COBRA Preprint Series

Year 2010 *Paper* 75

A Bayesian shared component model for genetic association studies

Juan J. Abellan [*] Carlos Abellan [†]

Juan R. Gonzalez [‡]

[*]Centre for Public Health Research (CSISP), Valencia, Spain, abellan_jua@gva.es

[†]Centre for Public Health Research (CSISP), Valencia, Spain

[‡]Centre for Research on Environmental Epidemiology (CREAL), Barcelona, Spain

http://biostats.bepress.com/cobra/ps/art75

A Bayesian shared component model for genetic association studies

Juan J. Abellan, Carlos Abellan, and Juan R. Gonzalez

Abstract

We present a novel approach to address genome association studies between single nucleotide polymorphisms (SNPs) and disease. We propose a Bayesian shared component model to tease out the genotype information that is common to cases and controls from the one that is specific to cases only. This allows to detect the SNPs that show the strongest association with the disease. The model can be applied to case-control studies with more than one disease. In fact, we illustrate the use of this model with a dataset of 23,418 SNPs from a case-control study by The Welcome Trust Case Control Consortium (2007) with 2,000 patients with diabetes type 1, 2,000 with diabetes type 2 and a control group with 3,000 individuals. We carry out a simulation study to assess the sensitivity and specificity of our model to detect SNPs with excess risk. Our results show that the method we propose here can be a very useful tool for this type of studies. The model has been implemented in the bayesGen library of the R statistical package.

A Bayesian shared component model for genetic association studies

Juan Jose Abellan[1,3,*], Carlos Abellan[1], Juan Ramon Gonzalez[2,3]

[1]Centre for Public Health Research (CSISP), Valencia, Spain
[2]Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain
[3]CIBER Epidemiología y Salud Pública (CIBERESP) Spain

Abstract

We present a novel approach to address genome association studies between single nucleotide polymorphisms (SNPs) and disease. We propose a Bayesian shared component model to tease out the genotype information that is common to cases and controls from the one that is specific to cases only. This allows to detect the SNPs that show the strongest association with the disease. The model can be applied to case-control studies with more than one disease. In fact, we illustrate the use of this model with a dataset of 23 418 SNPs from a case-control study by The Wellcome Trust Case Control Consortium (2007) with 2 000 patients with diabetes type 1, 2 000 with diabetes type 2 and a control group with 3 000 individuals. We carry out a simulation study to assess the sensitivity and specificity of our model to detect SNPs with excess risk. Our results show that the method we propose here can be a very useful tool for this type of studies. The model has been implemented in the *bayesGen* library of the *R* statistical package.

Keywords: Bayesian hierarchical models; Latent-variable models; Genotype; Phenotype.

1 Introduction

The main goal of genetic association (GA) epidemiological studies is to assess the potential relationship between genotype and phenotype, typically a disease. The genotype information of an individual is contained in the single nucleotide polymorphisms (SNP) from genes that might be related to the disease of interest. From a statistical point of view, the problem has been traditionally addressed with simple hypothesis testing. However, the number of SNPs that can be considered in a given case-control study can be really large (hundreds of thousands), especially considering the latest advances in laboratory techniques. Testing association between individual SNPs and disease becomes 'mass' hypothesis testing. This in turn poses the multiple-comparison problem in hypothesis testing so that individual p-values need to be corrected to keep the false discovery rate (FDR) under control. There are popular statistical tools that carry out this type of analysis (see e.g. Gonzalez et al., 2007). The appeal of this approach is its simplicity because tests are quick and easy to apply.

Another common characteristic in GAs is the relative small number of individuals compared to the number of SNPs. This large p small n problem prevents the use of standard statistical techniques such as logistic regression modes. More sophisticated regression-like methods have been proposed recently in the literature to overcome these limitations. These techniques address the problem by considering SNPs as independent variables and the disease group as response variable. The aim is to find the SNPs that best explain disease risk. Kooperberg et al. (2001) proposed a logic regression method that searches for presence/absence combinations of SNPs that best explain disease risk. Sha et al. (2004) suggest using Bayesian stochastic search variable selection in probit regression models. Their method is based on trans-dimensional models fitted with reversible jump Markov chain Monte Carlo (RJMCMC) techniques.

*Author for correspondence: abellan_jua@gva.es

Hoggart et al. (2008) propose a Bayesian-inspired stochastic search variable selection algorithm that uses a penalised maximum likelihood approach to estimate regression coefficients in logistic regression.

