Teresa H. Jelenkowska

Bayesian estimation of the intraclass correlation coefficients in the mixed linear model

Applications of Mathematics, Vol. 43 (1998), No. 2, 103–110

Persistent URL: http://dml.cz/dmlcz/134378

BAYESIAN ESTIMATION OF THE INTRACLASS CORRELATION COEFFICIENTS IN THE MIXED LINEAR MODEL

Teresa H. Jelenkowska, Lublin

(Received August 19, 1996)

Abstract. The method of determining Bayesian estimators for the special ratios of variance components called the intraclass correlation coefficients is presented. The exact posterior distribution for these ratios of variance components is obtained. The approximate posterior mean of this distribution is also derived. All computations are non-iterative and avoid numerical integration.

Keywords: Bayesian estimation, standard random linear model, posterior distribution, inverted multidimensional Dirichlet distribution, intraclass correlation coefficients

MSC 2000: 62A15, 62H12

1. Introduction

The mixed linear models can be of extreme value in the field of quality control, applied genetics, medicine, agriculture, etc. Considering the random models, the main problem is inference regarding the variance of the random effects, in particular the estimation of them.

Hill (1965) and (1967), Tiao and Tan (1965) and Stone and Springer (1965) were the first to develop variance component estimators based on the Bayesian approach to inference. Other papers devoted to the Bayesian approach to estimating variance components include Rajagopal and Broemeling (1983) and Gharraf (1979), Jelenkowska (1988) and Cook, Broemeling, Gharraf (1990). In the last paper the exact posterior probability distribution for the error variance was derived. In addition, the authors also derived the exact posterior mean and variance for an estimator of the remaining variance components. Using the exact posterior moments, they obtained approximate posterior probability distributions for the variance components.

In this paper, using this approximate posterior distribution for variance components we consider the problem of Bayesian estimation of some special ratios of the variance components called the intraclass correlation coefficients. Such coefficients are of importance in practice.

Section 2 describes the model and its assumptions. Section 3 provides an illustration of the posterior analysis which includes the approximate posterior distribution for the intraclass correlation coefficients and posterior means for them. These moments do not require numerical integration nor constrained optimization. Section 4 presents two numerical examples involving a one-way and a two-way nested design.

2. Model and assumptions

The general linear model with c variance components is given by

$$\mathbf{y} = X\theta + U_1\mathbf{b}_1 + \ldots + U_c\mathbf{b}_c, \tag{2.1}$$

where
$\mathbf{y}$ is an $(n \times 1)$ vector observations,
X is an $(n \times q)$ known design matrix of rank q,
θ is a $(q \times 1)$ vector of unobservable fixed effects,
U_i is an $(n \times m_i)$ known design matrix of rank m_i, $i = 1, \ldots, c$,
$\mathbf{b_i}$ is an $(m_i \times 1)$ vector of unobservable random variables.
In order to simplify somewhat the notation, let $\mathbf{b} = (\mathbf{b'}_1, \ldots, \mathbf{b'}_c)'$, $U = (U_1, \ldots, U_c)$.

Using this notation the model (2.1) can be written as

$$\mathbf{y} = X\theta + U\mathbf{b}.$$

If $c = 1$, the model (2.1) is reduced to the fixed model; if $c > 1$ we have the mixed linear model, and assuming $X_{n\times 1} = (1, \ldots, 1)'$ we receive the random model. Fixed, mixed, and random are terms which came from the sampling theory framework, and are not applicable to the Bayesian approach; nevertheless we will continue to use them, but within a Bayesian context.

Assume that the spaces generated by the columns of matrices $U_1, \ldots, U_c$ are mutually orthogonal and the vectors $\mathbf{b}_1, \ldots, \mathbf{b}_c$ are independent and normally distributed with zero means and covariance matrices $(\sigma_1^2 I_{m_1}, \ldots, \sigma_c^2 I_{m_c})$, respectively. The unknown parameters $(\sigma_1^2, \ldots, \sigma_c^2)$ are called the variance components.

