

Bayesian Modelling for Biological Annotation of Gene Expression Pathway Signatures

Haige Shen & Mike West¹

July 12, 2009

1 Introduction

Studies in high-throughput genomics often generate multiple gene expression *signatures* – lists of genes with associated numerical measures of change in gene expression relative to an experimental condition or outcome. A biological or environmental design factor in a controlled experiment generates a signature of response to that factor (Huang et al., 2003; Bild et al., 2006; Chen et al., 2008), while evaluation of expression related to a specific clinical outcome may generate a signature of the outcome in disease studies (West et al., 2001; Huang et al., 2003; Rich et al., 2005; Seo et al., 2007). Indeed, the concept of gene expression signatures as characterizing pathway status has emerged as central in Bayesian analyses in genomics and emerging systems biology in cancer and other areas. The development of sparse ANOVA and latent factor models to improve estimation of expression signatures in both experimental (*in vitro*) and observational (*in vivo*) contexts has been substantially motivated by this view (e.g. West, 2003; Lucas et al., 2006; Wang et al., 2007; Carvalho et al., 2008; Merl et al., 2009). Recent applied studies of deregulated pathways in cancer as well as other contexts, including basic biological pathway discovery and evaluation studies, studies of drug responsiveness and prognostic/predictive risk profiling, reflect this (e.g. Chen et al., 2008; Chang et al., 2009; Lucas et al., 2009; Merl et al., 2009).

Interpretation of identified gene expression signatures relies in part on comparison with biological databases that contain lists of putatively pathway-specific genes. The gene lists themselves, without real-biology connectivities explicitly described, simply represent biological pathways through the sets of genes named as participating in the biological processes the pathways play roles in. A core challenge is to assess the signatures against these databases to suggest potential pathway interpretations. Our focus here is a formal, novel Bayesian approach to this problem.

Identification of this problem led to the non-Bayesian gene set enrichment analysis (GSEA) method (Subramanian et al., 2005) and follow-on approaches (Newton et al., 2007). These methods aim to measure aggregate association between a full list of genes ranked by their association with an outcome - also referred to as a *phenotype* - and one or more given sets of genes. The underlying idea is to assess whether or not a specified “pathway” gene set is enriched with genes that score highly in association with the experimental outcome. Based on non-Bayesian testing and (sample or gene) randomization methods, these methods tend to lead to false positives, have difficulties in dealing with small sized gene sets, rely on an assumption that pathway database gene lists are error-free, and are restricted in applications to simple contexts of genes up/down regulated. On the latter point, we are particularly interested in understanding potential biological pathways

¹Haige Shen is Senior Biostatistician at Novartis Oncology, N.J., and Mike West is The Arts & Sciences Professor of Statistical Science at Duke University. *emails:* haigeshen@yahoo.com, mw@stat.duke.edu

underlying estimated latent factors applied to observational data sets (e.g. [Lucas et al., 2006](#); [Carvalho et al., 2008](#); [Lucas et al., 2009](#); [Merl et al., 2009](#)) and these existing methods simply do not apply to the forms of information summaries produced in such analyses.

Our Bayesian *probabilistic pathway annotation* (PROPA) model presented here addresses these broader questions. PROPA provides: (a) probabilistic assessments of phenotype-pathway concordance in terms of marginal likelihoods and posterior probabilities; (b) an ability to assessment of experimental results against many biological pathways simultaneously and in comparison with each other; (c) adaptation to uncertainties and potential errors in both experimentally defined gene-phenotype association measures *and* in biological databases; and (d) a general theoretical framework that allows the specific method to be extended to incorporate other forms of genomic data. Item (c) here also leads to an ability to suggest refinements to pathway gene lists. Simulation and breast cancer genomics examples illustrate these points. A core component of the annotation analysis involves evaluation of marginal likelihoods in models with high-dimensional parameters. For this, we develop a novel extension of variational methods (e.g. [Jordan et al., 1999](#); [McGrory and Titterton, 2007](#)) that, in addition to proving extremely effective in the pathway annotation analysis, is of broad interest and potential use in Bayesian model evaluation.

2 Context and Models

2.1 Notation and Framework

A biological study investigates the changes in gene expression on p genes due to an experimental *factor*. We are interested in (a) which genes are related to this factor in terms of the expression change, and (b) how does this factor relate to known, published gene lists representing annotation of biological pathways? The experiment leads to measures of association of the genes, in terms of expression changes, with the experimental factor; these are inputs to annotation analysis. We define terminology and notation as follows:

- $\mathcal{G} = \{1, \dots, p\}$, the full list of genes; in human studies, $p \sim 20 - 25,000$.
- A *pathway* is, simply, any specific subset of genes from \mathcal{G} .
- \mathcal{F} , an unknown list of genes whose expression changes are truly related to the experimental factor; we call \mathcal{F} the *factor pathway* to give it a definite name.
- $\Pi = \{\pi_g, g = 1, \dots, p\}$, a set of numerical measures of association of each of the genes, in terms of the expression change, with the experimental factor pathway \mathcal{F} .
- \mathcal{A} , a generic label for a *biological pathway*; \mathcal{A} is a simply an unknown list of genes. $A_j, j = 1, \dots, m$, a full set of known biological pathways.
- A , a generic label for a list of genes in a published, annotated biological database, putatively linked to a true, unknown biological pathway \mathcal{A} . We call A a *reference gene list* for pathway \mathcal{A} . $A_j, j = 1, \dots, m$, the set of reference gene lists corresponding to pathways \mathcal{A}_j .

We use the Molecular Signatures database (MSigDB C2 collection) (Broad Institute, 2007) to obtain $m \approx 1000$ reference gene sets A_j . These are, of course, incomplete and typically error-prone; A_j provides incomplete and noisy information on the pathway \mathcal{A}_j .

Based on the expression experiment, Π is known *data* to be used in assessing concordance of the unknown, underlying experimental factor pathway \mathcal{F} with candidate biological pathways \mathcal{A}_j , $j = 1, \dots, m$. We do this with models that compute the $j = 1, \dots, m$ posterior probabilities

$$Pr(\mathcal{F} = \mathcal{A}_j | \Pi, A_1, \dots, A_m) \propto Pr(\mathcal{F} = \mathcal{A}_j | A_1, \dots, A_m) p(\Pi | A_1, \dots, A_m, \mathcal{F} = \mathcal{A}_j). \quad (1)$$

Focus here on the likelihood terms $p(\Pi | A_1, \dots, A_m, \mathcal{F} = \mathcal{A}_j)$ as j moves across all the pathways; this is the overall measure from the experimental predictions Π that underlies pathway assessment, and can be applied whatever the chosen values of the $Pr(\mathcal{F} = \mathcal{A}_j | A_1, \dots, A_m)$.

Here Π may include essentially any measures, such as test statistics or other summaries of a statistical analysis of the experimental data. Different measures should be modelled differently within the overall framework. Here, our example measures are probabilities of differential expression. In designed experiments, π_g will be a posterior probability of differential expression of gene g related to an experimental intervention. In observational studies, π_g will be a posterior probability of a non-zero regression coefficient or loading on a latent factor in a sparse factor model of expression data (West, 2003; Lucas et al., 2006; Seo et al., 2007; Carvalho et al., 2008). So $\pi_g \in [0, 1]$ and larger values indicate stronger association with \mathcal{F} ; typically very many of the π_g will be very small, while those for genes associated with \mathcal{F} will be larger.

2.2 Statistical Model

Focus on a single, generic biological pathway $\mathcal{A} = \mathcal{A}_1$ and its reference gene list $A = A_1$, and consider relevant statistical models for the core component $p(\Pi | A_1, \dots, A_m, \mathcal{F} = \mathcal{A}_j)$.