1.1 Model motivation

In most GA studies, most of the SNPs considered have a similar pattern in cases and controls, and only a few, if any, are expected to show differences between the two groups, hence suggesting association with the disease of interest. In other words, cases and controls share allele frequencies for most SNPs. This leads us to propose here the use of shared component models as a new approach to address GA studies. These models are inspired in factor analysis (FA), a well-known statistical technique for dimension reduction in datasets with many variables that are believed to share a lot of information. In standard FA observed variables are assumed to be the outcome of linear combinations of unobserved (and possibly unobservable) latent variables called factors plus a residual or specific component unique to each variable. In the Bayesian framework both the common and specific factors are considered as random effects and are assigned prior distributions. An interesting characteristic of the Bayesian paradigm is that these priors may account for potentially complex autocorrelation structures if needed.

Bayesian shared component models have been used in a number of studies in several contexts in recent years. In finance, Aguilar and West (2000) apply a shared component model with a single common component in the context of currency exchange rates and introduce temporal autocorrelation in the common factor. In the context of socio-economic indices, Hogan and Tchernis (2004) apply this type of models to re-derive the Townsend index of material deprivation (Townsend et al., 1985) using the original four census variables but accounting for spatial autocorrelation. In epidemiology, these models have also arisen in a number studies in the context of spatial and spatio-temporal disease mapping to assess what is common in the geographical and temporal distribution of disease risk (see e.g. Knorr-Held and Best, 2001; Richardson et al., 2006). To the best of our knowledge, this is the first time that this type of models are used in GA studies.

The rest of the paper is organised as follows: in Section 2 we specify our Bayesian shared component model; Section 3 illustrates its application to a case study using data from the The Wellcome Trust Case Control Consortium (2007); in Section 4 we perform a simulation study to assess the sensitivity and specificity of our model to detect true SNP-disease associations; we conclude with some discussion in Section 5.

2 Model specification

The data in GA studies typically are counts X_{ijd} of the number copies of the variant allele in the SNP $j = 1, \ldots, p$ of individual $i = 1, \ldots, n$ in the disease group $d = 1, \ldots, D$, where say $d = 1$ is the control group. Since there are two alleles in each SNP, then $X_{ijd} \in \{0, 1, 2\}$. This corresponds to the so-called additive association model, which assumes that disease risk increases with each copy of the variant allele, so that, if a SNP is associated with the disease, an individual with two copies of the variant allele has double risk than an individual with just one copy. A simplification of these models are the dominant (recessive) models in which $X_{ijd} \in \{0, 1\}$ indicating presence of the dominant (recessive) allele.

In what follows, we will assume without loss of generality that we are in the additive scenario where $X_{ijd} \in \{0, 1, 2\}$. We will also assume that the SNPs to be analysed are in equilibrium according to the Hardy-Weinberg principle (Hardy, 1908; Weinberg, 1908). Last, we will assume that all the individuals in the same disease group have the same chance of having a variant allele in a given SNP. In other words,

$$X_{ijd} \sim \text{Binomial}(2, \pi_{jd}) \tag{1}$$

where π_{jd} is the probability of finding a copy of the variant allele in SNP j in individuals falling in disease group d. If the SNP j is not associated with the disease, then one would expect π_{jd} to be similar across disease groups. In contrast, if that SNP is related to one disease group, say d_0, the corresponding π_{jd_0} should be different from $\pi_{jd}, d \neq d_0$.

Let $Y_{jd} = \sum_{i=1}^{n_{jd}} X_{ijd}$ be the total number of variant alleles found in SNP j and disease group d, where n_{jd} is the number of individuals in disease group d with non-missing information on SNP j. Then, assuming independence between individuals, from Equation (1) we have

$$Y_{jd} \sim \text{Binomial}(2n_{jd}, \pi_{jd}) \tag{2}$$

We then decompose the variability of the π_{jd} into a component shared by all disease groups plus a specific component for each disease. More specifically, we assume

$$\text{logit}(\pi_{jd}) = \alpha_d + \delta_d \cdot \theta_j + \lambda_{jd} \tag{3}$$

where, α_d is a group-specific intercept, θ_j is the component shared by all disease groups (cases and controls), δ_d represent the loadings of the common component into each disease group d and λ_{jd} are the disease-specific components.

Expressions (2) and (3) specify the first layer of our Bayesian shared component model. Figure 1 shows a schematic representation of our model for the simple situation where there is just one disease group besides the control group.

