Biostatistics (2013), **0**, 0, *pp.* 1–29
doi:10.1093/biostatistics/article˙10˙12˙2013˙biostatistics

Meta-analysis of incomplete microarray studies

ALIX LEBOUCQ*, ANTHONY C. DAVISON, DARLENE R. GOLDSTEIN

EPFL SB MATHAA STAT, station 8, 1015 Lausanne

alix.leboucq@epfl.ch

Summary

Meta-analysis of microarray studies to produce an overall gene list is relatively straightforward when complete data are available. When some studies lack information, for example, having only a ranked list of genes instead of complete primary data, it is common to reduce all studies to ranked lists prior to combining them. Since this entails a loss of information, we consider a hierarchical Bayes modeling approach to combining studies using the type of information available in each study: the full data matrix, summary statistics, or ranks for each gene. The model uses an informative prior for the parameter of interest, which eases the detection of differentially expressed genes. Simulations show that the new approach can give substantial power gains compared to classical meta analysis and list aggregation. A large meta-analysis based on 11 published studies providing data of the types cited above is also performed and shows credible results by identifying genes known to be involved in ovarian cancer.

Key words: Bayesian hierarchical model; Gibbs sampler; Horseshoe prior; Microarray; Normal-gamma prior; Serous ovarian cancer; Spike and slab prior.

*To whom correspondence should be addressed.

1. Introduction

More and more studies use microarray data with mRNA measurement for thousands of genes simultaneously. They are used to detect genes differentially expressed between several groups of tissues. However, such data are highly variable and the results obtained are often non-reproducible or not robust to even mild perturbations (Ramasamy *and others*, 2008). This large variability in results may be due to improper analysis or insufficient control of false discoveries, increased by the small size of the samples compared to the number of genes. It is therefore of great interest to combine the results of different microarray studies that address the same question. Since the first meta-analysis of microarray data (Rhodes *and others*, 2002) was performed, many different methods have been developed to combine different studies. For example, one can combine p-values (Rhodes *and others*, 2002, 2004), or effect size changes (Choi *and others*, 2003), using a frequentist approach, use Bayesian hierarchical modeling as in Conlon *and others* (2007), or other methods, as described in Guerra and Goldstein (2009).

Raw data contain the fullest information one can obtain from a study and give control over preprocessing steps and analyses. Providing raw data is required by the MIAME (Brazma *and others*, 2001), guidelines for publication of microarray data, and also by numerous journals. However, publication of raw data is far from systematic; Larsson and Sandberg (2006) note that, on GEO (Edgar *and others*, 2002), raw data are available for only 34% of the samples, many of which did not meet the necessary quality standards. Jacob *and others* (2009) also point out that authors frequently publish only lists of significant genes, which are of limited utility for combining information across studies.

When raw data are not available and only ordered lists of genes, or ranks, are provided, one can use list aggregation methods (Lin, 2010). Borda (1781) method is one of the simplest, and combines the ranks of each gene in each study to obtain the final list of interesting genes. With $R_g = (R_g^{(1)}, \ldots, R_g^{(L)})$ denoting the ranks of gene g in all L studies, Borda's score for gene g is

$f(R_g)$, where the function f can be the arithmetic mean, which we will use in this paper, the median, or the geometric mean, among other possibilities. Borda scores are ordered to obtain the final list of interesting genes, where a small score implies high differential expression over all studies.

Even if combining similar types of information is relatively straightforward (Sutton *and others*, 2000; Guerra and Goldstein, 2009), each method has advantages and drawbacks. Combining decisions, i.e., whether a gene is differentially expressed or not, might lead to a dramatically reduced list of genes, as only a small overlap is likely to appear. This problem can also be encountered in combination of ranks, though ranks have the advantage of being robust. Combination of p-values loses information on the magnitude and direction of the differentials, i.e. whether the gene is over- or under-expressed. This is also the case for rank and vote aggregation. Combining effect sizes loses least information, but requires access to the raw data, or, less preferably, to the gene expression matrix, which, as mentioned previously, may be unavailable.

If the data to be combined are not of the same type, studies are either discarded, or data are transformed to the least informative support, in order to meet the requirements for combination (DeConde *and others*, 2006; Boulesteix and Slawski, 2009). Both can lead to a large loss of information and of power.

The purpose of this paper is to generalise the usual meta-analysis to combine results from studies providing different levels of information: full data; summary statistics such as z-scores, either full or partial; and incomplete lists of ranks. In Section 2 we describe a hierarchical Bayesian model developed for this purpose, under which each type of data is modeled using a single underlying parameter that measures the effect of each gene. A prior on this parameter is then carefully chosen to ease the detection of differentially expressed genes. Model efficiency is assessed using simulated microarray data in Section 3. Finally, a real data illustration is presented in Section 4, where our model is applied on serous ovarian cancer, in order to detect differentially

expressed genes between cancer and normal samples.

2. Bayesian hierarchical model

2.1 *Data types*

We aim to detect genes differentially expressed between two groups of samples, say, for concreteness cancer and normal. We assume that each gene g included in the analysis has a parameter γ_g representing its differential expression, but that this is not directly observed and instead, noisy realizations, $\beta_g^{(l)}$ are observed in study l, where $\beta_g^{(l)}$ follows a normal distribution centered at γ_g,

$$\beta_g^{(l)} \sim \mathcal{N}(\gamma_g, \sigma_\beta^2), \quad l = 1, \ldots, L. \tag{2.1}$$

Possible priors for the parameter γ_g will be discussed in Section 2.3, but first, we describe the four types of data that we consider:

1. *Raw data*: the most complete data one can obtain from a study, usually consisting of a matrix of gene expression values along with some clinical information about the study patients. If values are missing for technical reasons, we consider them to be missing completely at random, as the missing values depend neither on the observed nor the missing data.

2. *Full lists of z-scores or other statistics*: A list recording a value of a statistic for each gene. The statistic can be any result of a statistical analysis based on a gene expression matrix, but is most often a t-statistic or a p-value. Without loss of generality, we assume that the values are standardized and can be therefore considered to be z-scores.

3. *Partial z-scores*: This third type of data gives the value of some statistic for each of the top k genes. Here missingness is not at random, as only the most significant genes are observed. As for Type 2, we assume that the values can be transformed to z-scores.

4. *Partial list of ranks*: A list, often incomplete, of the "best" genes detected by a statistical

analysis based on a gene expression matrix. This list usually gives the top k most significant genes, with k often between 20 and 100.

We develop a hierarchical model for each of the previously defined data types. We consider one study of each type to simplify the representation of the model.

