Literature Review: Creating a Gene Expression Network Using Bayesian Hierarchical Modeling

A review of Luo and Zhao [2011] and related techniques

Rachel Lawrence and Una Boyle

April 30, 2014

Overview

In genomic studies, it is often a nontrivial issue to determine the structure of differential gene expression (DGE) networks—that is, the relationships between transcription levels of co-regulated genes, which may exist in pathways or participate in more complex interactions. This issue is of particular interest with respect to cancer research, in which identification of regulatory networks in distinct subclasses of cancerous tumors both aids in clinical classifications, determining the genes which most contribute to the disease state of the patient, and in finding novel treatments for disease [1]. We investigate "Bayesian Hierarchical Modeling for Signaling Pathway Interference from Single Cell Interventional Data" by Luo and Zhao, which puts forth an improved model for creating Gaussian graphical models to represent these regulatory pathways [2]. In specific, this paper details Luo and Zhao's work to create a directed graph with vertex set V = {proteins encoded by individual genes}, and edge set E such that $i \to j \in E$ between genes $i, j \in V$ represents a directional regulatory relationship in which i regulates j.

We aim to analyze and justify, on a step-by-step basis, each of the assumptions, inferences, and algorithms used by Zhao and Luo in their creation of three different models. We pay close attention especially to the prior distributions used to create the model and the motivations behind the assumptions, as well as details of the calculations used to draw conclusions about posterior distributions associated with the model.

Bayesian Hierarchical Models

The basis for Bayesian statistics is grounded in Bayes' rule,

$$\mathbb{P}(\theta|X) = \frac{\mathbb{P}(X|\theta) \cdot \mathbb{P}(\theta)}{\mathbb{P}(X)}$$

This simply means that the model requires the determination of likelihoods and prior distributions for the variables of interest, in order to calculate what we are really interested in—the posterior distribution $\mathbb{P}(\theta|X)$.

One of the central features of a Bayesian Hierarchical model is that it makes use of two separate ways of viewing its variables. The first is pooling, the assumption that all variables are sampled from identical prior distributions, which allows the model to borrow information from across similar variables by treating them all as samples from a single population. On the other hand, non-pooling interpretations assume that all variables are mutually independent, allowing a model using non-pooling methods to take into account the full range of information contained in the data. A hierarchical model is one that uses a combination of pooling and non-pooling, i.e. partial pooling. The choice of hyperparameters for the prior distribution, from which the pooled distribution is taken, can be adjusted within the hierarchical model in order to change the weighting of pooling versus non-pooling information.

Although the authors mention many commonly used statistical methods for creating these models, including Markov random fields and Bayesian networks (which we address in the *Discussion* section of this paper) the paper focuses on the use of a modified method based on “Dependency Networks” to create Gaussian Graphical Models. Dependency Networks are made up of a directed multigraph along with a set of conditional distributions for the variables represented by each node, such the a node has edges both to and from all other variables from which it is not independent. In our analysis of DGE networks, the nodes will represent individual protein expression levels, and a Dependency Network would depict which proteins affect the expression levels of others [3].

However, since these networks do not allow for inference of cause and effect—a necessary component of a model showing regulation among genes—the Bayesian hierarchical modeling approach described later must address some of these limitations. It does so via “experiments” based on controlled expression levels of certain genes, as well as a more robust series of linear models that can capture the differences in linear relationships between genes under these different controlled conditions.

The data consists of measured gene expression levels for genes i = {1 through P} in cell samples n = {1 through N_k} under experimental conditions k = {1 through K}. To complete the models, MCMC methods are used to sample the posterior distributions of variables that represent the likelihood of a relationship between two genes. The end-product graphs created by the hierarchical model were later compared, using simulations and case studies, to a restricted version of the model as well as an ordinary dependency network, confirming the model’s utility for creating improved DGE networks.

