Towards a Multidimensional Approach to Bayesian Disease Mapping. Supplementary material.

Martinez-Beneito, MA. Botella-Rocamora, P & Banerjee, S.

Description of the simulation study

This document shows the results of a simulation study carried out to illustrate the performance of some of the proposed models in the paper. Specifically, we pursued to assess two particular issues of these models: First, how DIC performs as model selection criterion to compare models within the proposed framework and, second, to assess the ability of the models to estimate the variance-covariance matrix between the log-risks of the different geographical patterns modelled.

All our simulated datasets involved 4 different observed counts for each municipality, supposedly corresponding to the observed deaths for two different diseases and both sexes. Regarding to the log-risks simulated, we compared 3 different settings: The first one assumes full independence between counts in each municipality; the second one assumes a separable dependence structure between diseases and sexes and finally the last one assumes a non-separable relationship between these two factors. More in detail, let us assume that the log-risks to be modelled $\mathbf{RR}$ corresponded, in this order to (disease 1-sex 1, disease 1-sex 2, disease 2-sex 1, disease 2-sex2) then the independent setting would assume:

$$vec(\log(\mathbf{RR})) \sim N_{4J}(\mathbf{0}_{4J}, 0.5^2\mathbf{I}_4 \otimes (\mathbf{D} - 0.95\mathbf{W}))$$

for $J = 540$, the number of municipalities in the Valencian Region, $\mathbf{W}$ the adjacency matrix of that region and $\mathbf{D}$ a diagonal matrix of elements $\mathbf{W1}_J$. Therefore, this setting reproduces four independent patterns with proper CAR spatial dependence of correlation parameter equal to 0.95, since only values of this parameter close to 1 reproduce substantial spatial dependence. Similarly, the separable setting assumes:

$$vec(\log(\mathbf{RR})) \sim N_{4J}\left(\mathbf{0}_{4J}, 0.5^2\begin{pmatrix} 1 & 0.5 & 0.5 & 0.25 \\ 0.5 & 1 & 0.25 & 0.5 \\ 0.5 & 0.25 & 1 & 0.5 \\ 0.25 & 0.5 & 0.5 & 1 \end{pmatrix} \otimes (\mathbf{D} - 0.95\mathbf{W})\right).$$

This setting assumes that, given the log-risks for a disease and sex, the corresponding log-risks for the alternative disease or sex has correlation 0.5 with the original pattern. Moreover, the effect of changing both disease and sex on the corresponding correlation between log-risks is multiplicative, i.e. changing both factors reduces the corresponding correlation to $0.25 = 0.5 \cdot 0.5$, what reproduces the separable correlation structure that

we pursued. Finally, the non-separable setting assumes:

$$vec(\log(\mathbf{RR})) \sim N_{4J}\left(\mathbf{0}_{4J}, 0.5^2 \begin{pmatrix} 1 & 0.5 & 0.5 & 0.75 \\ 0.5 & 1 & 0.75 & 0.5 \\ 0.5 & 0.75 & 1 & 0.5 \\ 0.75 & 0.5 & 0.5 & 1 \end{pmatrix} \otimes (\mathbf{D} - 0.95\mathbf{W})\right).$$

In this case, the correlation between patterns when changing either the disease or sex is once again 0.5, but changing both factors produces a correlation of 0.75 what makes the correlation structure non-separable. Note that the variance-covariance matrix chosen for this model is both symmetric and positive definite.

We generated 10 different data sets for each one of the settings above. The expected counts for each data set were generated as the product of the expected cases from the lung cancer/diabetes example in Section 5 of the paper and the corresponding simulated relative risks. The final data sets were generated as Poisson draws from these expected counts. For each one of these 30 datasets we run three different models: a model assuming independence between geographical patterns, a second one assuming a separable multivariate dependence structure and finally a model assuming a non-separable factorial relationship between disease and sex. For all these three settings proper CAR distributions were also assumed for modelling the spatial dependence of the data sets.

