Assessment of the INLA approach on gerarchic bayesian models for the spatial disease distribution: a real data application

Valutazione dell'approccio basato su INLA in modelli gerarchici bayesiani per la distribuzione spaziale di malattia: un'applicazione a dati reali

Paolo Girardi (1), Emanuela Bovo (2), Carmen Stocco (2), Susanna Baracco (2), Alberto Rosano (2), Daniele Monetti (2), Silvia Rizzato (2), Sara Zamberlan (2), Enrico Chinellato (2), Ugo Fedeli (1), Massimo Rugge (2,3)

Sommario The use of approximate methods as the INLA (Integrated Nested Laplace Approximation) approach is being widely used in Bayesian inference, especially in spatial risk model estimation where the Besag-York-Molliè (BYM) model has found a proper use. INLA appears time saving compared to Monte Carlo simulations based on Markov Chains (MCMC), but it produces some differences in estimates [1, 2]. Data from the Veneto Cancer Registry has been considered with the scope to compare cancer incidence estimates with INLA method and with two other procedures based on MCMC simulation, WinBUGS and CARBayes, under R environment. It is noteworthy that INLA returns estimates comparable to both MCMC procedures, but it appears sensitive to the a-priori distribution. INLA is fast and efficient in particular with samples of moderate-high size. However, care must to be paid to the choice of the parameter relating to the a-priori distribution.

Sommario *L'uso dei metodi basati sull'approssimazione di Laplace come INLA (Integrated Nested Laplace Approximation) è ampiamente utilizzato nell'inferenza Bayesiana, specialmente in modelli di rischio spaziale dove il modello di Besag-York-Molliè (BYM) ha trovato un uso appropriato. INLA permette un risparmio di tempo computazionale rispetto alle simulazioni Monte Carlo basate su Catene Markov (MCMC), ma produce alcune differenze nelle stime [1, 2]. Vengono considerati i dati del Registro dei Tumori del Veneto con lo scopo di di confrontare le stime ottenute con INLA rispetto a due procedure basata su MCMC, WinBUGS e CARBayes, svolte in ambiente R. E' importante notare che INLA restituisce stime comparabili ad entrambe le procedure MCMC, ma è sensibile alla distribuzione a priori. INLA è un metodo rapido ed efficiente, in particolare con campioni di elevata numerosità. Tuttavia, occorre prestare attenzione alla scelta del parametro relativo alla distribuzione a priori.*

(1) Sistema Epidemiologico Regionale, Azienda Zero, Padova.
mail: paolo.girardi@aulss6.veneto.it
(2) Registro Tumori del Veneto, Azienda Zero, Padova
(3) Dipartimento di Medicina DIMED, Università di Padova, Padova

Key words: BYM model, Cancer Registry, INLA, Laplace approximation, Bayesian methods

1 Introduction

In recent literature, the use of approximate methods in Bayesian inference has reported a great popularity. The Laplace approximation proposed by [Rue H., 2009] with the INLA acronym (Integrated Nested Laplace Approximation) has been adapted to the parameter estimations of an increasing number of statistical models; in addition, several papers have reported its use in wide range of real data applications. INLA offers the opportunity to perform Bayesian analyses through numerical integration avoiding extensive iterative computation; it usually implies a lower computational time respect to the classical Monte Carlo simulations based on Markov Chains (MCMC) with dedicated software (WinBUGS, OpenBUGS or JAGS). The major gain of INLA is the replacing of long chains used by MCMC methods to produce a-posteriori estimates of the coefficients distribution with a Laplace approximation of the a-posteriori distribution. Among hierarchical Bayesian models, the Besag-York-Mollièe (BYM) model [4] has became popular for the analysis of spatial distribution of occurrences in epidemiology (disease risk, mortality, etc...), in financial services (investments, prices) and in demography and sociology (deprivation index, unemployment rate, etc..). The availability of the INLA package for R software [5] has allowed an easy and friendly implementation of INLA for BYM models. However, recent publications show that INLA produces considerable differences in estimates [1, 2] and research on this topic remains already unexplored. The aim of the study is to compare risk estimates produced by INLA with those one of the MCMC simulations using a series of real data applications instead of simulations.