The Hierarchical Model

The overarching goals of creating the model were to take into account possible measurement error in data, to determine which genes were the most closely related, and to ultimately infer a causal relationship between regulator and regulated genes. In the next few sections (Equations 1-4), we shall discuss the assumptions made behind the equations that comprise the Hierarchical Model. Afterwards we derive Equation 5 and Equation 6, which are used for MCMC sampling.

Assumptions

Equation 1: Linearity

Luo and Zhao first put forth a series of assumptions upon which their statistical work would be based, the first being the assumption of linearity of the measured activity levels of each protein, giving us Equation 1:

$$\tilde{x}_{ink} = \alpha_{i0}^{(k)} + \sum_{j \neq i} \alpha_{ij}^{(k)} \tilde{x}_{jnk} + \epsilon_{ink}^{I}$$

This equation states that the *true* activity level, $\tilde{x}_{ink}$, of protein i in cell n and under condition k, can be expressed as a linear combination of the true values of each other protein's expression, with the addition of an error term, ϵ_{ink}^{I}, and a linearity constant, $\alpha_{i0}^{(k)}$. The error term serves to take into account a certain amount of noise "inherent" to the system. With this, another set of assumptions is made: ϵ_{ink} is defined to have a distribution $\sim N(0, (\sigma_i^I)^2)$ and are said to be independent of each other. It is standard to assume the errors in such a linear model take on a normal distribution [4]. Additionally, Luo and Zhao make the assumption that the intrinsic noise terms under each of the k conditions are independent because the experimental conditions are carried out separately from each other. Luo and Zhao also make the widely used assumption that $(\sigma_i^I)^2$ has prior distribution Gamma(γ_1, γ_2), and they vary these hyperparameters to gauge the sensitivity of the inference results at the end [5].

The assumption of linearity itself also requires some justification. While it is possible that the relationships between proteins' activity levels do not follow a perfect linear correspondence, a linear assumption is logical for a system that is typically thought to consist of two responses from protein: the expression level of a protein can either be regulated up or down in response to the expression level of another protein based on how related they are. It simplifies the model in a way that will allow for much less complicated analysis based on the magnitude of coefficients $\alpha_{ij}^{(k)}$, while still retaining the salient features of the true system.

Equation 2

Equation 2 follows from both equation 1 and the formula for the measured expression level of a protein, $x_{ink} = \tilde{x}_{ink} + \epsilon_{ink}^{I}$, where $\epsilon_{ink}^{I} \sim N(0, (\sigma^M)^2)$ is the measurement error. Like the error term in equation 1, the normal distribution assumption is a standard one to make. The measured errors are also independent of the true errors as the errors account for completely different phenomena; the measured error occurs when a mistake was made measuring the expression level of a protein while the true error occurs when there are fluctuations between expression levels of proteins. Lastly Luo and Zhao also make the widely used assumption again that $(\sigma_i^I)^2$ has prior distribution Gamma(γ_5, γ_6), and they vary these hyperparameters to gauge the sensitivity of the inference results at the end [5].

The missing algebraic steps in the derivation of equation 2 run as follows:

$$\tilde{x}_{ink} = \alpha_{i0}^{(k)} + \sum_{j \neq i} \alpha_{ij}^{(k)} \tilde{x}_{jnk} + \epsilon_{ink}^{I}$$

and

$$x_{ink} = \tilde{x}_{ink} + \epsilon^{M}_{ink}$$

$$= \alpha^{(k)}_{i0} + \sum_{j \neq i} \alpha^{(k)}_{ij} \tilde{x}_{jnk} + \epsilon^{I}_{ink} + \epsilon^{M}_{ink}$$

And now, substituting in for $\tilde{x}_{jnk} = x_{jnk} + \epsilon^{M}_{ink}$:

$$= \alpha^{(k)}_{i0} + \sum_{j \neq i} \alpha^{(k)}_{ij} (x_{jnk} + \epsilon^{M}_{ink}) + \epsilon^{I}_{ink} + \epsilon^{M}_{ink} \tag{2}$$

Distributing $\alpha^{(k)}_{ij}$ gives the second form of equation 2.

