Laplace Approximations for Variances of Estimators Based on Categorical Data in Presence of Unreported Items

Yves Thibaudeau, US Census Bureau
Yves Thibaudeau, US Census Bureau, Statistical Research, Washington, DC 20233-9100

Key Words: Method of Laplace, Hierarchical Log-Linear Model, Conjugate Prior.

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census bureau publications. This report is released to inform interested parties of research and to encourage discussion.

1. Introduction

The method of Laplace is an analytical tool to approximate intractable posterior expectations. It trades the problems of ambiguous convergence in time associated with asymptotic computational procedures (such as MCMC's) for an analytical challenge. The advantage of the method of Laplace is that, once the analytical difficulties are resolved, then, unlike asymptotic computational procedures, the method of Laplace can be applied repeatedly, and virtually instantaneously, with a quantifiable error. The natural question: is it worth the extra analytical burden?

The paper outlines a systematic approach to apply the method of Laplace to build inference based on categorical responses embedded in a hierarchical log-linear model, in presence of unreported items. As expected our approach is rather tedious, but it is also fairly general. We give an example from the 1998 dress rehearsal, where the inferential difficulties caused by the unreported items are typically addressed by supplying item imputations (Fay and Town, 1998). For this example we take advantage of the particular nature of the problem, e.g. some categorical variables are always reported. But, we think that this situation is representative of many "real" cases, and the example reveals the method of Laplace as a viable option to construct inference a postiori for a broad class of problems involving categorical data.

2. The method of Laplace for Approximating Posterior Expectations

Tierney and Kadane (1986), Tierney, Kass, and Kadane (1989), and Thibaudeau (1988) use the method of Laplace to approximate posterior expectations of analytical functions under intractable posterior distributions. Let $\rho(\alpha)$ be a prior density on the (multivariate) parameter α and let $L(\alpha;\Omega)$ be the likelihood of α given data Ω. Let $\beta(\alpha)$ be an analytic function of α. In the paper, $\beta(\alpha)$ is always a power of α. Then, let

$$\Lambda_1(\alpha;\Omega) = \log(\beta(\alpha)) + \log(\rho(\alpha)) + \log(L(\alpha;\Omega)) \qquad (1)$$

$$\Lambda_2(\alpha;\Omega) = \log(\rho(\alpha)) + \log(L(\alpha;\Omega)) \qquad (2)$$

The second order Laplace approximation is defined implicitly through the formula

$$E[\beta(\alpha)] = \left(\frac{\left\|-\Sigma_1\left(\hat{\hat{\alpha}}\right)\right\|}{\left\|-\Sigma_2(\hat{\alpha})\right\|}\right)^{-\frac{1}{2}} \times \exp\left\{\Lambda_1\left(\hat{\hat{\alpha}};\Omega\right) - \Lambda_2(\hat{\alpha};\Omega)\right\} \times (1 + O(n^{-2})) \qquad (3)$$

where $\hat{\hat{\alpha}}$ and $\hat{\alpha}$ are the maxima of $\Lambda_1(\alpha;\Omega)$ and $\Lambda_2(\alpha;\Omega)$ respectively. $\Sigma_1\left(\hat{\hat{\alpha}}\right)$ is the Hessian of $\Lambda_1(\alpha;\Omega)$ with respect to the derivand α, evaluated at $\hat{\hat{\alpha}}$, and a similar definition applies for $\Sigma_2(\hat{\alpha})$. $\| \ \|$ denotes the determinant.

We propose to use (3) to evaluate the posterior moment of multinomial parameters constrained by hierarchical log-linear models. We define the general form of the likelihood function and, implicitly, of the density associated with the natural conjugate prior, in the next section, and we exemplify the accuracy of (3) in later sections.

3. The Hierarchical Constrained Multinomial Likelihood

Let Δ represents a set of D categorical variables, with numbers of categories $M_1,\dots,M_D$ respectively. Then $M = \prod_{i\in\Delta} M_i$ is the number of categories in $\Delta^{\times}$, where $\Delta^{\times}$ represents the Cartesian product of the categorical variables in Δ. Let a multivariate category in $\Delta^{\times}$ be represented by the vector $\boldsymbol{v} = [v_1, v_2, \dots, v_D]$, where $v_j = 1,\dots,M_j$ indicates the category for categorical variable j. Also, let $\boldsymbol{p} = [p(\boldsymbol{v})]$ be the probability vector whose components characterize the events corresponding to the $\boldsymbol{v}$'s. Given these definitions we introduce operational notation to define the basic concepts of the paper. Let $\boldsymbol{v}_{(j)} = [v_1,\dots,v_j]$, and $\boldsymbol{v}^{(j)} = [v_{j+1},\dots,v_D]$. Thus $\boldsymbol{v}_{(j)}$ and $\boldsymbol{v}^{(j)}$ denote the vector $\boldsymbol{v}$ with ending and leading components truncated, respectively. In addition, let $|$ be the concatenating operator, and so $\boldsymbol{v} = \boldsymbol{v}_{(j)} \,|\, \boldsymbol{v}^{(j)}$. Then, define

