Are second-order approximations necessary to calculate prediction errors in empirical Bayes disease mapping?

Ugarte, M. D.[1], Militino, A. F.[1] and Goicoa, T.[1]

[1] Departamento de Estadística e Investigación Operativa, Universidad Pública de Navarra. Campus de Arrosadía, 31006 Pamplona, Spain.
E-mail: lola@unavarra.es

Abstract:
Mapping of small area mortality (or incidence) risks is a widely used technique in public health research. The standardized mortality ratio (SMR) is a common direct index of disease mortality, but it might be very imprecise in areas or counties with low population. Then, the use of models that borrow information from related (neighbouring) regions, smoothing the crude mortality risks, is inevitable. In this work, we focus on comparing prediction error estimators used in Empirical Bayes (EB) disease mapping with alternative proposals from the small area estimation literature, to check if second order approximations are necessary in the disease mapping framework. The well known Scottish lips cancer data is used for illustrative purposes. A simulation study is also conducted.

Keywords: Double Bootstrap; PQL; Relative Risks; Spatial Autocorrelation.

1 Introduction

Mapping disease or mortality risks is a useful technique to display the geographical variation of risks and to detect regions with extreme risk for planning health policies and allocating resources. Raw or direct measures, such as SMR, are not very reliable as they are usually highly variable in low populated areas or when the number of observed counts is very small. Then, disease mapping is a small area problem and models are essential to produce reliable estimates by borrowing information from neighbouring areas. There is a huge amount of research in small area estimation, but disease mapping has been approached somehow different from other small area applications because sampling is not involved. Typically, administrative data on number of deaths and related auxiliary variables are used, being the interest in the prediction of random effects instead of a mean or a total. More precisely, the focus is on making predictions of the relative risks, and assessing the prediction error, which is of capital relevance to build confidence intervals for the relative risks. Later, one may decide whether the regions exhibit extreme risks. The estimation of the prediction

errors is always a challenge in small area applications, and the difficulty increases as the underlying models become more and more complex. Disease mapping usually involves generalized linear mixed models (GLMM) requiring the prediction of random effects that are used to estimate relative risks. However, maximum likelihood estimation for GLMM with counts or proportions usually requires numerical integration to calculate the log-likelihood, and then, penalized-quasilikelihood (PQL), an approximation technique using a Laplace approximation to the integrated mixed model likelihood (see Breslow and Clayton, 1993), might be used instead. This technique provides adequate point estimates for the model parameters under spatially correlated data (Dean, Ugarte and Militino, 2004), but underestimates the prediction error because it ignores the uncertainty coming from the estimation of the variance components.
The main goal of this work is to compare prediction error estimators used in Empirical Bayes disease mapping with alternative proposals from the small area estimation literature, to check if second order approximations are necessary in this framework. We assess the performance of the different estimators in terms of relative bias and relative root mean squared error in a simulation study.

2 The Spatial Model

Let us suppose that the area under study is divided into I contiguous regions labelled $i = 1, \ldots, I$. Conditional on the random region effects r_i, the number of deaths in each area, C_i, is assumed to be Poisson distributed with mean $\mu_i = e_i r_i$, where r_i represents the unknown relative risks of mortality from a rare disease, and e_i are the expected number of deaths. Then,

$$C_i | r_i \sim Poisson(\mu_i = e_i r_i), \quad \log \mu_i = \log e_i + b_i,$$

where $b_i = \alpha + u_i = \log r_i$. The random component u_i models both intrinsic Gaussian autoregression, representing local spatially structured variation and the unstructured variation usually associated with covariates not included in the model. Here $\mathbf{u} \sim N(\mathbf{0}, \mathbf{D})$, where

$$\mathbf{D} = \sigma^2(\lambda \mathbf{Q}^- + (1 - \lambda)\mathbf{I}).$$

The matrix $\mathbf{Q}$ is determined by the neighbourhood structure with the ith diagonal element equal to the number of neighbours of the ith region and for $i \neq j$, $\mathbf{Q}_{ij} = -1$ if i and j are neighbours and 0 otherwise; $\mathbf{I}$ is the identity matrix and the parameter λ determines the relative weight between the spatial and the unstructured variation. When $\lambda = 1$, there is no unstructured heterogeneity, and the random effect u_i can be interpreted conditionally given $\mathbf{u}_{-i}$, the set of spatially structured random regions excluding the ith. That is $u_i | \mathbf{u}_{-i} \sim N(\bar{u}_{\delta_i}, \sigma_u^2/\delta_i)$, where $\bar{u}_{\delta_i}$ is the mean of

the random effects corresponding to regions in the neighbourhood of the ith, and δ_i is the number of regions forming the neighbourhood. Here, a neighbourhood consists of regions sharing a common boundary with a given region.