The first data type consists of a $p \times N$ data matrix Y, containing the gene intensities for the p genes, for n_1 cancer samples and $N - n_1$ normal control samples. We suppose that the first n_1 columns of the matrix contain the cancer samples. The intensities for each gene are denoted by Y_{gj} and are modeled using a parameter μ_g representing the baseline mean and a parameter $\beta_g^{(1)}$ which represents the differential expression of a gene g in the Type 1 study,

$$Y_{gj} \sim \mathcal{N}(\mu_g + \beta_g^{(1)} I_{\{j \leqslant n_1\}}, \sigma_g^2), \quad \sigma_g^{-2} \sim \text{Gamma}\,(b_1, b_2)\,, \quad j = 1, \ldots, N.$$

The second type of study provides a vector of z-scores, $Z = (Z_1, \ldots, Z_p)$, which result from performing gene-by-gene two-sample t-tests, or other similar tests, on a full data matrix, in order to compare the gene expression of cancer and control samples. In what follows the calculations are the same for each gene, so we omit the index g. The setting is the same as for the full data. Let, $\overline{Y}_1$, S_1^2 and $\overline{Y}_2$, S_2^2 denote the sample mean and sample variance of the cancer and normal groups respectively. The two-sample t-statistic is given by

$$T_{\text{obs}} = \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{N-n_1}\right) S_p^2}}, \quad S_p^2 = \frac{(n_1 - 1)S_1^2 + (N - n_1 - 1)S_2^2}{N - 2}.$$

Since $X = (N-2)S_p^2\delta^2 \sim \chi^2_{N-2}$, with $\delta^2 = \sigma^{-2}$ and $\nu = (N - n_1)n_1/N$, we first notice that, if N is large enough, $X/(N-2) \approx 1$, and then we obtain

$$T_{\text{obs}} \sim \beta^{(2)}\sqrt{\nu\delta^2} + t_{N-2} \,\dot{\sim}\, \mathcal{N}\left(\beta^{(2)}\sqrt{\nu\delta^2}, 1\right),$$

where $\beta^{(2)}$ represents differential expression of a gene in a Type 2 study. We can thus model the z-statistics for the Type 2 data as

$$Z_g \mid \beta_g^{(2)}, \delta_g^2 \sim \mathcal{N}\left(\beta_g^{(2)}\sqrt{\delta_g^2\nu}, 1\right), \quad \delta_g^2 \sim \text{Gamma}(f_1, f_2).$$

We now consider partial lists of z-scores, for which only the k largest absolute values are observed. For observed genes, we use the same model as for the Type 2 data, and for missing ones we impute them, borrowing information from the other studies. The z-scores decompose into $Z = (Z^{\mathrm{o}}, Z^{\mathrm{m}})$, for the observed genes and for the missing ones. We know that the observed Z^{o} are the largest k in absolute value of the entire vector. Let $|z_{\mathrm{ref}}|$ be the smallest observed score. Then, for any missing gene g belonging to the set $\{Z^{\mathrm{m}}\}$, $-|z_{\mathrm{ref}}| < Z_g < |z_{\mathrm{ref}}|$, and can be imputed by

$$Z_g \mid \beta_g^{(3)}, \delta_g \sim \mathcal{N}_{-|z_{\mathrm{ref}}|}^{|z_{\mathrm{ref}}|}\left(\beta_g^{(3)}\sqrt{\delta_g^2\nu}, 1\right), \quad \delta_g^2 \sim \mathrm{Gamma}(f_1, f_2),$$

where $\mathcal{N}_a^b$ is the truncated normal distribution on the interval $[a, b]$, and $\beta_g^{(3)}$ is the differential expression parameter for a gene g in the Type 3 study.

The last type of study consists of a partial list of ranks. For each gene g, a rank R_g is attributed by sorting some statistics, such as z-scores, which are unobserved. We introduce a latent variable u_g to model the summary statistic that leads to R, i.e., $R_g = \mathrm{rank}(|u_g|)$. Using the parameter $\beta_g^{(4)}$ to represent the differential expression of gene g in the Type 4 study, we have

$$u_g \sim \mathcal{N}\left(\beta_g^{(4)}\sqrt{\nu}, \sigma_{u,g}^2\right),$$

with $\nu = (N - n_1)n_1/N$. The variance $\sigma_{u,g}^2$ is gene-dependent, taking into account uncertainty about the methods used to obtain the ranks. Since the ranks are incompletely observed, the missing ranks are imputed by borrowing information from the other studies, as for the Type 3 data. We write $R = (R^{\mathrm{o}}, R^{\mathrm{m}})$, with corresponding $u = (u^{\mathrm{o}}, u^{\mathrm{m}})$, where the subscripts o and m denotes the sets of observed and missing genes. To impute the missing ranks, we first need to impute the missing $u_g \in \{u^{\mathrm{m}}\}$. Let R_{ref} be the smallest rank observed and let $|u_{\mathrm{ref}}|$ denote the corresponding latent variable. Any missing gene g would be ranked higher than R_{ref}, if it were observed, or said differently, $|u_g| \leqslant |u_{\mathrm{ref}}|$. Thus u_g is imputed according to

$$u_g \sim \mathcal{N}_{-|u_{\mathrm{ref}}|}^{|u_{\mathrm{ref}}|}\left(\beta_g^{(4)}, \sigma_{u,g}^2\right).$$

The missing ranks are then imputed by ordering the vector $|u|$.

2.2 *Combining all data types*

Now that we have modeled each type of data separately using hierarchical models, we can combine them using relation (2.1). As a summary, the full model combining all types of information is, for $g = 1, \ldots, p$ and $j = 1, \ldots, N$,

$$\begin{aligned} Y_{gj} \mid \mu_g, \beta_g^{(1)}, \sigma_g^2 &\sim \mathcal{N}(\mu_g + \beta_g^{(1)} I_{j \leqslant n_1}, \sigma_g^2), \quad \sigma_g^{-2} \mid b_1, b_2 \sim \text{Gamma}(b_1, b_2), \\ Z_g^{(i)} \mid \beta_g^{(i)}, \delta_g^2 &\sim \mathcal{N}\left(\beta_g^{(i)} \sqrt{\nu \delta_g^2}, 1\right), \; i = 2, 3, \quad \delta_g^2 \sim \text{Gamma}(f_1, f_2), \\ u_g \mid \beta_g^{(4)}, \sigma_u^2 &\sim \mathcal{N}\left(\beta_g^{(4)} \sqrt{\nu}, \sigma_{u,g}^2\right), \quad \sigma_{u,g}^{-2} \mid d_1, d_2 \sim \text{Gamma}(d_1, d_2), \\ \beta_g^{(i)} \mid \gamma_g, \sigma_\beta^2 &\sim \mathcal{N}(\gamma_g, \sigma_\beta^2), \quad i = 1, \ldots, 4, \quad \sigma_\beta^{-2} \mid e_1, e_2 \sim \text{Gamma}(e_1, e_2). \end{aligned} \tag{2.2}$$

Representation of the model as a directed acyclic graph and calculations of the posterior distributions for all parameters of (2.2) are available in the Supplementary Materials.

The parameter of interest γ_g indicates differential expression of the corresponding gene. A value close to zero indicates no differential expression, while a value far from zero indicates that the gene is differentially expressed between the two groups of samples, among all studies. As this parameter is decisive, we need to choose its prior carefully.

2.3 *Priors for the parameter of interest, γ*

We know that most of the genes included in the analysis are not significant, and therefore a large proportion of the γ's should be null. On the other hand, some of the genes are differentially expressed, and in order to be able to detect them easily, the corresponding γ should have a large value. Three different priors for the parameter γ are considered, all based on the idea of shrinking the parameters of the uninteresting genes towards zero, while leaving the parameters of the differentially expressed genes large enough to be easily detected. The first is the spike and slab prior (Mitchell and Beauchamp, 1988; George and McCulloch, 1993), which has been used in the

contexts of the detection of differentially expressed genes in microarrays by Ishwaran and Rao (2003, 2005) and of multiple hypothesis testing by Pang and Gill (2009). The second is the horseshoe prior, introduced by Carvalho *and others* (2010), and used in several contexts (Carvalho *and others*, 2009; Polson and Scott, 2010; Datta and Ghosh, 2012). Finally the normal-gamma prior (Griffin and Brown, 2010) will be introduced. Computations of the posterior densities for each of the models along with graphical representation of the priors can be found in the Supplementary Materials.