$$\Pi_1(\boldsymbol{v}) = \frac{p(\boldsymbol{v})}{p([1] \,|\, \boldsymbol{v}^{(1)})}$$

$$\Pi_{n+1}(\boldsymbol{v}) = \frac{\Pi_n(\boldsymbol{v})}{\Pi_n\left(\boldsymbol{v}_{(n)} \,|\, [1] \,|\, \boldsymbol{v}^{(n+1)}\right)} \tag{4}$$

With the notation in (4), we can present the central concepts of the paper.

Definition 1 – Hierarchical Constraint

Let $\boldsymbol{v}^* = [v_1^*, v_2^*, \dots, v_D^*]$ represents the same category in $\Delta^{\times}$ as $\boldsymbol{v}$ does, but allow for the order of the components of $\boldsymbol{v}^*$ to be permuted relative to $\boldsymbol{v}$. A hierarchical constraint on the probability vector $\boldsymbol{p}$ is a constraint of the form

$$\Pi_m(\boldsymbol{v}^*) = 1 \tag{5}$$

for all values of $\boldsymbol{v}^*$ such that $v_j^* \neq 1$ whenever $j = 1,\dots,m$. Agresti, (1990, pp. 149-150) gives insight on hierarchical constraints. In effect, (5) disallows interactions of order higher than the n-th order between the categories represented by $v_1^*,\dots,v_m^*$. Lower order interactions are allowed, provided another hierarchical constraint does not stipulate the contrary.

Definition 2 – Hierarchical Log-Linear Model

We say that Θ, the parameter space of $\boldsymbol{p}$, is a hierarchical log-linear model, whenever Θ is unambiguously defined by a set of hierarchical constraints as defined in (5), and the constraints $\sum_{\boldsymbol{v}\in\Delta^{\times}} p(\boldsymbol{v}) = 1$, and $p(\boldsymbol{v}) > 0$, for all $\boldsymbol{v} \in \Delta^{\times}$.

Definition 3 – Hierarchical Constrained Multinomial Likelihood

Let $\mathrm{Z} \subset \Psi$ be a subset of $E < D$ categorical variables, and let $\boldsymbol{u} = [u_1, u_2, \dots, u_E]$ represent a category in $\mathrm{Z}^{\times}$, the Cartesian product of the variables in Z. Let $\sigma(\Psi, \mathrm{Z}, \boldsymbol{u}) \subset \Psi^{\times}$ be the subset containing all the values of $\boldsymbol{v}$, such that $\boldsymbol{v}$ agrees with $\boldsymbol{u}$ on the categorical variables in Z. Let $N(\mathrm{Z}, \boldsymbol{u})$ be the number of occurrences of event $\boldsymbol{u} \in \mathrm{Z}^{\times}$. Then, $\boldsymbol{p}$ has a hierarchical constrained multinomial likelihood if its likelihood is

$$L(N;\boldsymbol{p}) = \prod_{Z \subset \Psi} \prod_{\boldsymbol{u} \in Z^{\times}} \left(\sum_{\boldsymbol{v} \in \sigma(\Psi, Z, \boldsymbol{u})} p(\boldsymbol{v}) \right)^{N(Z,\boldsymbol{u})} \qquad (6)$$

$$\boldsymbol{p} \in \Theta$$

where Θ is a hierarchical log-linear model. The representation in (6) accounts for situations where data are missing at random (Little and Rubin, p. 17). Then, only the variables represented by $\boldsymbol{u}$ are observed, for arbitrary instances of the subset Z.

