbol{\theta})$, and let $\tilde{\boldsymbol{\theta}}_j$ denote a value drawn from the posterior distribution on $\boldsymbol{\theta}$ given $\mathbf{y}$. Next, choose fixed quantiles $0 \equiv a_0 < a_1, \cdots < a_{K-1} < a_K \equiv 1$ and let $p_k = a_k - a_{k-1}$, $k = 1, \ldots, K$. Define $\mathbf{z}_i(\tilde{\boldsymbol{\theta}}_j)$ to be a vector of length K whose kth element is defined by

$$z_{i,k} = \begin{cases} 1 & \text{if } F(y_i \,|\, \tilde{\boldsymbol{\theta}}_j) \in (a_{k-1}, a_k] \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Finally, define

$$\mathbf{m}(\tilde{\boldsymbol{\theta}}_j) = \sum_{i=1}^{n} \mathbf{z}_i(\tilde{\boldsymbol{\theta}}_j).$$

That is, the kth component of $\mathbf{m}(\tilde{\boldsymbol{\theta}}_j)$ represents the number of observations that fell into the kth bin, where observations are binned according to the values of their conditional distribution function evaluated at $\tilde{\boldsymbol{\theta}}_j$.

Finally, letting

$$\mathbf{V}(\tilde{\boldsymbol{\theta}}_j) = \left(\frac{m_1(\tilde{\boldsymbol{\theta}}_j) - np_1}{\sqrt{np_1}}, \ldots, \frac{m_k(\tilde{\boldsymbol{\theta}}_j) - np_k}{\sqrt{np_k}} \right)', \tag{2}$$

then the Bayesian χ^2 statistic defined in Johnson (2004) can be expressed as

$$R^B(\tilde{\boldsymbol{\theta}}_j) = \mathbf{V}(\tilde{\boldsymbol{\theta}}_j)'\mathbf{V}(\tilde{\boldsymbol{\theta}}_j). \tag{3}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Under regularity conditions specified in Johnson (2004), the asymptotic sampling distribution of $R^B(\tilde{\boldsymbol{\theta}}_j)$ is that of a χ^2_{K-1} random variable. This sampling distribution is defined with respect to repeated sampling of both the observation vector $\mathbf{y}$ from $f(\cdot \,|\, \boldsymbol{\theta}_0)$ and $\tilde{\boldsymbol{\theta}}_j$ from the posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{y}$.

Although easy to compute using output from many MCMC algorithms, a complication that emerges when using values of $R^B(\tilde{\boldsymbol{\theta}}_j)$ as a model diagnostic arises from the fact that an arbitrary number of $R^B(\tilde{\boldsymbol{\theta}}_j)$ values can be generated from the posterior distribution. Although each $R^B(\tilde{\boldsymbol{\theta}}_j)$ value is marginally distributed as a χ^2_{K-1} random variable as $n \to \infty$, values of $R^B(\tilde{\boldsymbol{\theta}}_j)$ based on the same data vector $\mathbf{y}$ (i.e., the same posterior distribution) are correlated.

The summary diagnostic V_J proposed here combines the diagnostic information contained in the sampled values $R^B(\tilde{\boldsymbol{\theta}}_j)$ to arrive at a single summary diagnostic. It is defined as follows. Let $\tilde{\boldsymbol{\theta}}_j,\ i = 1, \cdots, J$ denote a random sample from the posterior distribution on $\boldsymbol{\theta}$ given $\mathbf{y}$ and define

$$\bar{\mathbf{V}} = \frac{1}{J}\sum_{i=1}^{J} \mathbf{V}(\tilde{\boldsymbol{\theta}}_j), \qquad \mathbf{B}_J = \mathbf{I}_K - \mathbf{q}\mathbf{q}' - \frac{1}{J}\sum_{i=1}^{J}\left(\mathbf{V}(\tilde{\boldsymbol{\theta}}_j) - \bar{\mathbf{V}}\right)\left(\mathbf{V}(\tilde{\boldsymbol{\theta}}_j) - \bar{\mathbf{V}}\right)'.$$

Then V_J is defined to be

$$V_J = \bar{\mathbf{V}}'\mathbf{B}_J^{+}\bar{\mathbf{V}}, \tag{4}$$

where $\mathbf{B}_J^{+}$ denotes the Moore-Penrose inverse of the matrix $\mathbf{B}_J$, $\mathbf{q}$ is the $K \times 1$ vector with components $\sqrt{p_k}$, and $\mathbf{I}_K$ is the $K \times K$ identity matrix.