2 Materials and methods

2.1 Veneto Cancer Registry

We consider all the cases of malignant cancers occurring in the year 2013 in the Veneto Region, one of the largest Region in Italy covering about five million of inhabitants. The area covered by the Veneto Cancer Registry includes the 96% of the territory (Figure 1). Every cancer case has been coded with the X version of International Classification of the Diseases (ICD-X) and has been aggregated at the municipality level (n=556 municipalities). In our comparison we consider 7 different primitive sites that have different number of cases: all the sites except skin, Hodgkin's lymphoma, myeloma and pancreas cancer among men; cancer of breast, cervix and cancer of esophagus among women.

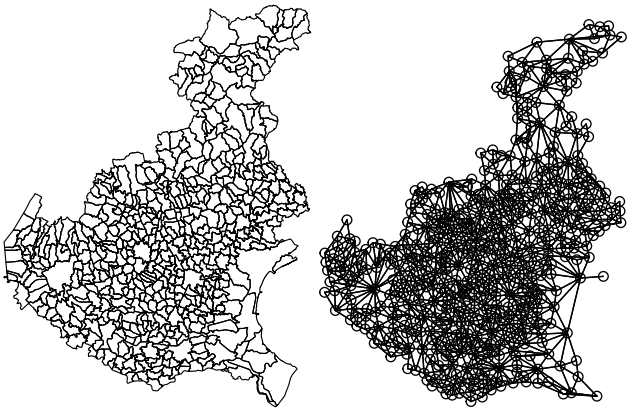


Figura 1 Boundaries and adjacency matrix of the 556 municipalities.

2.2 BYM model

The Standardized Incidence Ratios (SIR) have been estimated by means of a BYM model. The number of observed cases O_i is assumed to follow a Poisson distribution as

$$O_i \sim Poisson(\lambda_i) \tag{1}$$

where λ_i is the mean/variance parameter. Considering e_i the expected number of cases for the $i-th$ area calculated by an indirect standardization using the registry pool as reference, the estimated SIR is connected to the linear predictor η_i as follows

$$log(SIR_i) = log(\frac{\lambda_i}{e_i}) = \eta_i = (\alpha + \mu_i + \nu_i), \tag{2}$$

where α is the intercept quantifying the average incidence rate in all the 556 municipalities, while μ_i and ν_i are the correlated and uncorrelated spatial effects, following a normal distribution. While τ_ν is assumed to be distributed as a white noise ($\nu_i \sim N(0, \frac{1}{\tau_\nu})$), the μ_i distribution is modelled using an intrinsic conditional autoregressive structure (ICAR) as follow

$$\mu_i \sim N(\frac{1}{n_j}\sum_{\partial j} O_j, \frac{1}{n_j \tau_\mu}) \tag{3}$$

where O_j are the cases observed in ∂j which denotes the n_j municipalities bordering the *i-th* area, *i-th* area excluded. The precision parameters τ_μ and τ_ν follow a Gamma distribution. SIR has been estimated by INLA using R-INLA and by MCMC procedures using R2WinBUGS and CarBayes packages under R environment. For MCMC simulation, we took into account the results of 15.000 iterations discarding the first 5.000 as burn-in.

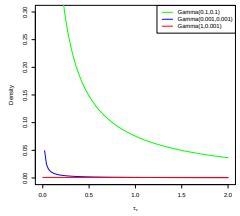


Figura 2 Distributions of the parameter τ_μ.

The study considers three different distributions for the precision of the spatial parameter τ_μ: $\Gamma(0.1,0.1)$, $\Gamma(0.001,0.001)$ and $\Gamma(1,0.001)$. The parameter of the precision related to the uncorrelated spatial effect has been fixed to be distributed as a $\Gamma(0.001,0.001)$.