Priors

Next, Luo and Zhao introduce conjugate priors for some of the variables. Conjugate priors are an invaluable tool when predicting the behavior of variables because they keep the the relationship between the likelihood function and the prior distribution simple [6]. The likelihood function is the model for the data we have observed, while the prior distribution is the model for unknown parameters, θ. To begin, the likelihood function and the prior distribution are related by Bayes' rule:

$$\mathbb{P}(\theta|X) = \frac{\mathbb{P}(X|\theta)\mathbb{P}(\theta)}{\int \mathbb{P}(X|\theta)\mathbb{P}(\theta) d\theta}$$

where X is the observed data, θ is the parameter whose distribution we desire, $\mathbb{P}(\theta)$ is the prior distribution, and $\mathbb{P}(X|\theta)$ is the likelihood function. The denominator serves as the normalizing constant for the posterior distribution, $\mathbb{P}(\theta|X)$. As more data is acquired, $\mathbb{P}(\theta)$ gets updated to better reflect the influence of θ on the data and give a better prediction of the posterior distribution.

To find the conjugate prior, the type of prior distribution (for example, a beta or gamma distribution) is deliberately chosen so that when we calculate the right hand side of Bayes' formula above, the posterior distribution is of the same type of distribution as the prior distribution. Then we say the prior distribution is the conjugate prior for the likelihood function. This calculation is not simple at all, which is why conjugate priors are so useful; in order to update the prior, the task is to find the θ that maximizes it.

From lecture, we know that optimization problems are extremely difficult and in practice, local maxima complicate the issue. In more simple situations, the EM algorithm may be used as it at least guarantees a local maxima, but otherwise, more complex algorithms are used. In this paper, Luo and Zhao only provide the end results for the values of θ's chosen for each prior, and do not describe the process of finding θ's as it is not crucial to conveying how the hierarchical model works, but it is still important to understand their role. To avoid digression, a simple example in which the Beta distribution is used as a prior has been included as [**A.3**] in the Appendix of this paper [7]. The Beta distribution is used extensively in Luo and Zhao's hierarchical model.

Equation 3

The first prior we encounter is equation 3. Here, Luo and Zhao use a combination of two different statistical distributions to create a probability density function that describes two possibilities for the coefficient of linear regression, $\alpha_{ij}^{(k)}$, relating two proteins i and j. The first possibility is that proteins i and j are not linearly related at all, a case which is accounted for by a point mass at zero denoting the discrete probability that $\alpha_{ij}^{(k)}$ equals zero. The other possibility is that there does exist a linear relationship between proteins i and j, for which Luo and Zhao use a normal distribution—a standard default assumption for a distribution which we expect to be roughly symmetric and centered at zero (since there is no reason to believe that genes are regulated up more often than down, or vice versa). Luo and Zhao call the variance of this normal distribution $(\sigma_{ij}^{\alpha})^2$, and let $(\sigma_{ij}^{\alpha})^{-2}$ have distribution Gamma(γ_3, γ_4). Again, they vary these hyperparameters to gauge the sensitivity of the inference results at the end [5]. Equation 3 also introduces a new variable, $w_{ij}^{(k)}$ the probability that $\alpha_{ij}^{(k)}$ is nonzero. Hence we get the following prior:

$$\alpha_{ij}^{(k)}|w_{ij}^{(k)}, \alpha_{ij}, \sigma_{ij}^{(\alpha)} \sim (1 - w_{i,j}^{(k)})\delta_0(\alpha_{i,j}^{(k)}) + w_{i,j}^{(k)} N(\alpha_{i,j}^{(k)}|\alpha_{i,j}, (\sigma_{i,j}^{\alpha})^2) \tag{3}$$

where i,j is the pooled distribution of $\alpha_{ij}^{(k)}$ over all k conditions, and $\sigma ij^{(\alpha)}$ is the standard deviation of $\alpha_{ij}^{(k)}$ over all k conditions. Similarly to $\alpha_{ij}^{(k)}$, α_{ij} has $N(a^{(i)}, \tau^{(i)})$ as a prior, where Luo and Zhao say $a^{(i)}$ and $\tau^{(i)}$ are manipulated to gauge the sensitivity of the inferred results. The choice of the mean and variance for the normal distribution in equation 3 is just based on the mean and variance of $\alpha_{ij}^{(k)}$ over all k conditions, and will predict future nonzero $\alpha_{ij}^{(k)}$.