Under certain regularity conditions stated below, the asymptotic distribution of V_J is that of a χ^2 random variable. Before stating this result formally, the following definitions are necessary.

COBRA
A Collection of Biostatistics Research Archive

Hosted by The Berkeley Electronic Press

Let $g(y \,|\, \mathbf{z}, \boldsymbol{\theta})$ denote the conditional distribution of y given $\mathbf{z}$ and $\boldsymbol{\theta}$, and define $\mathbf{J}^*$ to be the matrix whose (s,t)th component is

$$\mathbf{E}\left[\frac{\partial \log g(y \,|\, z, \theta)}{\partial \theta_s} \frac{\partial \log g(y \,|\, z, \theta)}{\partial \theta_t}\right].$$

Let $\tilde{\mathbf{J}} \equiv \mathbf{D}'\mathbf{D}$ denote the matrix with elements

$$\sum_{k=1}^{K} \frac{1}{p_k} \frac{\partial p_k}{\partial \theta_a} \frac{\partial p_k}{\partial \theta_b}.$$

Define $\mathbf{S} = \mathbf{D}(\tilde{\mathbf{J}} + \mathbf{J}^*)^{-1}$, $\mathbf{T} = \mathbf{S}\mathbf{D}'$, let $\boldsymbol{\alpha}(\tilde{\boldsymbol{\theta}}_j)$ denote the vector with components $(m_k(\tilde{\boldsymbol{\theta}}_j) - m_k(\hat{\boldsymbol{\theta}}))/\sqrt{np_k}$, and define $\mathbf{G} = \mathbf{D}(\tilde{\mathbf{J}} + \mathbf{J}^*)^{-1}\mathbf{D}'$.

Theorem 1. *Assume that the regularity conditions imposed in Theorem 1 of Johnson (2004) apply, and let the $\tilde{\boldsymbol{\theta}}_j$, $i = 1, \ldots, J$, denote independent values drawn from the posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{y} = (y_1, \ldots, y_n)$. Assume further that the random variables $\alpha(\tilde{\boldsymbol{\theta}}_j)$ and $\mathbf{V}(\tilde{\boldsymbol{\theta}}_j)'\mathbf{V}(\tilde{\boldsymbol{\theta}}_j)$ are uniformly integrable and that rank$(\mathbf{B}_J) \to$ rank$(\mathbf{I}_K - \mathbf{q}\mathbf{q}' - \mathbf{G})$. Then as $J, n \to \infty$, the statistic V_J converges in distribution to a χ^2 random variable with degrees of freedom equal to rank$(\mathbf{B}_J)$.*

Proof of Theorem 1:

The proof of Theorem 1 is based on the representation of $\mathbf{V}(\tilde{\boldsymbol{\theta}}_j)$ used in Johnson (2004) to demonstrate that a single value of $R^B(\tilde{\boldsymbol{\theta}}_j)$ has an asymptotic χ^2 distribution. That representation is based on the decomposition of the components of $\mathbf{V}(\tilde{\boldsymbol{\theta}}_j)$ as follows:

$$\frac{m_k(\tilde{\boldsymbol{\theta}}) - np_k}{\sqrt{np_k}} = \frac{m_k(\tilde{\boldsymbol{\theta}}) - m_k(\hat{\boldsymbol{\theta}})}{\sqrt{np_k}} - \frac{m_k(\boldsymbol{\theta}_0) - m_k(\hat{\boldsymbol{\theta}})}{\sqrt{np_k}} + \frac{m_k(\boldsymbol{\theta}_0) - np_k}{\sqrt{np_k}} \tag{5}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

