Published in final edited form as:

J Am Stat Assoc. 2009 ; 104(486): 597–607. doi:10.1198/jasa.2009.0025.

Log-Linear Models for Gene Association

Jianhua Hu, Adarsh Joshi, and Valen E. Johnson

Jianhua Hu is Assistant Professor of Biostatistics, Department of Biostatistics, University of Texas M. D. Anderson Cancer Center, 1400 Pressler Street, Houston, TX 77030 (E-mail: jhu@mdanderson.org). Adarsh Joshi is graduate student, Department of Statistics, Texas A&M University, College Station, TX 77030 (E-mail: adarsh@stat.tamu.edu). Valen E. Johnson is Professor of Biostatistics, Department of Biostatistics, University of Texas M. D. Anderson Cancer Center, 1515 Holcombe Boulevard, Unit 447, Houston, TX 77030 (E-mail: vejohanson@mdanderson.org)

Abstract

We describe a class of log-linear models for the detection of interactions in high-dimensional genomic data. This class of models leads to a Bayesian model selection algorithm that can be applied to data that have been reduced to contingency tables using ranks of observations within subjects, and discretization of these ranks within gene/network components. Many normalization issues associated with the analysis of genomic data are thereby avoided. A prior density based on Ewens' sampling distribution is used to restrict the number of interacting components assigned high posterior probability, and the calculation of posterior model probabilities is expedited by approximations based on the likelihood ratio statistic. Simulation studies are used to evaluate the efficiency of the resulting algorithm for known interaction structures. Finally, the algorithm is validated in a microarray study for which it was possible to obtain biological confirmation of detected interactions.

Keywords

Bayes factor; Deviance statistic; Ewens' sampling distribution; Gene interaction; Markov chain Monte Carlo

1. INTRODUCTION

Microarray studies and other high-throughput sequencing technology produce large amounts of biological data that contain potentially important information regarding the genesis of cancer and other diseases. Unfortunately, methods for extracting information from these data have not kept pace with the technologies used to collect it. Indeed, most statistical analyses of these data address the limited—but already formidable—problem of detecting the impact of single agents (e.g., a microarray probe cell's expression level) on a single trait (e.g., presence of breast cancer). However, the biological complexity of cell, tissue, and organ activities requires a more comprehensive understanding of how gene–gene, gene–protein, and gene–environment factors interact to cause the progression of cancer and other diseases. In this article, we propose new methodology to facilitate routine analysis of higher order interactions between measurements collected from high-throughput biological assays.

Previous approaches toward exploring interactions among explanatory variables have been based on the construction of regression models with a prespecified response variable. When the number of explanatory variables P included in the regression equation is small or moderate in size, it is possible to examine all possible interaction terms. Selection of a model can then be based on a number of model selection criteria (e.g., Akaike information criterion, C_p , or Bayesian information criterion), including Bayes factors and associated posterior model

probabilities. The Bayesian approach has the disadvantage of requiring careful specification of prior probabilities over the model space and prior densities on interaction parameters within models. In high-dimensional settings, however, it is not computationally feasible to consider all possible interactions. For example, many gene expression experiments yield in excess of 15,000 expression profiles. There are then 10^9 possible two-way interactions, and in excess of 10^{12} possible three-way interactions. A necessary focus of our methodology is thus a reduction in the dimension of the space of interaction models that we consider.

Recently, Bayesian methods have been used to formulate stochastic search algorithms for exploring high-dimensional model spaces in regression settings. An early contribution in this development was made by George and McCulloch (1993, 1997), who proposed a “stochastic search variable selection” procedure (SVSS) to determine promising subsets of explanatory variables. Lee et al. (2003) applied SVSS to probit models in which gene expression profiles were considered explanatory variables for the prediction of cancer. Subsequently, Chen et al. (2004) proposed a “conditional beta binomial” prior density for latent variables in an SVSS-type algorithm to model explicitly the presence of interaction terms. Liang et al. (2005) proposed mixtures of g-priors on regression coefficients to improve the computational efficiency of SVSS algorithms for linear regression, and also provided a discussion of the consistency and coherency of the resulting algorithms. Of more relevance to the approach taken here, Zhang and Liu (2007) constructed Bayes factors on groupings of single nucleotide polymorphism (SNP) data using a multinomial sampling model and a product Dirichlet prior model. They applied their model to datasets containing several hundred thousand SNP values to identify possible associations between subsets of SNPs. Posterior probabilities of interactions estimated from their model depend critically on a vaguely specified but very high-dimensional product Dirichlet prior model, which, as the authors note, limits the veracity of the estimated values.

Along somewhat different lines, Gaussian graphical models (GGMs; e.g., Lauritzen 1996, Jones and West 2005; Carvalho, Massam, and West 2007) have also been used to detect interactions among gene expression levels (Wu, Yong, and Kalpathi 2003; Dobra et al. 2004). We compare findings from our model with GGMs in the examples that follow.

Our proposal is based on four innovations. First, we reduce observed data structures to contingency tables. This reduction allows us to use classical log-linear models to explore and interpret interactions between arbitrary subsets of expression profiles without regard to a specified response variable. When a natural response variable is indicated, the connection between logistic regression models and log-linear models makes the extension of our methods to the logistic regression setting straightforward. Unlike several previous methods for “discretizing” genomic data (e.g., Potamias, Koumakis, and Moustakis 2004), our discretization method is based on within-subject rankings of expression levels. Because these ranked values are invariant to monotonic transformations applied to expression values at the subject level, this method of discretization eliminates many of the normalization artifacts often associated with gene expression experiments. In particular, the ranked values derived from the expression levels measured on a single microarray are invariant to monotonic transformations applied to expression levels measured from that chip.

The second innovation of our approach involves the use of Bayes factors based on the likelihood ratio statistic (Johnson 2005, 2008) to calculate the posterior odds between competing interaction models. As demonstrated in Hu and Johnson (2008), the use of Bayes factors based on test statistics simplifies the process of Bayesian model selection in linear and generalized linear models in several ways. Perhaps most importantly, this approach eliminates the need to specify prior distributions on high-dimensional parameter vectors. Instead, our model depends only on two scalar parameters. Criteria for setting the values of these parameters are discussed

in later sections, where we find that resulting estimates of posterior model probabilities are relatively insensitive to the specification of these values.

A related (and third) innovation of our model is the specification of a simple prior density on the space of cluster configurations. The particular form of the prior density is Ewens' sampling distribution, for which analytic expressions are available for prior expectations on the number of clusters. Such expressions provide important insights regarding plausible values of the single parameter associated with this distribution. We also describe a modification of this distribution that allows us to assign higher prior probability to clusters that include observed patient phenotypes.

Finally, we impose a constraint that variables can be included in only one cluster. This restriction essentially means that a variable A cannot be assumed to interact with both variables B and C unless there exists a three-way interaction between variables A , B , and C . This assumption is made primarily to improve the computational efficiency of our algorithm by eliminating the need to implement iterative estimation procedures at each parameter update. This allows the algorithm to be applied to high-throughput genomic data.

These innovations do not, of course, come without cost. One potentially important disadvantage of our method is that the discretization of continuous variables to generate high-dimensional contingency tables results in the loss of information. In addition, there is a possibility that the use of rank statistics in the construction of contingency tables might introduce associations between genes that were not present on the scale upon which data were originally recorded. Of course, similar concerns apply to any data transformation, and, as we demonstrate later, this discretization procedure ameliorates several problems associated with the normalization of expression values across experimental replicates. It also improves computational efficiency.