Results

Table 1 shows for each one of the settings considered (rows in the table) the number of times that every model (columns in the table) has achieved the lowest DIC when compared with the other alternative models run. As can be appreciated, the DICs determine the correct setting in a 76.7% of the data sets. It is also interesting to check how the DIC never points out towards a model which is less complex than the true one. Therefore, DIC is quite good detecting the complexity of data when it really exists (is quite sensitive in this sense) although it is no so specific when it points out to a complex model. Nevertheless, we find convenient to mention that the models considered are nested, i.e. independent settings can be reproduced within separable and non-separable models and separable settings can be reproduced within non-separable models. Therefore it is not so wrong that a more complex alternative model is determined as the best option in our study since that data set can be also reproduced with the complex alternative models. Nevertheless, we acknowledge that a higher penalization of complexity in DIC would seem convenient according to the results shown.

Table 1: Number of times that every model is selected as the best option according to DIC. Rows in the table stand for the simulated setting and columns for the model run.

	Independent	Separable	Non-separable
Independent	5	2	3
Separable	0	8	2
Non-separable	0	0	10

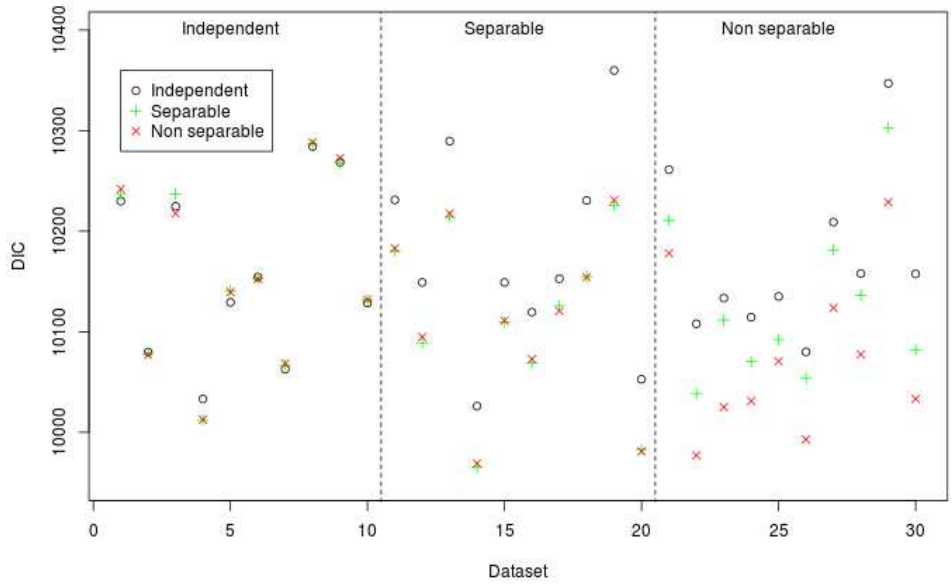

Figure 1: Comparison of DICs for the 10 models run in each setting.

Figure 1 shows the DIC obtained for each data set and model run. As can be appreciated differences in DIC are in general minor when we compare the true model generating the data with a more complex alternative. On the contrary, when the true model is compared with a simpler alternative differences in their DICs are much more evident in favour of the true model. So, when DIC points out towards a wrong model its difference with the true model is usually very mild.

We are now going to explore the ability of the models run to estimate the original variance-covariance matrix between geographical patterns. Figure 2 shows for the first of the non-separable simulated data sets the results retrieved for all three models. At this figure the red points correspond to the cells of the true variance-covariance matrix, the first 4 points corresponding to the first row of that matrix, the 4 next to the second row and so on. The vertical gray bars with the same x-coordinate than these points correspond