If $\boldsymbol{\beta}$ denotes the vector with components $(m_k(\boldsymbol{\theta}_0) - m_k(\hat{\boldsymbol{\theta}}))/\sqrt{np_k}$ and $\boldsymbol{\gamma}$ the vector with components $(m_k(\boldsymbol{\theta}_0) - np_k)/\sqrt{np_k}$, then from results given in Chernoff and Lehmann (1954) and Johnson (2004), it follows that

$$\begin{aligned} \boldsymbol{\alpha}(\tilde{\boldsymbol{\theta}}_j) &= \mathbf{T}\boldsymbol{\varepsilon}_j + \mathbf{S}\boldsymbol{\tau}_j + o_p(1) \\ \boldsymbol{\beta} &= \mathbf{T}\boldsymbol{\epsilon} + \mathbf{S}\boldsymbol{\eta} + o_p(1) \\ \boldsymbol{\gamma} &= \boldsymbol{\epsilon} + o_p(1), \end{aligned} \tag{6}$$

where the joint distribution of $(\boldsymbol{\varepsilon}_j{}', \boldsymbol{\epsilon}', \boldsymbol{\tau}_j{}', \boldsymbol{\eta}')'$ is multivariate normal with mean $\mathbf{0}$ and covariance matrix

$$\begin{bmatrix} \mathbf{I}_k - \mathbf{q}\mathbf{q}' & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_k - \mathbf{q}\mathbf{q}' & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{J}^* & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{J}^* \end{bmatrix}. \tag{7}$$

For a given posterior sample size J, define $\mathbf{W}_J$ to be the matrix whose ith row is $(\mathbf{T}\boldsymbol{\varepsilon}_j + \mathbf{S}\boldsymbol{\tau}_j - \mathbf{T}\boldsymbol{\epsilon} - \mathbf{S}\boldsymbol{\eta} + \boldsymbol{\epsilon})'$, and let

$$\mathbf{A} = \frac{1}{J}\mathbf{1}_J' \otimes \mathbf{I}_K, \tag{8}$$

where $\mathbf{1}_J$ is a $J \times 1$ vector of ones. It follows that $\bar{\mathbf{V}} = \mathbf{A}\mathbf{W}_J + o_p(1)$ where $\mathbf{A}\mathbf{W}_J$ is a multivariate normal random vector with mean $\mathbf{0}$ and covariance matrix

$$\mathbf{Cov}(\mathbf{A}\mathbf{W}_J) = \mathbf{I}_K - \mathbf{q}\mathbf{q}' - \frac{(J-1)}{J}\mathbf{G}. \tag{9}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

From the uniform integrability of $\boldsymbol{\alpha}(\tilde{\boldsymbol{\theta}}_j)$ and $\mathbf{V}(\tilde{\boldsymbol{\theta}}_j)'\mathbf{V}(\tilde{\boldsymbol{\theta}}_j)$, it follows that $\bar{\mathbf{V}}$ has the same asymptotic mean and variance as $\mathbf{AW}_J$.

Also from the uniform integrability assumptions, along with the weak law of large numbers, it follows that for a fixed value of $\mathbf{y}$,

$$\bar{\mathbf{V}} \xrightarrow{p} \boldsymbol{\epsilon} - \mathbf{T}\boldsymbol{\epsilon} - \mathbf{S}\boldsymbol{\eta}$$

and

$$\mathbf{B}_J \xrightarrow{p} \mathbf{I}_K - \mathbf{q}\mathbf{q}' - \mathbf{Var}(\mathbf{T}\boldsymbol{\varepsilon}_j + \mathbf{S}\boldsymbol{\tau}_j) = \mathbf{I}_K - \mathbf{q}\mathbf{q}' - \mathbf{G}.$$

Furthermore, from the assumption that $\text{rank}(\mathbf{B}_J) \to \text{rank}(\mathbf{I}_K - \mathbf{q}\mathbf{q}' - \mathbf{G})$, it follows from results exposed in Puri, Russell, and Mathew (1984; see also Stewart 1969) that $\mathbf{B}_J^+ \xrightarrow{p} (\mathbf{I}_K - \mathbf{q}\mathbf{q}' - \mathbf{G})^+$.