4. Homogeneous Associations

Let $m = 3$ in (5), for each permutation $\boldsymbol{v}^*$ of $\boldsymbol{v}$. Schafer (1997) refers to the corresponding hierarchical log-linear models as homogenous associations models (containing all two-way interactions, but no higher). We investigate homogeneous association models with the intention of developing a systematic method to apply the method of Laplace in this situation. In order to deduce general principles, we investigate the first non-trivial instance of this situation, that is the case $D = 4$. Then $\boldsymbol{p}$ in definition 2 is a $M_1 \times M_2 \times M_3 \times M_4$ probability vector, and its parameter space Θ is defined by setting $\Pi_3(\boldsymbol{v}^*) = 1$, for $\boldsymbol{v}^* = [v_1, v_2, v_3, v_4]$, $[v_1, v_3, v_2, v_4]$, $[v_1, v_4, v_2, v_3]$, $[v_2, v_3, v_1, v_4]$, $[v_2, v_4, v_1, v_3]$, $[v_3, v_4, v_1, v_2]$ in (5). To implement the method of Laplace we must identify a set of free parameters to represent $\boldsymbol{p}$ under the constraints that define Θ. Because of the dimensionality of Θ, $\boldsymbol{p}$ can be represented by $\sum_{\substack{i=1,4 \\ j=1,4 \\ i \neq j}} (M_i - 1) \times (M_j - 1) + \sum_{i=1,4} (M_i - 1)$ free parameters. To define a complete set of free parameters, we first define a set of free parameters to represent the marginal space generated by the first two categorical variables. We set

$$\psi_{a,b} = \sum_{\substack{\boldsymbol{v} \in \Delta^{\times} \\ v_1 = a \\ v_2 = b}} p(\boldsymbol{v}) \qquad (7)$$

Let ψ be the set containing $\psi_{1,1}, \dots, \psi_{M_1, M_2 - 1}$. Then ψ is a set of $M_1 \times M_2 - 1$ free parameters. We shall define additional free parameters to represent the parameter space of the conditional probability obtained by conditioning on the first two categorical variables. We can construct sets of free parameters by layers. The top layer corresponds to the parameter space of the probabilities associated with category 4 conditional on categories 1, 2, and 3, simultaneously. We define the free parameters of the top layer with the following formula:

$$\phi_{a,b,c,d} = \frac{p([a,b,c,d])}{\sum_{\substack{\boldsymbol{v} \in \Delta^{\times} \\ v_1 = a \\ v_2 = b \\ v_3 = c}} p(\boldsymbol{v})} \qquad (8)$$

We select the values of *a, b, c, d* in (8) to define a set of $M_4 - 1$ free parameters corresponding to the $M_4 - 1$ degrees of freedom associated with category 4, and to define sets of $(M_1 - 1)\times(M_4 - 1)$, $(M_2 - 1)\times(M_4 - 1)$, and $(M_3 - 1)\times(M_4 - 1)$ free parameters corresponding to the $(M_1 - 1)\times(M_4 - 1)$, $(M_2 - 1)\times(M_4 - 1)$, and $(M_3 - 1)\times(M_4 - 1)$ degrees of freedom associated with the second-order interactions between categories 4 and 1, between categories 4 and 2, and between categories 4 and 3, respectively. Let φ be the set of $\phi_{a,b,c,d}$'s for these values of *a, b, c, d* defining the free parameters. Then, for any values of *a, b, c, d* , $\phi_{a,b,c,d}$ can be expressed as a function of φ, in conjunction with the hierarchical constraints in (5).

The next layer in the construction of a complete set of free parameters corresponds to the probability associated with category 3, conditional on categories 1 and 2 simultaneously, for an arbitrary fixed value for category 4. Consider the following conditional probabilities:

$$\gamma_{a,b,c} = \frac{p([a,b,c,1])}{\sum_{\substack{v \in A^X \\ v_1 = a \\ v_2 = b \\ v_4 = 1}} p(v)} \qquad (9)$$

We choose values of *a, b, c* in (9) to define $M_3 - 1$ free parameters corresponding to the $M_3 - 1$ degrees of freedom associated with category 3, and to define sets of $(M_1 - 1)\times(M_3 - 1)$, and $(M_2 - 1)\times(M_3 - 1)$ free parameters corresponding to the $(M_1 - 1)\times(M_3 - 1)$, and $(M_2 - 1)\times(M_3 - 1)$ degrees of freedom associated with the second-order interactions between categories 3 and 1, and between categories 3 and 2, respectively. Let γ be the set of $\gamma_{a,b,c}$'s for the values of *a, b, c* yielding the free parameters. Then, the value of $\gamma_{a,b,c}$ for any selection of *a, b, c* can be expressed as a function of γ, in conjunction with the hierarchical constraints in (5). Note that in (9) category 4 is arbitrarily set to 1, since second-order interactions not involving category 4 do not depend on the values of category 4.

