

Bayesian Modeling for Biological Pathway Annotation of Genomic Signatures

Archival Version:

Longer Discussions, Additional Examples & Case Study Details

Haige Shen & Mike West

Department of Statistical Science
Duke University
Durham, NC 27708-0251

email: {*haige,mw*}@stat.duke.edu

April 28, 2008

Haige Shen is Senior Biostatistician at Novartis Research, N.J., and Mike West is The Arts & Sciences Professor of Statistical Science in the Department of Statistical Science at Duke University.

Abstract

We present Bayesian models and computational methods for the problem of matching predictions from molecular studies with known biological pathway databases - the problem of *pathway annotation* of summary results of an experiment or observational study. In areas such as cancer genomics, linking quantified, experimentally defined gene expression signatures with known biological pathway gene sets is essential to improving the understanding of the complexity of molecular pathways related to outcome. Our probabilistic pathway annotation (PROPA) analysis involves new models for formal assessment and rankings of pathways putatively linked to an experimental or observational phenotype, integrates qualitative biological information into the analysis, and generates coherent inferences on uncertainties about gene pathway membership that can inform the revision of pathway databases.

Our analysis relies on simulation-based computation in high-dimensional models, and introduces a novel extension of variational methods for computation of model evidence, or marginal likelihood functions, that are central to the comparison of multiple biological pathways. Examples highlight the methodology using both simulated and real data, and we develop detailed cases studies in breast cancer genomics involving hormonal pathways and pathway activities underlying cellular responses to lactic acidosis in breast cancer. The second study demonstrates the application of the method in decomposing the complexity of gene expression-based predictions about interacting biological pathway activation from both experimental (*in vitro*) and observational (*in vivo*) human cancer data.

KEYWORDS: biological pathway analysis, cancer genomics, factor regression models, gene expression signatures, gene set enrichment, marginal likelihood computation, Monte Carlo variational approximation, sparse factor analysis

1 Introduction

Experimental and observational studies in high-throughput genomics often generate multiple gene expression *signatures*, each signature being a list 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., 2003c,b; Bild et al., 2006; Chen et al., 2007), while evaluation of gene expression related to a specific clinical outcome or state may generate a signature as a biomarker of the outcome in disease studies (West et al., 2001; Huang et al., 2002, 2003a; Pittman et al., 2004; Seo et al., 2004; Rich et al., 2005; Seo et al., 2007). Interpretation and, often, follow-on biological studies rely on the comparison of such signatures with multiple, annotated biological pathway databases that contain lists of putatively pathway-specific genes based on cumulated biological research. A core challenge is then to assess the candidate signature gene sets and numerical summaries against these databases to suggest potential pathway interpretations and connections. Our focus here is a formal, novel statistical modeling approach to this problem.

The first statistical approach, and general identification of this problem area, led to the method of *gene set enrichment analysis* (GSEA) (Subramanian et al., 2005) and has generated some deeper statistical approaches more recently (Newton et al., 2007). GSEA aims to measure aggregate association between a full list of genes ranked by their association with an outcome - also referred to as a *phenotype* - and a set of genes in a predefined pathway gene set. The underlying idea is to assess whether or not the pathway gene set is enriched with genes that score highly in association with the experimental outcome, perhaps with a directional component that looks separately at genes positively versus negatively associated. GSEA was path-breaking and is now quite widely used. In our applied work, we are interested in broader questions and also in formal statistical inference on gene-pathway membership, and this has motivated a formal probabilistic framework that extends the basic thinking into a broader statistical approach. The resulting *probabilistic pathway annotation* (PROPA) methodology then also addresses a number of issues GSEA methods were not designed for, including the abilities to: (a) deliver formal probabilistic assessments of phenotype-pathway concordance, in terms of marginal likelihoods and posterior probabilities; (b) formally assess concordance of experimental results with several or many biological pathways simultaneously and in comparison with each other; (c) recognize that experimental inferences *and* established biological pathway databases are error prone, and allow for the identification and correction of errors of both kinds within the analysis; (d) utilize a range of direct numerical measures of association between genes and an experimental outcome as inputs; and (e) provide a more general framework that can be customized to apply to the outputs of gene expression, or other genomic studies of many forms. In addition to within-analysis robustness, item (c) here also leads to an

ability to suggest refinements to the pathway gene lists in established biological databases.

Our focus here is on applications in cancer genomics. While the primary aim of this paper is to highlight the area and applications, the statistical methodology has modeling and computational novelty. A core ingredient of biological pathway assessment is the evaluation of marginal likelihoods in Bayesian models fitted using MCMC methods. Marginal likelihood computations are common and often hard problems (e.g., Raftery et al. (2007) for a recent approach with discussion and many references to other approaches), especially in cases, such as here, of high-dimensional parameter spaces. Our favored approach involves a novel extension of variational methods that have been applied in other problems of marginal likelihood computation (e.g., Jordan et al. (1999); Corduneanu and Bishop (2001); McGrory and Titterton (2007)); in addressing this problem in our specific applied context, we have introduced an extension of existing variational methods that will apply in many other model contexts.

In Section 2 we lay out the basic context, define notation and the basic modeling approach. Section 3 describes the overall MCMC strategy for posterior simulation in an analysis focused on a single biological pathway, and the developments of computational methods for marginal likelihood computation to aid in comparisons of multiple biological pathways. This includes the innovations in variational methodology that are further detailed in the appendix. Section 4 explores examples to highlight the specification and use of the model. The first cancer genomics application in Section 5 concerns a detailed study of two well-known hormonal pathways in breast cancer. The second application in Section 6 concerns novel experimental data arising in studies of micro-environmental influences on gene expression from *in vitro* experiments, and connects these experimental findings to *in vivo* observational breast cancer data. Among other things, this case study demonstrates an overall strategy for *in vitro* to *in vivo* projection of gene expression patterns within which PROPA analysis plays key roles. We conclude in Section 7 with some summary comments.

2 Context and Models

2.1 Notation and Framework

A basic pathway annotation problem arising in DNA microarray experiments is as follows. An experimental study investigates the changes in gene expression on p genes in cultured cells due to an experimental *factor* - the factor may be an intervention to induce changes in the environment of cells under study, drug application, genetic modification, etc. We are interested in (a) which genes are related to this factor, and (b) how does this factor relate to known, published gene lists representing annotation of biological pathways? Statisti-

cal analysis of the experimental data generates numerical measures of association of the genes with the experimental factor that form the inputs to annotation analysis. We define terminology and notation as follows:

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

We use the Molecular Signatures database (MSigDB, Broad Institute) with currently thousands of reference gene lists A_j representing signalling and regulatory pathways, as well as simply published lists related to specific biological pathways; this provides information sets $A_j, j = 1 : m$. In our examples, $m \approx 1000$. These reference gene sets 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 $\mathcal{A}_j, j = 1 : m$. This assessment uses the database information $A_{1:m} \equiv \{A_j, j = 1 : m\}$. From our Bayesian perspectives, we do this with models that compute the $j = 1 : m$ posterior probabilities

$$Pr(\mathcal{F} = \mathcal{A}_j | \Pi, A_{1:m}) \propto Pr(\mathcal{F} = \mathcal{A}_j | A_{1:m}) p(\Pi | A_{1:m}, \mathcal{F} = \mathcal{A}_j). \quad (1)$$

We focus here on the likelihood terms $p(\Pi | A_{1:m}, \mathcal{F} = \mathcal{A}_j)$ as j moves across all the pathways, as this is the overall measure from the experimental predictions Π that feeds into pathway assessment, and can be applied whatever the chosen values of the priors $Pr(\mathcal{F} = \mathcal{A}_j | A_{1:m})$.

The numerical measures Π of association between genes and \mathcal{F} in the experimental context may be essentially any measures, such as test statistics or other summaries of a statistical analysis of the experimental data. Our example measures here are estimated probabilities of differential expression, or similar. In simple designed experiments, π_g will be an estimated posterior probability of a significant change in expression of gene g related to an experimental factor. In observational studies, as in our second case study on human breast cancer data, the π_g may be posterior probabilities of non-zero regression coefficients or loadings on latent factors in a sparse factor model of gene expression data (West, 2003; Lucas et al., 2006; Seo et al., 2007; Carvalho et al., 2008). In such contexts, typically very many of the π_g will be very small; those genes associated with \mathcal{F} will be larger.

We may include additional numerical measures, extending each π_g to a vector, but for the current work restrict to scalar values and use probabilities of expression changes with \mathcal{F} from analysis of experimental or observational data. Hence $\pi_g \in [0, 1]$ with larger values indicating stronger association with \mathcal{F} .

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 $p(\Pi|A_{1:m}, \mathcal{F} = \mathcal{A}_j)$. This is defined via the following two components.

2.2.1 Model for data Π assuming known pathway membership of genes

Assuming that \mathcal{F} is indeed \mathcal{A} , then observed values of π_g will be expected to be higher for genes $g \in \mathcal{A}$ than for gene $g \notin \mathcal{A}$. This suggests models that represent expected behavior of the data Π 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. The natural parametric choice is beta densities, and we consider $f_1(\pi) \equiv f_1(\pi|\alpha_1) = \text{Be}(\alpha_1, 1)$ and $f_0(\pi) \equiv f_0(\pi|\alpha_0) = \text{Be}(1, \alpha_0)$ with $\alpha_0, \alpha_1 > 1$ (Figure 1(a)). Such a specification is certainly consistent with the forms of histograms of π_g values generated in studies using sparse factor regression models in several areas (Lucas et al., 2006; Seo et al., 2007; Carvalho et al., 2008); see Figure 1(b).

All analyses here use independent reference priors for the α parameters, viz

$$p(\alpha_r) \propto \alpha_r^{-1}, \quad 1 < \alpha_r < a, \quad r = 0 : 1, \quad (3)$$

where a is a large, specified upper limit.

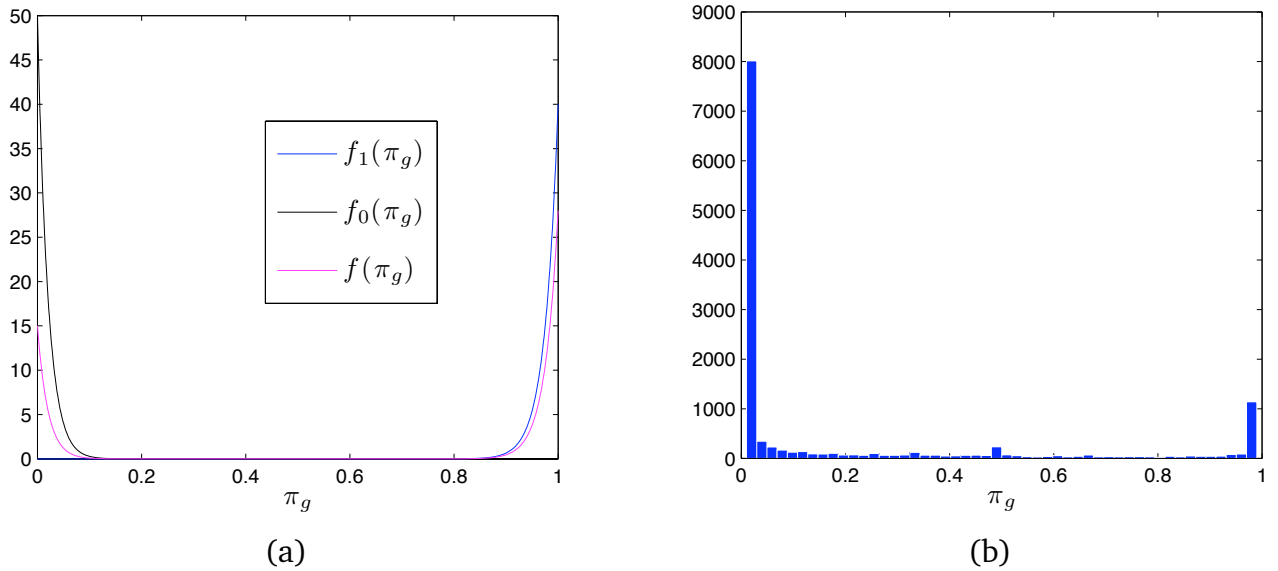


Figure 1: (a) Density function $f(\pi_g)$ modeled as a mixture of $f_1(\pi|\alpha_1) = \text{Be}(\alpha_1, 1)$ and $f_0(\pi|\alpha_0) = \text{Be}(1, \alpha_0)$. (b) Histogram of association probabilities Π generated in a real expression data analysis.

