

Selective Inference in Disease Mapping

Dolores Catelan, Annibale Biggeri

Dep. of Statistics “G. Parenti”, University of Florence;

Biostatistic Unit, ISPO, Florence;

catelan@ds.unifi.it

Abstract: The main goal of Disease Mapping is to investigate the geographical distribution of the risk of diseases. Spatially-structured priors were considered in all the proposed models in the literature to estimate relative risk surfaces. Selective inference on area-specific relative risks received little attention in the literature. We refer to selection and estimation of relative risks of areas at unusual (higher and/or lower) risk. Previous use of cross-validation posterior predictive distributions to detect outlying observation misses to address the selection effect in inference. In this work we review this issue in the context of hierarchical Bayesian models and we take advantage of a real example on the distribution of Lung cancer in Tuscany.

Keywords: Cross-validation predictive distributions, hierarchical Bayesian model, Disease Mapping.

1 Introduction

Disease mapping, i.e. the study of variability of disease occurrence on space, focused on relative risk surface estimation. Since the seminal paper of Clayton and Kaldor (1987) spatially-structured priors were considered in almost all the proposed models in the literature. However, inference on area-specific relative risks received little attention in the literature despite of the need to select areas (or regions) at unusual (high or low) risk. Stern and Cressie (2000) used cross-validation posterior predictive distributions to explore model fitting and identify outlying areas in disease mapping. The idea of cross-validation is to re-fit the model removing one observation in turn. The model is thus fitted to a subset of data Y_{-i} from which the i -th observation is dropped. The posterior predictive distribution $P(Y_i^{rep}|Y_{-i})$ for a replicate (Y_i^{rep}) of the i -th observation conditional to the remaining data Y_{-i} is then used for evaluation purposes. The extremeness is usually measured by some summaries over $P(Y_i^{rep}|Y_{-i})$, for example the posterior predicted p-values, $P(Y_i^{rep} \leq y_i|Y_{-i})$, or the conditional predictive ordinate, $p(Y_i^{rep} = y_i|Y_{-i})$. Marshall and Spiegelhalter (2003) noted that “...There are essentially two reasons why observations/regions may be divergent. First, the statistical assumptions underlying the model may be incorrect...[second], these regions could represent genuine ‘hot-spots’ of disease requiring further investigation.” Poor model fit is a reasonable explanation when a relevant number of observations/areas are identified as divergent while the presence of real hot-spots or outliers is the usual interpretation of few divergent ones.

Marshall and Spiegelhalter (2007) proposed a mixed approach to perform cross-validatory checks in disease mapping.

In this work we review this issue in the context of hierarchical Bayesian models and we take advantage of a real example on the distribution of Lung cancer in Tuscany.

2 Methods

Let Y_i be the number of observed cases in the i -th area ($i = 1, \dots, 287$) which follows a Poisson distribution with mean $E_i\theta_i$, where E_i is the expected number of cases under indirect standardization and θ_i the relative risk.

Besag et al. (1991) specified a random effect log linear model for the relative risk $\log(\theta_i) = u_i + v_i$. The heterogeneity random term u_i represents an unstructured spatial variability component assumed a priori distributed as Normal $(0, \lambda_u)$ where λ_u is the precision parameter modelled as Gamma. The clustering term v_i represents the structured spatial variability component assumed to follow a priori an intrinsic conditional autoregressive (ICAR) model. In other words, denoting S_i as the set of the areas adjacent to the i -th area, $v_i|v_{j \in S_i}$ is assumed distributed as Normal $(\bar{v}_i, \lambda_v n_i)$ where \bar{v}_i is the mean of the terms of adjacent areas to the i -th one (Besag and Kooperberg, 1995) and $\lambda_v n_i$ is the precision, which is dependent on n_i , the cardinality of S_i . Through these two random terms the BYM model shrinks the relative risk estimates both toward the local and the general mean.

The choice of a suitable combination of hyperparameters leads to different degrees of prior vagueness on the extent relative risk heterogeneity among areas.

For the Besag et al. (1991) model we took advantage of the proposal of Bernardinelli et al. (1995). The hyperpriors for the precision parameters were parameterized in terms of the ratio between the 95th percentile and the 5th percentile of the relative risk distribution.