Model for data Π assuming known pathway membership of genes

We first assume that π_g , given an associated pathway \mathcal{A} and its reference gene set A , is independent of the other $\{\pi_g\}_{k \neq g}$. Note that this *does not* assume lack of interaction or co-regulation among genes; that dependence should already be accounted for in the analysis that led to the Π . If $\mathcal{F} = \mathcal{A}$, then π_g will likely be higher for $g \in \mathcal{A}$ than for $g \notin \mathcal{A}$, suggesting models of the form

$$(\pi_g | g \in \mathcal{A}, \mathcal{F} = \mathcal{A}) \sim f_1(\pi_g) \quad \text{and} \quad (\pi_g | g \notin \mathcal{A}, \mathcal{F} = \mathcal{A}) \sim f_0(\pi_g) \quad (2)$$

where f_0, f_1 are densities on $[0, 1]$ with f_1 favoring high values of π_g and f_0 favoring lower values. A natural choice is beta densities: $f_1(\pi) \equiv f_1(\pi | \alpha_1) = \text{Be}(\alpha_1, 1)$ and $f_0(\pi) \equiv f_0(\pi | \alpha_0) = \text{Be}(1, \alpha_0)$ with $\alpha_0, \alpha_1 > 1$ (Fig. 1(a)). This picture is consistent with histograms of π_g values generated in sparse factor analyses (e.g. Carvalho et al., 2008; Wang et al., sent); see Figure 1(b) as an example. We have explored model robustness to the assumed form in other examples, including simulation examples, and have no major concerns about the beta forms being overly restrictive, though other

forms will be relevant in analyses with other definitions of Π . We use independent reference priors for the α parameters, viz $p(\alpha_r) \propto \alpha_r^{-1}$, $1 < \alpha_r$, ($r = 0, 1$).

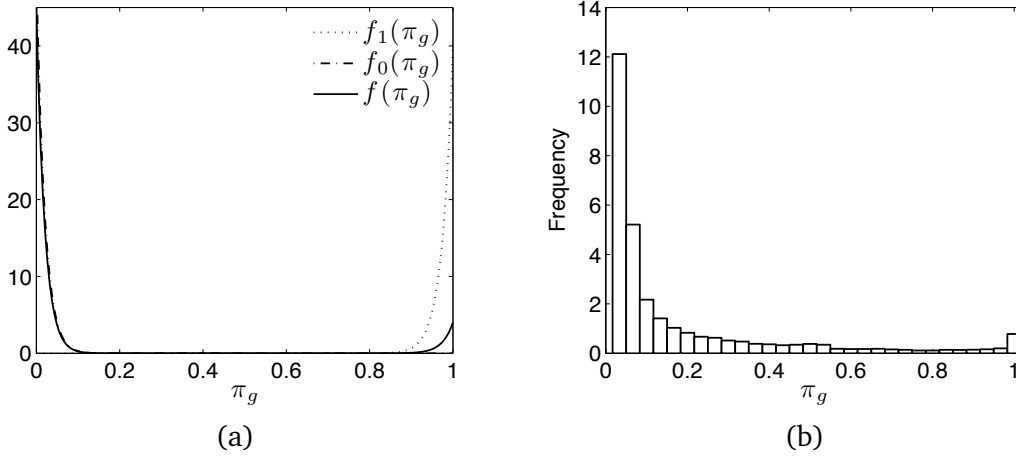


Figure 1: (a) $f(\pi_g)$ is a mixture of $f_1(\pi|\alpha_1) = \text{Be}(\alpha_1, 1)$ and $f_0(\pi|\alpha_0) = \text{Be}(1, \alpha_0)$; in this example, $f_0(\pi_g)$ is very close to $f(\pi_g)$. (b) Histogram of π_g of thousands of genes from a real expression data analysis.

Model for pathway membership of genes

We do not know which genes are in \mathcal{A} ; the reference gene set A provides data. If $g \in A$, that suggests $g \in \mathcal{A}$ although g may be a false-positive in the published list. Also, reference gene lists are subject to revision as new biological information arises, so genes $g \notin A$ may be members in future; hence, there may be false-negatives, i.e., genes $g \in \mathcal{A}$ but $g \notin A$.

Introduce indicators z_1, \dots, z_p such that, when $\mathcal{F} = \mathcal{A}$, $z_g = 1$ if $g \in \mathcal{A}$, and 0 otherwise. Call z_g the *pathway membership indicator* of gene g . We need probabilities over the z_g ; A provides relevant information. Assume conditionally independent Bernoulli models $Pr(z_g = 1|\beta_g) = \beta_g$, so that marginalization of equation (2) with respect to z_g yields the implied prior data distribution as a mixture of $f_1(\pi_g|\alpha_1)$ and $f_0(\pi_g|\alpha_0)$ weighted by β_g and $1 - \beta_g$; see Figure 1(a). To complete the model requires priors for the β_g , which we take as

$$\begin{aligned} (\beta_g|g \in A, \mathcal{F} = \mathcal{A}) &\sim \text{Be}(\phi_A r_A, \phi_A(1 - r_A)), \\ (\beta_g|g \notin A, \mathcal{F} = \mathcal{A}) &\sim \text{Be}(\phi_B r_B, \phi_B(1 - r_B)), \end{aligned} \quad (3)$$

with specified means $r_A, r_B \in (0, 1)$ and $\phi_A, \phi_B > 0$. Marginalizing over the β_g , we see that r_A is the *a priori* true positive probability for genes $g \in A$, while r_B is the false negative probability for $g \in \mathcal{A}$. Specification of r_A should depend on the expectation of the quality of reference gene sets. In a pathway gene set database, genes sets are curated from a variety of sources. We adopt a generic view that a published gene set A is a fairly good representation of the true pathway gene set \mathcal{A} but allow for errors, so take r_A relatively large, e.g., 0.7. For r_B , note that the number of genes in A , typically tens to a few hundreds, will usually be small compared to the full gene list \mathcal{G} , and a reasonable value of r_B should be at least less than the ratio of the number of signature genes

(genes with high probabilities of association with \mathcal{F}) to the total number of genes, e.g., 0.005. The specification of r_A and r_B is empirical and to some extent allows flexibility. The impact of r_A and r_B specification is demonstrated in an example below. The ϕ_A and ϕ_B constrain the variation range of the prior for the β_g around these means, and relatively small values provide robustness.

Annotated databases are incomplete and error prone. We can explore this using posterior pathway membership probabilities for each gene g , namely

$$\pi_g^* = Pr(g \in \mathcal{A} | \Pi, A, \mathcal{F} = \mathcal{A}), \quad (4)$$

with respect to the pathway \mathcal{A} . This is exemplified below. In any one example, there may well be genes *known* to lie in a specific biological pathway but that are not activated under a specific experimental condition. Such genes will be treated as false-positive members of a reference gene set in our model; they may, of course, appear differently under other experimental conditions.

2.3 Marginal Likelihood for Pathway Assessment

We use $\alpha_{0:1}$, $\beta_{1:p}$ and $z_{1:p}$ to denote $\{\alpha_0, \alpha_1\}$, $\{\beta_1, \dots, \beta_p\}$ and $\{z_1, \dots, z_p\}$, respectively, extending the use of this concise notation to other quantities as needed. The full model likelihood $p(\Pi | A, \mathcal{F} = \mathcal{A})$ can be expressed as

$$p(\Pi | A, \mathcal{F} = \mathcal{A}) = \int_{\alpha_{0:1}} \int_{\beta_{1:p}} \sum_{z_{1:p}} \mathcal{L}(\alpha_{0:1}, z_{1:p}) \prod_{g=1}^p p(z_g | \beta_g) p(\beta_g | A, \mathcal{F} = \mathcal{A}) p(\alpha_{0:1}) d\beta_{1:p} d\alpha_{0:1} \quad (5)$$

with

$$\mathcal{L}(\alpha_{0:1}, z_{1:p}) = \prod_{g=1}^p f_1(\pi_g | \alpha_1)^{z_g} f_0(\pi_g | \alpha_0)^{1-z_g}. \quad (6)$$

We can integrate analytically over $\beta_{1:p}, \alpha_{0:1}$ reducing the computation to summation over the 2^p values $z_{1:p}$; see Section 6.1 where we derive

$$p(\Pi | A, \mathcal{F} = \mathcal{A}) = \sum_{z_{1:p}} p(\Pi, z_{1:p} | A, \mathcal{F} = \mathcal{A}) \quad (7)$$

and where the quantity $p(\Pi, z_{1:p} | A, \mathcal{F} = \mathcal{A})$ can be evaluated at any chosen $z_{1:p}$. The sum above is a difficult numerical problem addressed in Section 3.2.

Another reduced form that is theoretically attractive but practically of little value results from marginalization over $z_{1:p}$ and $\beta_{1:p}$ conditional on $\alpha_{0:1}$, namely

$$p(\Pi | A, \mathcal{F} = \mathcal{A}) = \int_{\alpha_{0:1}} p(\Pi | \alpha_{0:1}, A, \mathcal{F} = \mathcal{A}) p(\alpha_{0:1}) d\alpha_{0:1}. \quad (8)$$

The integrand here can be evaluated, but only practicably when p is small; see Section 6.1.