2.2.2 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 $z_{1:p}$ such that, when $\mathcal{F} = \mathcal{A}$, $z_g = 1$ if $g \in \mathcal{A}$, and 0 otherwise. Call z_g the unknown *pathway membership indicator* of gene g . We need probabilities governing over the z_g and 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 specification requires priors for the β_g , which we take as

$$\begin{aligned} (\beta_g|g \in A, \mathcal{F} = \mathcal{A}) &\sim \text{Be}(\phi_A r_A, \phi_A(1 - r_A)), \\ (\beta_g|g \notin A, \mathcal{F} = \mathcal{A}) &\sim \text{Be}(\phi_B r_B, \phi_B(1 - r_B)), \end{aligned} \quad (4)$$

with specified means $r_A, r_B \in (0, 1)$ and $\phi_A, \phi_B > 0$. The values of r_A and r_B relate to the true positive rate and false negative rate: marginalising 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 $g \in A$. We expect r_A to be relatively large. The value of r_B depends on an assessment

of how many genes not in A are likely to be associated with \mathcal{F} . The number of genes in A , typically tens to a few hundreds, is small compared to the full gene list \mathcal{G} , and a reasonable value of r_B should be approximately equal to the ratio of the number of reference set genes to the total number of genes, for example, 0.005. The ϕ_A and ϕ_B constrain the variation range of the prior for the β_g .

Annotated biological databases are incomplete and error prone. The model provides opportunity to investigate this by estimating posterior pathway membership probabilities for each gene g , namely

$$\pi_g^* = Pr(g \in \mathcal{A} | \Pi, A, \mathcal{F} = \mathcal{A}), \quad (5)$$

with respect to the pathway of interest \mathcal{A} . This is accomplished from the MCMC analysis of the posterior distribution for all model quantities $z_{1:p}$, $\beta_{1:p}$ and α_0, α_1 .

2.3 Marginal Likelihood for Pathway Assessment

The full model likelihood $p(\Pi | A, \mathcal{F} = \mathcal{A})$ can be expressed as

$$p(\Pi | A, \mathcal{F} = \mathcal{A}) = \int_{\alpha_{0:1}} \int_{\beta_{1:p}} \sum_{z_{1:p}} \mathcal{L}(\alpha_{0:1}, z_{1:p}) \prod_{g=1}^p p(z_g | \beta_g) \prod_{g=1}^p p(\beta_g | A, \mathcal{F} = \mathcal{A}) p(\alpha_{0:1}) d\beta_{1:p} d\alpha_{0:1} \quad (6)$$

with

$$\mathcal{L}(\alpha_{0:1}, z_{1:p}) = \prod_{g=1}^p f_1(\pi_g | \alpha_1)^{z_g} f_0(\pi_g | \alpha_0)^{1-z_g}. \quad (7)$$

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 Appendix 1 that gives the form as

$$p(\Pi | A, \mathcal{F} = \mathcal{A}) = \sum_{z_{1:p}} p(\Pi, z_{1:p} | A, \mathcal{F} = \mathcal{A}) \quad (8)$$

where the quantity $p(\Pi, z_{1:p} | A, \mathcal{F} = \mathcal{A})$ can be evaluated at any chosen $z_{1:p}$. This sum is a marginal likelihood computation, and a difficult numerical problem that we discuss and solve below in Section 3.2.

Another reduced form that is theoretically attractive but practically of little value is derived by summation over the $z_{1:p}$ and integrations over $\beta_{1:p}$ conditional on $\alpha_{0:1}$, resulting in

$$p(\Pi | A, \mathcal{F} = \mathcal{A}) = \int_{\alpha_{0:1}} p(\Pi | \alpha_{0:1}, A, \mathcal{F} = \mathcal{A}) p(\alpha_{0:1}) d\alpha_{0:1}. \quad (9)$$

The integrand here can be evaluated, but only practicably when p is small (Appendix 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

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

subject to $1 < \alpha_r < a$, for $r = 0 : 1$. In practice, we take a large enough so that only the lower bound of 1 is used.

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 conditional pathway membership probabilities

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

Note that the posterior pathway membership probability π_g^* is just the posterior mean of ρ_g .

We have implemented efficient code for this MCMC and experienced generally fast mixing, and rapid, clean convergence across many examples. Evidently, there is rather weak dependence in the posterior among the z_g , induced by lack of knowledge of the $\alpha_{0:1}$, so that swift and clean 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 (6), and sets of such quantities $p(\Pi \mid A_{1:m}, \mathcal{F} = \mathcal{A}_j)$ in the practical context of assessing evidence for and against $\mathcal{F} = \mathcal{A}_j$ for a number or many pathways $j = 1 : m$.

In very small, unrealistic synthetic examples we can use quadrature methods to generate comparisons to other numerical approximations. We do this in the simulated example in Section 4, simply applying direct quadrature to the two-dimensional integral form of equation (9). Even with p very small, this method is limited since it requires evaluation of integrands on the density scale, rather than the log density scale, and quickly runs into floating-point overflow problem. Our problems have p in the tens of thousands and so quadrature is simply not relevant for real applications.

The reduced version of equation (8) has a closed form but involves summing over all 2^p values of $z_{1:p}$; with p in the tens of thousands in genome-wide expression data, numerical approximation is required. Since we use MCMC for simulation of the posterior defined by the summands, then methods of marginal likelihood computation using MCMC outputs are attractive. Having experimented with multiple such methods, all with their own pros and cons (Newton and Raftery, 1994; Chib, 1995; Meng and Wong, 1996) we adapted recent mean-field variational method (VM) approaches (Jaakkola and Jordan, 1997; Jordan et al., 1999; Corduneanu and Bishop, 2001; Beal, 2003; McGrory and Titterton, 2007) to this problem. Our studies have confirmed 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. In exploring this we extended the VM theory and methodology, quite generally, to also generate an *upper bound*, so that the two bounds together 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 $q(z_{1:p})$,

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

and

$$U(q) = \sum_{z_{1:p}} p(z_{1:p}|\Pi, A, \mathcal{F} = \mathcal{A}) \log[p(\Pi, z_{1:p}|A, \mathcal{F} = \mathcal{A})/q(z_{1:p})]. \quad (11)$$

The VM concept is to choose parametric densities $q_L(z_{1:p})$ and $q_U(z_{1:p})$ to optimise these bounds on $\log(p(\Pi|A, \mathcal{F} = \mathcal{A}))$. If each depends on a free parameter that can be varied to optimise the bounds, the problem is then the computational problem of finding the optimizing values of those *variational* parameters. The closer the variational density is to the actual posterior $p(z_{1:p}|\Pi, A, \mathcal{F} = \mathcal{A})$, the better will be the bound.

Mean-field VM methods adopt variational densities that are in factorized form. This is very natural here since the z_g are only weakly dependent under the posterior, and so approximation of the joint posterior with a product of estimated marginal posteriors, or similar, is likely to generate good bounds. In particular, this suggests that we choose each of q_L, q_U to be of the form

$$q(z_{1:p}|\gamma_{1:p}) = \prod_{g=1}^p \gamma_g^{z_g} (1 - \gamma_g)^{1-z_g} \quad (12)$$

where $\gamma_{1:p}$ is the vector of free variational parameters, and we will choose different values $\gamma_{1:p}$ for the upper and lower bounds. The computations of optimizing variational parameters is necessarily iterative, and good starting values are simply the posterior means of the z_g from the MCMC output.

3.3 Marginal Likelihood Computation: A Monte Carlo Variational Method

The implementation of the above ideas represents a specific case of a new VM approach that we refer to as Monte Carlo variational approximation (MCVA). The key innovations are to simultaneously compute upper and lower bounds using iterative methods reliant on the MCMC analysis already performed; details appear in Shen et al. (2008); essential results for the PROPA model here are noted.

3.3.1 Upper bound optimization

With q_U of the form in equation (12), it is trivially seen that the global minimum value of the upper bound in equation (11) 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 $\{z_{1:p}^i, i = 1 : I\}$ 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.

3.3.2 Lower bound optimization

Optimizing the lower bound is more of a challenge. The existing mean-field, lower bound variational methods typically build on the Monte Carlo EM (MCEM) 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). That work inspired a stochastic approximation version of the variational Bayesian method for our problem here, and that led to a more broadly applicable approach to the marginal likelihood bound computation as described in Shen et al. (2008). For our purposes here, the essential technical details are noted in Appendix 2. The path to solution starts by noting that the global optimizing value of $\gamma_{1:p}$ satisfies the set of p equations $f_{1:p}(\gamma_{1:p}) = 0$ where for each $g = 1 : p$

$$f_g(\gamma_{1:p}) = \sum_{z_{1:p}} (z_g - \gamma_g) [1 + \log q(z_{1:p} | \gamma_{1:p}) - \log p(\Pi, z_{1:p} | A, \mathcal{F} = \mathcal{A})]. \quad (13)$$

The iterative, numerical solution method successively approximates the solution to these equations using a combination of 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 given value of $\gamma_{1:p}$, while

the latter applies to successively update estimates of the optimizing vector $\gamma_{1:p}$. As detailed in Appendix 2, an iterative algorithm that builds on these components then defines a sequence of $\gamma_{1:p}$ vectors that converges with probability one to $\gamma_{1:p}^*$ satisfying $f_{1:p}(\gamma_{1:p}^*) = 0$; a finite run of the algorithm provides an iterative approximation to this optimizing value. Further, by Monte Carlo sampling $z_{1:p}^h \sim q(z_{1:p} | \gamma_{1:p}^*)$, ($h = 1 : H$), we can then evaluate a consistent Monte Carlo estimate of the optimal lower bound, namely

$$\bar{L} = H^{-1} \sum_{h=1}^H \{\log p(\Pi, z_{1:p}^h | A, \mathcal{F} = \mathcal{A}) - \log q(z_{1:p}^h | \gamma_{1:p}^*)\}. \quad (14)$$

4 Examples

4.1 Highlights of Modeling and Inference with ($p = 18, m = 17$)

A very small, synthetic example fixes ideas and demonstrates the accuracy of the marginal likelihood approximation. The simulated data set concerns $p = 18$ genes with association probabilities shown in Figure 2. Based on Π alone, the first five genes have high π_g so 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 rather uncertain. We consider $m = 17$ biological pathway reference gene sets, $A_{1:17}$, constructed as in Figure 2(a); reference set A_j is simply the first j genes in the ordered list of 18 genes. Analysis assumes hyperparameters $r_A = 0.8$, $r_B = 0.1$ and $\phi_A = \phi_B = 8$.