2.1 Identifability issues

The usual identifiability issues in FA apply to Bayesian shared-component models as well. In particular, the identifiability of scale of the loadings and the shared component in Equation (3). Since, for any constant $c \neq 0$

$$\delta_d \cdot \theta_j = c\delta_d \cdot \frac{\theta_j}{c}$$

then one needs to fix either one of δ_d or the variance of the common component σ^2_θ to ensure scale identifiability.

Another identifiability issue comes from the symmetry in the model formulation. Note that, in the simplest model of two groups, cases and controls, we split the variation of two variables into three components: θ_j common to both controls and cases, λ_{j1} specific to controls and λ_{j2} specific to cases. This decomposition if fine provided that there is enough signal in the data to identify all three terms, but problems may arise otherwise. To overcome this potential issue, an alternative is the asymmetric formulation in which the two variables are decomposed into two components: θ_j, which picks up the variability in the controls shared with cases and λ_{j2}, the differential effect of cases over controls (Figure 1). We opted for the latter formulation.

2.2 Prior distributions

We assigned flat normal prior distributions for the group-specific intercepts α_d. For the shared and specific components we chose zero-mean t distributions with 4 degrees of freedom and unknown variance parameters σ^2_θ and σ^2_d $(d > 1)$, respectively. The reason for this choice is the heavy probability tails of this distribution compared to the normal distribution, which can accommodate sizable departures from zero in the odds ratios of both the shared and disease-specific allele frequencies. For the loadings $\delta_d,\ d > 1$, we chose flat log-normal distributions whereas we set $\delta_1 = 1$ to avoid the scale identifiability issues mentioned above. The specific prior choice for the model parameter was as follows:

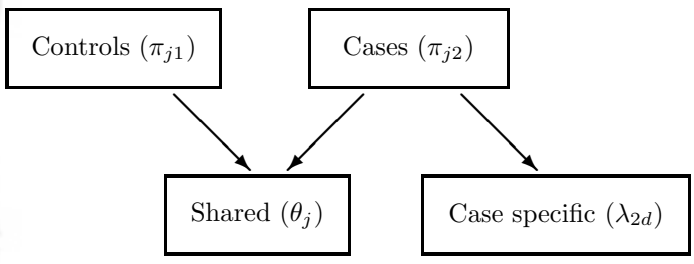

Figure 1: Schematic representation of the shared component model.

$$\begin{aligned} \alpha_d &\sim \text{Normal}(0, 1000) \\ \theta_j &\sim t_4(0, \sigma_\theta^2) \\ \lambda_{jd} &\sim t_4(0, \sigma_d^2) \quad (d > 1) \\ \log \delta_d &\sim \text{Normal}(0, 100) \quad (d > 1) \\ \delta_1 &= 1 \end{aligned} \tag{4}$$

To complete our model, we chose flat half-truncated normals as hyperpriors for the standard deviations of the random effects.

$$\sigma_\theta, \sigma_d \sim \text{Normal}(0, 100) \cdot \mathbb{1}_{(0,+\infty)} \tag{5}$$

Equations (2), (3), (4) and (5) represent the specific formulation of our shared component model for GA studies.

2.3 Model implementation

We built an R package called `BayesGen` with functions to fit the model in Equations (2) to (5). These functions use in turn Markov chain Monte Carlo (MCMC) simulation techniques as implemented in the free-software JAGS (`http://www-fis.iarc.fr/~martyn/software/jags/`), and more specifically in its R package `R2jags` (Su and Yajima, 2010). Since MCMC can take too long for large datasets with many SNPs, we also implemented a slightly different version of our model in R using approximate Bayesian inference instead of MCMC. Namely, we used the Integrated Nested Laplace Approximation (INLA) proposed by Rue et al. (2009) as implemented in the `INLA` package for R (Rue and Martino, 2009). The model based on INLA replaces the t distributions for Normals. This in principle could shrink sizeable SNP-disease risk associations more than the original model, but it runs much faster and it can be applied to genome-wide association studies.

3 Case study

To illustrate the use of our model we applied it to data from The Wellcome Trust Case Control Consortium (2007). Specifically, we considered 23 418 SNPs from chromosome 12 in 2 000 cases of diabetes type 1, 2 000 cases of diabetes type 2 and 3 000 controls. The original data therefore consisted in a matrix of variant allele frequencies $[X_{ijd}]$, $X_{ijd} \in \{0, 1, 2\}$, of dimension $7\,000 \times 23\,418$. Aggregating the counts of the variant alleles over the individuals in each disease group for each SNP gave rise to a new reduced matrix $[Y_{jd}]$ of dimension $23\,418 \times 3$.