The most serious disadvantage of our model involves the restriction to hierarchical models, which means that the model is not able to include any variable in more than one cluster. Thus, detecting an important interaction between variables A and B may not be possible when there is an even stronger interaction between variables B and C . We address this and related issues in Sections 2.2 and 5.

2. METHODS

2.1 Contingency Table Formation

To make our discussion more concrete, let us assume that interest focuses on the analysis of P gene expression profiles collected from each of J subjects. We also restrict attention to the formation of 2^P contingency tables, although extension to K^P ($K > 2$) tables is straightforward. We note that 2^P contingency tables are particularly relevant in studies that involve two patient phenotypes (e.g., diseased and healthy).

Let n_0 and n_1 denote the number of gene expression profiles that are to be assigned values of 0 and 1, respectively. Generally, it is convenient to take $n_0 = J - n_1 \approx J/2$, although n_0 and n_1 can be matched to the number of subjects in each of two treatment groups if there is a natural dichotomization of subjects into groups. Let $x_{i,j}$, $i = 1, \dots, P$ and $j = 1, \dots, J$ denote the expression profile for gene probe i and subject j . Then we can construct a 2^P contingency table from the expression profiles according to the following procedure:

1. Within each subject j , let $r_{i,j}$ denote the rank of $x_{i,j}$ among the expression levels $x_{1,j}, \dots, x_{P,j}$ observed for that subject.

2. For each i , divide subjects into two categories containing n_0 and n_1 members. Assign subjects with the n_0 smallest values of $r_{i,j}$ a value of 0, and assign the remaining subjects a value of 1. Let $y_{i,j}$ denote the dichotomous variable obtained from this procedure. This dichotomization of subjects according to values of $x_{i,j}$ produces a 2^P contingency table. This table is invariant to monotonic transformations of intensity values $x_{i,j}$ within subjects, and is robust to distributional assumptions regarding the values of expression profiles within genes. We thus avoid many of the normalization issues typically encountered in microarray analysis.

In the event that interest focuses on determining the relation of gene clusters with a predefined response variable, we augment the contingency table constructed with gene expression values with an additional variable representing this response. We assume that a K^P contingency table has been constructed from gene expression values when the response variable has K levels.

2.2 Bayes Factors Between Nested Interaction Models

Our purpose in constructing 2^P contingency tables for expression data is to cluster genes that exhibit high associations. To do so, we analyze the resulting contingency table within the context of a log-linear model. Clusters of associated expression levels are represented within this framework as variables that are included in two- or higher way interaction terms. As stated previously, we consider only hierarchical models (e.g., Bishop, Fienberg, and Holland 1975, p. 36), which means that a cluster of p associated variables is represented by a log-linear model containing a p -way interaction term and all lower order interaction terms. When a response variable has been included in the contingency table, we interpret the clusters in which this variable are included in post hoc fashion using well-known relations between log-linear models and logistic regression models with discrete covariates (e.g., Agresti 2002, 315 ff.). Modifications of the prior distribution on cluster configurations to handle the inclusion of a response variable are described in Section 2.3.

Classic tests for the significance of interaction terms in log-linear models are often based on the likelihood ratio statistic (e.g., Bishop, Fienberg, and Holland 1975). Under a variety of sampling schemes—including Poisson, multinomial, and product multinomial sampling—the distribution of the likelihood ratio statistic for nested models is approximately that of a central chi-squared distribution when the null hypothesis corresponding to the smaller model is true. In our case, the marginal table counts are fixed by construction, which means that the sampling distributions of cell counts follow multivariate hypergeometric distributions. However, for large J , the multivariate hypergeometric distribution can be approximated by a multinomial distribution (Haberman 1974). For computational convenience, we adopt this approximation and assume that the distribution of the likelihood ratio statistic is approximately chi-squared under our sampling scheme as well.

Given values of the likelihood ratio statistic between a pair of nested log-linear models, our strategy for exploring the model space on interactions is to transit between models according to their posterior model probabilities. We denote the space of “allowable” interaction models by $\mathcal{M}$, and denote models in this space by M_k , where k is a member of an index set $\mathcal{K}$. The definition of allowable interactions is provided later. To approximate the posterior probabilities of $M_k \in \mathcal{M}$, we use results from Johnson (2005, 2008) to obtain Bayes factors between nested models based on the likelihood ratio statistic.

The analytic form of the Bayes factor based on the likelihood ratio statistic can be defined as follows. Let d denote the degrees of freedom associated with the likelihood ratio statistic, let M_1 denote the null model, and let M_2 denote the alternative model, with $M_1 \subset M_2$. If the distribution of the noncentrality parameter under the alternative model is distributed a priori as a scaled chi-squared distribution with scale parameter c , then the marginal distribution of

the likelihood ratio statistic under the alternative model is a gamma distribution with shape parameter $d/2$ and scale parameter $1/[2(cJ + 1)]$. Thus, the Bayes factor in favor of the alternative model for a given value of the likelihood ratio statistic λ can be approximated by

$$BF(M_2|M_1) = (cJ+1)^{-d/2} \exp \left\{ \frac{cJ\lambda}{2(cJ+1)} \right\}. \quad (1)$$

The Bayes factor specified in (1), together with a prior distribution on the space of interaction models, can be used to specify the posterior distribution on the space of interaction models. Letting M_0 denote the independence model and $\pi(M_k)$ the prior probability assigned to an allowable interaction model M_k , it follows that the posterior probability of model M_k can be expressed

$$p(M_k|\mathbf{x}) = \frac{BF(M_k|M_0)\pi(M_k)}{\sum_{m \in \mathcal{K}} BF(M_m|M_0)\pi(M_m)}. \quad (2)$$

To make these notions more precise, consider a saturated log-linear model in the simple case of $P = 3$. Following the notation of Bishop, Fienberg, and Holland (1975), such a model can be expressed

$$\begin{aligned} \log m_{ijk} = & u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} \\ & + u_{23(jk)} + u_{123(ijk)}, \end{aligned} \quad (3)$$

where m_{ijk} is the expected cell count for level i of variable 1, level j of variable 2, and level k of variable 3. The overall mean is represented by the parameter u ; $u_{1(i)}$, $u_{2(j)}$, and $u_{3(k)}$ are the main effects of variables 1, 2, and 3, respectively; $u_{12(ij)}$, $u_{13(ik)}$, and $u_{23(jk)}$ are the two-way interactions between variables 1 and 2, 1 and 3, and 2 and 3, respectively; and $u_{123(ijk)}$ is the three-way interaction. Each of the seven subscripted u terms sum to zero over each lettered subscript (i , j , or k), ensuring that the number of independent parameters equals the number of elementary cells. A null model M_0 can be defined as the model containing only terms u , $u_{1(i)}$, $u_{2(j)}$, and $u_{3(k)}$. Suppose that model M_1 is defined according to

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)}, \quad (4)$$

and that λ is the likelihood ratio statistic (i.e., the difference in deviances obtained from models M_1 and M_0 under an assumption of simple multinomial sampling). Then the posterior odds in favor of model M_1 over M_0 for a fixed value of c can be obtained from (2). Notationally, we associate model M_1 in (4) with cluster configuration $\{(1, 2), (3)\}$ (assuming only three gene expression levels have been collected), and we define the cardinality of this configuration, $|M_1|$, to be 2 because it contains only two clusters.