3 Computation

3.1 MCMC Posterior Simulation

The following conditional distributions are immediate; in each, only the conditioning quantities required to specify the distribution are mentioned.

First, α_0 and α_1 are conditionally independent with truncated gamma conditionals; specifically, the two distributions are

$$\text{Ga} \left(\alpha_0 \mid \sum_{g=1}^p (1 - z_g), - \sum_{g=1}^p (1 - z_g) \log(1 - \pi_g) \right), \text{Ga} \left(\alpha_1 \mid \sum_{g=1}^p z_g, - \sum_{g=1}^p z_g \log \pi_g \right),$$

subject to $1 < \alpha_r, (r = 0 : 1)$.

Second, the β_g are conditionally independent with beta distributions $\text{Be}(a_g, b_g)$ depending on z_g . For $g \in A$, $a_g = z_g + \phi_A r_A$ and $b_g = (1 - z_g) + \phi_A (1 - r_A)$; for $g \notin A$, $a_g = z_g + \phi_B r_B$ and $b_g = (1 - z_g) + \phi_B (1 - r_B)$.

Third, the z_g are conditionally independent with probabilities on $z_g = 1$ of

$$\rho_g = \beta_g \alpha_1 \pi_g^{\alpha_1 - 1} / (\beta_g \alpha_1 \pi_g^{\alpha_1 - 1} + (1 - \beta_g) \alpha_0 (1 - \pi_g)^{\alpha_0 - 1}).$$

The posterior pathway membership probability π_g^* is the posterior mean of ρ_g .

Efficient code for this MCMC evidences generally fast mixing and rapid convergence across many examples. The rather low dependence among the z_g , induced by lack of knowledge of the $\alpha_{0:1}$, suggests swift convergence is to be expected even though $p \approx 20 - 25,000$.

3.2 Marginal Likelihood Computation: General Strategy

A core methodological issue is the evaluation of the determining marginal likelihood of equation (5), and sets of such quantities $p(\Pi \mid A_1, \dots, A_m, \mathcal{F} = \mathcal{A}_j)$ in the practical context of assessing evidence for and against $\mathcal{F} = \mathcal{A}_j$ for a number or many pathways $j = 1, \dots, m$.

In very small, unrealistic examples we can use quadrature methods to compare with other approximations. We do this in the simulated example in Section 4, simply applying direct quadrature to the two-dimensional integral form of equation (8). Even with p very small, this method is limited since it requires evaluation of integrands on the density scale and quickly runs into floating-point overflow problem. Quadrature is simply not relevant for real applications.

The reduced version of equation (7) has a closed form but involves summing over all 2^p values of $z_{1:p}$ so that numerical approximations are needed. Since we use MCMC, then methods of marginal likelihood computation using MCMC outputs are attractive. Having experimented with multiple such methods (Newton and Raftery, 1994; Chib, 1995), all found to be inapplicable due to either the floating-point overflow problem or difficulties in proposing good density functions to approximate the joint posterior distribution of model parameters, we adapted mean-field variational methods (VM) (Jordan et al., 1999; Corduneanu and Bishop, 2001; McGrory and Titterton,

2007). The VM approach naturally solves the floating-point overflow problem by using a summation of logarithmic terms to approximate the log marginal likelihood. Our studies confirm the utility of this approach, especially in this high-dimensional context. A VM method yields a *lower bound* on the target value of the marginal likelihood; our extensions include an *upper bound* so we can bracket the actual value.

For any two densities $q_L(z_{1:p}), q_U(z_{1:p})$ with the same support as $p(z_{1:p}|\Pi, A, \mathcal{F} = \mathcal{A})$, manipulating Jensen’s inequality easily yields

$$L(q_L) \leq \log(p(\Pi|A, \mathcal{F} = \mathcal{A})) \leq U(q_U)$$

where, for any such density $q(z_{1:p})$, the quantities

$$L(q) = \sum_{z_{1:p}} q(z_{1:p}) \log[p(\Pi, z_{1:p}|A, \mathcal{F} = \mathcal{A})/q(z_{1:p})] \quad (9)$$

and

$$U(q) = \sum_{z_{1:p}} p(z_{1:p}|\Pi, A, \mathcal{F} = \mathcal{A}) \log[p(\Pi, z_{1:p}|A, \mathcal{F} = \mathcal{A})/q(z_{1:p})] \quad (10)$$

bound the log marginal likelihood; see Section 6.2 for technical details.

The VM concept is to choose parametric *variational densities* $q_L(z_{1:p})$ and $q_U(z_{1:p})$ to optimize these bounds. If each depends on a free parameter that can be varied, the computational problem is optimizing these *variational* parameters. The closer a variational density is to $p(z_{1:p}|\Pi, A, \mathcal{F} = \mathcal{A})$, the better will be the bound. Mean-field VM methods use factorized variational densities. This is natural here since the z_g have low dependence under the posterior. Thus we use q_L, q_U of the form

$$q(z_{1:p}|\gamma_{1:p}) = \prod_{g=1}^p \gamma_g^{z_g} (1 - \gamma_g)^{1-z_g} \quad (11)$$

where the $\gamma_{1:p}$ are vectors of variational parameters to be chosen.

3.3 Marginal Likelihood Computation: A Variational Method

We refer to the implementation of the above ideas that rely on the MCMC analysis as *Monte Carlo variational approximation*. Full details appear in Ji et al. (2009); essential results for the PROPA model are noted here, with more details in Section 6.2

Upper Bound Optimization: With q_U of the form in equation (11), it is trivially seen that the global minimum value of the upper bound in equation (10) is achieved at $\gamma_{1:p} = \bar{z}_{1:p} = E(z_{1:p}|\Pi, A, \mathcal{F} = \mathcal{A})$, i.e., by setting the latent indicators $z_{1:p}$ equal to their posterior means. These means are estimated at values $\bar{z}_{1:p}$ based on the MCMC output $\{z_{1:p}^i, i = 1, \dots, I\}$ and the Monte

Carlo approximation to the optimal upper bound is simply

$$\bar{U} = I^{-1} \sum_{i=1}^I \{\log p(\Pi, z_{1:p}^i | A, \mathcal{F} = \mathcal{A}) - \log q(z_{1:p}^i | \bar{z}_{1:p})\}.$$

This is easily computed and, assuming MCMC convergence, \bar{U} converges almost surely to the true global minimum upper bound of the log marginal likelihood.

Lower Bound Optimization: Existing mean-field, lower bound variational methods typically build on the Monte Carlo EM algorithm (Celeux and Diebolt, 1992; Chan and Ledolter, 1995). By combining with a stochastic approximation step, convergence of a stochastic version of EM was established under mild conditions in Delyon et al. (1999). This inspired the novel variational method (Ji et al., 2009) that is applied here; full algorithmic details are in Section 6.3.

The global optimizing value of $\gamma_{1:p}$ satisfies the set of p equations $f_g(\gamma_{1:p}) = 0$, where for each $g = 1, \dots, p$,

$$f_g(\gamma_{1:p}) = \sum_{z_{1:p}} (z_g - \gamma_g) [1 + \log q(z_{1:p} | \gamma_{1:p}) - \log p(\Pi, z_{1:p} | A, \mathcal{F} = \mathcal{A})]. \quad (12)$$

An iterative procedure successively approximates the solution to these equations using Monte Carlo and stochastic approximation; the former allows us to estimate $f_g(\gamma_{1:p})$ by Monte Carlo over $z_{1:p}$ at any value of $\gamma_{1:p}$, while the latter applies to successively update estimates of the optimizing vector $\gamma_{1:p}$. As detailed in the Appendix, an iterative algorithm uses these ideas to define a sequence of $\gamma_{1:p}$ vectors that converges with probability one to $\gamma_{1:p}^*$ satisfying $f_g(\gamma_{1:p}^*) = 0$, $g = 1, \dots, p$; a finite run of the algorithm provides an iterative approximation to this optimizing value. By further Monte Carlo sampling $z_{1:p}^h \sim q(z_{1:p} | \gamma_{1:p}^*)$, ($h = 1, \dots, H$), we can then also evaluate a consistent estimate of the optimal lower bound,

$$\bar{L} = H^{-1} \sum_{h=1}^H \{\log p(\Pi, z_{1:p}^h | A, \mathcal{F} = \mathcal{A}) - \log q(z_{1:p}^h | \gamma_{1:p}^*)\}. \quad (13)$$

4 Evaluation and Illustrations

A “Small” Simulated Example ($p = 18, m = 17$) : In a synthetic example to fix ideas and demonstrate the marginal likelihood approximation, association probabilities on $p = 18$ genes (Fig. 2) show that the first five genes are likely members of \mathcal{F} , several genes with very low π_g are not likely to be in \mathcal{F} , while four genes with π_g near 0.5 are uncertain. Consider $m = 17$ biological pathway reference gene sets, A_1, \dots, A_{17} , constructed as in Figure 2(a); reference set A_j is the first j genes in the ordered list of 18 genes. Analyses use $r_A = 0.8$, $r_B = 0.1$ and $\phi_A = \phi_B = 8$.