The spike and slab prior (Mitchell and Beauchamp, 1988) uses a mixture of two normal distributions centered at zero, one with very small variance for the non differentially expressed genes, and the other with possibly large variance for differentially expressed genes,

$$\gamma_g \mid c_g, \tau_g^2 \sim \mathcal{N}\left(0, c_g\tau_g^2\right), \quad \tau_g^2 \mid a_1, a_2 \sim \text{Gamma}(a_1, a_2),$$
$$c_g \mid \alpha \sim (1-\alpha)\delta_{c^\star} + \alpha\delta_1, \quad \alpha \sim \mathcal{U}(0,1),$$

where δ_a is the Dirac function putting unit mass at a. The parameter c is binary and can either take the value of the hyperparameter $c^\star = 0.005$, very close to zero, which describes the spike of the distribution and shrinks most of the γ's to zero; or 1, leading to possibly large values of τ.

The horseshoe prior (Carvalho *and others*, 2010) uses half-Cauchy distributions, C^+, as follows:

$$\gamma_g \mid \lambda_g \sim \mathcal{N}(0, \lambda_g^2\tau^2), \quad \lambda_g \sim C^+(0,1), \quad \tau \sim C^+(0,1).$$

The name of the prior is due to the shape of the distribution of the parameter $\kappa_g = (1+\lambda_g^2)^{-1}$, which follows a Beta$(1/2, 1/2)$ distribution when $\lambda \sim C^+(0,1)$. The global shrinkage parameter τ gives a measure of the underlying sparsity of the data: a small τ indicates large shrinkage. The parameter λ_g is gene-dependent, yielding a highly adaptive prior. The posterior distribution of γ is obtained by conjugacy of the prior. However, sampling from the posterior distributions of λ_g and τ is not straightforward and we used the algorithm described in Scott (2011) (see Supplementary Materials for more details).

Finally, the normal-gamma prior, introduced by Griffin and Brown (2010) is defined as follows:

$$\gamma_g \mid \psi_g \sim \mathcal{N}(0, \psi_g), \quad \psi_g \mid \lambda, \tau \sim \text{Gamma}\left(\lambda, \frac{1}{2\tau^2}\right), \quad \lambda \sim \mathcal{E}(1),$$
$$\tau^{-2} \mid \lambda \sim \text{Gamma}\left(2, \frac{M}{2\lambda}\right), \quad M = \frac{1}{p}\sum_{g=1}^{p} \widehat{\gamma}_g^2, \quad \widehat{\gamma}_g = \frac{1}{L}\sum_{i=1}^{L} \beta_g^{(i)}.$$

Large shrinkage is obtained from small values of the parameter λ, whose prior is motivated by the fact that $\lambda = 1$ corresponds to the Bayesian Lasso, and therefore, the exponential prior gives more variability and flexibility around this value. The prior for τ^2 is chosen by noticing that the variance of ψ_g is $v_\psi = 2\lambda\tau^2$, on which we choose an inverse gamma prior. We obtain posterior densities by conjugacy of the priors for all parameters, except λ, for which a Metropolis–Hastings step is necessary.

2.4 *Practical consideration*

The model presented in the previous sections is fully Bayesian. However, some adjustments are needed for practical or computational reasons. First, we estimate the parameter μ_g representing the baseline mean in the Type 1 data, directly from the data for each gene. Indeed, Type 1 data are usually quite precise and they bring a lot of information, so we assume that this parameter can be quite well estimated by

$$\widehat{\mu}_g = \frac{1}{N - n_1} \sum_{j=n_1+1}^{N} Y_{gj}.$$

This empirical step also reduces the computational cost. The second parameter that is empirically estimated is δ_g^2, which appears in Type 2 and 3 data. Taking $\delta_g^2 \sim \text{Gamma}(f_1, f_2)$ leads to identifiability issues and considerably increases the computational cost, as it requires a Metropolis–Hastings step to sample from the posterior distribution. We therefore estimate the parameter δ_g^2 from the L_1 full datasets ($L_1 \geqslant 2$), as the inverse of the average gene variance over all Type 1 studies:

$$\widehat{\delta}_g^2 = \left(\frac{1}{L_1}\sum_{l=1}^{L_1} \widehat{\sigma}_g^{2,(l)}\right)^{-1}.$$

We also implemented a version of the model where a single δ^2 was sampled for all genes through a Metropolis–Hastings step, but it appeared to be less efficient than the empirical version, while also increasing the computational cost; thus we prefer to estimate δ_g^2 from the data.

The model was coded in R using C code for the Gibbs sampler part, to decrease computational time. We performed 31 500 iterations, discarding the 1500 first iterations and using thinning of 10 to reduce autocorrelation. The part concerning Type 4 data (ranks) is run 40 times more at each step to reduce the correlation between the draws. Indeed, for this particular type of data, each gene depends on its neighbours, making consecutive draws highly correlated. We obtain 3 000 independent realizations from the posterior distribution of the parameters. Values of the hyperparameters for the spike and slab prior (Section 2.3) are taken to be $a_1 = 5$ and $a_2 = 50$ as suggested by Ishwaran and Rao (2003). The values of the other priors are $b_1 = d_1 = e_1 = 5$ and $b_2 = d_2 = e_2 = 50$. Different values of the hyperparameters were tried but they did not affect the results. Computational time for $p = 200$ and $N = 50$, was of the order of 900 seconds for the spike and slab and the horseshoe priors, and 1 200 seconds for the normal-gamma prior. This last prior requires a larger computational time as one of its parameter is updated using a Metropolis–Hastings step.

2.5 *Discriminating between differentially expressed and non-differentially expressed genes*

Once values have been sampled from the posterior densities, we need a cutoff or a posterior quantity to decide whether a gene is differentially expressed or not. This problem was discussed in Carvalho *and others* (2010) and Scott (2009) for the horseshoe prior in the context of linear regression. These authors show that a simple thresholding rule, based on weights, leads to strong control of the false positives by automatically penalizing for multiple hypothesis testing (Scott, 2009), while maintaining high power. In the case of the horseshoe prior, the weight used, while not formally being the posterior probability of inclusion, can be used and interpreted informally

as such. Carvalho *and others* (2010) proposed to call g a signal if the corresponding weight is larger than 0.5, where the weights are identified as the quantity multiplying the response in the expression for the posterior mean of the parameter of interest. Datta and Ghosh (2012) show that this simple thresholding rule attains the risk of the Bayes oracle, introduced by Bogdan *and others* (2011), and Scott (2009) and Datta and Ghosh (2012) illustrate, by simulations, that this rule controls the number of false positives. In our context, we can use these results to construct a thresholding rule for our priors. Our parameter of interest is γ_g for all the priors, whose posterior mean is

$$\widehat{\gamma}_g = \mathbb{E}(\gamma_g \mid \lambda_g^2, \tau^2, \sigma_\beta^2, \beta_g) = \frac{\lambda_g^2 \tau^2}{\sigma_\beta^2 + L_g \lambda_g^2 \tau^2} \sum_{l=1}^{L_g} \beta_g^{(l)},$$

where L_g is the number of studies that contain gene g, rearrangement and we may write

$$\widehat{\gamma}_g = \left(\frac{L_g \lambda_g^2 \tau^2}{\sigma_\beta^2 + L_g \lambda_g^2 \tau^2} \right) \frac{1}{L_g} \sum_{l=1}^{L_g} \beta_g^{(l)}.$$

The quantity $w_{g,HS} = L_g \lambda_g^2 \tau^2 / (\sigma_\beta^2 + L_g \lambda_g^2 \tau^2)$ is a signal to noise variance ratio, $\lambda_g^2 \tau^2$ being the prior variance of γ_g and σ_β^2 / L_g being the variance of the average signal. The thresholding rule, inspired by Carvalho *and others* (2010) consists in calling gene g differentially expressed if $w_{g,HS} > 0.5$ and noise otherwise.