3 Prediction Error

PQL provides adequate point estimates for the model parameters, but it underestimates the variability of the random effects since it does not consider the uncertainty arising from the estimation of the variance parameters. More precisely,

$$\text{var}(b_i|\mathbf{C}) = E_{\boldsymbol{\zeta}|\mathbf{C}}[\text{var}(b_i|\mathbf{C},\boldsymbol{\zeta})] + \text{var}_{\boldsymbol{\zeta}|\mathbf{C}}[E(b_i|\mathbf{C},\boldsymbol{\zeta})], \tag{1}$$

where $\boldsymbol{\zeta} = (\alpha, \sigma^2, \lambda)$, and the PQL variance estimate $\widehat{\text{var}}(b_i|\mathbf{C},\hat{\boldsymbol{\zeta}})$ only approximates the first term. Ainsworth and Dean (2005) use the following first order variance estimator for the first and second terms in Equation (1)

$$\widehat{\text{var}}(b_i|\mathbf{C}) = \widehat{\text{var}}(\hat{\alpha}) + (\hat{\mathbf{D}} - \hat{\mathbf{D}}\mathbf{Z}'\hat{\mathbf{V}}^{-1}\mathbf{Z}\hat{\mathbf{D}}) + \Big(\frac{\partial\hat{\mathbf{u}}}{\partial\hat{\boldsymbol{\zeta}}}\Big)'\widehat{\text{var}}(\hat{\boldsymbol{\zeta}})\Big(\frac{\partial\hat{\mathbf{u}}}{\partial\hat{\boldsymbol{\zeta}}}\Big),$$

where $\mathbf{V}$ is the covariance matrix of the "working" vector used in the PQL method, $\mathbf{Z}$ is the design random effect matrix associated to the normal model in the PQL estimation process, and $\hat{\mathbf{u}}$ is the prediction of the random effects vector $\mathbf{u}$. An alternative measure of uncertainty is the mean squared error (MSE) given by

$$MSE[\hat{b}_i] = E[(\hat{b}_i - b_i)^2] = E[(\tilde{b}_i - b_i)^2] + E[(\hat{b}_i - \tilde{b}_i)^2], \tag{2}$$

where $\tilde{b}_i$ is the prediction of b_i assuming that the variance parameters are known and $\hat{b}_i$ is the corresponding prediction when the variance parameters are unknown. In this paper, different MSE estimators used in the small area literature are considered. Petrucci and Salvati (2006) develop a MSE estimator for spatially correlated data. They propose the following MSE estimator

$$\widehat{MSE}[\hat{b}_i] = g_1(\hat{\sigma}^2,\hat{\lambda}) + g_2(\hat{\sigma}^2,\hat{\lambda}) + 2g_3(\hat{\sigma}^2,\hat{\lambda}),$$

where the terms g_1, g_2 and g_3 captures all sources of variability. In addition, we also consider a bootstrap MSE estimator given by

$$\widehat{MSE}^*[\hat{b}_i] = \frac{1}{J}\sum_{j=1}^{B}(\hat{b}_i^{*(j)} - b_i^{*(j)})^2, \tag{3}$$

where J is the number of bootstrap populations, $\hat{b}_i^{*(j)}$ is the prediction of the ith random effect for the jth bootstrap population, and $b_i^{*(j)}$ is the corresponding true bootstrap random effect. Finally, we also consider a double

bootstrap estimator proposed by Hall and Maiti (2006) by combining Estimator (3) with

$$\widehat{MSE}^{**}[\hat{b}_i] = \frac{1}{J}\sum_{j=1}^{J}\frac{1}{K}\sum_{k=1}^{K}(\hat{b}_i^{**(jk)} - b_i^{**(jk)})^2. \quad (4)$$

Here, $\hat{b}_i^{**(jk)}$ and $b_i^{**(jk)}$ are defined as $\hat{b}_i^{*(j)}$ and $b_i^{*(j)}$ but applying the bootstrap twice.

4 Illustration

The different prediction errors estimators will be applied to the well known Scottish lips cancer data to build confidence intervals for the relative risks, and then, to detect regions with extreme risks. In addition, a simulation study is conducted to assess the performance of the different estimators in terms of relative bias and relative root mean squared error, to check if there is a significant gain in efficiency when using second order approximations.

Acknowledgements

This research has been partially supported by the Health Department of the Government of Navarra, Spain, (project Res 1878/2001) and by the Spanish Ministry of Science and Education (project MTM 2005-00511).

References

Ainsworth, L.M. and Dean, C.B. (2006). Approximate Inference for Disease Mapping. *Computational Statistics & Data Analysis*, **50**, 2552-2570.

Breslow, N. E. and Clayton, D. G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, **88**, 9-25.

Dean, C.B., Ugarte, M. D. and Militino, A. F. (2004). Penalized Quasi-Likelihood with Spatially Correlated Data. *Computational Statistics & Data Analysis*, **45**, 235-248.

Hall, P. and Maiti, T. (2006). On Parametric Bootstrap Methods for Small Area Prediction. *Journal of the Royal Statistical Society Series B*, **68**, 221-238.

Petrucci A. and Salvati, N. (2006). Small Area Estimation for Spatial Correlation in Watershed Error Assessment. *Journal of Agricultural, Biological and Environmental Statistics*, in press.