Because of computational constraints that result from the large number of interactions that are possible when P is large, we next restrict the class of permitted cluster configurations. Specifically, we exclude models in which any variable appears in two distinct clusters (e.g., a model including both interaction terms $u_{12(ij)}$ and $u_{13(ik)}$ without the encompassing higher order interaction/cluster (e.g., $u_{123(ijk)}$). We also exclude nonhierarchical models. Our

definition of hierarchical models is consistent with that used in Bishop, Fienberg, and Holland (1975), and implies that, for example, all descendent two-way interaction terms are included in a model that contains a three-way interaction term. Finally, because the number of subjects J available for detecting interactions is generally limited, we can restrict the order of interactions included in the model to be either three or four, depending on the value of J .

Extending the previous example to the case of $P = 4$, our exclusion rules imply that we allow models of the form

$$\log m_{ijkl} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{4(l)} + u_{12(ij)} + u_{34(lk)}, \quad (5)$$

models of the form (4), as well as the saturated model with all interaction terms. We do not consider cluster configurations of the form $\{(1, 2), (2, 3)\}$, which correspond to the log-linear model

$$\log m_{ijkl} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{4(l)} + u_{12(ij)} + u_{23(jk)}, \quad (6)$$

because variable 2 appears in two clusters without the encompassing $(1, 2, 3)$ cluster. Thus, it is possible that the restriction to hierarchical models might result in nondetection of a two-way interaction in (6), or the false detection of a three-way interaction. In the latter case, post hoc inspection of marginal tables can be used to clarify the association between variables detected within the same (hierarchical) cluster. This technique is illustrated in the examples that follow.

The restriction that we place on allowable model configurations does, however, provide two important advantages. First, when considering a test between a model that contains a specified cluster and a nested model obtained by the split of that cluster (e.g., $\{(1, 2, 3)\}$ vs. $\{(1, 2), (3)\}$), we are able to collapse the 2^P table across all variables not contained in the larger cluster to obtain the likelihood ratio statistic between the nested models. Thus, it is unnecessary to compute expected cell counts in all 2^P cells, which is generally not feasible. Second, and related to the first, direct estimates of expected cell counts are available for this class of models, which eliminates the need to perform iterative procedures to obtain the likelihood ratio statistic. Formulas for the direct estimates of cell counts required to compute likelihood ratio statistics are provided in, for example, Bishop, Fienberg, and Holland (1975).

2.3 A Prior on the Space of Interaction Models

The potentially large number of interaction terms included in our model necessitates the implementation of prior constraints on the total number of nonsingleton clusters. In the absence of a response variable and assuming exchangeability of all gene expression values, a convenient prior distribution for this purpose is provided by Ewens' sampling distribution (Ewens 1972). This distribution defines a probability distribution on the set of all partitions of a set. For a set with n elements, let a_1 denote the number of partitions (i.e., clusters) containing exactly one element; a_2 , the number of partitions containing exactly two elements; and so forth. Then for a given value of the parameter θ , the probability assigned by Ewens' sampling distribution to a partition of the set is specified as

$$Pr(a_1, \dots, a_n) = \frac{n!}{\theta(\theta+1) \cdots (\theta+n-1)} \prod_{j=1}^n \frac{\theta_j^a}{j^{a_j} a_j!}. \quad (7)$$

Here, $a_1, \dots, a_n$ is a sequence of nonnegative integers that satisfies

$$a_1 + 2a_2 + 3a_3 + \cdots + na_n = n. \quad (8)$$

In actual applications, the number of subjects available for inclusion in an analysis requires us to truncate Ewens' sampling distribution by imposing a constraint that $a_k = \cdots a_n = 0$ for a moderate value of k . Such a constraint limits the number of elementary cells used to construct likelihood ratio statistics to be, at most, 2^{k-1} cells, which in turn implies that the expected cell counts under the independence model is approximately $J/2^{k-1}$.

Tavare and Ewens (1995) provide analytic expressions for the expected number of singleton clusters and the total number of clusters K as a function of n and θ for Ewens' distribution. These expressions are

$$E(a_1) = \frac{n\theta}{n+\theta-1} \quad (9)$$

and

$$E(K) = \sum_{i=0}^{n-1} \frac{\theta}{\theta+1}. \quad (10)$$

From these equations, we are able to select values of θ to constrain the number of nonsingleton clusters expected a priori or, equivalently, to specify the prior expected proportion of singleton clusters.

Figure 1 depicts $E(a_1)$ and $r \equiv E(a_1)/E(K)$ for a range of θ values for three values of n used in the following examples. We note that the expected proportion of singleton clusters converges to 1 as θ increases. As a general strategy, we seek values of θ so that $r \approx 0.8$; such values are indicated by the horizontal line in Figure 1.

When a response variable is included, we modify Ewing's prior distribution on cluster configurations to account for a separate prior specification on the probability that the response variable is included in a nonsingleton cluster. If M_k denotes the cluster configuration obtained by deleting the response variable y from configuration M_k^* , and $\pi(M_k)$ denotes Ewing's prior density on configuration M_k , then we define $\pi^*(M_k^*)$ to be a modification of Ewen's prior density that can be expressed as

$$\pi^*(M_k^*) = \frac{\rho \pi(M_k)}{|M_k|} (1 - I_{\{y\}}) + (1 - \rho) \pi(M_k) I_{\{y\}}. \quad (11)$$

Here, $I_{\{y\}}$ indicates whether y appears in M_k^* as a singleton cluster. Note that equal prior probability is assigned to the inclusion of y in each of the $|M_k|$ gene clusters, and $(1 - \rho)$ is the probability that y is not associated with any gene cluster.

3. EXPLORATION OF THE PARAMETER SPACE

A complete enumeration of probabilities in (2) is not feasible even for moderate values of P . For this reason, we resort to a Markov chain Monte Carlo (MCMC) algorithm to identify cluster configurations that are assigned high posterior probability.

We take as our initial state of the Markov chain the null model containing only singleton cliques (i.e., the independence model). At each subsequent update of the chain, we first randomly choose to consider either a split of an existing cluster or a merge of two existing clusters. Depending on whether we are considering a split or merge, we calculate the likelihood ratio statistic between the existing and proposed state. Response variables, if included in the model, are treated the same as gene expression values in the calculation of the likelihood ratio statistic. Note that this calculation is based only on counts in a 2^k contingency table, where k is the maximum number of elements contained in any of the relevant clusters. Posterior odds of each state are then determined according to (2), and transitions between states are made accordingly.

More specifically, the steps in this MCMC algorithm can be described as follows:

- 1 Initialize the chain so that each variable forms its own cluster.
- 2 With probability P_s , randomly select a cluster c in the current model configuration—say, M_k —to split potentially. With probability $1 - P_s$, randomly select two clusters c_1 and c_2 to potentially merge.
- 3 If the cluster chosen to split contains only one variable, or if the two clusters considered for merge contain more than the allowed number of variables, record the current state and return to Step 2.
- 4 Let M' denote the proposed model state. If considering a split, define the proposal density $Q(M'|M_k)$ to be

$$Q(M'|M_k) = \frac{P_s}{|M_k|2^{1-|c|}},$$

and the (reverse) proposal density to be

$$Q(M_k|M') = \frac{(1 - P_s)}{\binom{|M_k|+1}{2}}.$$

These proposal densities reflect an equiprobable random split of a cluster c containing $|c|$ variables in model configuration M_k , which contains $|M_k|$ clusters, into one of $2^{|c|-1}$ configurations of two subclusters. If considering a merge of clusters c_1 and c_2 , the proposal densities are

$$Q(M'|M_k) = \frac{(1-P_s)}{\binom{|M_k|}{2}} \text{ and } Q(M_k|M') = \frac{P_s}{(|M_k|+1)2^{1-|c_1|+|c_2|}}.$$

- 5 Compute the Bayes factor BF based on the likelihood ratio statistic between the nested log-linear models defined by the current and candidate states according to (1).
- 6 Let the Ewens prior probability specified in (7) (or the modification of it to account for a response variable given in (11)) for the current and candidate states be denoted by π_k and π' , respectively. Transition to the candidate configuration with probability equal to

$$\min \left(1, BF \times \frac{Q(M_k|M')}{Q(M'|M_k)} \times \frac{\pi'}{\pi} \right) \text{ if a merge is considered or } \min \left(1, \frac{1}{BF} \times \frac{Q(M_k|M')}{Q(M'|M_k)} \times \frac{\pi'}{\pi} \right) \text{ if a split is considered.} \quad (12)$$

- 7 Record the current cluster configuration and return to Step 2 until a sufficient number of cluster configurations have been sampled.