The log marginal likelihood (shifted and scaled to $[0, 1]$ in Figure 2(b)) increases over $j = 1 : 5$ to a peak at $j = 5$ and then declines. This shows the evidence that biological pathways $\mathcal{A}_{4:5}$ corresponding to the reference gene sets $A_{4:5}$ are most strongly supported by the data Π . This is consistent with the original simulation design in that the first few genes are the signature genes of \mathcal{F} , having high π_g values. The variation of log marginal likelihood across the remaining reference gene sets is also reasonable given the values of the π_g across genes.

For each gene set, the optimal Monte Carlo variational upper and lower bounds of the log marginal likelihood are computed and displayed in Figure 3(a); also plotted are the exact values and quadrature based approximations using equation (9). In this “tiny p ” example, the exact and quadrature computations are feasible, and demonstrate the high accuracy of the upper and lower bound approximations. The approximation errors of the bounds (Figure 3(b)) are very small on the log likelihood scale, and certainly good enough to distinguish different biological pathways/models in this proof-of-principle example.

The simulation also illuminates how PROPA can aid in gene set refinement through the evaluation of posterior false-positive and false-negative probabilities gene-by-gene. The MCMC provides estimates of the posterior pathway membership probabilities π_g^* of equa-

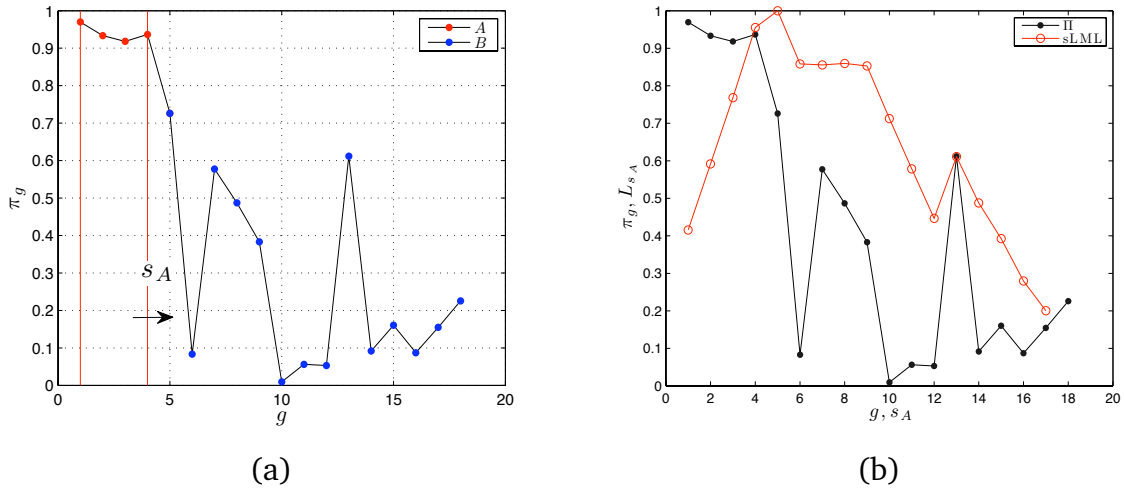


Figure 2: (a) Association probabilities π_g in the simulated data set. Genes in reference set A_j are genes $1 : j$ for each $j = 1 : 17$. (b) Standardized log marginal likelihood for each of the 17 pathways \mathcal{A}_j plotted with the π_j (here the x-axis simultaneously relates to genes and, by construction, the reference gene sets).

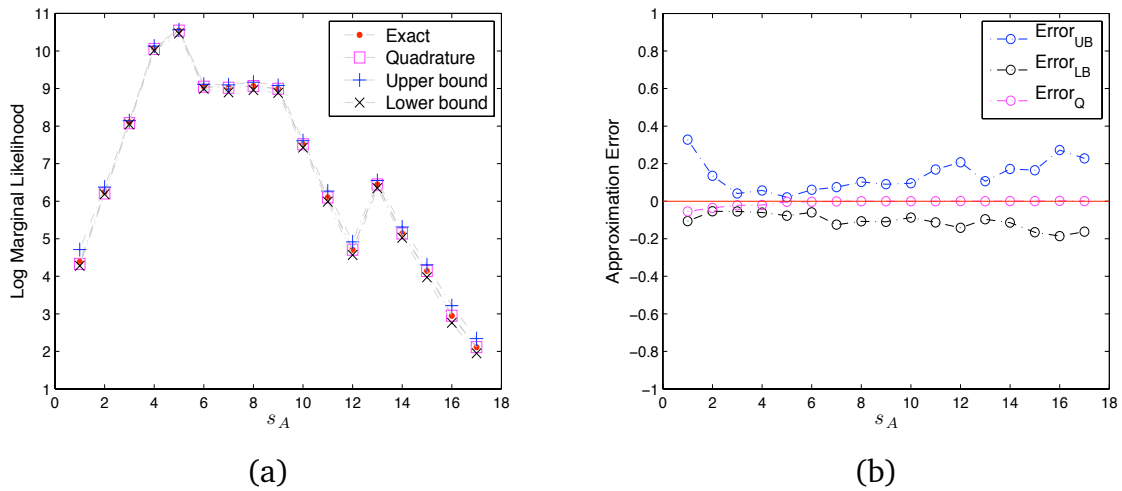


Figure 3: (a) Exact values, quadrature approximations, upper and lower bounds of log marginal likelihoods in the 17 pathway synthetic example. (b) Corresponding approximation errors of upper and lower bounds.

tion (5) to aid this. Focus on synthetic pathway \mathcal{A}_8 as an example; reference gene set A_8 is exactly the first 8 genes. For each gene g and each reference gene set analysis, 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 4. Genes in A_8 but with low π_g , and genes not in A_8 but with high π_g , are more likely to be regarded as false positives and false negatives, respectively, in terms of being members of \mathcal{A}_8 . Gene $g = 6$, a member of gene set A_8 , has membership evidence close to -20 dB, strongly suggesting it is not a member of the true pathway \mathcal{A}_8 (false positive). Gene 13 is not a member of A_8 , but it has membership evidence greater than 10 dB, which is substantial evidence that this gene is in fact a member of \mathcal{A}_8 (false negative).

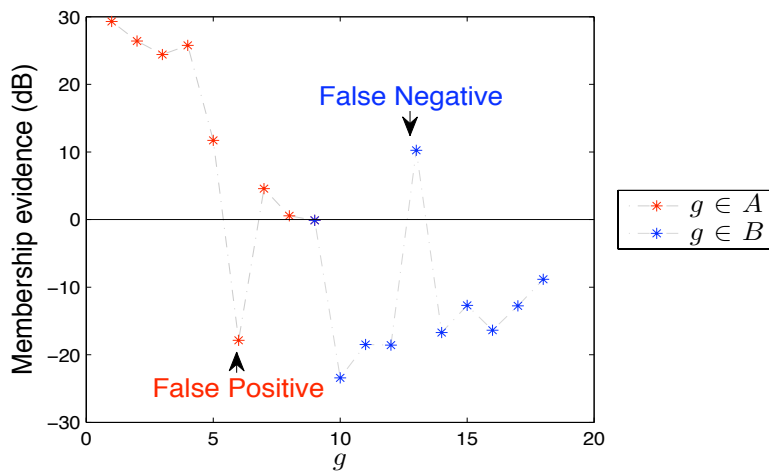


Figure 4: Pathway membership evidence for each gene g , in terms of log base 10 Bayes' factors for: against $g \in A_8$ in the 17 pathway synthetic example.

4.2 Marginal Likelihood Approximations with ($p = 100, m = 11$)

To further examine the performance of the marginal likelihood approximations, we simulated a larger data set, in which $p = 100$ and $m = 11$ gene sets were produced as above, now with $A_j = 1 : (14 + j)$ for $j = 1 : 11$; see Figure 5. Analysis assumes hyperparameters $r_A = 0.9$, $r_B = 0.05$ and $\phi_A = \phi_B = 8$. It is now impossible to compute the marginal likelihoods exactly, but quadrature can be applied (just). Applying the Monte Carlo VM method to generate bounds (Figure 6) is sufficient to distinguish the log marginal likelihoods of all the models, and the resulting PROPA identification of the true, known pathways is remarkably accurate.

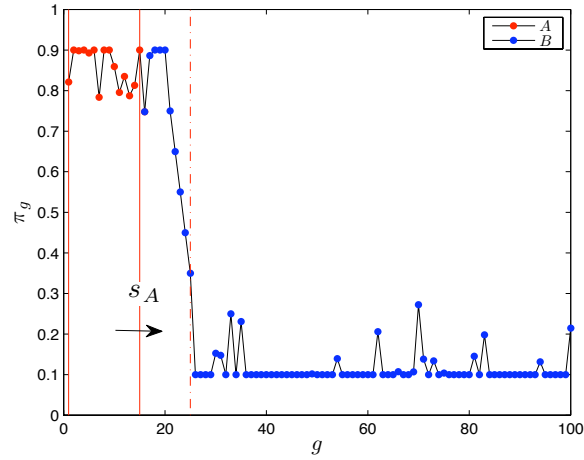


Figure 5: Association probabilities π_g in the simulated data set with 100 genes. Genes in reference set A_j are genes $1 : (14 + j)$ for each $j = 1 : 11$.

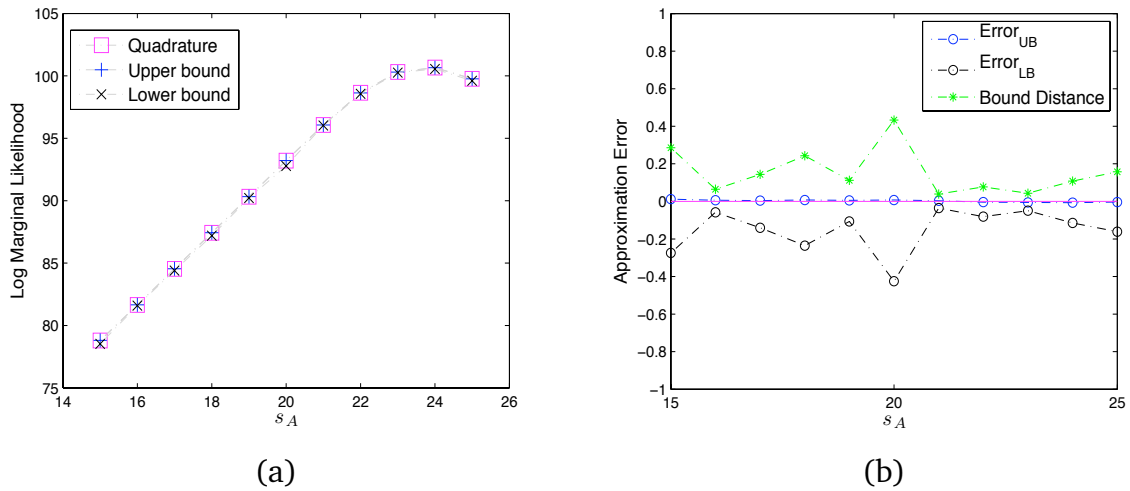


Figure 6: (a) Quadrature approximations, upper and lower bounds of log marginal likelihoods in the 100 gene, 11 pathway synthetic example. (b) Corresponding approximation errors of upper and lower bounds.