This thresholding rule can be applied in the same way to the other priors, using the appropriate form of the posterior weights, $w_{g,NG} = L_g \psi_g / (L_g \psi_g + \sigma_\beta^2)$, for the normal-gamma prior, and $w_{g,SAS} = L_g c_g \tau_g^2 / (L_g c_g \tau_g^2 + \sigma_\beta^2)$, for the spike and slab prior.

3. Numerical examples

3.1 *Simulations for microarray data*

We applied Model (2.2) with the different priors presented in Section 2.3 on simulated microarray data. We generated several studies according to the simulation design presented in Section 3.2. We consider five studies, two of Type 1 and one each of the other types. Each study consists of a

set of $p = 200$ genes, of which the first $g = 10$ are set to be differentially expressed, and $N = 50$ patients, of which $n_1 = 40$ are cancer samples. The model using the different priors of Section 2.3 is fitted to 500 simulated sets of such studies, using different levels of differential expression. The performance of the three priors is compared to Borda's method in Section 3.3. We also compare the performance of several ways to include studies in Section 3.4.

3.2 *Simulation design for microarray data*

The simulation design we used is inspired by DeConde *and others* (2006), who took the idea from Kooperberg *and others* (2005). We consider p genes, of which g are differentially expressed, with N samples belonging to $M = 2$ groups (cancer or normal) and L such studies. Let x_{ijml} denote the expression level of the ith gene for the jth sample belonging to the mth group (either cancer or control) and for study l, with $i = 1, \ldots, p$, $j = 1, \ldots, N$, $m = 1, 2$ and $l = 1, \ldots, L$. The expression level x_{ijml} is generated by decomposition into a mean expression μ_i, the differential expression parameter δ_{im} and noise Z_{ijml}, as

$$
\begin{aligned}
&x_{ijml} = \mu_i + \delta_{im} + Z_{ijml}, \quad \mu_i \sim \mathcal{U}(0,1),\\
&\delta_{im} = \begin{cases} a(2B_i - 1)G_i, & m = 1,\ i = 1, \ldots, g,\\ 0, & \text{otherwise.} \end{cases}\\
&B_i \sim Be(p), \quad G_i \sim \Gamma(5,1), \quad Z_{jl} \sim \mathcal{N}(0, \Sigma_l),
\end{aligned}
$$

where a is a differential expression parameter, usually chosen equal to 0.5, and Σ_l is a block matrix of size $p \times p$, having B blocks (in these simulations, we take $B = 1$ and $a \in [0.2; 2]$), diagonal elements σ_{il}^2 and off-diagonal elements $\sigma_i \sigma_j \rho_b$, where $\rho_b \sim \mathcal{U}(0.5, 1)$. The variance σ_{il}^2 is generated as $\sigma_{il}^2 = (0.3 - 0.02\mu_i)G_{il}$, where $G_{il} \sim \Gamma(5, 0.1)$. The variance parameter of a gene thus depends on the mean in a way that genes with smaller mean expression have a larger variance, and variance varies across studies.

The data matrix obtained for study l is

$$X = \left(\begin{array}{ccc|ccc} \mu_1 + \delta_{1,1} + Z_{1,1,1,l} & \cdots & \mu_1 + \delta_{1,1} + Z_{1,n_1,1,l} & \mu_1 + Z_{1,n_1+1,2,l} & \cdots & \mu_1 + Z_{1,N,2,l} \\ \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ \mu_g + \delta_{g,1} + Z_{g,1,1,l} & \cdots & \mu_g + \delta_{g,1} + Z_{g,n_1,1,l} & \mu_g + Z_{g,n_1+1,2,l} & \cdots & \mu_g + Z_{g,N,2,l} \\ \hline \mu_{g+1} + Z_{g+1,1,1,l} & \cdots & \mu_{g+1} + Z_{g+1,n_1,1,l} & \mu_{g+1} + Z_{g+1,n_1+1,2,l} & \cdots & \mu_{g+1} + Z_{g+1,N,2,l} \\ \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ \mu_p + Z_{p,1,1,l} & \cdots & \mu_p + Z_{p,n_1,1,l} & \mu_p + Z_{p,n_1+1,2,l} & \cdots & \mu_p + Z_{p,N,2,l} \end{array}\right).$$

This simulation design mimics the behaviour of real microarray data and is particularly realistic, unlike other simulation designs, which consist in generating independent normal distributions. This simulation design is completely independent of our model, which is not favoured in the simulations, so the comparison with Borda's method, presented in Section 3.3, should be fair. In the next sections, the simulation design can be used as such to obtain a Type 1 study. For Type 2 and 3 studies, a matrix X is generated and z-scores are obtained by applying limma (Smyth, 2005) from the bioconductor package in R. For Type 4 data, after generating another matrix X and applying limma, the absolute values of the z-scores are ordered to obtain the ranks.

3.3 *Comparison of the different methods and models*

We want to assess the performance of the different models in terms of detection of differentially expressed genes, for different degrees of differential expression. We also compare the performance of the models with Borda's method, which reduces all studies to ranks. We use Borda's method with the arithmetic mean as a comparison method because it is very simple, intuitive, not computationally intensive and has no tuning parameter. The median was also applied to combine the scores but it did not change the results of the simulations. There are many other methods for rank aggregation, but Lin (2010) shows that Borda's method with the arithmetic mean is best overall compared to other more complicated methods such as cross entropy Monte Carlo or Markov Chain methods. We also considered paired comparisons, using the Bradley and Terry (1952) model, but as its performance was worse than Borda's method, it was not compared with

our model.

For each value of the parameter a, we generated 100 datasets based on the simulation design presented in Section 3.2. For each dataset, we fitted the model using each of the three priors, and we also applied Borda's method. We recorded the values of the parameters $\widehat{\gamma}_g$ and $\widehat{w}_g$ for each prior and each gene g, along with the Borda scores. Figure 1 compares the priors and Borda's method in terms of the detection of differentially expressed genes; the absolute values of the posterior means of γ are ranked for each of the three priors and the number of true differentially expressed genes present in the top 10 is recorded; the same thing is done for Borda's method but using Borda scores. Our model detects the differentially expressed genes more easily than Borda's method, no matter which prior is chosen, and for all values of the differential expression parameter a.

It is also interesting to see how the priors perform in terms of false positives. For this analysis, we discard Borda's method, as it was identified as the worse method in Figure 1. Moreover, the only way to obtain a definition of a false positive for Borda's method is to use permutation p-values, which can be computationally intensive in high dimensional settings. In Figure 2, we present the receiver operating characteristic (ROC) curves, which plot the false positive rate against the true positive rate, for the three priors, for different values of a and for 500 simulations. Our models perform well in terms of ROC for values of $a \geqslant 0.3$, which happens to be the case in real genomics microarray data, where a realistic value for a is around 0.5. The different priors detect differentially expressed genes equally well in Figure 2, especially for small values of a, whereas for larger values, they all perform almost perfectly (plots not shown). The spike and slab prior seems to be best in each panel of Figure 2, closely followed by the normal-gamma prior, particularly when $a = 0.5$. The horseshoe prior performs less well, but it remains competitive and has high power. Since the normal-gamma prior requires a Metropolis–Hastings step to sample one of its parameters, we may prefer the spike and slab prior, which performs at least as well,

and is much faster to run.