The log marginal likelihood (shifted and scaled to $[0, 1]$ in Figure 2(b)) increases over $j = 1, \dots, 5$ to a peak at $j = 5$, suggesting pathways \mathcal{A}_4 and \mathcal{A}_5 are supported by the data Π . This is consistent with the simulation design in that the first few genes are the signature genes of \mathcal{F} , having high π_g values. The marginal likelihood across the remaining reference gene sets is also reasonable

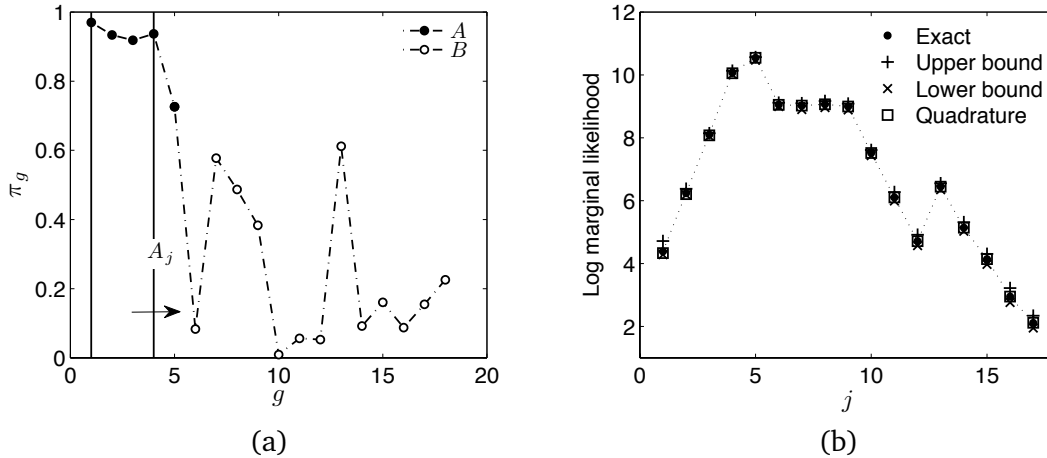


Figure 2: (a) π_g in the simulated data set with $p = 18$. Genes in reference set A_j are genes $1, \dots, j$ for each $j = 1, \dots, 17$. (b) Log marginal likelihood for each of the 17 pathways A_j .

given the values of the π_g . Figure 2(b) shows that the Monte Carlo variational upper and lower bounds agree well with the exact marginal likelihood values and quadrature based approximations using equation (8). In this “tiny p ” example, exact and quadrature computations are feasible, and demonstrate the accuracy of the upper and lower bound approximations. The spread between upper and lower bounds are small on the log likelihood scale ($\sim 0.05 - 0.2$) and certainly good enough to distinguish the different pathways/models.

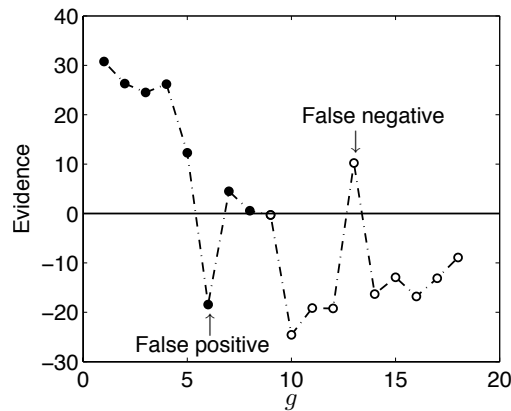


Figure 3: Pathway A_8 membership evidence for each gene g , in terms of log base 10 Bayes’ factors for: against $g \in A_8$, in analysis of simulated data with $p = 18$.

The MCMC provides estimates of posterior pathway membership probabilities π_g^* of equation (4) to aid false-positive/false-negative assessments. Focus on pathway A_8 ; reference set A_8 is exactly the first 8 genes. For each g and each reference gene set, compute π_g^* and convert to the corresponding *evidence* dB scale, i.e., the log base 10 Bayes’ factors on $g \in A_8$ versus $g \notin A_8$; see Figure 3. Genes in A_8 but with low π_g , and genes not in A_8 but with high π_g , might be regarded as false positives and false negatives, respectively. Gene $g = 6$, a member of gene set A_8 , has

membership evidence close to -20dB , strongly suggesting it is not a member of the true pathway \mathcal{A}_8 (false positive). Gene 13 is not a member of \mathcal{A}_8 , but it has membership evidence greater than 10dB , which is substantial evidence that this gene is in fact a member of \mathcal{A}_8 (false negative).

Marginal Likelihood Approximation with Real Data: “Large” $p = 19,645$: With realistically large p , the convergence of the iterative lower bound optimization can be slow. A slight modification of the bounding approach to address this is a compromise strategy with a pseudo-optimal lower bound; specifically, a bound as in equation (13) but now with $\gamma_{1:p}^*$ replaced by $\bar{z}_{1:p}$, the MCMC posterior mean of $z_{1:p}$. This uses the same variational density as in the optimal *upper* bound approximation. The rationale is that, when the factorized density q is a good approximation of the posterior for $z_{1:p}$, the optimal variational densities for upper and lower bounding will be similar; this has been seen in multiple examples. The pseudo-optimal lower bound is always less than the optimal, but is massively more attractive computationally when p is large.

We demonstrate this with real data on $p = 19,645$ genes and with $m = 15$ pathways whose reference gene sets come from the MSigDB C2 collection. The Π are probabilities of association between genes and a gene expression signature representing genes related to the responses of human mammary epithelial cells to lactic acidosis (Chen et al., 2008; Merl et al., 2009). Analysis assumes $r_A = 0.7$, $r_B = 0.005$, $\phi_A = 8$ and $\phi_B = 3$. Figure 4(a) shows upper and pseudo-optimal lower bounds of log marginal likelihoods for the 15 pathway gene sets. The distances between pairs of bounds are clearly small enough for practical usage; see Figure 4(b).

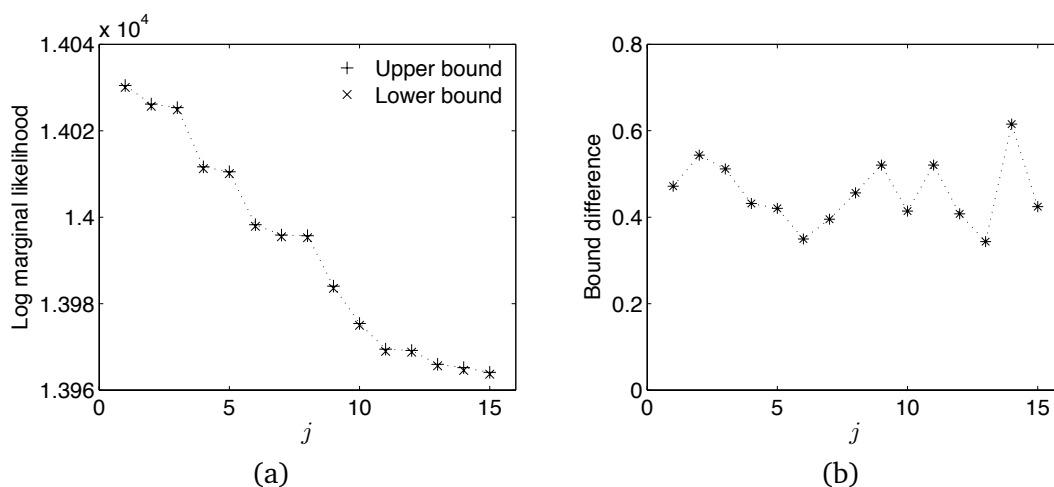


Figure 4: (a) Upper and quasi-lower bounds of log marginal likelihoods for each pathway $j = 1 : m, m = 15$ in study with $p = 19,645$ genes. (b) Upper minus lower bound for each pathway $j = 1 : 15$.

