

Correlated gamma process prior for Bayesian mapping disease

Processo Gamma Correlato a Priori per l'analisi della distribuzione geografica di malattie

Maura Mezzetti

Istituto Metodi Quantitativi

Università Bocconi, Milano

maura.mezzetti@uni-bocconi.it

Riassunto: Si propone un modello gerarchico Bayesiano a due stadi per l'analisi geografica del rischio relativo di malattia. Al primo livello si assume che il numero osservato di casi della malattia d'interesse si distribuisca in accordo a una distribuzione di Poisson, con valore atteso il prodotto tra il rischio relativo specifico per area e il numero atteso di casi sotto ipotesi di omogeneità. Al secondo livello, si assume che il rischio relativo specifico per area si distribuisca in accordo a una distribuzione Gamma, il cui valore atteso risulta la media dei rischi relativi delle regioni confinanti. L'aspetto innovativo della metodologia proposta è l'introduzione di due processi latenti che controllano il grado di correlazione tra rischi relativi di regioni adiacenti e permettono un'analisi del grado di dipendenza tra la diffusione di malattia.

Keywords: Correlated gamma process, mapping disease, hierarchical model.

1. Introduction

Interest in the analysis of geographical variation in rates of disease (or mortality) is increasing since it is fundamental in the formulation and validation of aetiological hypotheses. The classical approach consists of mapping area-specific standardized mortality ratios (SMRs) Breslow and Day (1985) which are maximum likelihood estimates of relative risk under a Poisson model of the counts. Since this approach does not take into account any spatial pattern in disease, several authors have chosen a Bayesian approach, whose flexibility allow to incorporate prior opinion about spatial pattern in disease risk.

I propose a Bayesian hierarchical model, starting with the empirical Bayesian approach proposed by Clayton and Kaldor (1987). Their method is to *shrink* the SMRs towards a local or a global mean where the amount of shrinkage is determined by the reliability of the data. Through a Bayesian approach, we can easily obtain a smooth estimates by defining a prior which attributes a high correlation to two regions, (or census tracts), that are close to each other. The proposed Markovian prior process for the standardized mortality rates (SMR), starts from the gamma process introduced by Wolpert and Ickstadt (1998) and from an idea of Nieto-Barajas and Walker (2001) introduced in the context of survival analysis. The process introduced reflects the prior opinion of dependence between the SMRs at adjacent regions (or counties) and also maintains the convenient conjugacy property of the gamma prior with the Poisson distribution. The main innovation is that the model yields estimates of smoothed SMRs, where the degree of smoothing is estimated itself from the data.

2. Model

Within a map of n regions, let y_i denote the observed counts of disease in the i -th region, e_i denote the expected counts in the i -th region and λ_i denote the relative risk in the region. We assume that the expected counts are known constants, obtained under the hypothesis of homogeneity (stratified by age and sex). The observed counts in the i -th region is assumed to be Poisson distributed with mean $\lambda_i e_i$:

$$\begin{aligned} y_i | \mu_i &\sim \text{Poisson}(\mu_i) \\ \mu_i &= \lambda_i e_i. \end{aligned}$$

The likelihood of $\lambda = (\lambda_1, \dots, \lambda_n)$ is given by:

$$L(\lambda | D) = \prod_i \frac{(\lambda_i e_i)^{y_i} \exp(-\lambda_i e_i)}{y_i!}, \quad (1)$$

$$l(\lambda) = \log(L(\lambda)) = \sum_i y_i \log(\lambda_i e_i) - \sum_i \lambda_i e_i, \quad (2)$$

where $D = (\{y_i, e_i\}, i = 1, \dots, n)$ indicate the data.

From (2), maximum likelihood estimation of SMRs are $\frac{y_i}{e_i}$. Bayesian approaches in this context combine two types of information: one provided in each area by the observed cases described by the likelihood in (2) and used to compute the SMRs, and the other one is the prior information on the relative risks specifying their variability in the overall map and summarized by the prior distribution $[\lambda]$. Different prior distributions have been proposed for λ_i , considering either spatial dependence or spatial independence, either including or not covariates.

We do not define a distribution for $\lambda_i | \lambda_{j \neq i}$, the idea from Nieto-Barajas and Walker (2001) is the introduction of a latent process u_i . Thus, the model becomes

$$y_i | \mu_i \sim \text{Poisson}(\mu_i) \quad (3)$$

$$\mu_i = \lambda_i e_i \quad (4)$$

$$\lambda_i | u_i, c_i \sim \text{gamma}(\alpha_i + u_i, \gamma_i + c_i) \quad (5)$$

$$u_i | \lambda_1, \dots, \lambda_n, c_i \sim \text{Poisson}(c_i \bar{\lambda}_i) \quad (6)$$

$$\bar{\lambda}_i = \frac{1}{\#ne(i)} \sum_{k \in ne(i)} \lambda_k \quad (7)$$

$ne(i)$ indicates the set of regions adjacent to the i -th one, $\#ne(i)$ its cardinality, thus, $\bar{\lambda}_i$ indicates the mean of relative risks in the adjacent areas. By computing expected values, we can more easily understand the underlying process:

$$\begin{aligned} E(\lambda_i | u_i, c_i) &= \frac{\alpha_i + u_i}{\gamma_i + c_i}, \\ E(u_i | \lambda_i, c_i) &= c_i \bar{\lambda}_i, \\ E(\lambda_i | \lambda_{j \neq i}, c_i) &= \frac{\alpha_i + c_i \bar{\lambda}_i}{\gamma_i + c_i}. \end{aligned}$$