We fitted the model in Equations (2) to (5) to the aggregated data $[Y_{jd}]$ using the JAGS based functions of BayesGen. Specifically, we ran two chains of 120 000 iterations. We discarded the first 30 000 as burn-in and kept every 90th to reduce autocorrelation in the chains. Inference is therefore based on (thinned) samples of size 2 000. We assessed convergence using visual checks and the Brooks-Gelman diagnostic. No symptoms of non-convergence were deected.

Table 1 shows the results obtained for the group-specific intercepts α_k and the factor loadings δ_k. The posterior distributions for δ_k seem to be quite concentrated and also located close to 1, which confirms that the two groups of diabetes have frequencies of the variant allele similar to the control group for the vast majority of the SNPs considered.

Looking into the specific component λ_{j1} for diabetes type 1, we found 60 'significant' SNPs (in the sense that their corresponding 95% credibility intervals exclude the value 0). Of those, 26 showed positive association with the disease and 34 exhibited negative association (i.e. the other allele is the one positively associated with disease). Figure 2 shows that the SNPs related to the disease are located in three different regions of the chromosome. For diabetes type 2, we found 3 'significant' SNPs all showing negative association, two in the same region, and the third one isolated in a different region (Figure 2). Table 2 shows a few of the SNPs associated to both types of diabetes (with at least one representative per region) and their estimated risks (ORs), 95% credibility intervals and exceedance posterior probabilities).

Group	Parameter	OR (95%CI)
Control	$\exp \alpha_1$	0.9561 (0.9225, 0.9815)
Diabetes 1	$\exp \alpha_2$	0.9559 (0.9222, 0.9811)
Diabetes 2	$\exp \alpha_3$	0.9559 (0.9224, 0.9814)
Control	δ_1	1
Diabetes 1	δ_2	1.0010 (1.0000, 1.0012)
Diabetes 2	δ_3	0.9900 (0.9983, 0.9995)

Table 1: Posterior median and 95% credibility intervals for some model parameters.

Disease	SNP	OR	95%CI	$P(OR > 1 \mid \text{data})$
Diab 1	rs11052423	1.14	(1.02, 1.26)	0.9965
Diab 1	rs705698	1.20	(1.08, 1.31)	> 0.9999
Diab 1	rs11171739*	1.29	(1.19, 1.39)	> 0.9999
Diab 1	rs17696736*	0.75	(0.69, 0.81)	< 0.0001
Diab 2	rs7132840	0.89	(0.80, 0.99)	0.018
Diab 2	rs1495377*	0.89	(0.80, 0.99)	0.019
Diab 2	rs10492267	0.14	(0.09, 0.21)	< 0.0001

Table 2: SNPs showing association with diabetes types 1 and 2. Posterior median of the odds ratio (OR), 95% credibility intervals, and posterior probability of excess risk $P(OR > 1 \mid \text{data})$. SNPs marked with an * were also reported as significant by The Wellcome Trust Case Control Study (2007).

Using 95% credibility intervals to declare a SNP as significantly associated to the disease implies setting the false discovery rate at 5%, which is too large considering the high number of SNPs in the case study. We therefore used a much higher coverage value for the posterior credibility intervals of the λ_{jd}, namely 99%. The limits of these intervals typically correspond to tail quantiles from the posterior distribution. It is difficult to estimate these quantiles from MCMC samples. Therefore, we computed these intervals using a normal approximation with parameters equal to the posterior mean and variance of the specific components. We also considered other another 'rule' to detect SNPs associated with the diseases based on the posterior probability that the OR is above 1 $P(OR > 1 \mid \text{data}) > p_c$, where p_c is a cutoff value that we set to 0.99 to keep the false discovery rate low.