5 Applications to Hormonal Pathways in Breast Cancer Genomics

5.1 ER Pathway

About two-thirds of diagnosed breast cancers show over-expression of ER, the estrogen-receptor gene. ER status (high/low) is a key prognostic factor in breast cancer (Deroo and Korach, 2006;

Moggs and Orphanieds, 2001). Our prior study of 153 primary breast tumor samples (Carvalho et al., 2008) records expression data and protein assay-based ER+/- status from immunohistochemical (IHC) staining. Analysis using BFRM (Wang et al., 2007) generated association probabilities $\Pi = \pi_{1:p}$ as well as the sign of association between expression and ER+/- status for $p = 8,764$ genes (unique Entrez gene IDs). The π_g , displayed in Figure 5(a), show that a substantial number of genes apparently associate with the experimental factor pathway \mathcal{F} , here known to be ER related. Figure 5(b) shows PROPA upper and pseudo-optimal lower bounds on log marginal likelihoods for the $m = 956$ MSigDB pathway gene sets. For some pathways, the distance between upper and lower bound is too large to reliably estimate the log marginal likelihood; for most of the top 20 or so pathways, however, the difference is very small and hence the evidence is reliably evaluated.

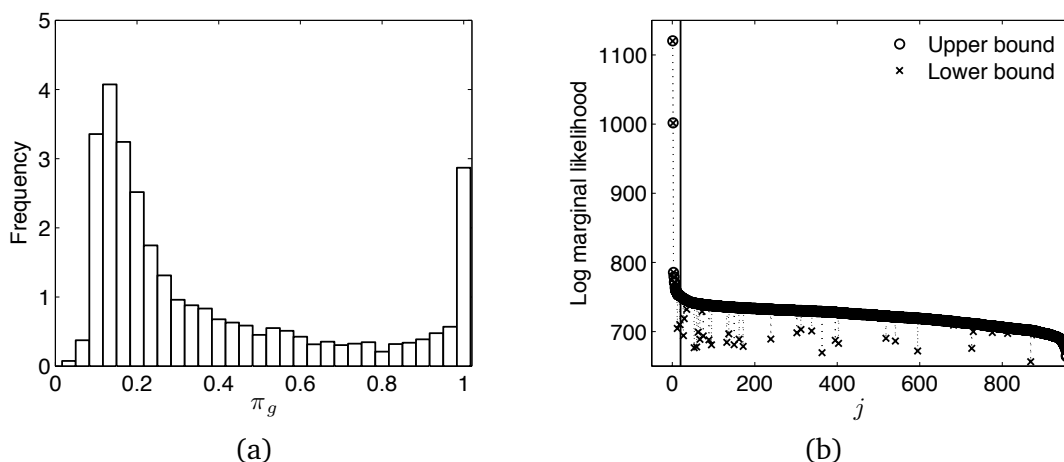


Figure 5: Breast cancer ER status study. (a) Histogram of association probabilities π_g . (b) Log marginal likelihood upper (\circ) and lower (\times) bounds for pathway gene lists $j = 1 : 956$ sorted in decreasing order; the line demarks the “top 20” pathways.

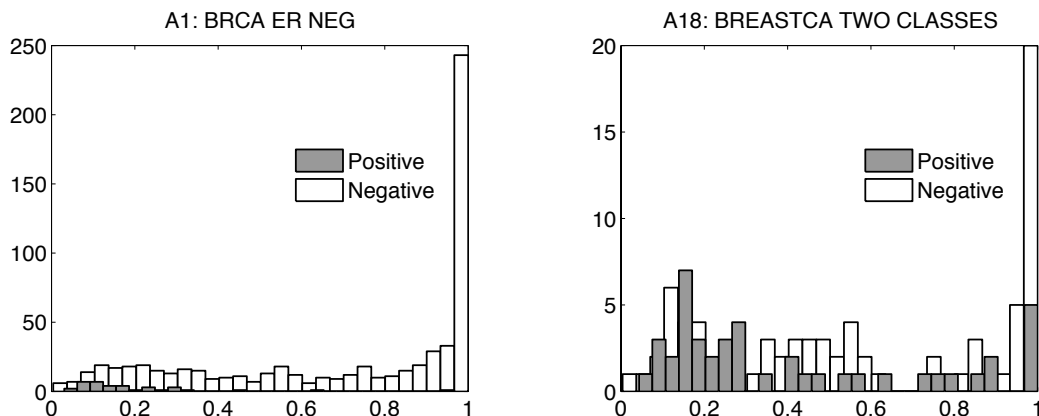


Figure 6: Breast cancer ER status study. Histograms of π_g for $g \in A_1$ and $g \in A_{18}$ of Table 1. Gray/white indicates genes whose expression levels are positively/negatively correlated with ER status, respectively.

RANK	PATHWAY	SIZE	Log(ML): UB	Log(ML): LB	UB-LB
1	BRCA ER NEG	692	1120.36	1120.01	0.35
2	BRCA ER POS	380	1001.61	1001.35	0.26
3	FLECHNER KIDNEY TRANSPLANT REJ/UP	81	785.59	785.51	0.08
4	LEE TCELLS2 UP	712	779.05	778.95	0.10
5	CARIES PULP UP	186	771.46	771.0	0.46
6	BRCA PROGNOSIS NEG	69	769.95	769.85	0.10
7	SERUM FIBROBLAST CELLCYCLE	83	763.65	763.51	0.13
8	CANCER UNDIFFERENTIATED META UP	65	760.71	760.57	0.14
9	VANTVEER BREAST OUTCOME/DOWN	58	758.1	757.98	0.12
10	FRASOR ER UP	29	757.88	757.85	0.03
11	CARIES PULP HIGH UP	83	757.51	757.32	0.20
12	IRITANI ADPROX VASC	146	755.68	704.48	51.20
13	UVB NHEK3 C7	50	753.38	753.35	0.03
14	LI FETAL VS WT KIDNEY DN	157	753.0	752.43	0.56
15	VANTVEER BREAST OUTCOME/UP	20	752.86	752.8	0.06
16	ZHAN MM CD138 PR VS REST	26	752.31	752.26	0.08
17	GREENBAUM E2A UP	25	752.31	752.23	0.05
18	BREASTCA TWO CLASSES	122	752.01	710.08	41.92
19	BRCA PROGNOSIS POS	26	751.88	751.85	0.04
20	MIDDLEAGE DN	13	751.5	751.47	0.04

Table 1: Summary of top ER-related pathways identified by PROPA

Table 1 summarises the top 20 pathway gene sets. The first two are breast tumor ER $-/+$ signatures defined by experimental microarray studies in [Van't Veer et al. \(2002\)](#), clearly validating the PROPA results. PROPA identifies several other pathway gene sets with defined links to breast tumor ER status. Patients with ER $-$ tumors generally have poorer prognoses than those with ER $+$ tumors, and there are several well-known risk-related signatures linked to this that involve intersecting gene sets ([Van't Veer et al., 2002](#); [Maynard et al., 1978](#)); these are well-represented among the highly scoring pathway gene sets, including A_6 , A_9 , A_{15} and A_{19} in Table 1. Further, ER $-$ breast tumors tend to be less well differentiated than ER $+$ s, consistent with the novel PROPA identification of the undifferentiated cancer signature A_8 . Figure 6 shows histograms of the π_g for $g \in A_1$ and $g \in A_{18}$. Genes in A_1 have expression associations with ER that are generally concordant, while those in A_{18} are heterogeneous. Unlike PROPA, existing annotation methods such as GSEA have difficulty dealing with the latter cases. Gene set A_{18} is clearly ER related, linked to a signature that discriminates between BRCA1 and BRCA2 forms of hereditary cancers ([Hedenfalk et al., 2001](#)). The relationships between ER and BRCA are complex and as yet poorly understood; the heterogeneity of the π_g in this case are symptomatic of this.