After an initial burn-in period, the simulated chain of model configurations can be used to perform approximate inference regarding the posterior probability that specific cluster configurations occur in the model.

In the following examples, we set $P_s = 0.5$. Other values of P_s might be considered to improve mixing, particularly when the number of variables is large.

4. SIMULATION STUDIES

To test the performance of the algorithm described in Section 3, we conducted a simulation study in which several strongly interacting variables were seeded into a contingency table containing a total of 20 binary variables with 120 joint observations on each. Interacting variables were organized into three clusters of different sizes and included clusters $\{(1, 2, 3, 4), (5, 6), (7, 8, 9)\}$ (see Table 1). In cluster (5, 6), the variables were assumed to be positively correlated and the deviance statistic was 58.22 in a test for their associated interaction term. The (7, 8, 9) cluster was generated so that two variables had very high positive correlation only when the third variable was at its lowest level. Note that cluster (7, 8) had a positive interaction, which was associated with a deviance statistic of 107.57. The value of the deviance statistic for a test for merging cluster (7, 8) and (9) was 120.75 on 3 degrees of freedom. The four-way interaction was generated along similar lines. No structure was imposed on the remaining variables, which were generated randomly.

Ewens' sampling distribution with $\theta = 33$ was used to impose a prior constraint that the expected proportion of singleton clusters would be 80%. Based on this prior and the data configuration just described, we estimated the posterior probability of clusters using five million iterations of the MCMC algorithm with $P_s = 0.5$. Table 2 shows the 10 clusters estimated to have the highest posterior probabilities when the scale parameter used to convert the deviance statistic to a Bayes factor was assigned values of $c = 1, 2$, and 3. All choices of c led to the assignment of high posterior probability to the three "true" clusters, although larger values of c tend to favor configurations with smaller clusters.

We investigated the convergence of the MCMC algorithm and the numerical accuracy of the probabilities displayed in Table 2 by rerunning the algorithm using different initial configurations. For high-probability clusters, we observed excellent agreement between the estimates of posterior probabilities obtained from several chains started from disparate configurations. Column 6 of Table 2 (headed by an asterisk) displays estimated posterior probabilities obtained from a chain initialized randomly to have five clusters each with four variables. The similarity of estimated posterior probabilities for the high-probability clusters from this chain and the chain initialized with all singletons suggests that five million MCMC updates are adequate to identify the high-probability clusters included in these data and to approximate their model inclusion probabilities, at least to within the accuracy of the proposed probability model. Five million updates required approximately 1.5 hours on a standard workstation.

As a cautionary note, we observe that this MCMC algorithm is unlikely to detect high-order interaction terms between a set of genes when all the lower order marginal interactions within the set are small. That is, the constituent clusters required for a merge operation are unlikely to form. In high-dimensional data, this problem seems insurmountable given the combinatorial challenge of searching $\binom{P}{k}$ combinations of clusters of size k for $k > 2$.

In the final two columns of Table 2, we present simulation results obtained by supplementing the original 20-gene experiment with an additional 980 randomly generated “genes.” In the expanded experiment, we set $\theta = 1,700$ again to limit the expected proportion of singleton clusters to 80%. Ten million iterations of the MCMC algorithm were performed to obtain the estimated posterior probabilities depicted in Table 2. As before, the three true clusters are detected with high posterior probability, although three spurious triples are assigned higher posterior probability than the true doubleton (5, 6). Our decision to sample 10 million values from the posterior distribution was again based on a comparison of the posterior probabilities assigned by several chains to the highest probability clusters. After that number of iterations, the estimates of these probabilities agreed to approximately within 0.01 for the high-probability clusters.

We next used the same data to compare results obtained from our algorithm with interactions detected using a GGM implemented in the R package *GeneNet*. To convert the discrete data simulated for our model to continuous values for use with *GeneNet*, we arbitrarily constructed a 120×20 matrix with entries generated randomly from an exponential distribution with mean equal to 1. Clearly, other continuous distributions (or mixtures of continuous distributions) might have been used to generate continuous values, but we felt that an exponential model provided a realistic reflection of the degree of non-normality typically encountered in gene expression data, without introducing extreme outliers that might be associated with heavier tailed distributions. After generating the continuous expression values, we rearranged the data so that the 60 smallest values in each column corresponded to zeros in the discrete data matrix, and the 60 largest values corresponded to discrete values of 1. The order of values within each group of 60 was random.

Table 3 provides a summary of the GGM fit obtained from *GeneNet*. We computed the empirical posterior probabilities and false discovery rate (FDR) probabilities listed in Table 3 using methodology described in Schafer and Strimmer (2005). From Table 3, we see that clusters (7, 8, 9) and (5, 6) were detected with high empirical posterior probability, whereas the four-way interaction represented by cluster (1, 2, 3, 4) was not, even though there were strong interactions between, for example, variables 2 and 4.

5. BREAST CANCER DATA

Our purpose in developing a statistical model for cluster configurations is to facilitate the analysis of gene expression data of the type presented in West et al. (2001). Those data consist of gene expression values measured on 7,129 genes sampled from 49 breast tumor tissue specimens. Twenty-five of these specimens were taken from patients diagnosed as estrogen receptor positive (ER+), whereas 24 samples were collected from estrogen receptor negative (ER-) patients. The raw gene expression data were normalized, background adjusted, and summarized using the robust multi-array analysis (RMA) procedure proposed by Irizarry et al. (2003).

There is a vast literature describing genetic biomarkers associated with breast cancer using microarray/SNP data, and a large number of genes are now suspected of having an association with breast cancer. However, the study of gene-gene interactions among these genes is limited, and one motivation for the current analysis is the exploration of interactions between previously implicated biomarkers of breast cancer. We therefore focus our study on 71 genes in which previous studies have indicated an association with some form of breast cancer (e.g., Cooper 2001; Spurdle et al. 2007).

As a first step in this analysis, we discretized the gene expression data using the method described in Section 2.1. Gene probes on each sample array were first ranked according to their expression values within the array. For each gene probe, we defined a binary variable for subject i to be 1 if the rank of that expression level was among the top 24 ranks observed for that gene probe, and 0 otherwise. Because only 49 samples were available for this study, we restricted our investigation to four-way interactions. Collapsed tables associated with four-way interactions thus had expected cell counts under the independence model of approximately three; such models clearly straddle the boundary defining the applicability of the chi-squared approximation to the likelihood ratio statistic.