Equation 4

As equation 3 depends on $w_{ij}^{(k)}$, we need a prior for this variable, too. The Beta distribution was chosen as the prior for $w_{ij}^{(k)}$ because $w_{ij}^{(k)}$ is the probability of one event occurring (proteins i and j having a nonzero linear relationship under condition k), so the likelihood function is a Bernoulli distribution. When the likelihood function is a Bernoulli distribution, it is standard to have the Beta distribution as the prior [7]. The pooled probability $w_{ij}^{(k)}$ over all k conditions is defined as w_{ij}, so the hyperparameters for this Beta distribution were chosen such that the mean and variance would evaluate to w_{ij} and $w_{ij}(1-w_{ij})/(v_{ij}+1)$. The mean and variance of a Beta distribution can be found through the equations below [8].

$$\begin{aligned}\mathbb{E}(X) =& \frac{\alpha}{\alpha+\beta}\\ var(X) =& \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}\end{aligned}$$

So the Beta distribution we desire for $w_{ij}^{(k)}$ is

$$w_{ij}^{(k)}|w_{ij}, v_{ij} \sim \text{Beta}(w_{ij}v_{ij}, (1 - w_{ij}v_{ij})$$

The prior for wij is also a Beta distribution, Beta(β_1, β_2), where Luo and Zhao say β_1 and β_2 are manipulated to gauge the sensitivity of the inferred results.

These equations and priors fully describe the model, and Luo and Zhao continue with some conclusions that can be drawn given these equations.

Integration

Equation 5

Equation 5 follows from equation 3, after integrating with respect to $w_{ij}^{(k)}$. We provide the missing derivation:

$$\begin{aligned}
\mathbb{P}(\alpha_{i,j}^{(k)}|w_{i,j},\alpha_{i,j},\sigma_{i,j}^{\alpha}) =& \mathbb{P}(\alpha_{i,j}^{(k)}|w_{i,j},\alpha_{i,j},\sigma_{i,j}^{\alpha},v_{i,j}) \\
& \text{because } \alpha_{i,j}^{(k)} \text{ is independent of } v_{i,j} \\
=& \int \mathbb{P}(\alpha_{i,j}^{(k)}|w_{i,j}^{(k)},v_{i,j},w_{i,j},\alpha_{i,j},\sigma_{i,j}^{\alpha})\mathbb{P}(w_{i,j}^{(k)}|\alpha_{i,j},\sigma_{i,j}^{\alpha},w_{i,j},v_{i,j})dw_{i,j}^{(k)} \\
& \text{because } \mathbb{P}(A|C) = \int \mathbb{P}(A|B,C)\cdot\mathbb{P}(B|C)dB \\
& \text{(see appendix } [\mathbf{A.4}] \text{ for derivation)} \\
=& \int \mathbb{P}(\alpha_{i,j}^{(k)}|w_{i,j}^{(k)},v_{i,j},w_{i,j},\alpha_{i,j},\sigma_{i,j}^{\alpha})\mathbb{P}(w_{i,j}^{(k)}|w_{i,j},v_{i,j})dw_{i,j}^{(k)} \\
& \text{because } w_{i,j}^{(k)} \text{ does not depend on } \alpha_{i,j} \text{ or } \sigma_{i,j}^{\alpha} \text{ from equation 3} \\
=& \int [(1-w_{i,j}^{(k)})\delta_0(\alpha_{i,j}^{(k)}) + w_{i,j}^{(k)}N(\alpha_{i,j}^{(k)}|\alpha_{i,j},(\sigma_{i,j}^{\alpha})^2)]\mathbb{P}(w_{i,j}^{(k)}|w_{i,j},v_{i,j})dw_{i,j}^{(k)} \\
=& \delta_0(\alpha_{i,j}^{(k)})\int \mathbb{P}(w_{i,j}^{(k)}|w_{i,j},v_{i,j})dw_{i,j}^{(k)} - \delta_0(\alpha_{i,j}^{(k)})\int w_{i,j}^{(k)}\mathbb{P}(w_{i,j}^{(k)}|w_{i,j},v_{i,j})dw_{i,j}^{(k)} \\
& + \delta_0(\alpha_{i,j}^{(k)}) + \int w_{i,j}^{(k)}N(\alpha_{i,j}^{(k)}|\alpha_{i,j},(\sigma_{i,j}^{\alpha})^2)\cdot\mathbb{P}(w_{i,j}^{(k)}|w_{i,j},v_{i,j})dw_{i,j}^{(k)} \\
=& \delta_0(\alpha_{i,j}^{(k)}) - \delta_0(\alpha_{i,j}^{(k)})\mathbb{E}(w_{i,j}^{(k)}) + \mathbb{E}(w_{i,j}^{(k)})N(\alpha_{i,j}^{(k)}|\alpha_{i,j},(\sigma_{i,j}^{\alpha})^2) \\
=& (1-w_{i,j})\delta_0(\alpha_{i,j}^{(k)}) + \mathbb{E}(w_{i,j})N(\alpha_{i,j}^{(k)}|\alpha_{i,j},(\sigma_{i,j}^{\alpha})^2) \\
=& (1-w_{i,j})\delta_0(\alpha_{i,j}^{(k)}) + w_{i,j}N(\alpha_{i,j}^{(k)}|\alpha_{i,j},(\sigma_{i,j}^{\alpha})^2) \qquad (5)
\end{aligned}$$