5.2 ErbB2 Pathway

ErbB2 is an epidermal growth factor receptor for which high levels of activity represents a substantial cancer risk factor. About 20-25% of breast cancers have over-expression of ErbB2, primarily due to gene amplification; this is the major cause of ErbB2 pathway deregulation in breast cancers ([Ménard et al., 2003](#); [Badache and Gonçalves, 2006](#)). Immunohistochemistry assays of protein levels measure ErbB2 status (+/-) on 146 of the primary breast tumor samples in ([Carvalho et al., 2008](#)) together with expression data. Analysis using Bayesian factor regression modelling method (BFRM) ([Wang et al., 2007](#)) generated posterior probabilities $\Pi = \pi_{1,p}$, as well as the sign of

association between gene expression and ErbB2 status for the set of $p = 8,764$ unique genes corresponding to Entrez gene IDs. The π_g are displayed in Figure 7(a) and show rather few genes are associated with the experimental factor pathway \mathcal{F} , here known to be the ErbB2 related. We curated the MSigDB gene lists to align with gene names based on the Entrez human gene database. Since the database does not include signatures explicitly linked to ErbB2, we curated two additional gene sets from the literature: first, a *molecular portrait* set of several genes in chromosome 17 linked to ErbB2 over-expression related to amplification (Perou et al., 2000; Sørlie et al., 2001); second, genes differentially expressed with-versus-without over-expression of the ErbB2 protein measured in data from tumors and cell lines, from Bertucci et al. (2004).

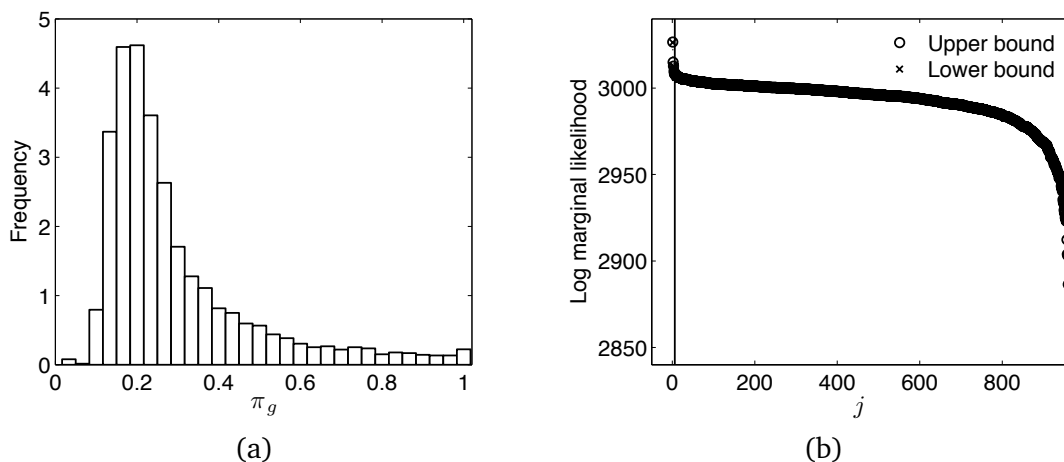


Figure 7: Breast cancer ErbB2 status study. (a) Histogram of association probabilities π_g . (b) Log marginal likelihood upper (\circ) and lower bounds (\times); pathways are sorted in order of decreasing upper bound and the vertical line indicates the top 6 pathways.

PROPA generated upper and pseudo-optimal lower bounds on log marginal likelihoods for each of the 958 gene sets appear in Figure 7(b). The decrease in marginal likelihoods notably diminishes after the first four or five pathways, suggesting stronger evidence of association between these few pathways and ErbB2. These pathways (Table 2) include the two ErbB2 signatures, ranked 1 and 4. These two curated gene sets are also identified by GSEA in the top “up-regulated” list but are ranked 4 and 6 by GSEA. Random set-based methods just fail to identify these two gene sets. The small numbers of genes in these sets limits GSEA and random set methods. In contrast, PROPA generally demonstrates good sensitivity and specificity when transcriptional evidence of phenotype-pathway association is relatively weak in the sense of small numbers of genes in the reference gene lists.

Table 3 gives information on the ErbB2 molecular portrait reference gene set \mathcal{A}_1 . This includes pathway membership inference via the π_g^* values and their corresponding log Bayes’ factors as well as the initial π_g . Six genes in the chromosomal regions 17q11-q12 and 17q21 have relatively high probabilities π_g of positive association with the experimental breast tumor ErbB2 factor pathway \mathcal{F} . The posterior membership probabilities of these genes confirm their membership in the molecular portrait biological pathway \mathcal{A}_1 . The other three genes with relatively low association probabilities are inferred by PROPA as false positive genes. Notably, gene MMP15 is located at 16q13-q21. It

RANK	PATHWAY	SIZE	Log(ML): UB	Log(ML): LB	UB-LB
1	ERBB2 overexpression cluster genes	9	3026.53	3026.29	0.24
2	HUMAN TISSUE KIDNEY	11	3014.87	3012.41	2.46
3	CROONQUIST IL6 STARVE UP	31	3012.64	3012.59	0.05
4	ERBB2 gene expression signature	24	3009.77	3009.73	0.04
5	HDAC1 COLON CUR16HRS DN	8	3008.42	3007.66	0.76
6	MMS HUMAN LYMPH LOW 4HRS DN	16	3007.84	3007.81	0.03

Table 2: Summary of top ErbB2-related pathways identified by PROPA

was included in the ErbB2 portrait gene set by a gene clustering analysis based on microarray data; we conclude that MMP15 should *not* be designated a member of the ErbB2 pathway. Several others genes not listed (G6PC, ERAL1, OMG, RPL19, CRKRS) are located in the regions 17q11-q12 and 17q21, and each has positive correlation with ErbB2 status. The Bayes' factors for pathway membership on these genes are greater than 34 dBs, indicating very strong if not decisive evidence for these genes being false negatives, i.e. they *are* members of the ErbB2 pathway.

g	SYMBOL	DESCRIPTION	GENE ID	CHR. LOC.	π_g	π_g^*	Log(BF)	CORR.
1	STARD3	START domain containing 3	10948	17q11-q12	0.99	1.00	10.07	+
2	GRB7	growth factor receptor-bound protein 7	2886	17q12	0.99	1.00	10.07	+
3	THRAP4	thyroid hormone receptor associated protein 4	9862	17q21.1	0.96	0.99	6.09	+
4	ERBB2	v-erb-b2 oncogene homolog 2	2064	17q11.2-q12	0.94	0.99	4.34	+
5	TRAF4	TNF receptor-associated factor 4	9618	17q11-q12	0.90	0.92	1.60	+
6	FLOT2	flotillin 2	2319	17q11-q12	0.88	0.72	0.12	+
7	PCGF2	polycomb group ring finger 2	7703	17q12	0.57	0.00	-16.19	+
8	MMP15	matrix metalloproteinase 15	4324	16q13-q21	0.34	0.00	-30.78	+
9	SMARCE1	SWI/SNF related regulator of chromatin	6605	17q21.2	0.21	0.00	-42.19	-

Table 3: Genes in the ErbB2 molecular portrait gene set

6 Theoretical and Algorithmic Details

6.1 PROPA Model Marginal Likelihood

Refer to the marginal likelihood function shown in (5). Integrating out $\beta_{1:p}$ and $\alpha_{0:1}$ results in

$$p(\Pi, z_{1:p}|A, \mathcal{F} = \mathcal{A}) = c(\Pi, z_{1:p}) \prod_{g=1}^p \left[\left(\frac{r_A}{\pi_g} \right)^{z_g} \left(\frac{1-r_A}{1-\pi_g} \right)^{1-z_g} \right]^{I(g \in A)} \left[\left(\frac{r_B}{\pi_g} \right)^{z_g} \left(\frac{1-r_B}{1-\pi_g} \right)^{1-z_g} \right]^{I(g \notin A)}$$

where

$$c(\Pi, z_{1:p}) = \gamma_{1:p}(\nu_1) \gamma_{1:p}(\nu_0) \lambda_1^{-\nu_1} \lambda_0^{-\nu_0} (1 - \Psi(1; \nu_0, \lambda_0)) (1 - \Psi(1; \nu_1, \lambda_1))$$

with $\nu_1 = \sum_{g=1}^p z_g$, $\nu_0 = \sum_{g=1}^p (1 - z_g)$, $\lambda_1 = -\sum_{g=1}^p (z_g \log \pi_g)$, $\lambda_0 = -\sum_{g=1}^p (1 - z_g) \log(1 - \pi_g)$, and where Ψ are gamma cdfs. Then the marginal likelihood is

$$p(\Pi|A, \mathcal{F} = \mathcal{A}) = \sum_{z_{1:p}} p(\Pi, z_{1:p}|A, \mathcal{F} = \mathcal{A}),$$

where each summand can be evaluated.