4.3 Marginal Likelihood Approximations with $p = 19,645$

In the simulation studies above, both the lower and the upper bound optimization methods have good performance in terms of accuracy and double-sided bounding provides sufficient information to facilitate pathway comparisons. With larger 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 (14) 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 defines the optimal upper bound approximation. The rationale is that, when the factorized variational density q is a good approximation of the joint posterior distribution of $z_{1:p}$, the variational densities corresponding to optimal upper and lower bounding are likely to converge to similar if not the same values; this has been experienced in multiple examples. Clearly, the value of the pseudo-optimal lower bound is always less than the lower bound described, but is massively more attractive computationally when p is very large.

We demonstrate this with a real data set on $p = 19,645$ genes and with $m = 15$ pathways whose reference gene sets come from the MSigDB (Broad Institute) database. The Π are probabilities of association between genes and the lactic acidosis cancer micro-environmental factor in human mammary epithelial cell cultures (from Section 6). Figure 7(a) presents the optimal upper bounds and pseudo-optimal lower bounds of log marginal likelihoods for the 15 pathway gene sets. The distances between pairs of bounds are shown in Figure 7(b); the pseudo-optimal lower bounds are very close to the optimal upper bounds, and such bounds are tight enough to discriminate the evidence for different models.

5 Case Study: Hormonal Pathways in Breast Cancer

5.1 Breast Cancer ER Pathway Annotation

Estrogen-receptor α ($ER\alpha$) is the primary mediator of estrogenic actions in breast cancer. About two-thirds of breast cancers show over-expression of $ER\alpha$ at the time of diagnosis. Both basic science and clinical data indicate the value of $ER\alpha$ level as an important predictor of breast cancer prognosis and outcomes (Deroo and Korach, 2006; Moggs and Orphanides, 2001).

Our study of 153 primary breast tumor samples (Carvalho et al., 2008) records expression data and $+/-ER$ levels from immunohistochemical (IHC) staining. Analysis using BFRM generated posterior probabilities $\Pi = p_{i_{1:p}}$, as well as the sign of association (positive or negative) between gene expression and ER status for the set of $p = 8,764$ unique genes corresponding Entrez gene IDs. The π_g are displayed in Figure 8(a)), showing a

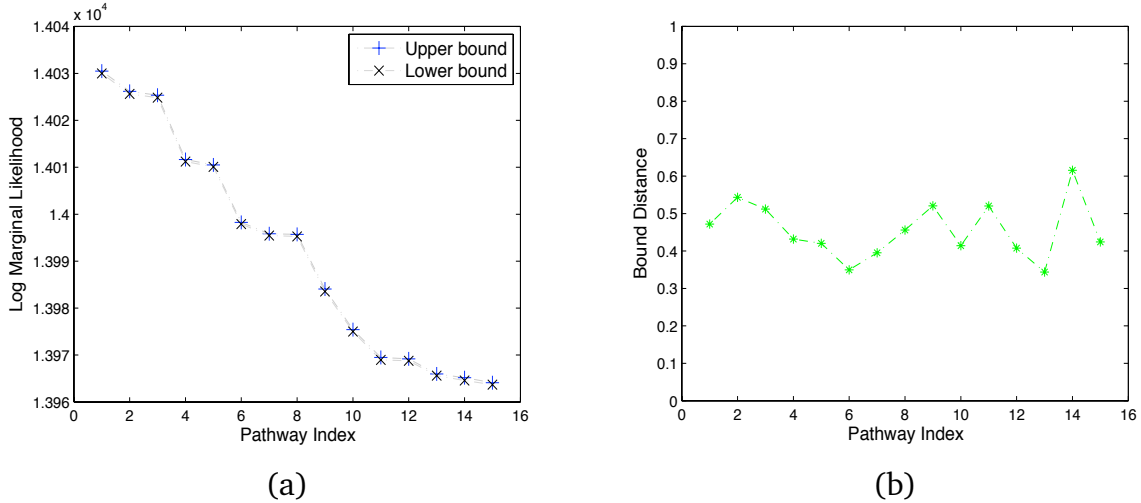


Figure 7: (a) Upper and quasi-lower bounds of log marginal likelihoods in the real data analysis with $p = 19,645$ genes and $m = 15$ pathways. (b) Corresponding difference between upper and lower bounds.

substantial number of genes apparently associated with the experimental factor pathway \mathcal{F} , here known of course to be the ER pathway. We re-curated the set of $m = 956$ gene sets from the MSigDB C2 collection (MSigDB, Broad Institute) to align with gene names based on the Entrez human gene database. PROPA analysis then generated upper and pseudo-optimal lower bounds on log marginal likelihoods for each of the 956 reference gene sets, and these are plotted, with pathways arranged in decreasing order of the upper bound, in Figure 8(b). For some pathways, the distance between the upper bound and the lower bound is too large to give a reliable approximation for the log marginal likelihood, but for most of the top 20 pathways, the bound distances are very small and reliably estimate the evidence.

The top 25 biological pathways are in Table 1, and histograms of a subset of the π_g for g in some of these reference gene sets these gene sets appear in Figure 9. The first two gene sets are breast tumor ER negative and positive signatures defined by van't Veer et al. (2002) through microarray analysis of a set of primary breast tumors. Besides the ER signatures, PROPA has also identified some other pathway signatures whose linkage to breast tumor ER status have been confirmed by previous research. The identification of the breast cancer prognosis signatures defined by van't Veer et al. (2002) (A_6 , A_{12} , A_{18} and A_{21}) agrees with the clinical research conclusion that patients with ER-negative tumors generally have worse prognosis than those with ER-positive tumors (Maynard et al., 1978). Compared with ER-positive breast cancers, ER-negative cancers are more likely to be poorly differentiated. This rationale to the finding of the undifferentiated cancer

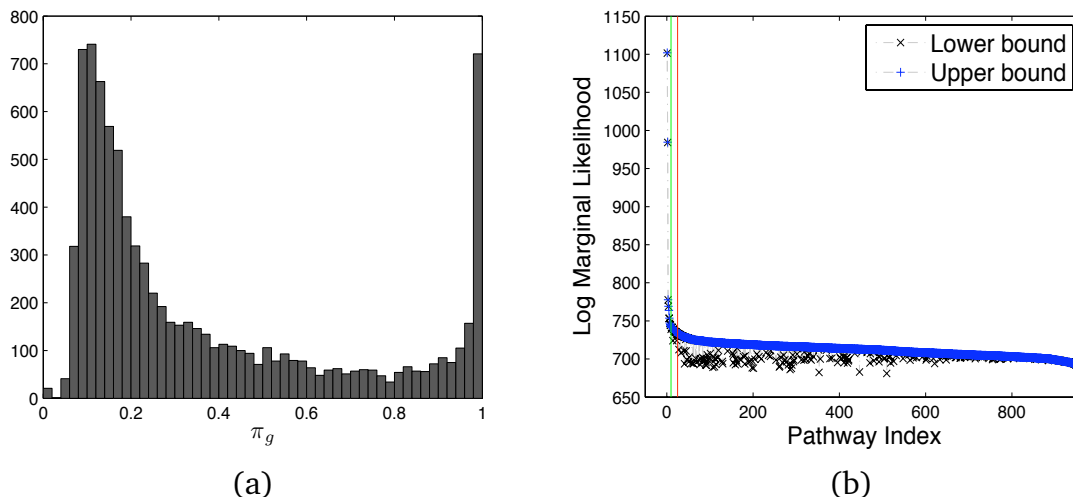


Figure 8: Breast cancer ER status study. (a) Histogram of association probabilities π_g . (b) Pathway log marginal likelihood upper bounds (+) and lower bounds (o); pathways are sorted in a decreasing order of log marginal likelihood; pathways on the left side of the red line are the top 25 pathways.

signature (A_9) in the pathway annotation analysis of breast cancer ER phenotype.

The same data set is analyzed by using GSEA (Table 1) for comparison. The pathway annotation results provided by PROPA and GSEA are generally consistent. However, PROPA detects the gene set representing Myb pathway (A_{15}) Lei et al. (2004), which is critically biologically related to breast tumor ER regulation (Hodges et al., 2003). GSEA cannot identify this relationship because it performs one-way tests. The ability of PROPA to identify gene sets comprised of both up-regulated and down-regulated genes offers advantage when gene sets are complicated and the expression regulation direction information is not available.

5.2 Breast Cancer ErbB2 Pathway Annotation

ErbB2 is a hormone in the same transmembrane receptor family as epidermal growth factor receptor (EGFR), and 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).

IHC recorded ErbB2 status (-/+) is available on 146 of the primary breast tumor samples (Carvalho et al., 2008). The same approach as for ER above is applied; the ErbB2 association probabilities in Figure 10(a) show that, relative to ER, only a very small set of