The main problem in this paper consists in estimating a function of variance components of the form

$$\left(\frac{\sigma_1^2}{\sigma^2}, \frac{\sigma_2^2}{\sigma^2}, \ldots, \frac{\sigma_c^2}{\sigma^2}\right), \quad \text{where} \quad \sigma^2 = \sum_{i=1}^{c} \sigma_c^2. \tag{2.2}$$

These ratios are called intraclass correlation coefficients. Estimating the vector (2.2), we can make inference about the relative contribution of each source of variation to the total variance.

3. Posterior distribution of the intraclass correlation coefficients

In Section 2 we have assumed that

$$\mathbf{b} \sim N(0, \sigma_i^2 I_{m_i}), \quad \sigma_i^2 > 0,\ i = 1, \dots, c. \tag{3.1}$$

The problem of estimating the ratios of variance components

$$X_1 = \frac{\sigma_1^2}{\sigma^2},\ \dots,\ X_c = \frac{\sigma_c^2}{\sigma^2} \tag{3.2}$$

will now be considered. For an exact Bayesian analysis, the posterior distribution of $(X_1, X_2, \dots, X_c)$ is needed.

As a starting point for the posterior independent σ_i^2 we can use the approximate posterior distribution of variance components derived by Cook, Broemeling and Gharraf (1990) as

$$p(\sigma_1^2, \dots, \sigma_c^2 | \mathbf{y}) \propto \prod_{i=1}^{c} f_{i\chi^2}\Big(\frac{\sigma_i^2}{S_i}\ \Big|\ \nu_i\Big), \tag{3.3}$$

where $f_{i\chi^2}$ is the inverted chi-square density function with degree of freedom

$$\nu_i = 4 + \frac{(\nu-4)T_i^2}{V_i}, \quad i = 2, \dots, c, \quad \nu_1 = \nu = n - m - q,$$

and

$$S_i = \frac{T_i}{m_i(\nu-2)}\Big[2 + \frac{(\nu-4)T_i}{V_i}\Big] \quad i = 2, \dots, c,$$
$$S_1 = S = \mathbf{y}' R_b \mathbf{y} - \widehat{\theta}' P_\theta \widehat{\theta},$$

where

$$R_b = I_n - U(U'U)^- U', \quad \widehat{\theta} = (X' R_b X)^- X' R_b Y$$

and

$$T_i = S \operatorname{tr} \Delta_{ii} + (\nu-2)\widehat{\mathbf{b}'}_i \widehat{\mathbf{b}}_i,$$
$$V_i = S^2 \overset{2}{\operatorname{tr}} \Delta_{ii} + (\nu-2) \operatorname{tr} \Delta_{ii}^2 + 2(\nu-2)(\nu-4) S \widehat{\mathbf{b}'}_i \Delta_{ii} \widehat{\mathbf{b}}_i.$$

The matrix L_i is defined so that

$$\widehat{\mathbf{b}} = L_i\widehat{\mathbf{b}}, \quad \widehat{\mathbf{b}} = (U'R_\theta U)'U'R_\theta Y, \quad R_\theta = I_n - X(X'X)^- X'$$

and

$$\Delta_{ii} = L_i(U'R_\theta U)^- L_i', \quad i = 2, \dots, c.$$

Next, we will make the transformation

$$\begin{aligned} \sigma_1^2 &= \sigma_1^2, \\ \sigma_2^2 &= \sigma_1^2(1 - X_2 - \dots - X_c)^{-1}X_2, \\ &\vdots \\ \sigma_c^2 &= \sigma_1^2(1 - X_2 - \dots - X_c)^{-1}X_c. \end{aligned} \tag{3.4}$$

The Jacobian of this transformation is

$$\left|J\Big(\frac{\sigma_1^2, \dots, \sigma_c^2}{\sigma_1^2, X_2, \dots, X_c}\Big)\right| = (\sigma_1^2)^{c-1}X_1^{-c}, \text{ where } X_1 = 1 - X_2 - \dots - X_c. \tag{3.5}$$