The alternative expression derived by summation over the $z_{1:p}$ and integration over $\beta_{1:p}$ conditional on $\alpha_{0:1}$ is

$$p(\Pi|A, \mathcal{F} = \mathcal{A}) = \int_{\alpha_{0:1}} p(\Pi|\alpha_{0:1}, A, \mathcal{F} = \mathcal{A})p(\alpha_{0:1})d\alpha_{0:1}$$

where

$$p(\Pi|\alpha_{0:1}, A, \mathcal{F} = \mathcal{A}) = \prod_{g=1}^p p(\pi_g|\alpha_{0:1}, A, \mathcal{F} = \mathcal{A}).$$

The terms here are

$$p(\pi_g|\alpha_{0:1}, A, \mathcal{F} = \mathcal{A}) = \begin{cases} r_A f_1(\pi_g|\alpha_1) + (1 - r_A) f_0(\pi_g|\alpha_0), & g \in A, \\ r_B f_1(\pi_g|\alpha_1) + (1 - r_B) f_0(\pi_g|\alpha_0), & g \notin A. \end{cases}$$

6.2 Marginal Likelihood Upper and Lower Bound Theory

For model M and data D , marginal likelihood in a general form is

$$p(D|M) = \int_{\Theta} p(\theta, D|M)d\theta,$$

with $\theta = \{\theta_1, \dots, \theta_K\} \in \Theta$ representing model parameters. In PROPA, data D is Π , model M is specified with $A, \mathcal{F} = \mathcal{A}$, and parameters are $z_{1:p}$ in the reduced form.

For any density function $q(\theta; \gamma)$ parameterized by $\gamma = \{\gamma_1, \dots, \gamma_J\} \in \Gamma$ and with the same support as the posterior for θ , $p(\theta|D, M)$, Jensen's inequality

$$\log p(D|M) \geq \int_{\Theta} q(\theta; \gamma) \log \frac{p(\theta, D|M)}{q(\theta; \gamma)} d\theta$$

provides a lower bound for log marginal likelihood. Maximization of this lower bound corresponds to minimization of the Kullback-Leibler divergence of model parameter posterior density $p(\theta|D, M)$ from the variational density $q(\theta; \gamma)$.

If $q(\theta; \gamma) = p(\theta|D, M)$, we can rewrite

$$\log p(D|M) = \int_{\Theta} p(\theta|D, M) \log p(D, \theta|M) d\theta - \int_{\Theta} p(\theta|D, M) \log p(\theta|D, M) d\theta.$$

Combining this expression with Gibbs' inequality

$$-\int_{\Theta} p(\theta|D, M) \log p(\theta|D, M) d\theta \leq -\int_{\Theta} p(\theta|D, M) \log q(\theta; \gamma) d\theta$$

leads to

$$\log p(D|M) \leq \int_{\Theta} p(\theta|D, M) \log \frac{p(\theta, D|M)}{q(\theta; \gamma)} d\theta,$$

which provides an upper bound on the log marginal likelihood.

6.3 Monte Carlo Variational Algorithm

The Monte Carlo variational method using stochastic approximation to generate estimates of the lower bound of marginal likelihoods in the PROPA model has the key steps below. The resulting algorithm is easy to implement, and its convergence can be guaranteed as described, in more general contexts, in [Ji et al. \(2009\)](#). In essentials here, it is first easy to see that the global, lower bound optimizing value of $\gamma_{1:p}$ satisfies $f_g(\gamma_{1:p}) = 0$, $g = 1, \dots, p$ for the function defined in equation (12). The method is based on the observations that:

1. $f_g(\gamma_{1:p})$, $g = 1, \dots, p$ is an expectation with respect to $z_{1:p} \sim q(z_{1:p}|\gamma_{1:p})$. Monte Carlo averaging can efficiently estimate this expectation at any value of $\gamma_{1:p}$; in our model this simply involves generating repeat Monte Carlo sample of p independent Bernoulli variates; and
2. the resulting Monte Carlo estimate of $f_g(\gamma_{1:p})$, $g = 1, \dots, p$ can be used to derive updated values of $\gamma_{1:p}$ using stochastic approximation ([Robbins and Monro, 1951](#)).

The algorithmic implementation of these ideas is as follows:

- Begin at iterate $t = 0$ with values of $\gamma_{1:p} = \bar{z}_{1:p}$, the approximate posterior means from the MCMC posterior sample.
- At any later iterate $t \geq 1$ based on current values $\gamma_{1:p}^{(t-1)}$, generate a random sample of $z_{1:p}$ from $q(z_{1:p}|\gamma_{1:p}^{(t-1)})$;
- Compute the implied Monte Carlo estimate of $f_g^{(t-1)}(\gamma_{1:p}^{(t-1)})$, $g = 1, \dots, p$ replacing the sum in equation (12) with the Monte Carlo average over the samples of $z_{1:p}$;
- Update via the stochastic approximation form

$$\gamma_{1:p}^{(t)} = \gamma_{1:p}^{(t-1)} + s^{(t)} f_{1:p}^{(t-1)}(\gamma_{1:p}^{(t-1)})$$

where $s^{(t)}$ is a chosen sequence of weights whose sum over $t \geq 1$ diverges but for which the sum of squared values is finite, e.g., $s^{(t)} = c/t$ for some constant $c > 0$.

This is an example of a general algorithm for which it can be shown ([Robbins and Monro, 1951](#); [Ji et al., 2009](#)) that $\gamma_{1:p}^{(t)}$ converges with probability one to $\gamma_{1:p}^*$ satisfying $f_g(\gamma_{1:p}^*) = 0$, $g = 1, \dots, p$, providing an iterative approximation of the lower bound optimizing value. Terminate the iterates

at some finite step assuming $\gamma_{1:p}^* \approx \gamma_{1:p}^{(t)}$, draw a final, large Monte Carlo sample $z_{1:p}^h$, ($i = 1, \dots, H$), from $q(z_{1:p} | \gamma_{1:p}^*)$, and then evaluate the Monte Carlo estimate of lower bound

$$\bar{L} = H^{-1} \sum_{h=1}^I \{\log p(\Pi, z_{1:p}^h | A, \mathcal{F} = \mathcal{A}) - \log q(z_{1:p}^h | \gamma_{1:p}^*)\}.$$

This is a consistent estimate of the optimal lower bound assuming the stochastic approximation estimate has converged (Ji et al., 2009).

7 Summary Comments

PROPA is a formal, fully Bayesian framework for matching experimental signatures of structure or outcomes in gene expression – represented in terms of weighted gene lists – to multiple biological pathway gene sets from curated databases. In the setting here, gene weights are explicit gene-factor phenotype association probabilities. The analysis delivers estimated marginal likelihood values over pathways for each factor phenotype, allowing quantitative assessment and ranking of pathways putatively linked to the phenotype as well as gene-specific posterior membership probabilities. We develop a novel Monte Carlo variational method for estimating marginal likelihoods for model comparisons, and evaluate and illustrate the model with simulated and cancer genomic data.

For the future, there is a key need for improved quality of biological pathway databases, an area that PROPA can contribute to as we have exemplified. Open methodological issues include specification of model priors across pathways, and the use of alternative, multiple numerical summaries of the relationships between genes and experimental phenotypes. Advances in these areas will enhance the contributions of Bayesian reasoning in biological pathway studies. Software for practitioners is also key; current PROPA code, with examples, is freely available at the url below.

Acknowledgments: We are grateful to Ashley Chi, Joe Lucas and Chunlin Ji for discussions and input, and to Quanli Wang for computational contributions. This work was partly supported by NSF (DMS-0102227, DMS-0342172) and NIH (U54-CA-112952-01). Any opinions, findings and conclusions or recommendations expressed in this work are those of the authors and do not necessarily reflect the views of the NSF or NIH.