We applied the MCMC algorithm described in Section 3 to obtain 10 million samples from the posterior distribution on cluster configurations. The Ewen's prior parameter was set to $\theta = 119$, which implies that the prior proportion r of non-singleton clusters is no larger than 0.2. As before, several values of c were evaluated. Results from this algorithm are displayed in Table 4, where the 10 clusters estimated to have the highest posterior probabilities are depicted. The posterior probabilities displayed in column 5 were estimated from a chain that was started from an all-singleton configuration; these estimates are consistent with values obtained using several other starting configurations. For example, column 6 of Table 4 (headed by an asterisk) displays estimated posterior probabilities obtained from a chain initialized randomly with 10 clusters each containing four variables and all remaining variables in singleton clusters ($c = 2$).

From this table we see that the two-gene clusters (*BRCA1*, *JNK*) and (*HSF1*, *P53a*) were detected with high posterior probabilities and would appear in the median probability model for all values of c . In addition, cluster (*ESR1*, *GATA3*) exhibits a high posterior probability for $c = 1, 2$. For $c = 3$, (*ESR1*, *GATA3*) is assigned a posterior probability of 0.457, and also appears in a fairly high-probability (0.438) three-variable subcluster (*GSTP1a*, *ESR1*, *GATA3*).

At this same value of c , cluster (*AR*, *FOXAI*) is assigned posterior probability 0.476, whereas for $c = 2, 3$, the three-gene cluster (*GSTP1a*, *AR*, *FOXAI*) is assigned posterior probabilities of 0.389 and 0.350, respectively. When not in this three-gene cluster, the subcluster (*AR*, *FOXAI*) appears as a doublet with probabilities 0.287 and 0.393 for $c = 1, 2$, respectively. Thus, the total posterior probability that (*AR*, *FOXAI*) appears in a model exceeds 50% for all values of c . Taken together, these observations suggest that there may be a more complicated interaction structure between the five genes *GSTP1a*, *ESR1*, *AR*, *GATA3*, and *FOXAI*.

Collapsed three-way contingency tables for the two clusters (*GSTP1a*, *ESR1*, *GATA3*) and (*GSTP1a*, *AR*, *FOXA1*) are depicted in Table 5. From these tables, we see that gene *GSTP1a* plays a similar role in both clusters. Loosely speaking, the three-gene clusters (*GSTP1a*, *ESR1*, *GATA3*) and (*GSTP1a*, *AR*, *FOXA1*) appear to compete for gene *GSTP1a*, which under our model is allowed to appear in only one interaction model.

For reasons discussed later, we also wish to examine the posterior probability that genes *ESR1*, *GATA3*, and *FOXA1* form an interaction cluster. To eliminate competition between this cluster and the triplets discussed earlier, we arbitrarily refit the model after excluding genes *GSTP1a* and *AR*. We then reran the MCMC algorithm on the remaining 69 genes after resetting $\theta = 116$ (again, to bound the expected prior proportion of nonsingleton clusters by 0.2).

Results from this algorithm are displayed in Table 6. As suspected, cluster (*ESR1*, *GATA3*, *FOXA1*) appears with high posterior probability for each choice of c , with posterior probability ranging between 0.742 and 0.911. The three-way contingency table displayed in Table 7 suggests that both genes *ESR1* and *GATA3* coexpress at the low expression level when gene *FOXA1* is downregulated, and coexpress at the high expression level when gene *FOXA1* is upregulated. Again, the two gene–gene interactions (*BRCA1*, *JNK*) and (*HSF1*, *P53a*) were also assigned high posterior probability and would appear in the median probability model for each value of c .

Another interesting three-gene cluster is (*KRAS*, *CYR61*, *ATF*), which is also assigned relatively high posterior probability for $c = 2, 3$. (This cluster was also assigned high probability in Table 4.) When this three-variable cluster is absent from a model, the subcluster (*CYR61*, *ATF*) is consistently assigned high posterior probability.

The interaction pattern between genes *KRAS*, *CYR61*, and *ATF* can be better understood by an examination of their collapsed contingency table, which is also provided in Table 7. This table demonstrates that the association between genes *CYR61* and *ATF* is highest when gene *KRAS* is not highly expressed, and diminishes as the expression level of gene *KRAS* increases. This three-way interaction does not naturally decompose into submodels containing only lower order interaction terms.

To determine whether the gene clusters identified by our model are biologically important, we used the PubMed search engine (PubMed 2007) to identify functions associated with identified gene clusters. As it happens, most of the gene–gene interactions detected by our algorithm have been previously reported in the bioinformatics literature.

In cluster (*BRCA1*, *JNK*), *BRCA1* is a well-known gene associated with breast cancer that acts as a transcription factor (Narod et al. 1993). Harkin et al. (1999) pointed out that expression of *JNK* is induced by activation of *BRCA1*.

The high coexpression pattern between *HSF1* and *P53a* has been described in Quenneville et al. (2002).

Identification of the three-gene cluster (*GSTP1a*, *ESR1*, *AR*) is more interesting. Panet-Raymond et al. (2000) found a significant interaction between ER *ESR1* and *AR*. Although we have not found a previous, independent verification of a biological link between *GSTP1a* and *ESR1*, the two genes are both involved in a biological pathway involved in carcinogen/estrogen metabolism. Indeed, the Gene Network Database (Genetwork 2007) suggests that the two genes are indirectly correlated through $ESR1 \rightleftharpoons MNAT1 \rightleftharpoons PSMC4 \rightleftharpoons GSTP1$. Nevertheless, the interaction between these three genes, or indeed among the larger group of genes (*GSTP1a*, *ESR1*, *AR*, *GATA3*, *FOXA1*), lacks formal validation from laboratory procedures.

Interactions of genes in the cluster (*ESR1*, *GATA3*, *FOXA1*) is well known and motivated us to determine whether this interaction would be detected if competing genes *GSTP1a* and *AR* were removed from the model. ER alpha (*ESR1*) plays a key role in the development and progression of breast cancer. It also plays a role in regulating the growth of some breast cancer cells in the presence of estrogen, and may inhibit growth in the presence of antiestrogens like tamoxifen. Ferguson and Davidson (1997) demonstrated that normal control of ER gene expression/function is lost in breast cancer. The *GATA3* protein is able to bind to *ESR1*, thereby altering its function (Lacroix and Leclercq 2004; Eeckhoutte et al. 2007). The transcription factor Forkhead-box A1 (*FOXA1*), a member of the FOX class of transcription factors, has been implicated in the pathogenesis of several types of cancers. Extensive literature (e.g., Bertucci et al. 2000; Lacroix and Leclercq 2004) has confirmed the strong interaction among these three genes.

The three-gene cluster (*KRAS*, *CYR61*, *ATF*) was also assigned high posterior probability by our algorithm. One of this cluster's genes, *CYR61*, is a secreted, cysteine-rich heparin-binding protein encoded by a growth factor-inducible early gene that acts as an extracellular, matrix-associated signaling molecule promoting the adhesion of endothelial cells through interaction with integrin $\alpha_V\alpha_3$. Decrease in *CYR61* is associated with the growth of cancer cells in, for example, endometrial cancer cells. Studies also show that mutation often occurs within gene *KRAS*, which encodes an oncogene-associated protein. Chien et al. (2004) pointed out that low levels of *CYR61* can be induced by aberrant expression levels of transcription factors such as *KRAS* (e.g., during carcinogenesis of endometrial adenocarcinomas). On the other hand, *ATF* belongs to the *ATF/CREB* family, whose members can "homodimerize" to repress transcription from promoters containing *ATF*-binding sites. *ATF3* can be induced by stress, such as exposure to carbon tetrachloride (CCl_4), and may have a key role in modulating cellular response to radiation and chemical exposure. We note that the two genes *CYR61* and *ATF3* are also indirectly connected through reactions $\text{CYR61} \rightleftharpoons \text{FGFR1} \rightleftharpoons \text{MAPK10} \rightleftharpoons \text{TP53} \rightleftharpoons \text{ATF3}$.