3.4 *What do we gain from including all studies?*

In order to highlight the gain of power from the inclusion of several types of data rather than only considering full raw datasets, we performed 500 simulations, generating two Type 1 datasets, and one dataset of each of the other types, and included the different studies as follows:

1. only the two Type 1 studies (full raw data);
2. the two Type 1 studies and the full list of z-scores;
3. the two Type 1 studies and the partial list of ranks;
4. the two Type 1 studies and the partial list of z-scores;
5. the two Type 1 studies, the full list of z-scores and the partial list of ranks; and
6. all the studies.

We fitted the model using the spike and slab prior to these data and plotted the ROC curves for several values of the differential expression parameter a in Figure 3. Using all the information available results in a clear increase of power compared to the case where only full data are combined.

More formally, we tested whether the differences between the black (full studies only) and the pink (all data included) ROC curves are significantly different, using a non-parametric test introduced by DeLong *and others* (1988). We used the `roc.test` from the *pROC* R package (Robin *and others*, 2011), and found that all tests are highly rejected, as presented in Supplementary materials for different values of a. This result indicates a significant improvement of the power of the model when all studies are included.

4. Real data

We now integrate results from studies of the four types defined in Section 2.1, in order to find differentially expressed genes between serous ovarian cancer and normal samples. Electronic search of the online databases (Pubmed, GEO, ArrayExpress) identifies 11 studies (Mok *and others*, 2009; Yoshihara *and others*, 2009; Lili *and others*, 2013; Cancer Genome Atlas Research Network, 2011; Welsh *and others*, 2001; Warrenfeltz *and others*, 2004; Meinhold-Heerlein *and others*, 2007; Zhang *and others*, 2005; Bignotti *and others*, 2006; Martoglio *and others*, 2000; Donninger *and others*, 2004) comparing patients with serous ovarian cancer and healthy patients, and being of one of the four types that we consider. The study conducted by Lili *and others* (2013) included also stroma samples that we did not consider in this analysis. In what follows, the studies are denoted by the first three letters of the name of the first author, except for TCGA, which comes from the Cancer Genome Atlas team, and was downloaded via the `curatedOvarianData` R package (Ganzfried *and others*, 2013). Of these datasets, MOK, YOS, LIL and TCGA, have full information (Type 1); WEL provides a full list of z-scores (Type 2); MAR and DON provide only a partial list of z-scores (Type 3); and WAR, MEI, ZHA and BIG provide a partial list of ranks (Type 4), see Table 1. Studies were conducted on different platforms, and so include different probes and therefore different genes. Moreover, as some of them only provide partial lists, only a fraction of these genes are visible. The distribution of the genes among the studies, given in Table 2, shows that no gene appears in all studies and some genes only appear in very few studies.

Since the union of all genes for all the studies includes more than 27 000 unique genes, some gene selection is essential before fitting the hierarchical model. We performed gene selection in two steps, aiming to reduce the gene set to about 5 000 genes. The data were separated into full (Types 1 and 2) and partial (Types 3 and 4), and we conducted gene selection on these two sets separately. Based on the idea that the partial data include only genes that were found interesting at the study level, it is important to include all the genes from these studies, as they are the best

potential candidates for differential expression. The union of all genes provided by partial studies comprised 1 201 genes. For the set of full data, we selected only genes common to all studies. This intersection comprised 4 343 genes. The final set of genes consisted of the union of the genes coming from the partial studies and the full studies and comprised 5 141 genes distributed as in Table 2.

We fitted the hierarchical model using the spike and slab prior to the 11 studies on the reduced gene set, with 31 500 iterations performed in order to obtain 3 000 independent samples from the posterior distribution. This took 25 days, due to the large number of iterations of the Gibbs sampler and the number of genes. We found a total of 296 differentially expressed genes ($\widehat{w} > 0.5$), among which 68 of the corresponding values of $\widehat{w}$ had 95% confidence intervals not containing $w = 0.5$. Table 3 presents the names of the top 100 genes, the corresponding values of $\widehat{\gamma}$, which gives information about the direction of the differential expression, and $\widehat{w}$. A more complete list of differentially expressed genes is in the Supplementary Materials. Bold genes from Table 3 are known to be active in ovarian cancer (Jacob *and others*, 2009, for instance), which shows that our model produces a credible list of differentially expressed genes.

5. Conclusion and discussion

Our Bayesian hierarchical model is useful in the context of genomic data analyses, where many studies are available but not all provide raw data. Published results then become useless unless raw data are available. The model allows the integration of various data types, avoiding loss of information, which is common when performing meta-analysis or list aggregation, while maintaining, or even increasing, the power of detection of differentially expressed genes. Our model is adaptive, but is not heavily influenced by the choice of the hyperparameters. Choosing priors that shrink the parameters of interest close to zero for uninteresting genes eases the detection of the differentially expressed genes, contributing to the high power of our model. Comparisons of

our model with different priors and Borda's method show a clear improvement, increasing the number of correctly identified genes by 15% on average. As Borda's method with arithmetic mean was found to perform well compared to other rank aggregation methods (Lin, 2010), our model is a serious competitor.

The gain of power obtained by including all possible studies rather than only those providing raw data is highly significant. Moreover, the simulation design used in this article is realistic by mimicking the behavior of real microarray data, which reinforces the results that we obtained, and indicates that our model is efficient for real microarray data. We emphasize that the simulations are performed from an independent design and not from the model, meaning that no model is favoured in the simulations. Applications on a set of 11 real datasets also show promising results by retrieving many genes known to be involved in ovarian cancer. The criterion for detection of differentially expressed genes is a simple thresholding rule which is easy to apply using the posterior output. One of its main advantages is the absence of multiplicity correction, unlike that needed in the frequentist approach.

One drawback of our model for application on microarray data is its large computational requirements. It is infeasible to fit the model to the union of all genes from all studies, so gene selection is essential. Even if our gene selection is not based on differential expression, we might discard important genes. Including only the best candidate genes is unrealistic and is not appropriate to the chosen priors, which require a large proportion of uninteresting genes to be included for the priors to be valid. The real data example of Section 4 performs a gene selection independent of the model, as we only select genes corresponding to union or intersection of groups of genes. However, it would be interesting to be able to fit the model to the full set of genes.

In this paper, all studies contribute equally to the model and to the detection of differentially expressed genes. However, one might put more confidence in full studies, for example, where the analysis as well as the preprocessing steps are entirely reproducible, rather than in lists of

interesting genes that are not reproducible. Due to the high adaptivity of our model, we can easily associate weights to each of the studies to be combined, reflecting prior confidence in each.

We also assume that the genes are independent, which is not true in practice, though it is a common assumption in the literature, even for single study analysis. Correlations between the genes are included in the simulation design (Section 3.2), and do not seem to affect the efficiency of the model. However we believe the model would perform even better if the elements on which it is applied are independent. One could define modules, sets of genes that are correlated or are related in some sense, and apply the model on these elements. This trick would solve the gene selection and the dependence problems simultaneously. Indeed, using modules results in a huge dimension reduction, by grouping genes together, and it also results in independence, as genes that do not belong to the same module would be considered as independent. With the computational time therefore reduced, our model would be even more interesting and competitive with other methods.

6. Software

Software in the form of R and C code, is available on request from the corresponding author (alix.leboucq@epfl.ch).

7. Supplementary Materials

The reader is referred to the on-line Supplementary Materials for technical appendices, additional figures and details about simulations and real data.