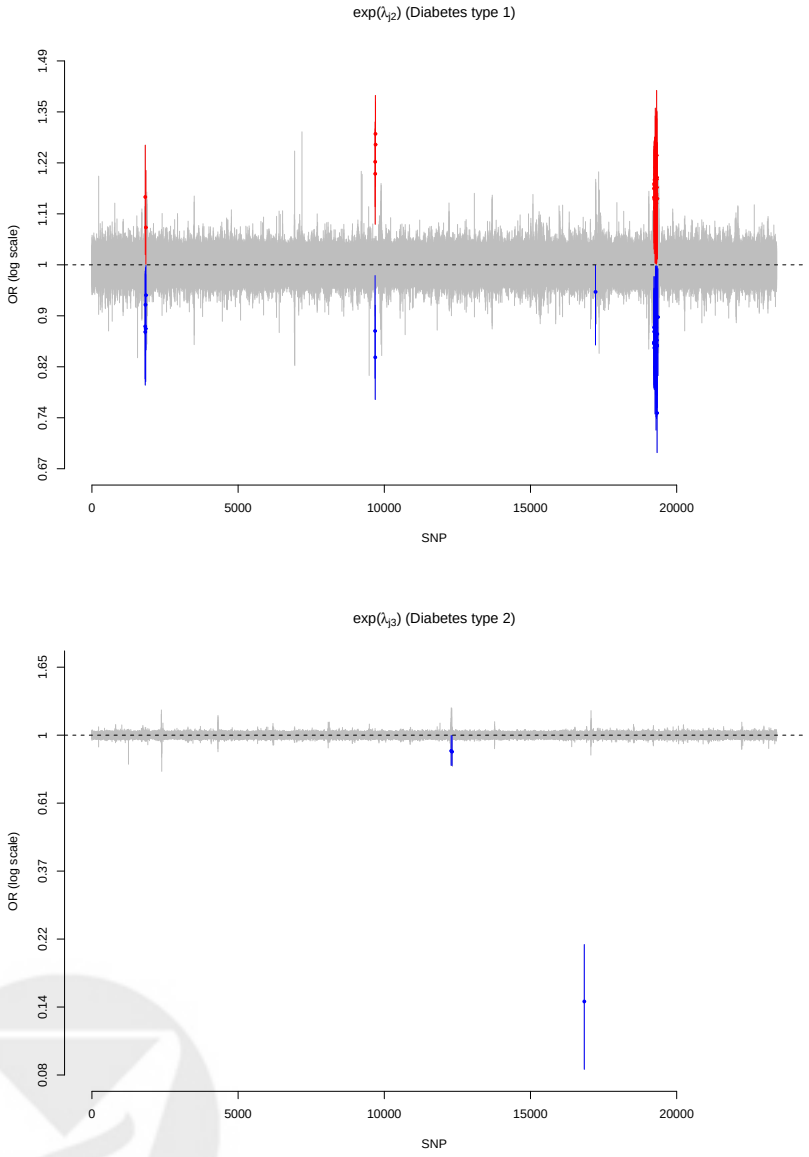

Figure 2: Posterior medians and 95% credibility intervals for the odds ratios of the specific components for diabetes types 1 (top) and 2 (bottom). Significant SNPs are coloured in red (blue) if the association is positive (negative).

4 Simulation study

To analyze the sensitivity and specificity of our model, we simulated data under four different scenarios varying the number of individuals, the number of significant SNPs and the values of the odds ratios (ORs).

Specifically, we simulated datasets of variant allele frequencies, $[X_{ijd}]$, $X_{ijd} \in \{0, 1, 2\}$, with $i = 1, \ldots, N$ individuals, $j = 1, \ldots, 1000$ SNPs and $d = 2$ groups, just case and controls. By changing N and the number of SNPs with significant association to the disease, we created the following four scenarios:

Case 1.1: Dataset with N=500 individuals in each disease group and 25 of the 1 000 SNPs are significant with the following OR distribution: four SNPs with very high risk, OR=3; nine with high risk, OR=2; and the remainder 12 with low risk, OR=1.2. The very high risk is not too realistic, but we include it rather as a benchmark.

Case 1.2: Dataset with N=500 individuals per disease group and just two of the 1 000 SNPs significant with ORs set to 3 and 1.2.

Case 2.1: Similar to case 1.1 but with N=100 individuals, i.e: 1 000 SNPs where 25 of them are significant with the same OR distribution mentioned above, four SNPs with OR=3; nine with OR=2; and twelve with OR=1.2.

Case 2.2: Similar to 1.2 but again with N=100 individuals.

The first scenario has a relatively high number of individuals and significant SNPs, ranging from low risk to very high risk; the second one has again a large number of individuals but a low number of significant SNPs; the third one has a rather moderate number of individuals a many significant SNPs and the last one has a moderate number of individuals and a low number of significant SNPs.

We simulated 100 datasets from each scenario and fitted the shared component model to all the replicates using the `BayesGen` R package. We then used the two different rules mentioned above to detect SNPs with significant association to the disease. Rule 1 is the one based on the posterior probability and Rule 2 uses the credibility intervals.