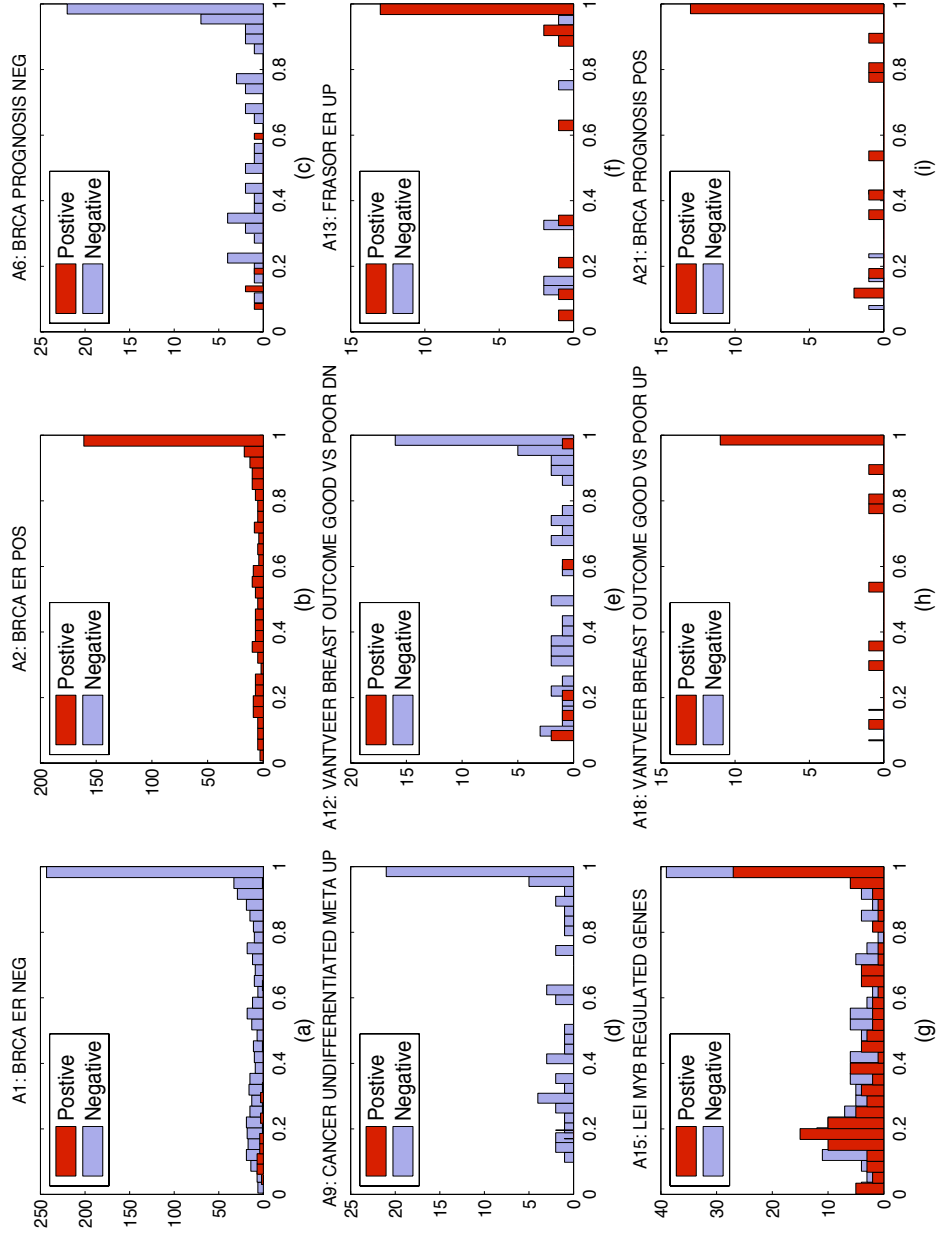


Figure 9: Breast cancer ER status study. For each gene set, genes with expression positively correlated with ER status are shaded dark, those negatively correlated with ER are unshaded.

Rank	Pathway	Size	LogML (UB)	LogML (LB)	UB-LB	GSEA			
						Rank	P-value	FDR (q-val)	Corr
1	BRCA ER NEG	692	1101.9	1101.47	0.44	1	0	0.02	-
2	BRCA ER POS	380	984.54	984.19	0.35	1	0	0.02	+
3	LEE TCELLS2 UP	712	777.72	777.66	0.06	47	0.008	0.21	-
4	FLECHNER KIDNEY TRANSPLANT REJN	81	769.23	769.11	0.11	40	0.035	0.21	-
5	CARIES PULP UP	186	763.79	753.12	10.67	67	0.045	0.21	-
6	BRCA PROGNOSIS NEG	69	753.65	753.4	0.26	4	0	0.21	-
7	SERUM FIBROBLAST CELLCYCLE	83	747.15	746.94	0.21	22	0.036	0.23	-
8	TARTE PLASMA BLASTIC	295	744.86	744.71	0.15	29	0.019	0.21	-
9	CANCER UNDIFFERENTIATED META UP	65	744.36	744.15	0.22	26	0.027	0.22	-
10	LI FETAL VS WT KIDNEY DN	157	742.33	739.01	3.31	121	0.095	0.28	-
11	CARIES PULP HIGH UP	83	741.89	741.45	0.45	136	0.095	0.3	-
12	VANTVEER BREAST OUTCOME/DOWN	58	741.7	741.48	0.21	3	0.002	0.15	-
13	FRASOR ER UP	29	741.37	741.34	0.03	6	0	0.23	+
14	CIS XPC UP	131	740.51	723.93	16.58	144	0.315	0.81	+
15	LEI MYB REGULATED GENES	302	739.04	738.96	0.08	438	0.512	0.67	-
16	RUTELLA HEMATOGENESIS DIFF	505	738.79	738.78	0.01	192	0.119	0.34	-
17	UVB NHEK3 C7	50	736.74	736.67	0.07	46	0.012	0.21	-
18	VANTVEER BREAST OUTCOME/UP	20	736.44	736.37	0.07	3	0	0.05	+
19	GREENBAUM E2A UP	25	735.93	735.84	0.09	97	0.089	0.25	-
20	ZHAN MM CD138 PR VS REST	26	735.86	735.79	0.07	31	0.014	0.22	-
21	BRCA PROGNOSIS POS	26	735.48	735.45	0.03	2	0	0.64	+
22	MIDDLEAGE DN	13	735.15	735.13	0.02	50	0.019	0.2	-
23	IRITANI ADPROX LYMPH	121	735.05	724.14	10.92	132	0.025	0.29	-
24	LEE TCELLS3 UP	63	734.16	733.98	0.18	15	0.014	0.27	-
25	KLEIN PEL DN	57	734.01	733.6	0.41	43	0.03	0.23	-

Table 1: Summary of the top 25 ER-related pathways identified by PROPA

genes is associated with the experimental factor pathway \mathcal{F} , here known of course to be the ErbB2 pathway.

The 956 human pathway gene sets drawn from MSigDB do not include pathway signatures explicitly linked to breast tumor ErbB2 status. To validate the effectiveness of PROPA, we curated two gene sets from the literature: the first gene set, which we will call the *molecular portrait* of ErbB2-positive breast tumors, consists of several genes that are mainly located at the chromosome 17 and have been identified as a cluster corresponding to ErbB2 over-expression (Perou et al., 2000; Sørlie et al., 2001); the second gene set is curated from the ErbB2 gene expression signature defined by Bertucci et al. (2004), and includes genes differentially expressed with-versus-without over-expression of ErbB2 protein as measured in data from a range of tumors and cell lines

The bounds on marginal likelihoods appear in Figure 10(b) shows the optimal upper bound and corresponding lower bound of log marginal likelihood for each pathway given by PROPA. The first four or five pathways may be of particular interest and are summarized in Table 2. Our two ErbB2 signatures are identified by PROPA as the top one and fourth of the $m = 958$ pathway signatures. In Figure 11(a) and (c), the π_g of genes in the two sets are categorized by the sign of the correlation with ErbB2 status.

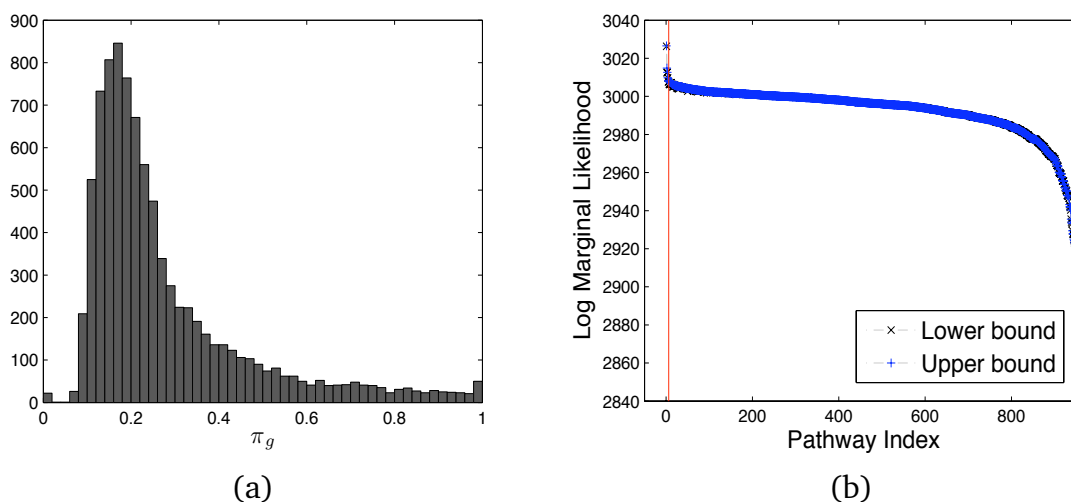


Figure 10: Breast cancer ErbB2 status study. (a) Histogram of association probabilities π_g . (b) Pathway log marginal likelihood upper bounds (+) and lower bounds (o); pathways are sorted in a decreasing order of log marginal likelihood; pathways on the left side of the red line are the top 6 pathways.

The same data set and pathway genes sets were analyzed by GSEA; GSEA concluded that *none* of the 958 gene sets are enriched in ErbB2 (FDR q -value less than 25%). The two curated ErbB2 signature gene sets are identified by GSEA in the top up-regulated gene

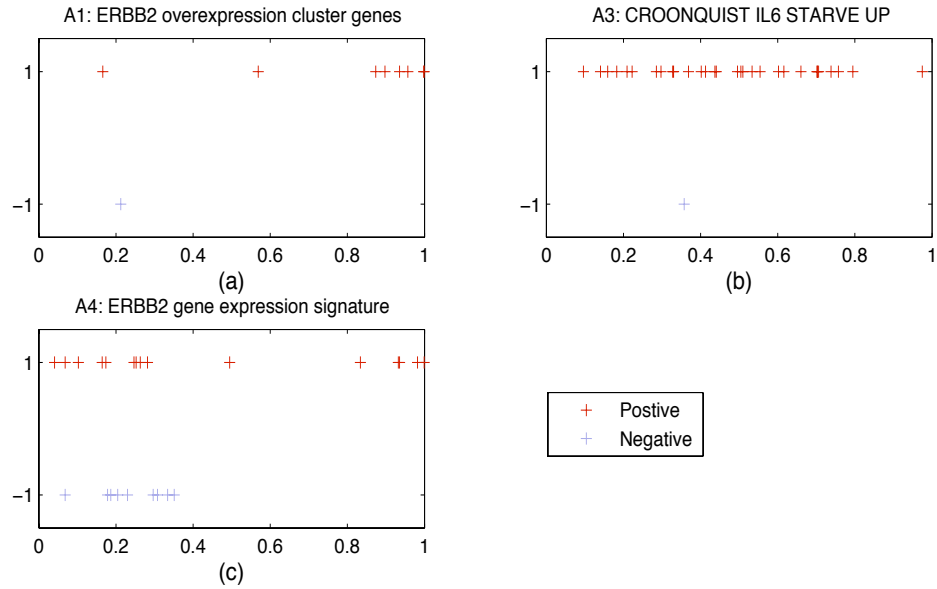


Figure 11: Association probability plots of the ErbB2-related pathway gene sets identified by PROPA. In each plot, the x-axis is probability, and y-axis has two states, -1 and 1 representing negative and positive correlation with ErbB2, respectively. For each gene set, genes positively correlated with tumor ErbB2 status are marked $+$, those negatively correlated are \circ .

Rank	Pathway	Size	LogML (UB)	LogML (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: Six pathways identified by PROPA as most related to breast tumor ErbB2 status

set list (top four and six) ranked according to the normalized enrichment scores (NES). However, neither of these are considered strong significant by GSEA. In this example, compared with GSEA, PROPA presents better sensitivity and specificity when transcriptional evidence of phenotype-pathway association are relatively weak in the sense of small numbers of genes in the reference gene lists.