Acknowledgments

We thank V. Heinzelmann and F. Jacob for their knowledge and expertise on the list of differentially expressed genes, and E. Boggis for helpful discussion about the normal-gamma prior.

References

Bignotti, E., Tassi, R. A., Calza, S., Ravaggi, A., Romani, C., Rossi, E., Falchetti, M., Odicino, F. E., Pecorelli, S. and Santin, A. D. (2006). Differential gene expression profiles between tumor biopsies and short-term primary cultures of ovarian serous carcinomas: identification of novel molecular biomarkers for early diagnosis and therapy. *Gynecologic Oncology* **103**(2), 405–416.

Bogdan, M., Ghosh, J. K. and Tokdar, S. T. (2011). Asymptotic bayes-optimality under sparsity of some multiple testing procedures. *The Annals of Statistics* **39**(3), 1551–1579.

Borda, J. C. (1781). Mémoire sur les élections au scrutin. *Histoire de l'Académie des Sciences*.

Boulesteix, A. L. and Slawski, M. (2009). Stability and aggregation of ranked gene lists. *Briefings in Bioinformatics* **10**(5), 556–568.

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* **39**(3/4), 324–345.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P. *and others*. (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genetics* **29**, 365–371.

Cancer Genome Atlas Research Network. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature* **474**(7353), 609–615.

Carvalho, C. M., Polson, N. G. and Scott, J. G. (2009). Handling sparsity via the horseshoe. *International Conference on Artificial intelligence and Statistics* **5**, 73–80.

Carvalho, C. M., Polson, N. G. and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97**(2), 465–480.

Choi, J. K., Yu, U., Kim, S. and Yoo, O. J. (2003). Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics* **19**(90001), 84–90.

Conlon, E., Song, J. and Liu, A. (2007). Bayesian meta-analysis models for microarray data: a comparative study. *BMC Bioinformatics* **8**(1), 80.

Datta, J. and Ghosh, K. (2012). Asymptotic properties of Bayes risk for the horsehoe prior. *Bayesian Analysis* **7**(4), 771–792.

DeConde, R. P., Hawley, S., Falcon, S., Clegg, N., Knudsen, B. and Etzioni, R. (2006). Combining results of microarray experiments: a rank aggregation approach. *Statistical Applications in Genetics and Molecular Biology* **5**(1), article 15.

DeLong, E. R., DeLong, D. M. and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**(3), 837–845.

Donninger, H., Bonome, T., Radonovich, M., Pise-Masison, C. A., Brady, J., Shih, J. H., Barrett, J. C. and Birrer, M. J. (2004). Whole genome expression profiling of advance stage papillary serous ovarian cancer reveals activated pathways. *Oncogene* **23**, 8065–8077.

Edgar, R., Domrachev, M. and Lash, A. E. (2002). Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research* **30**(1), 207–210.

Ganzfried, B. F., Riester, M., Haib-Kains, B., Risch, T., Tyekucheva, S., Jazic, I., Wang, X. V., Ahmadifar, M., Birrer, M. J., Parmigiani, G., Huttenhower, C. *and others*. (2013). `curatedOvarianData`: clinically annotated data for the ovarian cancer transcriptome. *Database* **2013**, bat013.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**(423), 881–889.

Griffin, J. E. and Brown, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis* **5**(1), 171–188.

Guerra, R. and Goldstein, D. R. (editors). (2009). *Meta-analysis and Combining Information in Genetics and Genomics*. Chapman & Hall/CRC.

Ishwaran, H. and Rao, J. S. (2003). Detecting differentially expressed genes in microarrays using Bayesian model selection. *Journal of the American Statistical Association* **98**(62), 438–455.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab gene selection for multigroup microarray data. *Journal of the American Statistical Association* **100**(471), 764–780.

Jacob, F., Goldstein, D. R., Fink, D. and Heinzelmann-Schwarz, V. (2009). Proteogenomic studies in epithelial ovarian cancer: established knowledge and future needs. *Biomarkers in Medicine* **3**(6), 743–756.

Kooperberg, C., Aragaki, A., Strand, A. D. and Olson, J. M. (2005). Significance testing for small microarray experiments. *Statistics in Medicine* **24**(15), 2281–2298.

Larsson, O. and Sandberg, R. (2006). Lack of correct data format and comparability limits future integrative microarray research. *Nature Biotechnology* **24**(11), 1322–1323.

Lili, L. N., Matyunina, L. V., Walker, L., Benigno, B. B. and McDonald, J. F. (2013). Molecular profiling predicts the existence of two functionally distinct classes of ovarian cancer stroma. *BioMed Research International* **2013**, 1–9.

Lin, S. (2010). Rank aggregation methods. *Wiley Interdisciplinary Reviews: Computational Statistics* **2**(5), 555–570.

Martoglio, A-M., Tom, B. D. M., Starkey, M., Corps, A. N., Charnock-Jones, D. Stephen and Smith, S. K. (2000). Changes in tumorigenesis and angiogenesis related gene transcript abundance profiles in ovarian cancer detected by tailored high density cDNA arrays. *Molecular Medicine* **6**(9), 750–765.

Meinhold-Heerlein, I., Bauerschlag, D., Zhou, Y., Sapinoso, L. M., Ching, K., H. Frierson, Jr, Bräutigam, K., Sehouli, J., Stickeler, E., Könsgen, D., Hilpert, F., von Daisenberg, C. S., Pfisterer, J., Bauknecht, T., Jonat, W., Arnold, N. *and others*. (2007). An integrated clinical-genomics approach identifies a candidate multi-analyte blood test for serous ovarian carcinoma. *Clinical Cancer Research* **13**(2), 458–466.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association* **83**(404), 1023–1032.

Mok, S. C., Bonome, T., Vathipadiekal, V., Bell, A., Johnson, M. E., Wong, K.-K., D.-C-Park, Hao, K., Yip, D. K. P., Donninger, H., Ozbon, L., Samini, G., Brady, J., Randonivich, M., Pise-Masison, C.A., Barett, J. C., Wong, W. H., Welch, W. R., Berkowitz, R. S. *and others*. (2009). A gene signature predictive for outcome in advanced ovarian cancer identifies a survival factor: Microfibril-associated glycoprotein 2. *Cancer Cell* **16**, 521–532.

Pang, X. and Gill, J. (2009). Spike and slab prior distributions for simultaneous Bayesian hypothesis testing model selection and prediction or nonlinear outcomes. Downloaded February 2012.

Polson, N. G. and Scott, J. G. (2010). Shrink globally, act locally: sparse Bayesian regularization and prediction. *Bayesian Statistics* **9**, 76.

Ramasamy, A., Mondry, A., Holmes, C. C. and Altman, D. G. (2008). Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Medicine* **5**(9), e184.

Rhodes, D. R., Barrette, T. R., Rubin, M. A., Ghosh, D. and Chinnaiyan, A. M. (2002). Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Research* **62**(15), 4427–4433.

Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A. and Chinnaiyan, A. M. (2004). Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proceedings of the National Academy of Sciences* **101**(25), 9309–9314.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C. and Müller, M. (2011). `pROC`: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**(1), 77.

Scott, J. G. (2009). Bayesian adjustment for multiplicity [Ph.D. Thesis]. Department of Statistical Science, Duke University.

Scott, J. G. (2011). Bayesian estimation of intensity surfaces on the sphere via needlet shrinkage and selection. *Bayesian Analysis* **6**(2), 307–327.

Smyth, G. K. (2005). Limma: linear models for microarray data. In: Gentleman, R., Carey, V., Dudoit, S. and R. Irizarry, W. Huber (editors), *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. New York: Springer, pp. 397–420.