3 Results

The main characteristics of the selected cancer sites are reported in Table 1. The sites are ordered decreasing the number of observed cases. A high number of male cases (15'416) is registered taking into account all the cancer sites except the skin; conversely, a low number of cervix cancers among women is reported, equal to 200, less than 1 per municipality. The average number of cases for municipality is always lower than the variance estimates indicating an over-dispersion. The p-values associated to the Moran's I test applied to empirical SIR ($\frac{o_i}{e_i}$) support the spatial independence for the distribution of each considered primitive site.

Cancer site	**Total cases**	Average	Variance	Moran's I test (p-value)
All sites (men)	15'416	27.7	5063.8	0.430
Breast (women)	4'372	7.9	535.5	0.185
Pancreas (men)	535	1.0	8.10	0.065
Cervix cancer (women)	200	0.4	1.5	0.796
Myeloma (men)	199	0.4	1.2	0.758
Hodgkin's lymphoma (men)	101	0.2	0.6	0.483
Esophagus (women)	59	0.1	0.4	0.163

Tabella 1 Characteristics of the selected cancer sites and p-value associated to Moran's I test for the empirical SIR.

SIR estimates are calculated for each selected cancer site varying the distribution of the precision parameter τ_μ computing the Pearson correlation index between INLA and MCMC-based estimates. The results are reported in Table 2. The correlation indices ranges from 0.344 in esophageal cancer with a-priori distribution $\tau_\mu \sim \Gamma(1,0.001)$, which indicates a poor agreement between INLA and CARBayes estimates, to 0.998/0.996 relatively to all male cancer sites with $\tau_\mu \sim \Gamma(0.1,0.1)$ resulting in a perfect overlapping between INLA and WinBUGS/CARBayes methods. The degree of agreement between INLA and MCMC procedures depends on: 1) the a-priori distribution of the variance of spatial component; 2) the number of incident cases. Overall, the best agreement (all r' Pearson indices $>$0.9) is obtained choosing a $\Gamma(0.1,0.1)$ for the τ_μ. As reported in Table 3 INLA returns estimates faster than MCMC procedure (about 15/20 times).

Cancer site	MCMC procedures	Distribution of τ_μ $\Gamma(0.1,0.1)$	$\Gamma(0.001,0.001)$	$\Gamma(1,0.001)$
All cancers (men)	WinBUGS / CarBayes	0.998 / 0.996	0.992 / 0.987	0.990 / 0.985
Breast (women)	WinBUGS / CarBayes	0.997 / 0.995	0.994 / 0.988	0.992 / 0.983
Pancreas (men)	WinBUGS / CarBayes	0.997 / 0.995	0.987 / 0.949	0.983 / 0.976
Cervix cancer (women)	WinBUGS / CarBayes	0.986 / 0.945	0.966 / 0.947	0.961 / 0.872
Myeloma (men)	WinBUGS / CarBayes	0.910 / 0.966	0.925 / 0.963	0.948 / 0.969
Hodgkin's lymphoma (men)	WinBUGS / CarBayes	0.981 / 0.917	0.763 / 0.850	0.930 / 0.882
Esophagus (women)	WinBUGS / CarBayes	0.955 / 0.935	0.858 / 0.937	0.802 / 0.344

Tabella 2 Correlation index between INLA and MCMC-based methods on SIR estimates by distribution of τ_μ.

Although r's pearson index indicates a high agreement between INLA estimates compared to the MCMC-based methods, the graphical analysis permits to verify the presence of marked differences in the estimated risks (Figure 3), for example, relatively to the Myeloma SIR. The difference is marked considering the spatial distribution of the esophageal cancer incidence among women obtained by a comparison between INLA and CARBayes procedures (Fig. 4) that in Table 2 reports a weak agreement.

Procedures	$\Gamma(0.1,0.1)$	$\Gamma(0.001,0.001)$	$\Gamma(1,0.001)$
INLA	5.50	4.04	9.7
WinBUGS	87.25	90.99	85.93
CARBayes	68.7	64.1	68.9

Tabella 3 Computation time for INLA, WinBUGS and CARBayes esophageal SIR estimates.