Figures 3 and 4 show ROC curves obtained for Rule 1 and Rule 2 applied to the four scenarios using JAGS and INLA, respectively. For a specificity of 90% the sensitivity varies between 70% and 85% depending on the scenario. In general we can see that the larger the number of individuals the higher the sensitivity, which in turn is similar for the two actual numbers of significants SNPS considered. We can also observe that both rules seem to perform similarly, though in most cases Rule 1 shows slightly higher sensitivity than Rule 2. The model that uses Normal distributions and is fitted with INLA (Figure 4) seems also to perform slightly better than the one using t_4 distributions and fitted with JAGS (Figure 3).

5 Discussion

We proposed a new approach to GA studies based on Bayesian shared component models. We illustrated its utility on a subset of data from The Wellcome Trust Case Control Consortium (2007) comprising approximately 23 000 SNPs from 7 000 individuals from three groups, two with diabetes of type 1 and 2, respectively, and one control. The method proved useful to tease out what is shared by controls and cases from what is specific to cases from each disease. The model is hierarchical so it should be straightforward to add further structure at gene or region level if needed.

We considered the additive model of association between SNPs and disease, but our shared component approach can easily be adapted to the dominant or recessive models where $X_{ijd} \in \{0, 1\}$ by simply assuming $X_{ijd} \sim \text{Bernoulli}(\pi_{jd})$, which upon aggregation over individuals leads to $Y_{jd} \sim \text{Binomial}(n_{jd}, \pi_{jd})$

The models show reasonable values of sensitivity for a specificity of 90%, though these are somewhat higher for the model that considers Normal distributions for the shared and specific components than for the model that uses t_4 distributions. However, in GA studies the sensitivity typically is set to much higher values to ensure a low rate of false positives, which decreases the sensitivity.

We considered t distributions with four degrees of freedom for the specific components because its heavy tails can esaily accommodate large departures from zero for the λ_{jd}. Other priors however can be considered instead to detect associations between SNPs and disease. A simple alternative is the double exponential distribution, so that $\lambda_{jd} \sim \text{DoubleExponential}(\mu, b)$. Another option, from gene expression studies, are the spike and slab priors that assume $\lambda_{jd} \sim p \cdot \text{Normal}(0, \sigma_1^2) + (1-p) \cdot \text{Normal}(0, \sigma_2^2)$ with σ_1 small, and σ_2 large.

Both models have been implemented in an R package called `BayesGen`. However, there is a massive difference in the computation time between fitting the model with the standard MCMC-based approach and the approximate Bayesian inference based on INLA. For our case study the former took three days whereas the former took just a few minutes on the same server.

In conclusion, we provided a novel application of Bayesian shared component models that adds to the methodological toolbox for GA studies. We suggested two different versions of the model and also implemented them using two different inferential approaches in R to make them available to the scientific community.

Acknowledgments

This study makes use of data generated by the Welcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from `http://www.wtccc.org.uk`. We thank Prof. Havard Rue for help with the implementation in R of a version of the shared component model proposed here that uses INLA. JA was was partly supported by Grants GVPRE/2008/010 and AP-055/09 from Generalitat Valenciana.

Conflict of Interest: None declared.

References

Aguilar, O. and West, M. (2000). Bayesian dynamic factor models and portfolio allocation. *Journal of Business and Economic Statistics*, 18:338–357.

Gonzalez, J. R., Armengol, L., Sole, X., Guino, E., Mercader, J. M., Estivill, X., and Moreno, V. (2007). SNPassoc: an R package to perform whole genome association studies. *Bioinformatics*, 23(5):644–645.

Hardy, G. H. (1908). Mendelian proportions in a mixed population. *Science*, 28:49–50.

Hogan, J. W. and Tchernis, R. (2004). Bayesian factor analysis for spatially correlated data, with application to summarizing area-level material deprivation from census data. *Journal of the American Statistical Association*, 99:314–324.

Hoggart, C. J., Whittaker, J. C., De Iorio, M., and Balding, D. J. (2008). Simultaneous analysis of all snps in genome-wide and re-sequencing association studies. *PLoS Genet*, 4(7):e1000130.

Knorr-Held, L. and Best, N. G. (2001). A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society - A*, 164:73–85.

Kooperberg, C., Ruczinski, I., LeBlanc, M. L., and Hsu, L. (2001). Sequence analysis using logic regression. *Genet Epidemiol*, 21 Suppl 1:S626–31.

Richardson, S., Abellan, J. J., and Best, N. (2006). Bayesian spatio-temporal analysis of joint patterns of male and female lung cancer risks in yorkshire (UK). *Statistical Methods in Medical Research*, 15:385–407.