Software: <http://www.stat.duke.edu/research/software/west/propa>

References

- Badache, A. and A. Gonçalves (2006). The ErbB2 signaling network as a target for breast cancer therapy. *Journal of Mammary Gland Biology and Neoplasia* 11, 13–25.
- Bertucci, F., N. Borie, C. Ginestier, A. Groulet, E. Charafe-Jauffret, J. Adélaïde, J. Geneix, L. Bachelart, P. Finetti, A. Koki, F. Hermitte, J. Hassoun, S. Debono, P. Viens, V. Fert, J. Jacquemier, and D. Birnbaum (2004). Identification and validation of an ERBB2 gene expression signature in breast cancers. *Oncogene* 23(14), 2564–2575.
- Bild, A. H., G. Yao, J. T. Chang, Q. Wang, A. Potti, D. Chasse, M. Joshi, D. Harpole, J. M. Lancaster, A. Berchuck, J. A. Olson Jr. , J. R. Marks, H. K. Dressman, M. West, and J. R. Nevins (2006, January). Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* 439, 353–357.
- Broad Institute (2007). Molecular signatures data base. <http://www.broad.mit.edu/gsea/msigdb/>.
- Carvalho, C. M., J. E. Lucas, Q. Wang, J. Chang, J. R. Nevins, and M. West (2008). High-dimensional sparse factor modelling: Applications in gene expression genomics. *Journal of the American Statistical Association* 103(484), 1438–1456.
- Celeux, G. and J. Diebolt (1992). A stochastic approximation type EM algorithm for the mixture problem. *Stochastics and Stochastics Reports* 41, 127–146.
- Chan, K. S. and J. Ledolter (1995). Monte Carlo EM estimation for time series models involving observations. *Journal of the American Statistical Association* 90, 242–252.
- Chang, J., C. M. Carvalho, S. Mori, A. Bild, Q. Wang, M. West, and J. R. Nevins (2009). Decomposing cellular signaling pathways into functional units: A genomic strategy. *Molecular Cell* 34, 104–114.
- Chen, J. L., J. E. Lucas, T. Schroeder, S. Mori, J. Nevins, M. Dewhirst, M. West, and J. T. Chi (2008). The genomic analysis of lactic acidosis and acidosis response in human cancers. *PLoS Genetics* 4(12), e1000293.
- Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* 90, 773–795.
- Corduneanu, A. and C. M. Bishop (2001). Variational Bayesian model selection for mixture distributions. In T. Jaakkola and T. Richardson (Eds.), *Artificial Intelligence and Statistics*, pp. 27–34. Morgan Kaufmann.
- Delyon, B., M. Lavielle, and E. Moulines (1999). Convergence of a stochastic approximation version of the EM algorithm. *Annals of Statistics* 27(1), 94–128.

- Deroo, B. J. and K. S. Korach (2006, March). Estrogen receptors and human disease. *J. Clin. Invest.* 116(3), 561–570.
- Hedenfalk, I., D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, O. P. Kallioniemi, B. Wilfond, A. Borg, J. Trent, M. Raffeld, Z. Yakhini, A. Ben-Dor, E. Dougherty, J. Kononen, L. Bubendorf, W. Fehrle, S. Pittaluga, S. Gruvberger, N. Loman, O. Johannsson, H. Olsson, and G. Sauter (2001). Gene-expression profiles in hereditary breast cancer. *New England Journal of Medicine* 344, 539–48.
- Huang, E., S. Chen, H. Dressman, J. Pittman, M. H. Tsou, C. F. Horng, A. Bild, E. S. Iversen, M. Liao, C. M. Chen, M. West, J. R. Nevins, and A. T. Huang (2003). Gene expression predictors of breast cancer outcomes. *The Lancet* 361, 1590–1596.
- Huang, E., S. Ishida, J. Pittman, H. Dressman, A. Bild, M. D’Amico, R. Pestell, M. West, and J. R. Nevins (2003). Gene expression phenotypic models that predict the activity of oncogenic pathways. *Nature Genetics* 34, 226–230.
- Ji, C., H. Shen, and M. West (2009). Monte carlo variational approximation of marginal likelihoods. Technical report, Department of Statistical Science, Duke University.
- Jordan, M., Z. Ghahramani, T. Jaakkola, and K. Saul (1999). An introduction to variational methods for graphical models. *Machine Learning* 37(2), 183–233.
- Lucas, J., C. Carvalho, Q. Wang, A. Bild, J. R. Nevins, and M. West (2006). Sparse statistical modelling in gene expression genomics. In P. M. K. A. Do and M. Vannucci (Eds.), *Bayesian Inference for Gene Expression and Proteomics*, pp. 155–176. Cambridge University Press.
- Lucas, J. E., C. M. Carvalho, J.-T. A. Chi, and M. West (2009). Cross-study projections of genomic biomarkers: An evaluation in cancer genomics. *PLoS One* 4(2), e4523.
- Lucas, J. E., C. M. Carvalho, and M. West (2009). A Bayesian analysis strategy for cross-study translation of gene expression biomarkers. *Statistical Applications in Genetics and Molecular Biology* 8(1), Article 11.
- Maynard, P. V., C. J. Davies, R. W. Blamey, C. W. Elston, J. Johnson, and K. Griffiths (1978). Relationship between oestrogen-receptor content and histological grade in human primary breast tumours. *Brit. J. Cancer* 38(6), 745–748.
- McGrory, C. A. and D. M. Titterington (2007). Variational approximations in Bayesian model selection for finite mixture distributions. *Computational Statistics & Data Analysis* 51(11), 5352–5367.
- Ménard, S., S. M. Pupa, M. Campiglio, and E. Tagliabue (2003). Biologic and therapeutic role of her2 in cancer. *Oncogene* 22, 6570–6578.
- Merl, D., J. L.-Y. Chen, J. T. Chi, and M. West (2009). Integrative analysis of cancer gene expression studies using Bayesian latent factor modelling. *Annals of Applied Statistics* -, -.

- Merl, D., J. E. Lucas, J. R. Nevins, H. Shen, and M. West (2009). Trans-study projection of genomic biomarkers using sparse factor regression models. In A. O'Hagan and M. West (Eds.), *The Handbook of Applied Bayesian Analysis*. Oxford University Press.
- Moggs, J. G. and G. Orphanides (2001). Estrogen receptors: Orchestrators of pleiotropic cellular responses. *EMBO Rep.* 2(9), 775–781.
- Newton, M. A., F. A. Quintana, J. A. den Boon, S. Sengupta, and P. Ahlquist (2007). Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *Annals of Applied Statistics* 1(1), 85–106.
- Newton, M. A. and A. E. Raftery (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society, Series B* 56, 3–48.
- Perou, C. M., T. Sorlie, M. B. Eisen, M. van deRijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lonning, A. Borresen-Dale, P. O. Brown, and D. Botstein (2000). Molecular portraits of human breast tumours. *Nature* 406(6797), 747–752.
- Rich, J. N., C. Hans, B. Jones, E. S. Iversen, R. E. McLendon, B. K. Rasheed, A. Dobra, H. K. Dressman, D. D. Bigner, J. R. Nevins, and M. West (2005). Gene expression profiling and genetic markers in glioblastoma survival. *Cancer Research* 65, 4051–4058.
- Robbins, H. and S. Monro (1951). A stochastic approximation method. *Annals of Mathematical Statistics* 22, 400–407.
- Seo, D. M., P. J. Goldschmidt-Clermont, and M. West (2007). Of mice and men: Sparse statistical modelling in cardiovascular genomics. *Annals of Applied Statistics* 1(1), 152–178.
- Sørli, T., C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. Lønning, and A. Børresen-Dale (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. , USA* 98(19), 10869–10874.
- Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lauder, and J. P. Mesirov (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* 102(43), 15545–15550.
- Van't Veer, L. J., H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerckhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415, 530–536.
- Wang, Q., C. M. Carvalho, J. E. Lucas, and M. West (2007). BFRM: Bayesian factor regression modelling. *Bulletin of the International Society for Bayesian Analysis* 14, 4–5.

- Wang, Q., C. M. Carvalho, J. E. Lucas, and M. West (2007-present). BFRM: Software for sparse Bayesian factor regression modelling. <http://www.stat.duke.edu/research/software/west/bfrm/index.html>.
- West, M. (2003). Bayesian factor regression models in the “large p, small n” paradigm. In J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West (Eds.), *Bayesian Statistics 7*, pp. 723–732. Oxford University Press.
- West, M., C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. R. Marks, and J. R. Nevins (2001). Predicting the clinical status of human breast cancer utilizing gene expression profiles. *Proceedings of the National Academy of Sciences* 98, 11462–11467.