2.1 Cross-validation predicted p-values

Divergence from the hierarchical null models is assessed via posterior predictive distribution. The posterior predictive distribution is:

$$P(Y^{rep}|Y) = \int P(Y^{rep}|Y, \theta)P(\theta|Y)d\theta = \int P(Y^{rep}|\theta)P(\theta|Y)d\theta$$

assuming conditional independence of Y^{rep} and Y given the parameters. This is too confident since the data are used twice, for deriving posteriors and for obtaining replicates (Plummer 2008). To control for excess in optimism the posterior predictive distribution is replaced by the cross-validation (leave-one-out) posterior predictive distributions:

$$P(Y^{rep}|Y_{-i}) = \int P(Y^{rep}|\theta)P(\theta|Y_{-i})d\theta$$

Cross validation posterior predicted distributions are computationally prohibitive. Several approximations have been proposed. A mixed approach was given by Marshall and Spiegelhalter (2007). At each Montecarlo iteration a replicate value for the random parameters for the i -th observation is generated and then used to generate a replicate observation Y_i^{rep} . This approach is called mixed because random effects are drawn from their predictive distribution and not from the posterior.

A measure of divergence can be the cross validation posterior predicted p values defined, using mid-p for a discrete response, as:

- if $Y_i > E_i$: $Pr(Y_i^{rep} > Y_i^{obs}|Y_{-i}) + \frac{1}{2}Pr(Y_i^{rep} = Y_i^{obs}|Y_{-i})$
- if $Y_i < E_i$: $Pr(Y_i^{rep} < Y_i^{obs}|Y_{-i}) + \frac{1}{2}Pr(Y_i^{rep} = Y_i^{obs}|Y_{-i})$

where Y_i is the observed and E_i the expected number of cases in the i -th area.

The need of post-processing of any model-based p-values was discussed by Ohlssen et al. (2007).

3 Results

Lung cancer death certificates were considered for males resident in the 287 municipalities of the Tuscany Region (Italy) for the period 1995-1999. Data were made available by the Regional Mortality Register. A set of reference rates (Tuscany, 1971-1999) have been used to compute the expected number of cases for each municipality, following indirect standardization and classifying the population by 18 age classes (0-5, ..., 85 or more).

We explored several choices of hyperprior parameters for the Besag et al. model. These choices are expressed as prior 90% centile range of relative risk among areas. They represent different beliefs about the background variability of disease risk. Each choice produced a different nested set of divergent observations. The priors defined by the hyperparameters are very informative. In some sense, we deliberately specified a series of constrained bad-fitting models, which represents a series of beliefs on the role of confounders in modifying the baseline risk among areas. A vague (non informative) null with leave-one out (leave-a-group out) cross-validation did not work in our Disease mapping context.

4 Conclusion and Discussion

This approach does not correspond to a Bayesian version of hypothesis testing because a mixture model is not specified. One consequence is that posterior probabilities may not protect to multiple testing. Post-processing of cross-validation

posterior predictive p-values was used by Spiegelhalter. Tri-level Bayesian model was proposed by Catelan et al (2010) in the context of Disease Mapping. Similar approaches to hierarchical modelling of the null are described in Ohlssen et al. (2007). The authors argued that fitting null model by leave-one out cross-validation may be sufficient to detect divergent observations. We disagree with this point, as we show in the results section. In Disease mapping hierarchical modelling of the null can be reached by specifying informative null priors. Prior predictive, posterior predictive and partial predictive distribution can be discussed also in this context.

References

- Bernardinelli, L., Clayton, D. and Montomoli, C.** (1995). Bayesian Estimates of disease maps : how important are priors? *Stat Med*, **14**, 2411–31.
- Besag, J., York, J., and Mollié, A.** (1991). Bayesian Image Restoration, with Two Applications in Spatial Statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, **43**, 1-59.
- Catelan, D., Lagazio, C., and Biggeri, A.** (2010). A hierarchical Bayesian approach to multiple testing in disease mapping. *Biometrical Journal*, **52**, 784-97.
- Clayton, D. and Kaldor, J.** (1987). Empirical Bayes Estimates of Age-Standardized Relative Risks for Use in Disease Mapping. *Biometrics*, **43**, 671-81.
- Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter D.** (2000): WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, **10**, 4, 325-37.
- Marshall, C.A. and Spiegelhalter, D.J.** (2003). Approximate cross-validated predictive checks in disease mapping models. *Stat Med*, **22**, 1649–60.
- Marshall, C.A. and Spiegelhalter, D.J.** (2007). Identifying outliers in Bayesian hierarchical models: a simulation-based approach. *Bayesian Analysis*, **2**, 409-44.
- Ohlssen, D.I., Sharples, L.D. and Spiegelhalter, D.J.** (2007). A Hierarchical Modelling Framework for Identifying Unusual Performance in Health Care Providers. *Journal of the Royal Statistical Society, Series A*, **170**, 865-90.
- Plummer, M.** (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics*, **9**, 523-39.
- Stern, H.S. and Cressie, N.** (2000). Posterior predictive model checks for disease mapping models *Stat Med*, **19**, 2377–97.