Sutton, A. J., Abrams, K. R., Jones, D. R., Jones, D. R., Sheldon, T. A. and Song, F. (2000). *Methods for Meta-Analysis in Medical Research*. London: Wiley.

Warrenfeltz, S., Pavlik, S., Datta, S., Kraemer, E. T., Benigno, B. and McDonald, J. F. (2004). Gene expression profiling of epithelial ovarian tumours correlated with malignant potential. *Molecular Cancer* **3**, 27–43.

Welsh, J. B., Warrinkar, P. P., Sapinoso, L. M., Kern, S. G., Behling, C. A., Monk, B. J., Lockhart, D. J., Burger, R. A. and Hampton, G. M. (2001). Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. *Proceedings of the National Academy of Sciences* **98**(3), 1176–1181.

Yoshihara, K., Tajima, A., Komata, D., Yamamoto, T., Kodama, S., Fujiwara, H., Suzuki, M., Onishi, Y., Hatae, M., Sueyoshi, K., Fujiwara, H., Kudo, Y., Inoue, I. *and others*. (2009). Gene expression profiling of advanced-stage serous ovarian cancers distinguishes novel subclasses and implicates ZEB2 in tumor progression and prognosis. *Cancer Science* **100**(8), 1421–1428.

Zhang, X., Feng, J., Cheng, Y., Yao, Y., Ye, X., Fu, T. and Cheng, H. (2005). Characterization of differentially expressed genes in ovarian cancer by cDNA microarrays. *International Journal of Gynecological Cancer* **15**(1), 50–57.

8. Figures and tables

[*Received ?, ?; revised ?, ?; accepted for publication ?,?*]

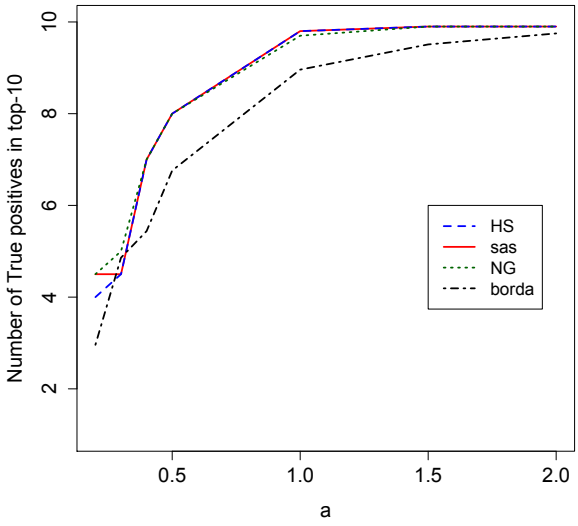
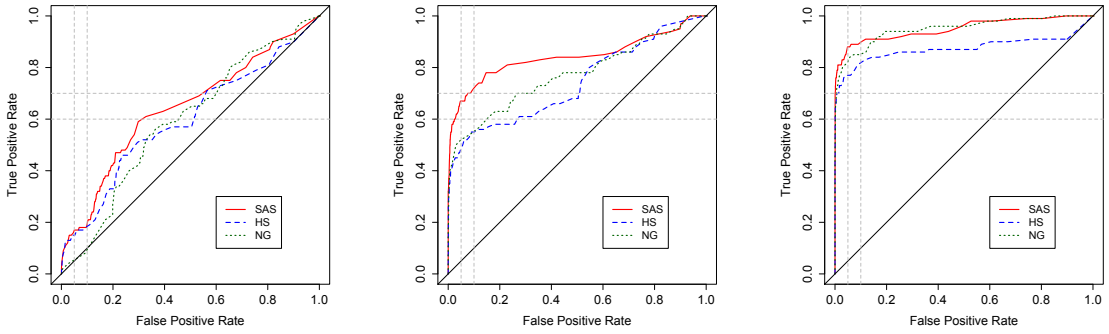


Fig. 1. Comparison of the model with the three different priors: horseshoe (HS), spike and slab (SAS) and normal gamma (NG) priors, and Borda's method. Here we compare the number of true differentially expressed genes in the top 10 genes based on the value of $\widehat{\gamma}$ for our model, or the Borda scores for Borda's method. In the simulation design, 10 genes were selected to be differentially expressed, which is why we look only at the top 10.

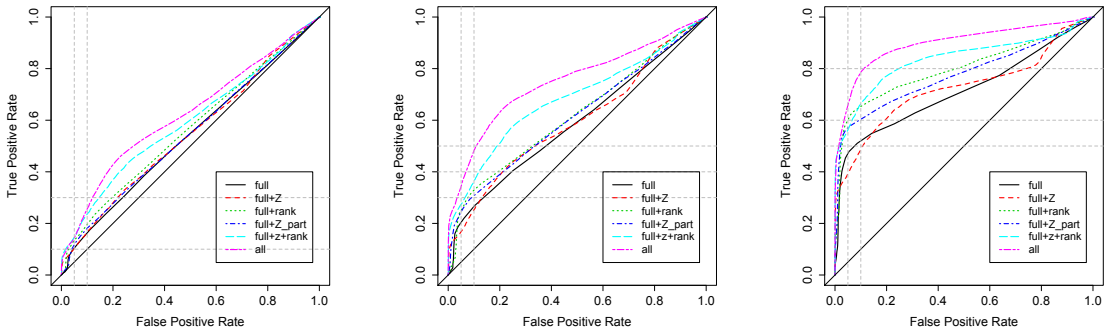


(a) $a = 0.2$ (b) $a = 0.3$ (c) $a = 0.5$

Fig. 2. ROC curves for prior comparisons. Power of the model for each of the three priors: spike and slab (SAS), horseshoe (HS) and normal-gamma (NG) priors, for several values of the differential expression parameter a.

Table 1. Summary of the 11 studies included in the meta-analysis. SOC denotes serous ovarian cancer.

Abbreviation	Study	Data source	Platform	Samples
Type 1: Full data				
MOK	Mok *and others* (2009)	GEO GSE18520	Affymetrix U133 plus 2.0	53 SOC+10 controls 20827 unique genes
YOS	Yoshihara *and others* (2009)	GEO GSE12470	Agilent Human 1A	43 SOC+10 controls 16546 unique genes
LIL	Lili *and others* (2013)	GEO GSE38666	Affymetrix U133 plus 2.0	18 SOC+12 controls 21049 unique genes
TCGA	Cancer Genome Atlas Research Network (2011)	`curatedOvarianData` (R package)	Affymetrix U133A	570 SOC+8 controls 12 981 unique genes
Type 2: z-scores				
WEL	Welsh *and others* (2001)	Website	Affymetrix HuGene FL	27 SOC+4 controls 5280 unique genes
Type 3: partial z-scores				
MAR	Martoglio *and others* (2000)	Article	Human Genome mapping	4 SOC+5 controls 33 unique genes
DON	Donninger *and others* (2004)	Supplementary file	Affymetrix U133 Plus 2.0	37SOC+6 controls 995 unique genes
Type 4: partial list of ranks				
WAR	Warrenfeltz *and others* (2004)	Article	CMT-GAPS slide	31 SOC+5 controls 15 unique genes
MEI	Meinhold-Heerlein *and others* (2007)	Article	Affymetrix U133A	67 SOC+9 controls 43 unique genes
ZHA	Zhang *and others* (2005)	Article	united genes holding	4 SOC+13 controls 39 unique genes
BIG	Bignotti *and others* (2006)	Supplementary file	Affymetrix U133A	19SOC+15 controls 117 unique genes



(a) $a = 0.2$ (b) $a = 0.3$ (c) $a = 0.5$

Fig. 3. Comparison of the power of several data combinations for two full studies, one full list of z-scores, one partial list of ranks and one partial list of z-scores. The different combinations are as follow (in the same order as the legend): *black*: only full studies, *red*: the full studies and the full list of z-scores, *green*: the full studies and the partial list of ranks, *blue*: the full studies and the partial list of z-scores, *cyan*: the full studies, the full z-scores and the partial list of ranks, *pink*: all the studies. The ROC curves are plotted for several values of the differential expression parameter a.