Making the transformation from $(\sigma_1^2, \dots, \sigma_c^2)'$ to $(\sigma_1^2, X_2, \dots, X_c)'$ and using (3.3), (3.4) and (3.5), the joint posterior density function for $(\sigma_1^2, X_1, \dots, X_c)'$ is

$$p(\sigma_1^2, X_1, \dots, X_c) \propto \frac{X_1^{\frac{1}{2}\sum_{i=1}^{c}\nu_i} \prod_{i=1}^{c}\big(\frac{S_i}{X_i}\big)^{\frac{\nu_i}{2}+1}}{(\sigma_1^2)^{\frac{1}{2}\sum_{i=1}^{c}\nu_i+1} \prod_{i=1}^{c}\Gamma\big(\frac{\nu_i}{2}\big)S_i} \exp\Big[-\frac{1}{2\sigma_1^2}\Big(S_1 + \sum_{i=2}^{c}\frac{S_i}{X_i}X_1\Big)\Big]. \tag{3.6}$$

Integrating (3.6) with respect to σ_1^2, the joint posterior distribution of $(X_1, X_2, \dots, X_c)'$ is given by

$$p(X_1, X_2, \dots, X_c) \propto \frac{\Gamma\big(\frac{1}{2}\sum_{i=1}^{c}\nu_i\big)}{\prod_{i=1}^{c}\Gamma\big(\frac{\nu_i}{2}\big)S_iX_1^c} \prod_{i=2}^{c}\big(\frac{S_iX_1}{X_iS_1}\big)^{\frac{\nu_i}{2}+1}\Big(1 + \sum_{i=2}^{c}\frac{S_iX_1}{X_iS_1}\Big)^{-\frac{1}{2}\sum_{i=1}^{c}\nu_i}. \tag{3.7}$$

Thus, using the transformation

$$\begin{aligned} w_2 &= \frac{S_2X_1}{X_2S_1}, \quad 0 < w_2 < \frac{S_2}{S_1}, \\ w_3 &= \frac{S_3X_1}{X_3S_1}, \quad 0 < w_3 < \frac{S_3}{S_1}, \\ &\vdots \\ w_c &= \frac{S_cX_1}{X_cS_1}, \quad 0 < w_c < \frac{S_c}{S_1}, \end{aligned}$$

the Jacobian of which is

$$\left| J\Big(\frac{X_1, X_2, \ldots, X_c}{w_2, \ldots, w_c}\Big) \right| = \Big(\frac{X_1}{S_1}\Big)^c \prod_{i=2}^{c} S_i w_i^{-2},$$

one can receive the joint posterior distribution for $w_2, \ldots, w_c$ as

$$p(w_1, \ldots, w_c) \propto \frac{\Gamma\big(\frac{1}{2}\sum_{i=1}^{c}\nu_i\big)}{\prod_{i=1}^{c}\Gamma\big(\frac{\nu_i}{2}\big)} \prod_{i=2}^{c} W_i^{\frac{\nu_i}{2}-2}\Big(1+\sum_{i=2}^{c} W_i\Big)^{-\frac{1}{2}\sum_{i=1}^{c}\nu_i}. \tag{3.8}$$

As can be seen, the expression (3.8) has the form of the inverted Dirichlet distribution (Tiao and Tan (1965)).

This leads to the following theorem:

Theorem 1. *The posterior distribution of the vector* $(w_2, \ldots, w_c)'$ *is the joint inverted Dirichlet density, given by* (3.8) *with the vector of parameters* $\big(\frac{\nu_2}{2}, \ldots, \frac{\nu_c}{2}, \frac{\nu_1}{2}\big)$ *for* $\nu_i > 0$ *and* $i = 1, \ldots, c$.

So, the first marginal moments of $(w_1, \ldots, w_c)'$ can be written (see Press, 1972, p. 136) in the form

$$E(W_i) = \frac{\nu_i}{\nu_1 - 2}.$$

Now we can write

$$X_1 = \frac{r_1}{r}, \quad X_2 = \frac{r_2}{r}, \ldots, X_c = \frac{r_c}{r},$$

where

$$r_1 = 1,\ r_2 = \frac{\sigma_2^2}{\sigma_1^2}, \ldots, r_c = \frac{\sigma_c^2}{\sigma_1^2}, \quad \text{and} \quad r = \sum_{i=1}^{c} r_i.$$

Thus,

$$\begin{aligned} \hat{r}_1 &= 1, \\ \hat{r}_2 &= \frac{(\nu_1 - 2)S_2}{\nu_2 S_1}, \\ &\vdots \\ \hat{r}_c &= \frac{(\nu_1 - 2)S_c}{\nu_c S_1}. \end{aligned}$$