Equation 6

Finally, in equation 6, the posterior distribution of $w_{ij}^{(k)}$ is calculated given a data sample of $\alpha_{ij}^{(k)}$. The posterior distribution given here is just a modification of equation 4: the difference is the addition of the information from $\alpha_{ij}^{(k)}$, which tells us whether a nonzero correlation exists between genes i and j under condition k.

Using the definition of the Beta distribution, this information is added one of the two hyperparameters of the distribution as an indicator function—adding 1 to the first parameter and 0 to the second if $\alpha_{ij}^{(k)}$ is nonzero to denote a "successful" Bernoulli trial, and the opposite in the case that $\alpha_{ij}^{(k)} = 0$ (a "failure").

Now Luo and Zhao were able to sample from the posterior distributions of w_{ij} and $\alpha_{ij}^{(k)}$ and then sample from $w_{ij}^{(k)}$ using equation 6. All of the posterior distributions concerned are proper because all of the prior distributions used so far have been proper [9]. A proper prior means that the denominator in Bayes' formula (the normalizing constant) is always finite.

Inferences

Causality and Edge Directions

One of the crucial inferences required to create a graph of genetic interactions is not only whether two genes are related (and thus, are connected by an edge), but also the direction of that edge. Because the edge direction indicates a cause-and-effect relationship between two genes—that is, $X \rightarrow Y$ denotes that X regulates Y—creating a directed graph including this information is a crucial step in understanding the workings of the expression network. However, a mere correlation between two variables cannot provide any information about causation between the variables; for this, it is necessary instead to perform an "experiment" to determine whether varying one parameter while keeping all else constant directly influences another. In order to create this experimental data, Luo and Zhao use the conditions k in which a single gene's expression is controlled by experimenters.

To determine the direction of the (i, j) edge, there first must exist at least one condition k under which either i or j is perturbed, i.e. some condition such as "Inhibit i" which indicates the direct, controlled manipulation of that gene's expression in the laboratory. If there exist no such conditions, there is no way to observe the necessary experimental results, and so no directed edge can be inferred from even the strongest linear relationship between proteins i and j.