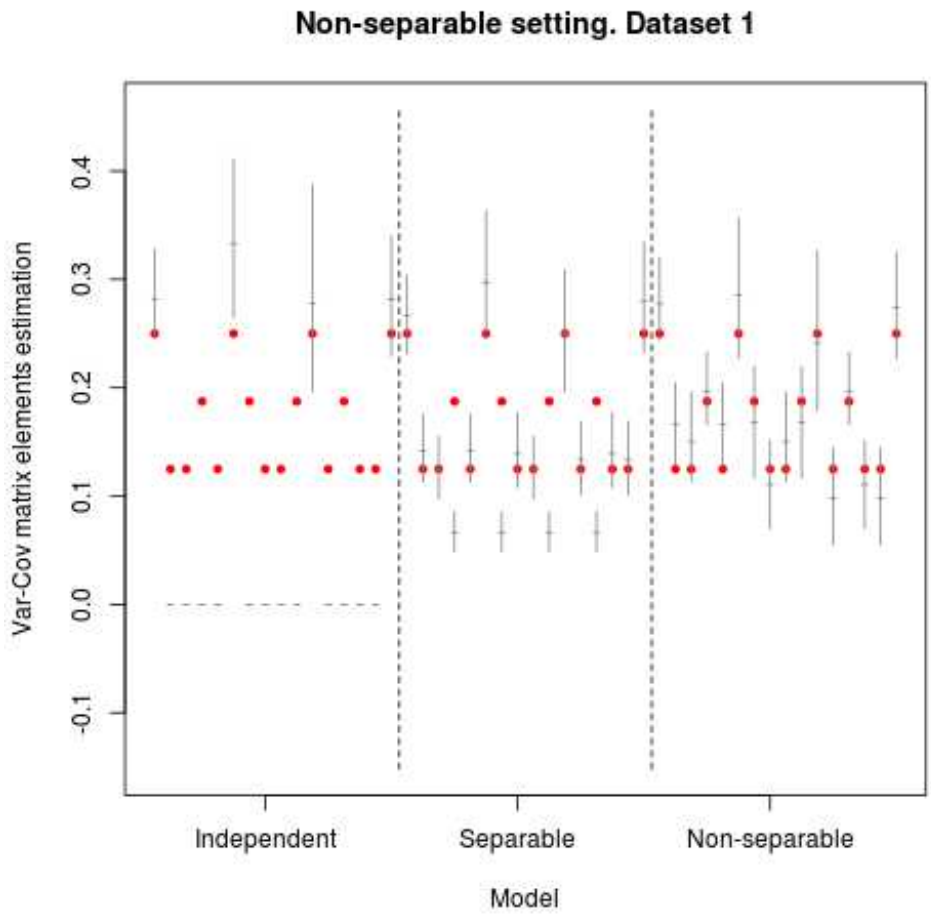

Figure 2: Variance-covariance estimates for the first non-separable data set and the different models run.

to the estimate of the corresponding cell in the variance-covariance matrix. Specifically, that bar correspond to the 80% central posterior credible interval of that covariance and the central horizontal bars denote the posterior median for each one of them. Obviously, the independent and separable models unsurprisingly are not able to reproduce the non-separable covariance pattern, meanwhile the non-separable model reproduces it quite well. The annex archives `MatVarCovInd.pdf`, `MatVarCovSep.pdf` and `MatVarCovNonSep.pdf` shows these same figures for all 30 simulated data sets.

Table 2 shows the number of times that the 80% posterior credible intervals of the estimated variances and covariances contain the corresponding true values. The non-separable model shows excellent and similar coverage rates for all three settings. In contrast, the independent and separable models only attain a reasonable coverage for

those settings that they are supposed to fit well. Nevertheless, for these settings, the empirical coverage attained is higher than the corresponding hypothetical value of 80%. However, it should be born in mind that some of the covariances are necessarily well estimated by these models. For example, the independent model always estimates a covariance of 0 between geographical patterns. So, this model estimates properly all the covariances in the independent setting. If we limit ourselves in this case to the variance estimates, leaving aside the covariances, the 80% posterior CIs contain the true value in 33 out of 40 times, i.e. 82.5% of the total. So, the behaviour of this model also seems correct in this setting. Something similar happens with the separable and non-separable data sets for this model. Leaving aside the covariance estimates which are always wrong for this model we have that the coverage rates for these settings are, respectively, 80% and 87.5%. So once again the performance of the independent model seems reasonable even though the setting under study is different to that corresponding to that model.

Table 2: Number of times that the 80% CIs contain the true values for each setting and model. Rows in the table stand for the simulated setting and columns for the model run.

	Independent	Separable	Non-separable
Independent	0.96	0.84	0.76
Separable	0.20	0.89	0.80
Non-separable	0.22	0.52	0.79