The process does not have the same autoregressive first order parameter in all the map. In a way, the parameter c_i controls *the strength* of the autoregressive structure, (I remind that

I am assuming that observations are space dependent, not time dependent). The greater is the value of c_i , the closer is the process to a first order autoregressive structure. Furthermore, we observe that $c_i = 0$ corresponds to an hypothesis of independence between adjacent regions. We introduce an other level on the hierarchical model:

$$c_i | \xi_i \sim \text{Exp}(\xi_i).$$

The introduction of this level represent the main innovation of the process: the *strength* of the correlation between adjacent regions is now a random process. Basically, the main idea is that we learn from the data about the spatial structure between regions.

After the computation of the full conditionals, I implemented the sampling algorithm, through a hybrid-Metropolis algorithm when necessary.

3. Application

The rates of lip cancer from 56 counties in Scotland have been analyzed several times, since Clayton and Kaldor (1987) introduced this example. I started from this example since it allows a comparison with other methodologies, and then consider other cancer sites, in particular lung and testis cancer, where the evidence of spatial aggregation is less evident.

Concerning the definition of the prior distribution for the parameters ξ_i we follow the next three schemes:

- a. ξ_i constant for the entire region, $c_i \sim \text{Exp}(\xi_i)$, and let the prior mean of the latent stochastic process c_i to be constant for each region.
- b. ξ_i as a function of the geographical position, the closer the region to the center, the greater the prior expected value for ξ_i .
- c. ξ_i is proportional to the number of neighbouring regions (this prior assumption is similar to that of Bernardinelli *et al.* (1995)).

The aim of this procedure is twofold, on one side the estimation of the area-specific relative risk and on the other hand the posterior estimation of the process c_i . As expected, the posterior estimation for the random process c_i , according to prior *a*, is significantly correlated with the number of neighbouring regions of each region. Higher the number of the regions at the border, stronger is the autoregressive structure. Higher the value of ξ_i , smoother are the values of SMRs. The Scottish counties are highly variable in size, shape and spatial distribution. Number of neighbours vary between 1 and 11. Thus since the appropriateness of intrinsic autoregressive is ambiguous, estimating the correlation between neighbouring regions from the data seem reasonable.

The comparison of the posterior estimations of the process c_i respect to the different cancer sites reveals that the hypothesis of spatial pattern is plausible only for the lip cancer. The posterior estimations are close to zero when referred to lung and testis cancer. Furthermore, lower the number of observed cases, stronger the autoregressive structure tends to be (lip cancer is a rare disease).

We assume that some explanatory variables can explain the spread of the SMRs geographically. Increased exposure to sunlight has been implicated in the excess occurrence of lip cancers among rural populations, producing high rates in people who work outdoors. For each county, one covariate measuring the percentage of the population engaged in agriculture, fishing, and forestry was measured. This covariate is suspected to

be related to sunlight exposure. At the present time, the covariates are not included in the model, only in a *a posteriori* interpretation of the estimates. The posterior estimation of c_i according to the three different prior assumptions tend to confirm the effect of the covariate, they tend to be higher when adjacent regions have the same values of covariates, and lower when adjacent regions have different behavior in terms of sunlight exposure. A slightly similar effect can be found between testis cancer and a socio economic status indicator. Comparing the posterior estimation of c_i from one side confirm the presence of the known covariate, on the other hand could be itself an hint of some unmeasured covariate with an unclear association with the disease. Furthermore, covariates information can help itself in definition of a prior distribution for the process c_i .

4. Conclusion

The idea of using a Bayesian hierarchical model is to impose a plausible structure of spatial relatedness on the unknown relative risks by modeling them collectively as a spatial stochastic process. The idea is to use a Markovian first order autoregressive model on the standardized mortality ratios SMR. A Gamma process result particularly appealing since its coniugacy with the Poisson process. The main innovation is the introduction of two latent processes that let the data control the strength of the autoregressive process. Our process allow on one side the estimation of area specific relative risk and on the other hand to investigate on the nature of spatial structure. Identification of usual low or high risk can help the actions to be taken, since may be used also in providing a 'clean' map of the disease risk to allow better resource allocation and risk assessment.

References

- Bernardinelli L., Clayton D. and Montomoli C. (1995) Bayesian estimates of disease maps: How important are priors?, *Statistics in Medicine*, 14, 2411–2431.
- Breslow N. and Day N. (1985) *Biostatistics: Statistics in Biomedical, Public Health and Enviromental Sciences*, Elsevier Science Publishers V.B, Noth Holland.
- Clayton D. and Kaldor J. (1987) Empirical bayes estimates of age-standardized relative risks for use in disease mapping, *Biometrics*, 43, 671–681.
- Nieto-Barajas L. and Walker S. (2001) Markov gamma and beta processes for modelling hazard rates, *Scandinavian Journal of Statistics (to appear)*.
- Wolpert R. and Ickstadt K. (1998) Poisson/gamma random field models for spatial statistics, *Biometrika*, 85, 251:267.