Finally, the estimators of the relative contributions of the variance components to the total variance of $\mathbf{y}$ are

$$\widehat{X}_1 = \frac{\hat{r}_1}{\hat{r}}, \ldots, \hat{X}_c = \frac{\hat{r}_c}{\hat{r}}, \text{ where } \hat{r} = \sum_{i=1}^{c} \hat{r}_i. \tag{3.9}$$

4. Numerical examples

The relative contribution of variance components to the total variance described in the earlier sections will now be illustrated by two data sets. Both data sets are generated by known parameters and taken from Box and Tiao (1973). The first data follow a one-way balanced design. The statistical model used to generate the data was

$$y_{ij} = \theta + b_i + b_{ij}, \tag{4.1}$$

where θ represents the population mean, b_i, $i = 1, \ldots, 6$ represents a random effect due to the particular batches that are to be selected, and b_{ij}, $j = 1, \ldots, 5$ denotes the experimental error for sample j taken from batch i. Further, it will be assumed that

$$b_i \sim N(0, \sigma_2^2) \quad \text{and} \quad b_{ij} \sim N(0, \sigma_1^2). \tag{4.2}$$

The variance of y_{ij} is $\sigma_2^2 + \sigma_1^2$. Our interest will be focused on the estimation of the intraclass correlation coefficients

$$X_1 = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \quad \text{and} \quad X_2 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}. \tag{4.3}$$

The data in Table 1 represent a balanced design with six batches and five samples for each batch.

Batch					
1	**2**	**3**	**4**	**5**	**6**
7.298	5.220	0.110	2.212	0.282	1.722
3.846	6.556	10.386	4.852	9.014	4.782
2.434	0.608	13.434	7.092	4.458	8.106
9.566	11.788	5.510	9.288	9.446	0.758
7.990	−0.892	8.166	4.980	7.198	3.758

Table 1. Generated Data from Box and Tiao (1973) p. 247.

The true values of the variances are known and are $\sigma_1^2 = 16$ and $\sigma_2^2 = 4$. Consequently,

$$X_1 = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} = 80\,\%, \quad X_2 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} = 20\,\%.$$

The ANOVA estimates are $\hat{\sigma}_1^2 = 12.34$ and $\hat{\sigma}_2^2 = 0.67$, so that the intraclass correlation coefficients are $\frac{\hat{\sigma}_1^2}{\hat{\sigma}_1^2 + \hat{\sigma}_2^2} = 94.85\,\%$ and $\frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2 + \hat{\sigma}_2^2} = 5.15\,\%$.

The approximate Bayesian estimators for these special ratios of variance components are

$$\hat{x}_1 = 81.34\,\% \quad \text{and} \quad \hat{x}_2 = 18.66\,\%.$$

It is interesting to note that the Bayesian estimators are closer to the true valued than the standard estimators.

The second example describes a two-way nested design. Table 2 contains data taken from Box and Tiao (1973), page 282. The observations were generated from the table of random normal deviates using the three component model. The model used to generated the data was

$$y_{ijk} = \theta + b_i + b_{ij} + b_{ijk}, \; i = 1, \ldots, I; \; j = 1, \ldots, J; \; k = 1, \ldots, K, \tag{4.4}$$

where y_{ijk} are the observations, θ is a common location parameter, b_i, b_{ij} and b_{ijk} are three different kinds of random effects. The random effects (b_i, b_{ij}, b_{ijk}) are all independent and

$$b_i \sim N(0, \sigma_3^2), \quad b_{ij} \sim N(0, \sigma_2^2), \quad \text{and} \quad b_{ijk} \sim N(0, \sigma_1^2). \tag{4.5}$$

It follows in particular that $Var(y_{ijk}) = \sigma^2 = \sigma_1^2 + \sigma_2^2 + \sigma_3^2$ so that the parameters $(\sigma_1^2, \sigma_2^2, \sigma_3^2)$ are the variance components.