$\&$	Symbol	Description	Gene ID	Chr. Loc.	π_g^*	$\log(\text{BF})$	π_g	Corr.
1	STARD3	START domain containing 3	10948	17q11-q12	1	10.07	0.99	+
2	GRB7	growth factor receptor-bound protein 7	2886	17q12	1	10.07	0.99	+
3	THRAP4	thyroid hormone receptor associated protein 4	9862	17q21.1	0.99	6.09	0.96	+
4	ERBB2	v-erb-b2 oncogene homolog 2	2064	17q11.2-q12	0.99	4.34	0.94	+
5	TRAF4	TNF receptor-associated factor 4	9618	17q11-q12	0.92	1.60	0.90	+
6	FLOT2	flotillin 2	2319	17q11-q12	0.72	0.12	0.88	+
7	PCGF2	polycomb group ring finger 2	7703	17q12	0	-16.19	0.57	+
8	MMP15	matrix metalloproteinase 15	4324	16q13-q21	0	-30.78	0.34	+
9	SMARCE1	SWI/SNF related regulator of chromatin	6605	17q21.2	0	-42.19	0.21	-

Table 3: Genes in the ErbB2 molecular portrait gene set

Table 3 gives information on the ErbB2 molecular portrait reference gene set 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 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. Meanwhile, several genes (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 34 dBs, indicating very strong if not decisive evidence for these genes being false negatives, i.e. they are members ErbB2 pathway.

6 Case Study: Lactic Acidosis in Breast Cancer

6.1 Lactic Acidosis Study

This study demonstrates the central role of PROPA in a cancer research strategy that makes use of the heterogeneity in a large set of tumor samples to dissect the complex pathway activities involved in cancerous cellular response to a biological intervention. The biological

focus is lactic acidosis (LA), a combined measure of lactate and acidity in the environment cells grow in. Our studies of LA in cell culture experiments and breast tumors (Chen et al., 2007) generated gene expression signatures of two kinds: first, *in vitro* defined signatures of the response of normal mammary epithelial cells (HMECs) to exposure to high levels of LA; second, a set of signatures from a latent factor analysis of a heterogeneous sample of breast tumor samples (Lucas et al., 2007; Chen et al., 2007), in which the genes used to define the initial factor model were taken from the *in vitro* LA signature gene set. In each case, analysis using BFRM (Carvalho et al., 2008; Wang et al., 2007) generated association probabilities for input to PROPA analysis. An overall schematic outline of the strategy of analysis appears in Figure 12. Here we explore pathway annotations of the two types of signature.

6.2 *In Vitro* Lactic Acidosis Response Annotation

The *in vitro* experiment compared expression data from 6 control cell samples with 6 samples grown in an LA rich culture. A sparse analysis of variance (Lucas et al., 2007) generated association probabilities π_g on over 20,000 genes. Using the same database of $m = 956$ reference gene sets, the most highly ranked PROPA pathways (Table 4) reveal relationships between the LA transcriptional response and several common biological processes and traits of cancer development. Histograms of genes in each of the top 12 reference gene sets appear in Figure 13. The activation of the genes in A_1 , A_7 and A_4 under high LA is consistent with the fact that nutrient starvation, hypoxia and lactic acidosis are commonly coexisting/interacting conditions in tumor micro-environments (Peng et al., 2002; Manalo et al., 2005). We infer that LA down-regulates gene sets that characterize transcriptional activities of wound healing - A_2 , A_3 and A_8 (Chang et al., 2004); tumor angiogenesis - A_5 and A_{12} (Croonquist et al., 2003); and immune regulation - A_{10} (Lee et al., 2004). Together these indicate that lactic acidosis may play roles in a number of processes that inhibit cancer progression. Further, we infer that LA up-regulates genes in A_6 , A_{11} and A_9 , and the corresponding pathways appear to define undifferentiated cancers, correlated with poor prognosis (Rhodes et al., 2004), poor breast cancer outcomes (van't Veer et al., 2002) and gastric cancer drug resistance (?)¹, respectively; this suggests that increased LA may itself engender reduced tumor aggressiveness. The association of LA with these pathways indicates that the molecular mechanisms by which LA modulates cellular behavior is directly, and beneficially, predictive of clinical cancer phenotypes and progression.

A number of these genes appear, in terms of the π_g values, to be unrelated to the LA response, perhaps due to the specificity of the experiment; PROPA allows us to explore

¹HAIGE: gastric cancer drug resistance reference is needed!!!

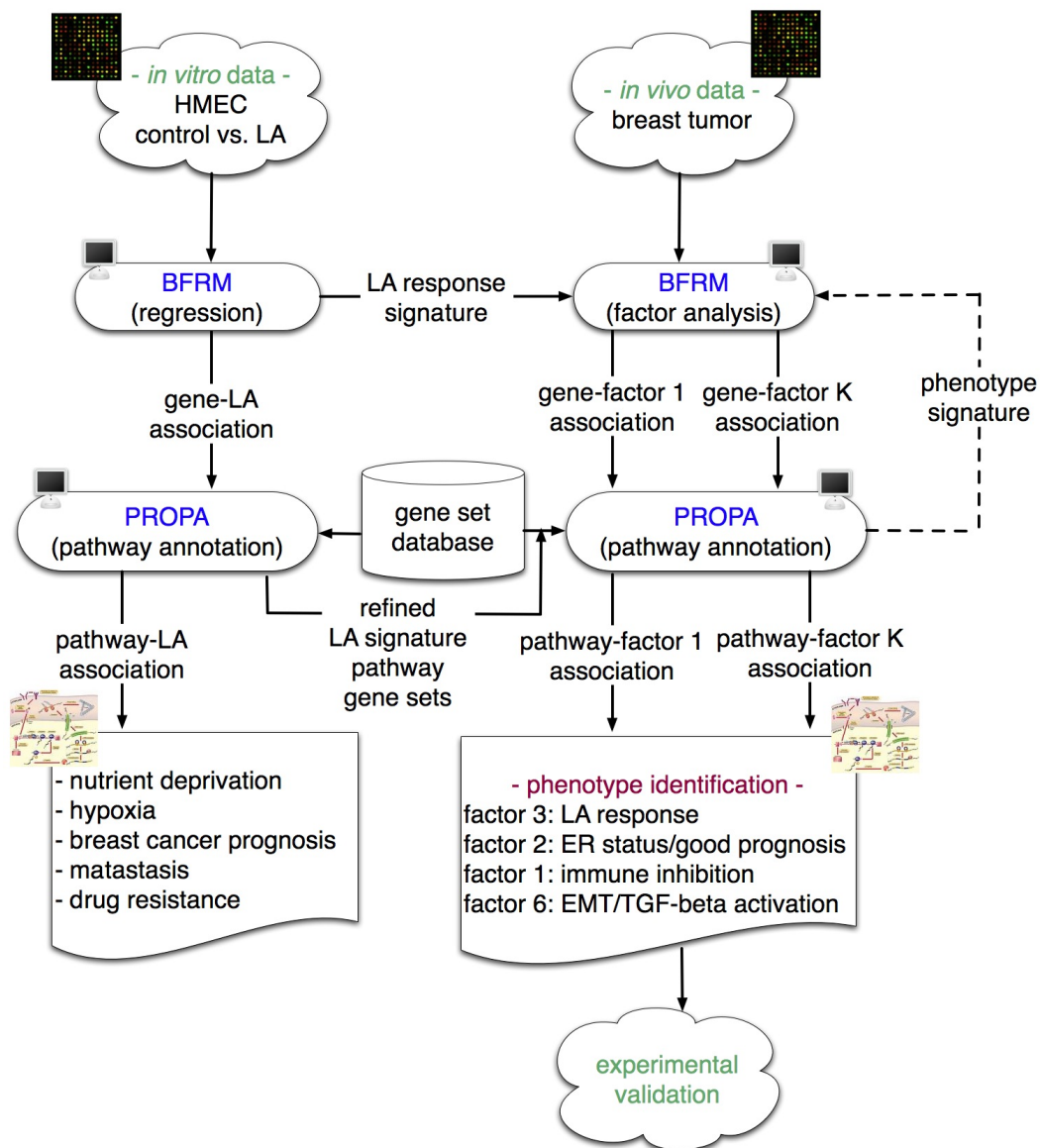


Figure 12: Lactic acidosis in cancer. Schematic framework for integrating *in vitro* pathway identification with an elaborated, *in vitro* profile of multiple statistical factors underlying a signature when projected and analysed in human tumor data sets. PROPAs maps expression signatures and multiple sets of profiled factors to the available biological pathway databases.

this via the posterior pathway membership probabilities π_g^* . This identifies candidate true-positives, namely genes within the corresponding gene set that have high π_g^* – here we simply threshold the log Bayes’ factors to identify those genes with pathway membership evidence greater than 20dB, and define revised gene sets

$$A_{LA,j} = \{g : g \in A_j, 10 \log_{10} BF_g > 20\}, \quad j = 1 : 12.$$

These novel *LA pathway signatures* are added to the database for future analyses.

Rank	Pathway	Size	LogML (UB)	LogML (LB)	UB-LB
1	PENG GLUTAMINE DN	250	14031.01	14030.31	0.7
2	SERUM FIBROBLAST CELLCYCLE	134	14026.46	14025.87	0.59
3	CHANG SERUM RESPONSE UP	145	14025.38	14024.98	0.4
4	MANALO HYPOXIA DN	77	14012.37	14011.97	0.4
5	CROONQUIST IL6 STARVE UP	32	14010.64	14010.24	0.4
6	CANCER UNDIFFERENTIATED META UP	66	13998.43	13998.40	0.03
7	PENG LEUCINE DN	141	13996.20	13995.88	0.31
8	SERUM FIBROBLAST CORE UP	199	13995.72	13995.5	0.21
9	DOX RESIST GASTRIC UP	44	13984.33	13984.22	0.11
10	LEE TCELLS3 UP	100	13976.11	13975.43	0.68
11	VANTVEER BREAST OUTCOME/DOWN	65	13969.68	13969.14	0.54
12	CROONQUIST IL6 RAS DN	23	13969.55	13968.99	0.57

Table 4: Top 12 PROPA pathways related to lactic acidosis

6.3 *In Vivo* Lactic Acidosis Response in Human Breast Cancers

The LA pathway signature genes were used to initialize an evolutionary factor analysis (Carvalho et al., 2008) of gene expression data from 251 breast tumor samples (Miller et al., 2005) involving roughly 18,000 genes. This factor analysis generated several estimated factors in the breast samples that reflect patterns of association among genes in the LA signature as well as additional genes with related expression patterns (Chen et al., 2007). This is an example of the strategy of *signature factor profiling analysis* (Lucas et al., 2008). Each statistical factor has a vector of gene-factor loading probabilities Π representing the posterior probabilities in the data analysis of a non-zero loading on the factor for each gene. These are inputs to PROPA analysis to explore pathway annotation of each of the factors, i.e., to study the potentially broader set of pathways that LA relates to in human cancers, compared to those in the narrower, controlled context of *in vitro* cultured cells.