Table 2. Distribution of the genes by studies for the entire gene set and after gene selection. The table show the number of genes appearing in k studies, $k = 1, \ldots, 11$, before and after gene selection.

Number of studies	1	2	3	4	5	6	7	$\geqslant 8$
Number of genes (before)	7228	4655	4623	5936	4208	400	12	1
Number of genes (after)	778	19	1	0	3940	390	12	1

Table 3. Top-100 list of differentially expressed genes. Genes are ordered according to the value of $\widehat{w}$, from the most to the least differentially expressed. The estimates $\widehat{w}$ are obtained from the fit of our model to the 11 studies selected for the analysis. The estimate of $\widehat{\gamma}$ is also given for each gene and indicates the direction of differential expression. Bold genes are known to have a role in ovarian cancer.

Rank	Genes	$\widehat{w}$	sd($\widehat{w}$)	$\widehat{\gamma}$	sd($\widehat{\gamma}$)
1	CP	0.99	$< 10^{-2}$	3.07	0.38
2	**TOP2A**	0.99	$< 10^{-2}$	2.60	0.38
3	BCHE	0.99	0.01	-2.49	0.41
4	NEK2	0.99	0.01	2.48	0.38
5	TTK	0.99	0.01	2.48	0.38
6	CENPA	0.99	0.02	2.45	0.38
7	SPP1	0.99	0.02	2.31	0.39
8	**MELK**	0.99	0.02	2.35	0.39
9	PRAME	0.99	0.02	2.41	0.39
10	ADH1B	0.99	0.02	-2.33	0.41
11	KIAA0101	0.99	0.02	2.41	0.40
12	NMU	0.99	0.02	2.30	0.39
13	IGFBP6	0.99	0.03	-2.22	0.39
14	**KLK6**	0.99	0.02	2.24	0.39
15	EVI1	0.99	0.01	2.67	0.45
16	**CLDN3**	0.99	0.02	2.38	0.41
17	SST	0.99	0.02	2.27	0.41
18	**FOLR1**	0.99	0.03	2.22	0.39
19	**WFDC2**	0.99	0.03	2.30	0.40
20	UBE2C	0.99	0.03	2.22	0.38
21	CD24	0.99	0.03	2.47	0.45
22	HMGA2	0.99	0.04	2.21	0.43
23	SPARCL1	0.99	0.04	-2.00	0.38
24	KIF2C	0.99	0.04	2.08	0.40
25	ELF3	0.99	0.05	2.11	0.41
26	**MAL**	0.99	0.05	2.07	0.42
27	CDKN2A	0.98	0.05	2.09	0.41
28	**PAX8**	0.98	0.05	2.05	0.41
29	FOXM1	0.98	0.05	2.05	0.40
30	CENPF	0.98	0.06	2.02	0.40
31	TNNT1	0.98	0.06	2.01	0.42
32	**CLDN4**	0.98	0.06	2.05	0.42
33	HMMR	0.98	0.07	1.99	0.42
34	LCT	0.98	0.07	1.94	0.41
35	KIF11	0.98	0.07	1.92	0.41
36	TRIM31	0.98	0.08	1.92	0.43
37	CDC20	0.98	0.08	1.88	0.42
38	**TACSTD1**	0.98	0.08	2.61	0.63
39	**CCNB1**	0.97	0.09	1.87	0.44
40	PTTG1	0.97	0.10	1.93	0.48
41	PRSS8	0.97	0.10	1.87	0.46
42	CCNE1	0.97	0.11	1.80	0.45
43	RRM2	0.96	0.11	1.84	0.47
44	EHF	0.96	0.11	1.90	0.48
45	**S100A1**	0.96	0.11	1.82	0.47
46	ATP6V1B1	0.95	0.13	1.75	0.50
47	KLK7	0.95	0.13	1.72	0.48
48	ABCA8	0.95	0.13	-1.77	0.49
49	ALDH1A1	0.95	0.13	-1.70	0.47
50	SCNN1A	0.94	0.14	1.73	0.51
51	C7	0.94	0.14	-1.69	0.52
52	PDHA2	0.94	0.15	1.70	0.52
53	ANXA8	0.94	0.17	2.22	0.74
54	GLDC	0.94	0.15	1.66	0.49
55	LAMP3	0.94	0.16	1.72	0.54
56	KRT7	0.93	0.16	1.57	0.50
57	CKS2	0.93	0.16	1.61	0.50
58	AOX1	0.93	0.16	-1.63	0.53
59	EZH2	0.93	0.16	1.60	0.49
60	MYH11	0.93	0.16	-1.63	0.51
61	NY-REN-7	0.92	0.20	2.41	0.94
62	CCNA2	0.91	0.18	1.51	0.53
63	TK1	0.91	0.18	1.50	0.52
64	MAD2L1	0.91	0.18	1.51	0.54
65	ACTG2	0.91	0.18	-1.53	0.56
66	SCGB2A1	0.91	0.19	1.65	0.64
67	**MUC1**	0.90	0.19	1.52	0.57
68	SLC2A1	0.90	0.19	1.46	0.53
69	SULT1C2	0.89	0.20	1.47	0.57
70	MGP	0.87	0.21	-1.37	0.55
71	THBD	0.87	0.21	-1.38	0.57
72	IGF2BP3	0.86	0.22	1.36	0.58
73	SPINT2	0.86	0.22	1.36	0.58
74	ZIC1	0.86	0.22	1.38	0.60
75	CGN	0.85	0.24	1.44	0.66
76	RNASE4	0.85	0.22	-1.31	0.57
77	EFEMP1	0.85	0.22	-1.33	0.60
78	APOA1	0.85	0.23	1.31	0.59
79	DXYS155E	0.84	0.28	2.22	1.21
80	TFAP2A	0.83	0.23	1.25	0.57
81	NDP52	0.83	0.29	2.21	1.27
82	FRY	0.82	0.24	-1.25	0.62
83	DEFB1	0.79	0.25	1.12	0.59
84	TYMS	0.79	0.25	1.12	0.56
85	GRPR	0.79	0.25	1.15	0.60
86	MYBL2	0.79	0.25	1.13	0.58
87	MAOB	0.79	0.25	-1.11	0.56
88	CNN1	0.77	0.25	-1.08	0.57
89	CLDN7	0.77	0.26	1.14	0.64
90	BLM	0.77	0.25	1.09	0.59
91	CXCL10	0.76	0.25	1.07	0.61
92	KIF23	0.76	0.25	1.04	0.56
93	KLF4	0.76	0.25	-1.03	0.56
94	CDKN3	0.75	0.26	1.03	0.57
95	HTR3A	0.75	0.25	1.02	0.57
96	EPCAM	0.75	0.27	1.09	0.67
97	GNG11	0.75	0.26	-1.01	0.55
98	KIF14	0.74	0.25	1.00	0.55
99	NDN	0.74	0.25	-0.99	0.54
100	RAD54L	0.74	0.25	0.98	0.55