Rue, H. and Martino, S. (2009). *INLA: Functions which allow to perform a full Bayesian analysis of structured additive models using Integrated Nested Laplace Approximation*. R package version 0.0.

Rue, H., Martino, S., and 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*, 71:319–392.

Sha, N., Vannucci, M., Tadesse, M. G., Brown, P. J., Dragoni, I., Davies, N., Roberts, T. C., Contestabile, A., Salmon, M., Buckley, C., and Falciani, F. (2004). Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. *Biometrics*, 60(3):812–819.

Su, Y.-S. and Yajima, M. (2010). *R2jags: A Package for Running jags from R*. R package version 0.02-02.

The Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–678.

Townsend, P., Simpson, D., and Tibbs, N. (1985). Cardiovascular risks factors and the neighbourhood environment. *International Journal of Health Services*, 15:637–663.

Weinberg, W. (1908). Über den Nachweis der Vererbung beim Menschen. *Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg*, 64:368–382.

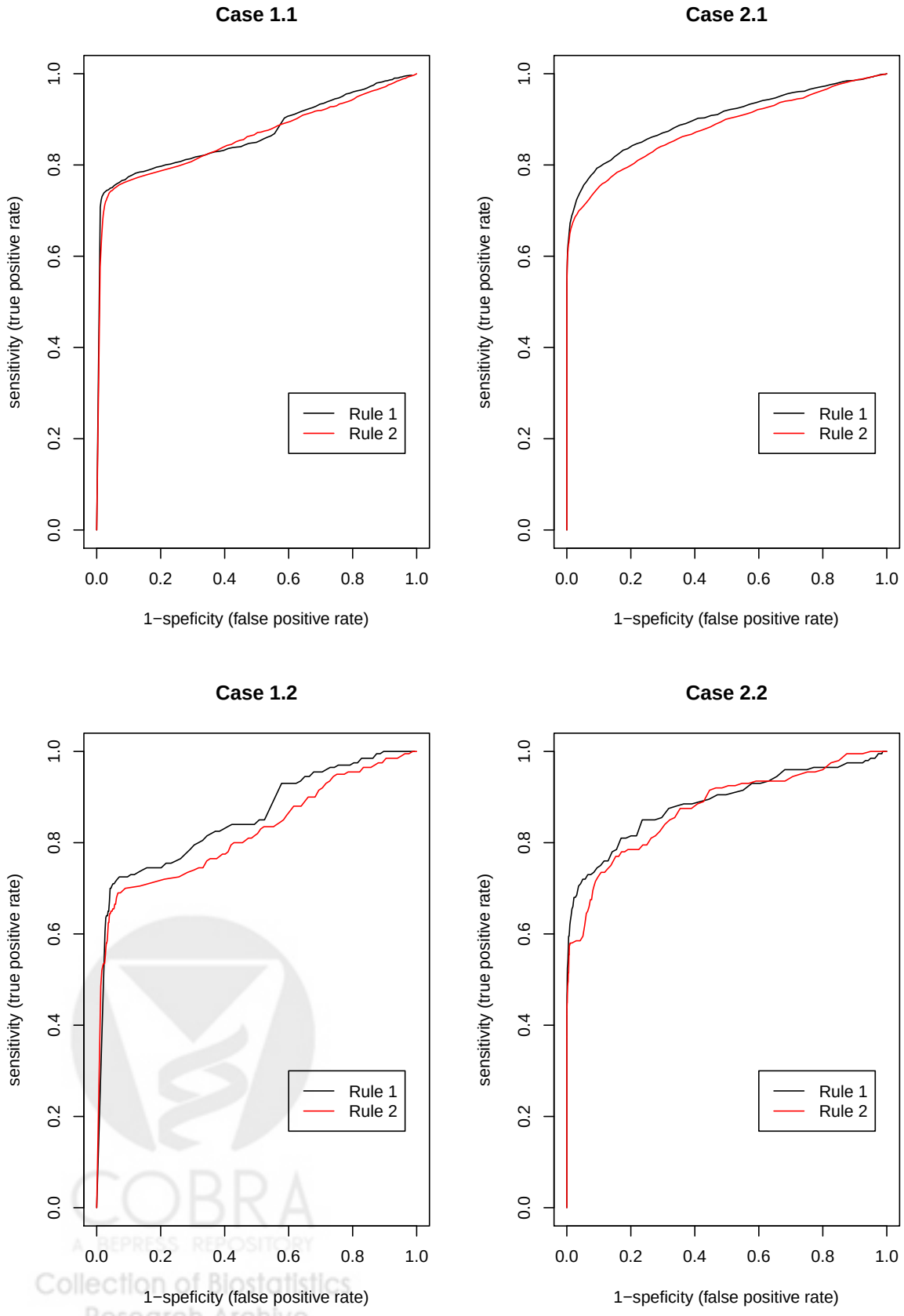

Figure 3: Using JAGS to fit share component model, this are the ROC curves associated with Rule 1 and Rule 2 for the four scenarios.