Because

$$\mathbf{Var}(\boldsymbol{\epsilon} - \mathbf{T}\boldsymbol{\epsilon} - \mathbf{S}\boldsymbol{\eta}) = \mathbf{I}_K - \mathbf{q}\mathbf{q} - \mathbf{G},$$

the proof follows from standard results regarding the distribution of quadratic forms (e.g., Rao [1973], page 188).□

Several comments are in order regarding the regularity conditions stated in Theorem 1. The condition that $\text{rank}(\mathbf{B}_J) \to \text{rank}(\mathbf{I}_K - \mathbf{q}\mathbf{q}' - \mathbf{G})$ is required to guarantee the convergence of $\mathbf{B}_J^+$ to $(\mathbf{I}_K - \mathbf{q}\mathbf{q}' - \mathbf{G})^+$. As Chernoff and Lehmann (1954) show, the eigenvalues of the matrix $\mathbf{I}_K - \mathbf{q}\mathbf{q}' - \mathbf{G}$ include $K - s - 1$ ones, one zero, and s values between 0 and 1. Heuristically, if the information in each bin's raw data regarding the value of each of the components of the parameter $\boldsymbol{\theta}$ is greater than the information in the corresponding binned data, the s eigenvalues falling between 0 and 1 will be strictly greater than 0 and the degrees of freedom of the combined

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/mdandersonbiostat/paper1

χ^2 will be $K-1$, as long as this is also the rank of B_J. More technically, this will occur when rank $\left[(\mathbf{J}^* + \tilde{\mathbf{J}})^{-1} - (\mathbf{J}^* + \tilde{\mathbf{J}})^{-1}\tilde{\mathbf{J}}(\mathbf{J}^* + \tilde{\mathbf{J}})^{-1}\right] = K$.

The theorem also extends to samples $\tilde{\boldsymbol{\theta}}_j$ drawn from ergodic MCMC algorithms, as well as to data $\mathbf{y}$ that are conditionally independent but not necessarily identically distributed, provided that the conditions of Corollary 1 of Johnson (2004) are satisfied.

The theorem does not, however, extend immediately to discrete data when values of $R^B(\tilde{\boldsymbol{\theta}}_j)$ are defined with respect to fixed bins on the sampling space of $\mathbf{y}$ as specified in Corollary 2 of Johnson (2004). In that case, it is not clear how to obtain a consistent estimate of the asymptotic covariance matrix of $\mathbf{V}(\tilde{\boldsymbol{\theta}}_j)$ because the vector $\mathbf{q}$ is then not constant.

References

Chernoff, H. and Lehmann, E.L. (1954), "The use of maximum likelihood estimates in χ^2 tests for goodness of fit," *Annals of Mathematical Statistics*, **25**, 579–586.

Pearson, K. (1900), "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling", *Philosophy Magazine*, 5th Series, **50**, 157–172.

Puri, M.L., Russell, C.T., and Mathew, T. (1984), "Convergence of generalized inverses with applications to asymptotic hypothesis testing," *Sankhya: The Indian Journal of Statistics*, **46**, Series A, 277–286.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Spiegelhalter, D., Best, N., Carlin, B. and van der Linde, A. (2002), "Bayesian measures of model complexity and fit," *Journal of the Royal Statistical Society, Series B*, **64**, 583–616.

Spiegelhalter, D., Thomas, A. and Best, N. (2000), *WinBUGS Version 1.3 Users Manual*, Cambridge: Medical Research Council Biostatistics Unit. (Available from http://www.mrc-bsu.cam.ac.uk/bugs.)

Stewart, G.W. (1969), "On the continuity of the generalized inverse," *SIAM Journal of Applied Mathematics*, **17**, 33–45.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/mdandersonbiostat/paper1