PROPA first identifies the lactic acidosis respond factor, factor 3 (the index in the original paper Chen et al. (2007)), characterized by the signature pathway gene sets defined in the previous *in vitro* analysis. This factor bonds the entire tumor data analysis to the

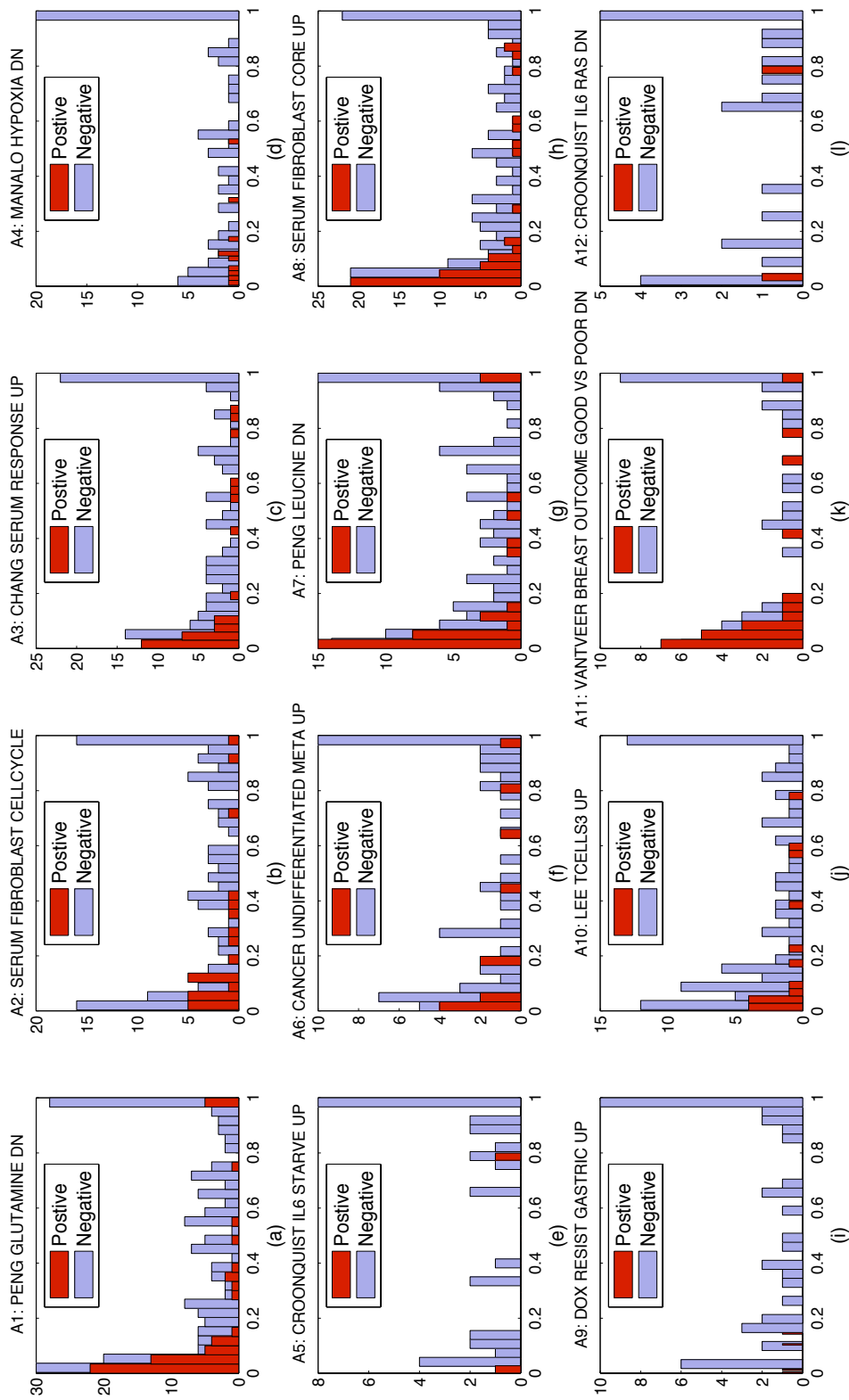


Figure 13: Lactic acidosis *in vitro* study. For the top 12 pathways A_j identified by PROPA, histograms of the π_g for each reference gene set $g \in A_j$ are coded to indicate genes positively correlated with lactic acidosis levels (dark shading) and genes negatively related to lactic acidosis (no shading).

primary lactic acidosis response. Additional PROPA identified pathways link lactic acidosis as a potential signal to breast cancer ER, immune and TGF- β induced EMT pathways. The annotation of factor 2 connects ER pathway activation with other oncogenic pathways, including Ras, p110- α , E2F3, Myc and β -catenin pathways, through the lactic acidosis response signature. The detection of immune function pathways in factor 7 and 9 provides evidence for the role of lactic acidosis in immune cell function perturbation during cancer development. All of these findings relate well to known biology (Pethe and Shekhar, 1999; Fry, 2001; Sears et al., 2000; Leone et al., 2001; Fischer et al., 2007).

A more intriguing finding is the association between LA and the TGF- β induced epithelial-to-mesenchymal transition (EMT) pathway (Zavadil et al., 2004) in factor 6; this PROPA discovery generated a concrete biological hypothesis that is currently under experimental investigation. EMT is a process involved in wound healing whereby fully differentiated epithelial cells undergo transition to a mesenchymal phenotype giving rise to fibroblasts and myofibroblasts (Vincent-Salomon and Thiery, 2003). TGF- β is a family of multifunctional cytokines that plays an important role in the regulation of epithelial cell growth, differentiation and apoptosis. It is known that TGF- β inhibits epithelial cell growth in early stage of breast tumorigenesis, and induces EMT in a later stage of carcinogenesis. Hence the possibility of regulation of the effects of LA on TGF- β -induced EMT may explain the link of LA with cancer prognosis.

7 Summary Comments

PROPA is a formal model-based 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 canonical setting here, the gene weights are explicit gene-factor phenotype association probabilities. The formal probabilistic model delivers estimated marginal likelihood values over pathways for each factor phenotype, allowing for a quantitative assessment and ranking of pathways putatively linked to the phenotype as well as refinement of pathway databases through posterior membership probabilities. Our examples with simulated and real data sets, and our case studies in two core areas of cancer genomics, highlight the use and application of the methodology and its potential as a tool in integrative genomic studies.

In connection with the computational aspects of PROPA analysis, we make use of a novel Monte Carlo variational method for estimating marginal likelihoods for model comparisons, here to compare biological pathways. The method is applicable more generally, and our examples and case studies illustrate its use and effectiveness.

Among many open questions and directions for further development, we mention the need for improved quality of biological pathway databases, an area that PROPA can con-

tribute to as we have exemplified; methodological issues related to the specification of model priors across pathways, the use of alternative, multiple forms of numerical summary of the relationships between genes and experimental factor phenotypes, and of course an interest in developing our current software for public use. Such directions should help aid in the contributions of more relevant, model-based statistical reasoning to the broader pathway annotation enterprise in modern biological studies.

Acknowledgments

We are grateful to Ashley Chi, Joe Lucas and Chunlin Ji of Duke University for discussions and important input. We acknowledge support of the National Science Foundation (grants DMS-0102227 and DMS-0342172) and the National Institutes of Health (grants NHLBI P01-HL-73042-02 and NCI 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.

Appendix 1

Refer to the marginal likelihood function shown in (6). Integrating out $\{\beta_g\}_{g \in \mathcal{G}}$, α_0 and α_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

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

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 Φ 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 integrations 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

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

Appendix 2

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 Shen et al. (2008). In essentials here, it is first easy to see that the global, lower bound optimizing value of $\gamma_{1:p}$ satisfies $f_{1:p}(\gamma_{1:p}) = 0$ for the function defined in equation (13). The method is based on the observations that:

1. $f_{1:p}(\gamma_{1:p})$ 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 $f_{1:p}(\gamma_{1:p})$ can be used to drive 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 $f_{1:p}^{(t-1)}(\gamma_{1:p}^{(t-1)})$ replacing the sum in equation (13) 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 more general algorithm for which it can be shown (Robbins and Monro, 1951; Shen et al., 2008) that $\gamma_{1:p}^{(t)}$ converges with probability one to $\gamma_{1:p}^*$ satisfying $f_{1:p}(\gamma_{1:p}^*) = 0$, 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 $z_{1:p}^h$, ($i = 1 : H$), from $q(z_{1:p} | \gamma_{1:p}^*)$, and then evaluate the Monte Carlo estimate of the optimal lower bound

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

This is a consistent estimate of the maximum lower bound assuming the stochastic approximation estimate has converged (Shen et al., 2008).

References

- Badache, A. and Gonçalves, A. “The ErbB2 signaling network as a target for breast cancer therapy.” *Journal of Mammary Gland Biology and Neoplasia*, 11:13–25 (2006).
- Beal, M. “Variational Algorithms for Approximate Bayesian Inference.” Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London (2003).
- Bertucci, F., Borie, N., Ginestier, C., Groulet, A., Charafe-Jauffret, E., Adélaïde, J., Geneix, J., Bachelart, L., Finetti, P., Koki, A., Hermitte, F., Hassoun, J., Debono, S., Viens, P., Fert, V., Jacquemier, J., and Birnbaum, D. “Identification and validation of an ERBB2 gene expression signature in breast cancers.” *Oncogene*, 23(14):2564–2575 (2004).
- Bild, A. H., Yao, G., Chang, J. T., Wang, Q., Chasse, A. P. D., Joshi, M., Harpole, D., Lancaster, J. M., Berchuck, A., Olson Jr., J. A., Marks, J. R., Dressman, H. K., West, M., and Nevins, J. R. “Oncogenic pathway signatures in human cancers as a guide to targeted therapies.” *Nature*, 439:353–357 (2006).
- Carvalho, C., Lucas, J., Wang, Q., Chang, J., Nevins, J., and West, M. “High-dimensional sparse factor modelling: Applications in gene expression genomics.” *Journal of the American Statistical Association*, (in press) (2008).
- Celeux, G. and Diebolt, J. “A stochastic approximation type EM algorithm for the mixture problem.” *Stochastics and Stochastics Reports*, 41:127–146 (1992).
- Chan, K. S. and Ledolter, J. “Monte Carlo EM estimation for time series models involving counts.” *Journal of the American Statistical Association*, 90:242–252 (1995).