We can define p strictly as a function of the free parameters in ψ, φ, and γ. We can generalize the layered approach of this section to define a complete set of free parameters for any hierarchical log-linear model. Then, based on the parametrization given by the free parameters , we can implement the method of Laplace.

5. Example

We give an example of inference by proxy. At the 1998 dress Rehearsal of Census 2000 in Sacramento (Kostatnich, 1999), race and tenure were requested from each householder. For our example race has four categories (White, Black, Asian, Other) and tenure has two (Owner, Renter). Race and tenure are unreported by for approximately 3 % and 7 % of the householders respectively. The traditional approach to deal with unreported items is to substitute them with the corresponding items for a neighbor who acts as a proxy. This method is called the hot-deck. The method of Laplace allows us to infer on the unreported items conditional on the values of the proxy, without actually imputing these values. Note that it is assumed that for each householder there are recorded proxy values. We center our attention on the corresponding 4 by 2 by 4 by 2 table defined by the items of the householders and their proxy values, for tract X in Sacramento, which has 1583 householders.

Let proxy race, proxy tenure, race, and tenure be categorical variables 1, 2, 3, 4. We specify the homogeneous associations log-linear model to define Θ, the parameter space of $\boldsymbol{p}$, the vector of the multinomial probabilities. We parameterize $\boldsymbol{p}$ in terms of the free parameters in ψ, φ, and γ, as defined in the previous section, and so ψ contains 7 free parameters, φ contains 8, and γ contains 15. Since we assume that the proxy values are always recorded, the joint likelihood of ψ, φ, γ, or equivalently the posterior density under a uniform prior distribution can be writen as

$$L(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{S},\boldsymbol{T};\psi,\varphi,\gamma)$$

$$\begin{aligned}
= &\prod_{w_1,w_2}\left(\psi_{w_1,w_2}\right)^{Q(w_1,w_2)} \\
\times &\prod_{w_1,w_2,w_3,w_4}\left(p\left(w_3,w_4 \mid w_1,w_2\right)\right)^{R(w_1,w_2,w_3,w_4)} \\
\times &\prod_{w_1,w_2,w_3}\left(\sum_{w_4} p\left(w_3,w_4 \mid w_1,w_2\right)\right)^{S(w_1,w_2,w_3)} \\
\times &\prod_{w_1,w_2,w_4}\left(\sum_{w_3} p\left(w_3,w_4 \mid w_1,w_2\right)\right)^{T(w_1,w_2,w_4)}
\end{aligned} \quad \textbf{(14)}$$

In (14) $p\left(w_3,w_4 \mid w_1,w_2\right)$ is the conditional probabillity of observing race w_3, and tenure w_4, conditional on having observed proxy race w_1, and proxy tenure w_2. $\boldsymbol{Q},\boldsymbol{R},\boldsymbol{S},\boldsymbol{T}$ are the sets of counts for the multivariate categories represented by the arguments of the exponents on the RHS. Note that $p\left(w_3,w_4 \mid w_1,w_2\right)$ is strictly a function of φ and γ. So we write

$$\begin{aligned}
&L(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{S},\boldsymbol{T};\psi,\varphi,\gamma) \\
&\quad = H(\boldsymbol{Q};\psi)\times K(\boldsymbol{R},\boldsymbol{S},\boldsymbol{T};\varphi,\gamma)
\end{aligned} \quad \textbf{(15)}$$

Our ojective is to infer on the cases with unreported race and/or unreported tenure. In that respect, $p\left(w_3,w_4 \mid w_1,w_2\right)$ tells us everything we can hope to learn with this model. For instance, let $X^{\text{ten}}_{w_1,w_2,w_3}$ be the number of householders who did not report tenure, but who reported race w_3, and proxy race w_1, and proxy tenure w_2 were observed. Let $Y^{\text{ten}}_{w_4|w_1,w_2,w_3}$ be the number of householders among those whose tenure is in fact w_4. Then

$$E\left[Y^{\text{ten}}_{w_4|w_1,w_2,w_3}\right]$$

$$= X^{\text{ten}}_{w_1,w_2,w_3} \times E\left[\frac{p\left(w_3,w_4 \mid w_1,w_2\right)}{\sum_{w_4} p\left(w_3,w_4 \mid w_1,w_2\right)}\right]$$