The preceding analysis focused on detecting interactions among genes. As described in Section 3, a slight modification on the prior probabilities assigned to clusters enables us to examine the association between gene expression values and tissue phenotype. To illustrate this capability, we next included ER status into the gene expression contingency table and reran the modified MCMC algorithm. Parameter values were fixed at $\theta = 119$, $c = 2$, and $\rho = 0.5$. A value of $\rho = 0.5$ corresponds to the assignment of even prior odds to the event that ER status is associated with a gene association cluster. Resulting posterior probability estimates of the occurrence of high-probability clusters are shown in the rightmost column of Table 4. As expected, the association of ER status with gene *ESR1* is quite strong, and this gene appears in most of the clusters associated with ER status. The strongest association between ER status and an single cluster of genes occurs with the cluster (*ESR1*, *GATA3*, *FOXA1*); this grouping is assigned approximate posterior probability 0.739. West et al. (2001) previously documented the marginal association between ER status and each of these three genes; our model facilitates examination of higher order interactions.

Table 8 provides a cross-tabulation of the 49 tissue samples by ER status and genes *ESR1*, *GATA3*, and *FOXA1*. Because not all cells are occupied, it is not possible to estimate the logistic regression coefficients of all components of the three-way interaction of these genes in predicting ER status. However, the strong effect of having all three genes at either high or low expression levels is apparent from the table. When all three genes are at the low expression level, 19 out of 20 specimens are ER-. When all three are highly expressed, 20 out of 20 specimens are ER+.

To facilitate comparison with other methods, we also applied the GGM algorithm of Schafer and Strimmer (2005) to these data. Results extracted from the *GeneNet* algorithm are displayed in Table 9. The pairwise associations reflected in this table are largely consistent with the clusters detected using our algorithm. The GGM identifies (*CYR61*, *ATF*) as a potential important association, but fails to identify the association of these gene probes with *KRAS*. Similarly, correlations between (*ESR1*, *GATA3*) and (*AR*, *FOXAI*) are identified, but there is no indication that these two pairs of variables might be related to each other and gene *GSTP1a*. On the other hand, the GGM model identifies several pairwise correlations that were not detected by the cluster model. We were not able to identify these pairwise correlations in a cursory search of PubMed, although this does not necessarily mean that these interactions are not biologically significant.

Finally, we note that the similarity of results obtained using our method and GGM for genes known to be associated suggests that little information was lost in the discretization procedure used in the preprocessing step of our method.

6. DISCUSSION

Throughout this article, our application focus has been the analysis of gene expression levels obtained from microarray data. Within this context, interaction terms detected by our method were confirmed as having a biological basis. In the future, we hope to extend this methodology to other applications, particularly to the analysis of SNP data and the inclusion of environmental and disease status variables.

The examples illustrated in this article were performed using “convenience code” written in either MATLAB or R. To extend the proposed methodology to larger datasets, and in particular to SNP data, we are currently developing extensions of the MCMC algorithm that will allow the algorithm to run in parallel on multiple processors. This will be accomplished by periodic random partition and reassignment of existing clusters to distinct processors. Merge and split operations will then be performed independently within the variables assigned to each processor. Maintenance of detailed balance in the “parallelized” version of the algorithm will be possible because of our restriction to hierarchical cluster configurations. Codes written in high-level computer languages will be converted and optimized to run in lower level languages (e.g., C).

Additional research is required before the model can be expanded to handle more general interaction structures. In particular, we are currently developing methodology to accommodate nonhierarchical models. To the extent that such expansion proves computationally feasible, it will allow us to handle more coherently higher order interactions of the type detected in the breast cancer data.

Finally, we plan to develop methodology that will allow us to exploit more efficiently the particular form of the interaction structures detected in the log-linear models in the follow-on specification of network models between expressed pathways.

Acknowledgments

This research work is supported in part by the National Science Foundation (grant DMS-0706818) and the National Institutes of Health (grant R01 RGM080503A). The authors thank David Banks, an anonymous associate editor, and a referee for a number of comments that significantly improved the content of this article.

References

Agresti, A. Categorical Data Analysis. Hoboken, N.J: Wiley; 2002.

- Bertucci F, Houlgatte R, Benziene A, Granjeaud S, Adelaide J, Tagett R, Loriod B, Jacquemier J, Viens P, Jordan B, Birnbaum D, Nguyen C. Gene Expression Profiling of Primary Breast Carcinomas Using Arrays of Candidate Genes. *Human Molecular Genetics* 2000;9:2981–2991. [PubMed: 11115842]
- Bishop, YMM.; Fienberg, SE.; Holland, PW. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, Mass: MIT Press; 1975.
- Carvalho CM, Massam H, West M. Simulation of Hyper-inverse Wishart Distributions in Graphical Models. *Biometrika* 2007;94:647–659.
- Chen, W.; Ghosh, D.; Raghunathan, T.; Kardia, S. A Bayesian Method for Finding Interactions in Genomic Studies. The University of Michigan Department of Biostatistics Working Paper Series . 2004. www.bepress.com/umichbiostat/paper48
- Chien W, Kumagai T, Miller CW, Desmond JC, Frank JM, Said JW, Koeffler HP. Cyr61 Suppresses Growth of Human Endometrial Cancer Cells. *Journal of Biological Chemistry* 2004;79:53087–53096. [PubMed: 15471875]
- Cooper CS. Applications of Microarray Technology in Breast Cancer Research. *Breast Cancer Research* 2001;3:158–175. [PubMed: 11305951]
- Dobra A, Han C, Jones B, Nevins J, Yao G, West M. Sparse Graphical Models for Exploring Gene Expression Data. *Journal of Multivariate Analysis* 2004;90:196–212.
- Eeckhoutte J, Keeton EK, Lupien M, Krum SA, Carroll JS, Brown M. Positive Cross-Regulatory Loop Ties GATA-3 to Estrogen Receptor Alpha Expression in Breast Cancer. *Cancer Research* 2007;67:6477–6483. [PubMed: 17616709]
- Ewens W. The Sampling Theory of Selectively Neutral Alleles. *Theoretical Population Biology* 1972;3:87–112. [PubMed: 4667078]
- Ferguson AT, Davidson NE. Regulation of Estrogen Receptor Alpha Function in Breast Cancer. *Critical Reviews in Oncogenesis* 1997;8:29–46. [PubMed: 9516085]
- Genetwork. 2007 . <http://humgen.med.uu.nl/lude/genenetwork/>
- George EI, McCulloch RE. Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association* 1993;88:881–889.
- George EI, McCulloch RE. Approaches for Bayesian Variable Selection. *Statistica Sinica* 1997;7:339–374.
- Haberman, SJ. *The Analysis of Frequency Data*. Chicago, Ill: University of Chicago Press; 1974.
- Harkin DP, Bean JM, Miklos D, Song YH, Truong VB, Englert C, Christians FC, Ellisen LW, Maheswaran S, Oliner JD, Haber DA. Induction of GADD45 and JNK/SAPK-Dependent Apoptosis Following Inducible Expression of BRCA1. *Cell* 1999;28:575–586. [PubMed: 10367887]
- Hu J, Johnson VE. Bayesian Model Selection Using Test Statistics. *Journal of the Royal Statistical Society, Ser B, JRSS-B* 2008;71:143–158.
- Irizarry, RA.; Hobbs, B.; Collin, F.; Beazer-Barclay, YD.; Antonellis, KJ.; Scherf, U.; Speed, TP. *Biostatistics. Vol. 4*. Oxford; England: 2003. Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data; p. 249-264.
- Johnson VE. Bayes Factors Based on Test Statistics. *Journal of the Royal Statistical Society, Ser B* 2005;67:689–701.
- Johnson VE. Properties of Bayes Factors Based on Test Statistics. *Scandinavian Journal of Statistics* 2008;35:354–368.
- Jones B, West M. Covariance Decomposition in Undirected Gaussian Graphical Models. *Biometrika* 2005;92:779–786.
- Lacroix M, Leclercq G. About GATA3, HNF3A, and XBP1: Three Genes Co-expressed With the Oestrogen Receptor-Alpha Gene (ESR1) in Breast Cancer. *Molecular and Cellular Endocrinology* 2004;219:1–7. [PubMed: 15149721]
- Lauritzen, SL. *Graphical Models*. Oxford: Clarendon Press; 1996.
- Lee, KE.; Sha, N.; Dougherty, ER.; Vannucci, M.; Mallick, B. *Bioinformatics. Vol. 19*. Oxford; England: 2003. Gene Selection: A Bayesian Variable Selection Approach; p. 90-97.
- Liang, F.; Paulo, R.; Molina, G.; Clyde, MA.; Berger, JO. ISDS Discussion Paper. Vol. 05–12. Durham, NC: Duke University; 2005. Mixtures of g-Priors for Bayesian Variable Selection.