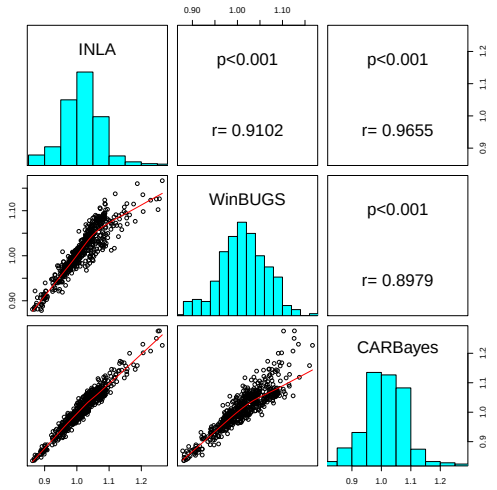
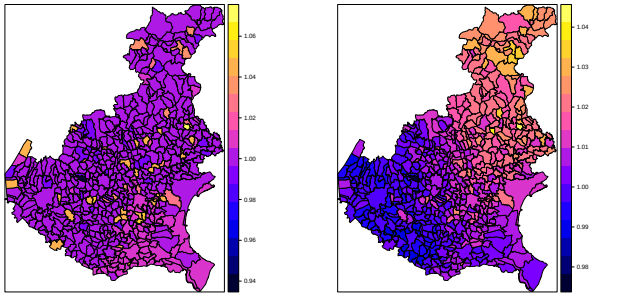


Figura 3 Esophagus SIR distribution among women with $\tau_\mu \sim \Gamma(0.1,0.1)$ estimated with INLA and with two MCMC-based procedures (WinBUGS and CARBayes).

4 Conclusions

In presence of non-informative a-priori distributions, INLA and MCMC procedures reported different estimates, even more clean-cut considering low sample size. INLA confirms to be a fast and efficient method for spatial risk estimation and, in general, for hierarchical Bayesian models [6, 7]. However, in order to avoid an



(a) INLA SIR estimates $\tau_\mu \sim \Gamma(1, 0.001)$

(b) CARBayes MCMC SIR estimates $\tau_\mu \sim \Gamma(1, 0.001)$

Figura 4 Distribution of Esophageal SIR estimated by INLA and CARBayes.

over-smoothing of the risks and/or excessive imprecision of the estimates, particular attention must to be paid to the choice of the a-priori distribution for the variance of the spatial component. Further analyses are required in order to assess the comparability of INLA and MCMC estimates looking at the distribution of the uncorrelated spatial parameter and at the presence of spatial dependence.

Riferimenti bibliografici

1. De Smedt, T., Simons, K., Van Nieuwenhuyse, A., Molenberghs, G. (2015). Comparing MCMC and INLA for disease mapping with Bayesian hierarchical models. Archives of Public Health, 73(1), O2.
2. Carroll, R., Lawson, A. B., Faes, C., Kirby, R. S., Aregay, M., Watjou, K. (2015). Comparing INLA and OpenBUGS for hierarchical Poisson modeling in disease mapping. Spatial and spatio-temporal epidemiology, 14, 45-54.
3. Rue, H., Martino, S., Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the royal statistical society: Series b (statistical methodology), 71(2), 319-392.
4. Besag, J., York, J., Mollie, A. Bayesian image restoration with two applications in spatial statistics (with discussion) Ann Inst Stat Math. 1991; 43: 1–59. doi: 10.1007. BF00116466
5. Lindgren, F., Rue, H. (2015). Bayesian spatial modelling with R-INLA. Journal of Statistical Software, 63(19).
6. Blangiardo, M., Cameletti, M., Baio, G., Rue, H. (2013). Spatial and spatio-temporal models with R-INLA. Spatial and spatio-temporal epidemiology, 4, 33-49.
7. Bilancia, M., Demarinis, G. (2014). Bayesian scanning of spatial disease rates with integrated nested Laplace approximation (INLA). Statistical Methods & Applications, 23(1), 71-94.