	$j = 1$		$j = 2$	
I	$k = 1$	$k = 2$	$k = 1$	$k = 2$
1	2.004	2.713	0.603	0.252
2	4.342	4.229	3.344	3.057
3	0.869	−2.621	−3.896	−3.696
4	3.531	4.185	1.722	0.380
5	2.579	4.271	−2.101	0.651
6	−1.404	−1.003	−0.775	−2.202
7	−1.676	−0.208	−9.139	−8.653
8	1.670	2.426	1.834	1.200
9	2.141	3.527	0.462	0.665
10	−1.229	−0.596	4.471	1.606

Table 2 Generated Data from Box and Tiao (1973), p. 282

The parameters used in generating the data were $\theta = 0$, $\sigma_1^2 = 1.0$, $\sigma_2^2 = 4.0$, $\sigma_3^2 = 2.25$, $I = 10$, $J = k = 2$. Hence $\frac{\sigma_1^2}{\sigma^2} = 13.8\,\%$, $\frac{\sigma_2^2}{\sigma^2} = 55.2\,\%$, $\frac{\sigma_3^2}{\sigma^2} = 31.0\,\%$. The ANOVA estimators are $\hat{\sigma}_1^2 = 1.04$, $\hat{\sigma}_2^2 = 5.62$, $\hat{\sigma}_3^2 = 3.6$, $\hat{\sigma}^2 = 10.26$. Within the sampling theory framework such estimates of the relative contributions of $(\sigma_1^2, \sigma_3^2, \sigma_3^2)$ to the total variance σ^2 are the ratios $\frac{\hat{\sigma}_1^2}{\hat{\sigma}^2} = 10.1\,\%$, $\frac{\hat{\sigma}_2^2}{\hat{\sigma}^2} = 54.7\,\%$, $\frac{\hat{\sigma}_3^2}{\hat{\sigma}^2} = 35.1\,\%$.

The Bayesian estimators for these special ratios of variance components are

$$\widehat{X}_1 = 10.07\,\%, \quad \widehat{X}_2 = 59.55\,\%, \quad \widehat{X}_3 = 30.38\,\%.$$

It is interesting to note that the Bayesian estimator $\widehat{X}_1$ is consistent with ANOVA point estimate and the estimator $\widehat{X}_3$ is very near to the true value of 31.0 %.

References

[1] *Box, G.E.P. and Tiao G.C.*: Bayesian Inference in Statistical Analysis. Massachusetts: Addison Wesley, 1973.

[2] *Broemeling, L.D.*: Bayesian Analysis of Linear Models. Marcel Dekker, New York, 1985.

[3] *Cook, P., Broemeling, L.D. and Gharraf, M.*: A Bayesian analysis of the mixed linear model. Communications in Statistics *A19, No 3* (1990), 987–1002.

[4] *De Groot, M.H.*: Optimal Statistical Decisions. McGraw-Hill Co, New York, 1970.

[5] *Gharraf, M.*: General Solution to Making Inference about the Parameters of Mixed Linear Models. Ph.D. dissertation, Oklahoma State University, Stillwater, Oklahoma, 1979.

[6] *Hill, B.M.*: Inference about variance components in the one-way model. Jour. of the American Statistical Association *60* (1979), 806–825.

[7] *Hill, B.M.*: Correlated errors in the random model. Journal of the American Statistical Association *62* (1967), 1387–1400.

[8] *Jelenkowska, T.H.*: Bayesian estimation of the variance components in a general random linear model, Bayesian Statistics 3. Oxford Sci. Publ. Oxford Univ. Press, New York (1988), 653–656.

[9] *Press, J.S.*: Applied Multivariate Analysis. Holt, Reinhart and Winston, Inc., 1972.

[10] *Rajagopal, M. and Broemeling L.D.*: Bayesian inference for the variance components of the general mixed model. Communications in Statistics *12, No 6* (1983), 701–724.

[11] *Searle, S.R.*: A Summary of Recently Developed Methods of Estimation of Variance Components. Technical Report, BU-388, Biometrics Unit., Cornell University, Ithaca, New York, 1991.

[12] *Stone, M. and Springer, B.G.F.*: A paradox involving quasi-prior distributions. Biometrika *52* (1965), 623–627.

[13] *Tiao, G.C. and Tan, W.Y.*: Bayesian analysis of random-effect models in the analysis of variance—I: Posterior distribution of variance components. Biometrika *52* (1965), 37–53.

Author's address: *Teresa H. Jelenkowska*, Department of Applied Mathematics, University of Agriculture, Akademicka 13, 20-934 Lublin, Poland, e-mail: `jelen@ursus.ar.lublin.pl`.