- Narod S, Lynch H, Conway T, Watson P, Feunteun J, Lenoir G. Increasing Incidence of Breast Cancer in Family With BRCA1 Mutation. *Lancet* 1993;24:1101–1102. [PubMed: 8097005]
- Panet-Raymond V, Gottlieb B, Beitel LK, Pinsky L, Trifiro MA. Interactions Between Androgen and Estrogen Receptors and the Effects on Their Transactivational Properties. *Molecular and Cellular Endocrinology* 2000;25:139–150. [PubMed: 11000528]
- Potamias, G.; Koumakis, L.; Moustakis, V. *Methods and Applications of Artificial Intelligence*. Berlin: Springer; 2004. Gene Selection via Discretized Gene-Expression Profiles and Greedy Feature-Elimination; p. 3025
- Pubmed. National Library of Medicine and National Institutes of Health . 2007.
www.ncbi.nlm.nih.gov/sites/entrez
- Quenneville LA, Trotter MJ, Maeda T, Tron VA. P53-Dependent Regulation of Heat Shock Protein 72. *British Journal of Dermatology* 2002;146:786–791. [PubMed: 12000374]
- Schafer, J.; Strimmer, K. *Bioinformatics*. Vol. 21. Oxford; England: 2005. An Empirical Bayes Approach to Inferring Large-Scale Gene Association Networks; p. 754-764.
- Spurdle AB, Chang J, Byrnes GB, Chen X, Dite GS, McCredie MRE, Giles GG, Southey MC, Chenevix-Trench G, Hopper JL. A Systematic Approach to Analysing Gene–Gene Interactions: Polymorphisms at the Microsomal Epoxide Hydrolase EPHX and Glutathione S-transferase GSTM1, GSTT1, and GSTP1 Loci and Breast Cancer Risk. *Cancer Epidemiology, Biomarkers & Prevention* 2007;16:769–774.
- Tavare, S.; Ewens, WJ. The Ewens Sampling Formula. In: Johnson, NL.; Balakrishanan, N., editors. *Multivariate Discrete Distributions*. New York: Wiley; 1997.
- West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson JAJr, Marks JR, Nevins JR. Predicting the Clinical Status of Human Breast Cancer by Using Gene Expression Profiles. *Proceedings of the National Academy of Sciences of the United States of America* 2001;98:11462–11467. [PubMed: 11562467]
- Wu, X.; Yong, Y.; Kalpathi, RS. 3rd ACM SIGKDD Workshop on Data Mining in Bioinformatics. Troy, NY: Rensselaer Polytechnic Institute; 2003. Interactive Analysis of Gene Interactions Using Graphical Gaussian Model; p. 63-69.
- Zhang Y, Liu JS. Bayesian Inference of Epistatic Interactions in Case–Control Studies. *Nature Genetics* 2007;39:1167–1173. [PubMed: 17721534]

Figure 1.

The expected proportion of singleton clusters among the expected number of all the clusters versus the hyperparameter θ in Ewens' probability distribution.

Table 1
Collapsed table for three strongly interacting clusters: (1, 2, 3, 4), (5, 6), and (7, 8, 9).

Cluster (5, 6)									
$x_5 = 0$ 1	$x_6 = 0$	1							
	50	10							
	10	50							
Cluster (7, 8, 9)									
	$x_9 = 0$	$x_9 = 1$							
$x_7 = 0$ 1	$x_8 = 0$	1	$x_8 = 0$	1					
	54	2	0	2	2				
	2	2	1	2	54				
Cluster (1, 2, 3, 4)									
	$x_3 = 0, x_4 = 0$	$x_3 = 1, x_4 = 0$				$x_3 = 0, x_4 = 1$		$x_3 = 1, x_4 = 1$	
$x_1 = 0$ 1	$x_2 = 0$	1	$x_2 = 0$	1	$x_2 = 0$	1	$x_2 = 0$	1	$x_2 = 0$
	43	1	0	1	6	0	6	1	0
	1	1	1	6	1	1	1	6	1
43									

Table 2
The 10 (nonsingleton) clusters with the highest posterior probabilities versus different choices of c in a simulation study with 120 observations.

Rank	20 Genes					1,000 Genes				
	$c = 1$	$c = 2$	*	$c = 3$		$c = 2$		$c = 2$		
1	1,2,3,4	1	1	1,2,3,4	1	7,8,9	1	7,8,9	0.982	
2	7,8,9	1	1	7,8,9	1	1,2,3,4	1	1,2,3,4	0.975	
3	5,6	0.995	0.998	5,6	0.999	465,737,871		465,737,871	0.955	
4	14,15	0.176	0.160	14,15	0.146	587,743,809		587,743,809	0.920	
5	12,19	0.091	0.080	12,19	0.074	500,503,877		500,503,877	0.882	
6	16,17	0.055	$<1 \times 10^{-4}$	12,20	1×10^{-4}	5,6		5,6	0.757	
7	12,18	0.050	$<1 \times 10^{-4}$	13,17	1×10^{-4}	35,347,859		35,347,859	0.725	
8	13,16	1×10^{-4}	$<1 \times 10^{-4}$	17,18	1×10^{-4}	230,885		230,885	0.646	
9	14,20	1×10^{-4}	$<1 \times 10^{-4}$	12,13	$<1 \times 10^{-4}$	49,89		49,89	0.629	
10	12,13	1×10^{-4}	0.005	16,19	$<1 \times 10^{-4}$	189,380		189,380	0.625	

NOTE: The rightmost column contains the result with 1,000 genes. All the other columns contain the results with 20 genes.

Table 3

Summary of results from the GGM.