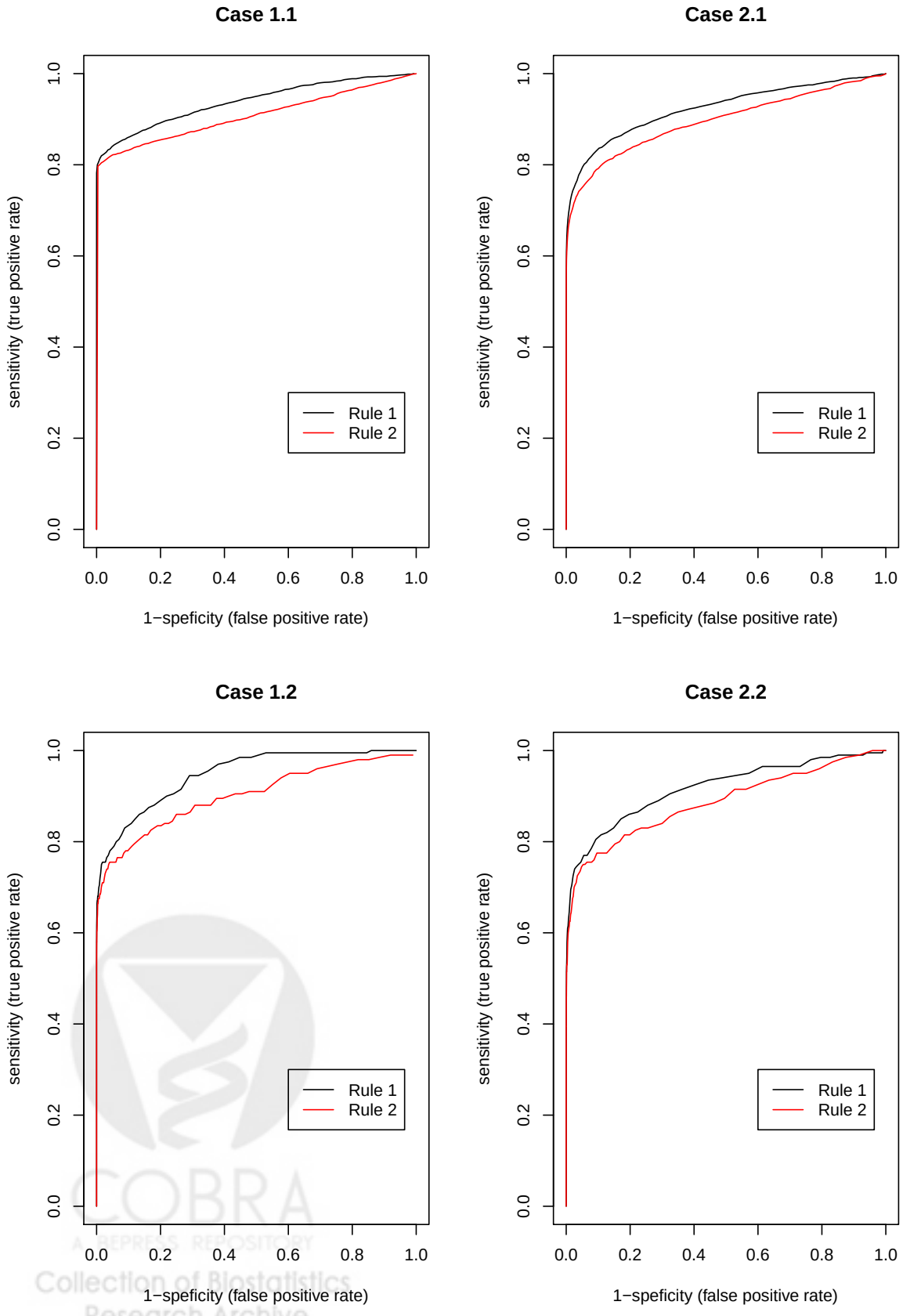

Figure 4: Using INLA to fit share component model, this are the ROC curves associated with Rule 1 and Rule 2 for the four scenarios.