- Chang, H. Y., Sneddon, J. B., Alizadeh, A. A., Sood, R., West, R. B., Montgomery, K., Chi, J., van de Rijn, M., Botstein, D., and Brown, P. O. "Gene expression signature of fibroblast serum response predicts human cancer progression: Similarities between tumors and wounds." *PLoS Biol.*, 2:e7 (2004).
- Chen, J. L., Lucas, J. E., Schroeder, T., Mori, S., Nevins, J., Dewhirst, M., West, M., and Chi, J. A. "Genomic analysis of response to lactic acidosis in human cancers." *Department of Statistical Science, Duke University, Discussion Paper*, (Submitted for publication) (2007).
- Chib, S. "Marginal likelihood from the Gibbs output." *Journal of the American Statistical Association*, 90:773–795 (1995).
- Corduneanu, A. and Bishop, C. M. "Variational Bayesian model selection for mixture distributions." In Jaakkola, T. and Richardson, T. (eds.), *Artificial Intelligence and Statistics*, 27–34. Morgan Kaufmann (2001).
- Croonquist, P. A., Linden, M. A., Zhao, F., and Van Ness, B. G. "Gene profiling of a myeloma cell line reveals similarities and unique signatures among IL-6 response, N-ras-activating mutations, and coculture with bone marrow stromal cells." *Blood*, 102:2581–2592 (2003).
- Delyon, B., Lavielle, M., and Moulines, E. "Convergence of a stochastic approximation version of the EM algorithm." *Annals of Statistics*, 27:94–128 (1999).
- Deroo, B. J. and Korach, K. S. "Estrogen receptors and human disease." *J. Clin. Invest.*, 116:561–570 (2006).
- Fischer, K., Hoffmann, P., Voelkl, S., Meidenbauer, N., Ammer, J., Edinger, M., Gottfried, E., Schwarz, S., Rothe, G., Hoves, S., Renner, K., Timischl, B., Mackensen, A., Kunz-Schughart, L., Andreesen, R., Krause, S. W., and Kreutz, M. "Inhibitory effect of tumor cell-derived lactic acid on human T cells." *Blood*, 109:3812–3819 (2007).
- Fry, M. J. "Phosphoinositide 3-kinase signalling in breast cancer: How big a role might it play?" *Breast Cancer Res*, 3:304–312 (2001).
- Hodges, L. C., Cook, J. D., Lobenhofer, E. K., Li, L., Bennett, L., Bushel, P. R., Aldaz, C. M., Afshari, C. A., and Walker, C. L. "Tamoxifen functions as a molecular agonist inducing cell cycle-associated genes in breast cancer cells." *Mol. Cancer Res.*, 1:300–311 (2003).
- Huang, E., Chen, S., Dressman, H., Pittman, J., Tsou, M., C.F. Horng, A. B., Iversen, E., Liao, M., Chen, C., West, M., Nevins, J., and Huang, A. "Gene expression predictors of breast cancer outcomes." *The Lancet*, 361:1590–1596 (2003a).

- Huang, E., Ishida, S., Pittman, J., Dressman, H., Bild, A., M. D'Amico, R. P., West, M., and Nevins, J. "Gene expression phenotypic models that predict the activity of oncogenic pathways." *Nature Genetics*, 34:226–230 (2003b).
- Huang, E., West, M., and Nevins, J. "Gene expression profiles and predicting clinical characteristics of breast cancer." *Hormone Research*, 58:55–73 (2002).
- . "Gene expression phenotypes of oncogenic pathways." *Cell Cycle*, 2:415–417 (2003c).
- Jaakkola, T. S. and Jordan, M. I. "Bayesian logistic regression: A variational approach." In *Proceedings on the 1997 Conference on Artificial Intelligence and Statistics*, 283–294. Fort Lauderdale, FL (1997).
- Jordan, M., Ghahramani, Z., Jaakkola, T., and Saul, K. "An introduction to variational methods for graphical models." *Machine Learning*, 37:183–233 (1999).
- Lee, M. S., Hanspers, K., Barker, C. S., Korn, A. P., and McCune, J. M. "Gene expression profiles during human CD4+ T cell differentiation." *International Immunology*, 16:1109–1124 (2004).
- Lei, W., Rushton, J. J., Davis, L. M., Liu, F., and Ness, S. A. "Positive and negative determinants of target gene specificity in Myb transcription factors." *J. Biol. Chem.*, 279:29519 – 29527 (2004).
- Leone, G., Sears, R., Huang, E., Rempel, R., Nuckolls, F., Fielda, S. J., Thompson, M. A., Yang, H., Fujiwara, Y., Greenberg, M. E., Orkin, S., DeGregoria, J., Smith, C., and Nevins, J. R. "Myc requires distinct E2F proteins to induce S phase and apoptosis." *Mol. Cell Biol.*, 8:105–114 (2001).
- Lucas, J., Carvalho, C., Merl, D., and West, M. "In-vitro to in-vivo factor profiling in expression genomics." In Dey, D., Ghosh, S., and Mallick, B. (eds.), *Bayesian Modelling in Bioinformatics*. Taylor-Francis (2008).
- Lucas, J., Carvalho, C., Wang, Q., Bild, A., Nevins, J. R., and West, M. "Sparse statistical modelling in gene expression genomics." In K.A. Do, P. M. and Vannucci, M. (eds.), *Bayesian Inference for Gene Expression and Proteomics*, 155–176. Cambridge University Press (2006).
- Lucas, J. E., Carvalho, C. M., Chen, L., Chi, J., and West, M. "Bench-to-bedside and cross-study projections of genomic biomarkers: An evaluation in breast cancer genomics." *Department of Statistical Science, Duke University, Discussion Paper*, (Submitted for publication) (2007).

- Manalo, D. J., Rowan, A., Lavoie, T., Natarajan, L., Kelly, B. D., Ye, S. Q., Garcia, J. G., and Semenza, G. L. “Transcriptional regulation of vascular endothelial cell responses to hypoxia by HIF-1.” *Blood*, 105:659–669 (2005).
- Maynard, P., Davies, C. J., Blamey, R., Elston, C. W., Johnson, J., and Griffiths, K. “Relationship between oestrogen-receptor content and histological grade in human primary breast tumours.” *Brit. J. Cancer*, 38:745–748 (1978).
- McGrory, C. A. and Titterton, D. M. “Variational approximations in Bayesian model selection for finite mixture distributions.” *Computational Statistics & Data Analysis*, 51:5352–5367 (2007).
- Ménard, S., Pupa, S. M., Campiglio, M., and Tagliabue, E. “Biologic and therapeutic role of HER2 in cancer.” *Oncogene*, 22:6570–6578 (2003).
- Meng, X. L. and Wong, W. H. “Simulating ratios of normalizing constants via a simple identity: A theoretical exploration.” *Statistica Sinica*, 6:831–860 (1996).
- Miller, D. L., Smeds, J., George, J., Vega, V. B., Vergara, L., Ploner, A., Pawitan, Y., Hall, P., Klaar, S., Liu, E. T., and Bergh, J. “An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival.” *Proc. Nat. Acad. Sci. U.S.A.*, 102:13550–13555 (2005).
- Moggs, J. G. and Orphanides, G. “Estrogen receptors: Orchestrators of pleiotropic cellular responses.” *EMBO Rep.*, 2:775–781 (2001).
- MSigDB. “Molecular Signatures Data Base.” <http://www.broad.mit.edu/gsea/msigdb/> (Broad Institute).
- Newton, M. A., Quintana, F. A., den Boon, J. A., Sengupta, S., and Ahlquist, P. “Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis.” *Annals of Applied Statistics*, 1:85–106 (2007).
- Newton, M. A. and Raftery, A. E. “Approximate Bayesian inference with the weighted likelihood bootstrap.” *Journal of the Royal Statistical Society, Series B*, 56:3–48 (1994).
- Peng, T., Golub, T. R., and Sabatini, D. M. “The Immunosuppressant Rapamycin Mimics a Starvation-Like Signal Distinct from Amino Acid and Glucose Deprivation.” *Mol. Cell. Biol.*, 22:5575–5584 (2002).
- Perou, C. M., Sorlie, T., Eisen, M. B., van deRijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslén, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lonning, P. E., Borresen-Dale, A., Brown, P. O., and Botstein, D. “Molecular portraits of human breast tumours.” *Nature*, 406:747–752 (2000).

- Pethe, V. and Shekhar, P. V. M. “Estrogen inducibility of c-Ha-ras transcription in breast cancer cells.” *J. Biol. Chem.*, 274:30969–30978 (1999).
- Pittman, J., Huang, E., Dressman, H., Horng, C., Cheng, S., Tsou, M., Chen, C., Bild, A., Iversen, E., Huang, A., Nevins, J., and West, M. “Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes.” *Proc. Nat. Acad. Sci. U.S.A.*, 101:8431–8436 (2004).
- Raftery, A., Newton, M., Satagopan, J., and Krivitsky, P. “Estimating the integrated likelihood via posterior simulation using the harmonic mean identity (with discussion).” In Bernardo, J., Bayarri, M., Berger, J., David, A., Heckerman, D., Smith, A., and West, M. (eds.), *Bayesian Statistics 8*, 1–45. Oxford University Press (2007).
- Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A., and Chinnaiyan, A. M. “Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression.” *Proc. Nat. Acad. Sci. U.S.A.*, 101:9309–14 (2004).
- Rich, J., Jones, B., Hans, C., Iversen, E., Roger McClendon, A. R., Bigner, D., Dobra, A., Holly Dressman, J. N., and West, M. “Gene expression profiling and genetic markers in glioblastoma survival.” *Cancer Research*, 65:4051–4058 (2005).
- Robbins, H. and Monro, S. “A stochastic approximation method.” *Annals of Mathematical Statistics*, 22:400–407 (1951).
- Sears, R., Nuckolls, F., Haura, E., Taya, Y., Tamaia, K., and Nevins, J. R. “Multiple Ras-dependent phosphorylation pathways regulate Myc protein stability.” *Genes and Development*, 14:2501–2514 (2000).
- Seo, D., Wang, T., Dressman, H., Herderick, E., E.S. Iversen, C. D., Vata, K., Milano, C., Rigat, F., J. Pittman, J. N., West, M., and Goldschmidt-Clermont, P. “Gene expression phenotypes of atherosclerosis.” *Arteriosclerosis, Thrombosis and Vascular Biology*, 24:1922–1927 (2004).
- Seo, D. M., Goldschmidt-Clermont, P. J., and West, M. “Of mice and men: Sparse statistical modelling in cardiovascular genomics.” *Annals of Applied Statistics*, 1:152–178 (2007).
- Shen, H., Ji, C., and West, M. “Monte Carlo variational approximation of marginal likelihoods.” *Department of Statistical Science, Duke University, Discussion Paper*, (Submitted for publication) (2008).
- Sørli, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O.,

- Botstein, D., Lønning, P., and Børresen-Dale, A. “Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.” *Proc. Natl. Acad. Sci. U.S.A.*, 98:10869–10874 (2001).
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lauder, E. S., and Mesirov, J. P. “Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.” *Proc. Nat. Acad. Sci. U.S.A.*, 102:15545–15550 (2005).
- van’t Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. “Gene expression profiling predicts clinical outcome of breast cancer.” *Nature*, 415:530–536 (2002).
- Vincent-Salomon, A. and Thiery, J. P. “Host microenvironment in breast cancer development: Epithelial–mesenchymal transition in breast cancer development.” *Breast Cancer Research*, 5:101–106 (2003).
- Wang, Q., Carvalho, C., Lucas, J., and West, M. “BFRM: Bayesian factor regression modelling.” *Bulletin of the International Society for Bayesian Analysis*, 14:4–5 (2007).
- West, M. “Bayesian factor regression models in the “large p, small n” paradigm.” In Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M. (eds.), *Bayesian Statistics 7*, 723–732. Oxford University Press (2003).
- West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., R. Spang, H. Z., Marks, J., and Nevins, J. “Predicting the clinical status of human breast cancer utilizing gene expression profiles.” *Proc. Nat. Acad. Sci. U.S.A.*, 98:11462–11467 (2001).
- Zavadil, J., Cermak, L., Soto-Nieves, N., and Böttinger, E. P. “Integration of TGF- β /Smad and Jagged1/Notch signalling in epithelial-to-mesenchymal transition.” *The EMBO Journal*, 23:1155–1165 (2004).