Rank	Doubleton	<i>p</i> -Value	<i>q</i> -Value	Empirical posterior probability
1	5,6	<0.0001	<0.0001	0.999
2	7,9	<0.0001	0.0009	0.987
3	7,8	0.0001	0.0068	0.987
4	1,2	0.0001	0.0068	0.767
5	1,3	0.0002	0.0610	0.767
6	8,9	0.0027	0.0822	0.432
7	2,4	0.0054	0.1447	<0.0001
8	12,16	0.0063	0.1643	<0.0001
9	2,8	0.0140	0.3039	<0.0001
10	1,4	0.0151	0.3208	<0.0001

NOTE: Pairs of nodes that are estimated to have high probability are listed in the second column of the table. The third column provides the *p*-value for the test of zero partial correlation. The fourth column provides the FDR-based *q*-value, which accounts for multiplicity in testing. The final column provides an estimate of the empirical posterior probability that the indicated nodes are connected in the GGM.

Table 4
Nonsingleton breast cancer gene clusters assigned highest posterior probability.

	c = 1	c = 2		*	c = 3		Association with ER		
1	BRCA1,JNK	0.597	BRCA1,JNK	0.694	0.683	BRCA1,JNK	0.763	ESR1,GATA3,FOXA1	0.739
2	HSF1,P53a	0.575	HSF1,P53a	0.635	0.627	HSF1,P53a	0.645	ESR1,AR	0.141
3	GSTP1b,AKT2	0.432	ESR1,GATA3	0.532	0.538	KRAS,CYR61,ATF	0.479	GSTP1a,ESR1,AR	0.097
4	ESR1,GATA3	0.422	GSTP1b,AKT2	0.500	0.467	AR,FOXA1	0.476	ESR1	0.014
5	CDKN2B,TSG101	0.304	AR,FOXA1	0.393	0.389	GSTP1b,AKT2	0.474	ESR1,BCL2	0.002
6	AR,FOXA1	0.287	GSTP1a,AR,FOXA1	0.389	0.382	ESR1,GATA3	0.457	AR	0.002
7	ESRRA,ERBB2	0.283	KRAS,CYR61,ATF	0.345	0.391	GSTP1a,ESR1,GATA3	0.438	AR,BCL2	0.001
8	P53b,WNT10b	0.283	P53b,WNT10b	0.332	0.353	P53b,WNT10b	0.385	GSTT1,ESR1	0.0007
9	ERK2,NRAS	0.264	CDKN2B,TSG101	0.323	0.310	GSTP1a,AR,FOXA1	0.350	GADD45,ESR1	0.0006
10	CYR61,ATF	0.263	ERK2,NRAS	0.314	0.325	ERK2,NRAS	0.348	ESR1,KIT	0.0005

NOTE: The rightmost column shows the association between ER status and gene clusters. All the other columns reflect posterior probabilities of interactions among genes.

Table 5
Several three-gene clusters from the breast cancer expression profiles.

Cluster (<i>GSTP1a</i> , <i>ESR1</i> , <i>GATA3</i>)				
$x_{GSTP1a} = 0$		$x_{GSTP1a} = 1$		
$x_{ESR1} = 0$	$x_{GATA3} = 0$	1	$x_{GATA3} = 0$	1
	4	1	0	0
	1	19	1	4
Cluster (<i>GSTP1a</i> , <i>AR</i> , <i>FOXA1</i>)				
$x_{GSTP1a} = 0$		$x_{GSTP1a} = 1$		
$x_{AR} = 0$	$x_{FOXA1} = 0$	1	$x_{FOXA1} = 0$	1
	4	2	0	0
	1	18	1	4

Table 6
Clusters assigned highest posterior probabilities by models defined using different values of *c* excluding genes *GSTP1a* and *AR* for the breast cancer data.

Rank	<i>c</i> = 1			<i>c</i> = 2			<i>c</i> = 3		
1	<i>ESR1</i>	<i>GATA3</i>	<i>FOXA1</i>	0.742	<i>ESR1</i>	<i>GATA3</i>	<i>FOXA1</i>	<i>GATA3</i>	<i>FOXA1</i>
2	<i>BRCA1</i>	<i>JNK</i>		0.559	<i>BRCA1</i>	<i>JNK</i>		<i>JNK</i>	0.753
3	<i>HSF1</i>	<i>P53a</i>		0.550	<i>HSF1</i>	<i>P53a</i>		<i>P53a</i>	0.673
4	<i>GSTP1b</i>	<i>AKT2</i>		0.468	<i>GSTP1b</i>	<i>AKT2</i>		<i>CYR61</i>	0.517
5	<i>CDKN2B</i>	<i>TSG101</i>		0.338	<i>KRAS</i>	<i>CYR61</i>	<i>ATF</i>	<i>AKT2</i>	0.460
6	<i>ERK2</i>	<i>NRAS</i>		0.318	<i>P53b</i>	<i>WNT10b</i>		<i>WNT10b</i>	0.398
7	<i>ESRRA</i>	<i>ERBB2</i>		0.301	<i>ESRRA</i>	<i>ERBB2</i>		<i>ERBB2</i>	0.359
8	<i>P53b</i>	<i>WNT10b</i>		0.282	<i>ERK2</i>	<i>NRAS</i>		<i>NRAS</i>	0.341
9	<i>CYR61</i>	<i>ATF</i>		0.281	<i>CDKN2B</i>	<i>TSG101</i>		<i>TSG101</i>	0.277
10	<i>IFI27</i>	<i>TNFa</i>		0.235	<i>CYR61</i>	<i>ATF</i>		<i>TNFa</i>	0.268

Table 7
Collapsed contingency tables for high-probability three-gene interaction models when genes *GSTP1a* and *AR* are removed from the model.

Hu et al.

Cluster (<i>ESR1</i> , <i>GATA3</i> , <i>FOXA1</i>)				
$x_{FOXA1} = 0$		$x_{FOXA1} = 1$		
$x_{ESR1} = 0$	$x_{GATA3} = 0$	1		$x_{GATA3} = 0$
	20	1	0	4
	1	3	1	0
				20
Cluster (<i>KRAS</i> , <i>CYR61</i> , <i>ATF</i>)				
$x_{KRAS} = 0$		$x_{KRAS} = 1$		
$x_{CYR61} = 0$	$x_{ATF} = 0$	1		$x_{ATF} = 0$
	17	0	0	4
	1	8	1	4
				12

Table 8
Cross-tabulation of gene expression levels for genes *ESR1*, *GATA3*, and *FOXA1* against ER status.

<i>FOXA1</i>	0												1																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																															
<i>GATA3</i>	0												1												1																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																			
<i>ESR1</i>	0												1												0												1												1																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																											
ER	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	

Table 9

Summary of results from GGM.

Rank	Doubleton	<i>p</i> -Value	<i>q</i> -Value	Empirical posterior probability
1	<i>ESR1,GATA3</i>	<0.0001	0.089	0.825
2	<i>AR,FOXA1</i>	0.0001	0.139	0.825
3	<i>ESRRA,ERBB2</i>	0.0002	0.147	0.635
4	<i>GSTM1,GADD45</i>	0.0006	0.245	0.635
5	<i>BRCA1,BCL2</i>	0.0006	0.256	0.635
6	<i>CYR61,ATF</i>	0.0007	0.269	0.635
7	<i>BARD1,ERBB2</i>	0.0010	0.287	0.635
8	<i>HSF1,P53a</i>	0.0010	0.291	0.635
9	<i>FOLH1,IL10</i>	0.0011	0.292	0.426
10	<i>FOLH1,DCC</i>	0.0016	0.3504	0.426

NOTE: Pairs of nodes that are estimated to have high probability are listed in the second column of the table. The third column provides the *p*-value for the test of zero partial correlation. The fourth column provides the FDR-based *q*-value, which accounts for multiplicity in testing. The final column provides an estimate of the empirical posterior probability that the indicated nodes are connected in the GGM.