Similar formulae are available for variances and covariances of the $Y^{\text{ten}}_{w_4|w_1,w_2,w_3}$'s and they involve computation based on the condtional posterior means, variances, and covariances of the $p\left(w_3,w_4 \mid w_1,w_2\right)$'s. The good news is that we can use the method of Laplace in (3) to approximate the conditional posterior means and variances of the $p\left(w_3,w_4 \mid w_1,w_2\right)$'s. Our task is simplified because $H(\boldsymbol{Q};\psi)$ in (15) factors out on the numerator and the denominator in (3), and thus it can be ignored. As it turns out, for the approximation of the condtional posterior means and second order moments, the integrands on both the numerator and denominator in (3) are always of the form $K(\boldsymbol{R},\boldsymbol{S},\boldsymbol{T};\varphi,\gamma)$ in (15). We can get approximation of the posterior means and variances by systematically finding the maximum of functions of the form $K(\boldsymbol{R}, \boldsymbol{S}, \boldsymbol{T};\ \varphi, \gamma)$, with respect to φ, and γ. This can always be done with the EM algorithm (Dempster, Laird, Rubin, 1977). We obtain an approximation for the variance by taking the difference between the Laplace approximation of the second order moment, and the square of the Laplace approximation of the mean.

6. Preliminary Results

We give preliminary results for the Sacramento inference by proxy example. As a reference point we also evaluate posterior expectations using 10,000 iterations of Bayesian iterative proportional fitting (BIPF), (Schafer, 1997, p. 357). We assume that for a householder in tract X, only the race and tenure of the neighbor is known. We approximate the mean of the posterior probability that the householder is respectively a White owner, a Black owner, and an owner, conditional on the information that the neighbor is a White renter. We also approximate the standard errors of these posterior conditional probabilities. The results, under a uniform prior, are shown in table 1.

Table 1 - Posterior Means and Standard Errors for Householder Race-Tenure Probabilities when the Neighbor is a White Renter

	Mean	Mean	S.E.	S.E.
House-holder	**Laplace**	**BIPF**	**Laplace**	**BIPF**
White Owner	.0613	.0646	.00710	.00714
Black Owner	.00659	.00638	.00170	.00184
Owner	.0817	.0856	.00907	.00868
Renter	.918	.914		

7. Discussion

We have represented a hierarchical log-linear model with a set of Hierarchical constraints in order to delineate an explicit non-degenerate parameterization for the model. We use the parameterization to implement the method of Laplace, as suggested in (3). The particular form of the likelihood in (14) and (15), allows us to compute the maxima $\hat{\hat{\alpha}}$ and $\hat{\alpha}$ in (3) with the EM algorithm. Then we can apply the method of Laplace to approximate a posterior expectation, whenever the expectant $\beta(\alpha)$ implicit in (3) is a power of the conditional probability for a category given another category, and the conditional category is always observed. This in turn allows us to approximate the posterior variance by subtracting two power-moment approximations.

At this point, the results are encouraging (table 1). But we hope to develop direct Laplace approximations for the posterior variance, e.g. not by subtracting two power-moment approximations, which can increase the relative error. It appears that such an endeavor requires using maximization techniques other than the EM algorithm and thus further complicates the problem. Nevertheless, it may be worth pursuing, as it could lead to a systematic procedure to build inference based on categorical data in presence of unreported items for an important class of problems, without the disadvantages of computational simulation based on asymptotics.

8. References

Agresti, A. (1990). *Categorical Data Analysis*, Wiley-Intersience

Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). "Maximum Likelihood from Incomplete Data Via the EM Algorithm," *Journal of the Royal Statistical Society*, Series B, Vol. 39, 1-22.

Fay, R. E., Town, M. K. (1998). "Variance Estimation for the 1998 Census Dress Rehearsal," *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Kostanich, D. L. (1999). DSSD Census 2000 Dress Rehearsal Memorandum Series #A, US Bureau of the Census.

Little, R. J. A., Rubin, D. B. (1987). *Statistical Analysis with Missing Data*, Wiley.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*, Chapman & Hall.

Thibaudeau, Y. (1988). "Approximating the Moments of a Multimodal Posterior Distribution with the Method of Laplace," Ph.D. Dissertation, Carnegie Mellon University.

Tierney, L., Kadane, J. B. (1986). "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association*, **81**, 82-86.

Tierney, L., Kass, R. E., Kadane, J. B. (1989). "Fully Exponential Laplace Approximations to Expectations and Variances of Nonpositive Functions," *Journal of the American Statistical Association*, **84**, 710-717.