To infer edge directions, we use the posterior mean $\hat{w}_{ij}^{(k)}$—that is, the mean of the posterior of w_{ij} over all k, for all pairs of i and j. We also define a *stream* from i to j as the set of all of the values $\{\hat{w}_{ij}^{(k]}, ...\}$ for every value of k; its reverse is the set of values $\{\hat{w}_{ji}^{(k)}, ...\}$ again for all values of k. We explain the 3 possible remaining cases that Luo and Zhao used:

Case 1 For the simplest case, suppose that i or j are perturbed only under one condition k. To determine the direction of the regulatory relationship between the two, Luo and Zhao compare the maximum $(\hat{w}_{ij}^{(k)})$, over all values of k, to $\hat{w}_{ij}^{(k')}$, supposing that i is the manipulated variable. If the difference between the terms is higher than some threshold, defined as u_3, then $(j \rightarrow i)$ is likely the correct direction–controlling i independently of j apparently dissolves the connection between j and i by manipulating the responding gene by hand even as the regulatory gene changes. However, if the difference between the terms is instead very small (less than the, as yet undetermined, threshold), this indicates that the linear relationship between i and j is instead largely maintained when i is manipulated. This indicates that i is instead the regulatory gene, causing j to display a direct response however i is manipulated.

To check this result, Luo and Zhao then repeat the comparison on the opposite stream, from j to i: $max_k(\hat{w}_{ji}^{(k)})$ and $\hat{w}_{ji}^{(k')}$. If the inference from both stream directions is the same, this is taken as sufficient to create a directed edge from the regulatory gene to the responding gene, indicating a directional causal relationship between expression of the two genes. However, if the inference in both directions come to different conclusions, the method can infer no direction for the edge.

Cases 2 and 3 The other cases follow according to similar logic, but more information allows for an analysis with more internal verification before a conclusion is drawn. If the same protein i is controlled over several experiments, then each value of k which perturbs i can yield its own directional result as in case 1; if all of the directions across each of the experimental conditions agree, that direction is inferred for the ij stream. Meanwhile, if both i and j are controlled under multiple different experimental conditions, then the streams for different values of k can directly be compared: $|(\hat{w}_{ji}^{(k)})_1 - (\hat{w}_{ji}^{(k)})_2|$. In this case, if the difference between pairwise posterior means is on the same side of the threshold u_3 for every pair of conditions, the direction is determined. Subsequently, in both cases, if the directions inferred from both the ij and ji streams are the same (or if one is inconclusive and the other is inferred), the $(i.j)$ edge is added in that direction; otherwise, the edge is still not inferred.

Analysis

It is interesting that Luo and Zhao chose this method for determining direction; it only takes two "disagreeing" values of k—potentially out of a very long list of experimental conditions—in order to render the direction inconclusive; however, an inconclusive stream can be overridden by one in the other direction. One can imagine a situation in which this would seem illogical: supposing that the ij direction led to the inference of an $i \to j$ edge, but the ji direction pointed to a $j \to i$ edge in all cases except one. If there were a large number of streams being examined, it would be hard to argue that this edge should definitively belong in one direction or the other; especially not more-so than one in which all but one stream agreed in each direction.

Although this concern was not addressed by the paper, it was likely considered to be of lesser importance given the particular data Luo and Zhao were looking to analyze. In this dataset, it is notable that only 9 different perturbations (that is, experimental conditions k) were used, and thus, having more than two streams manipulating the same gene was not a concern. Due to the limited nature of typical genetic experiments, it is likely that this issue similarly fails to be a concern in most cases, but perhaps would need to be addressed for a more expansive set of experimental conditions.

The Restricted Hierarchical Model

The Restricted Hierarchical Model proposed by Luo and Zhao is a slightly simplified version of the original hierarchical model. The difference is that the model constrains that $w_{ji} = w_{ij}$ for all proteins i and j, and similarly $w_{ji}^{(k)} = w_{ij}^{(k)}$ for all conditions k. This means that using this model requires us to assume that whether or not there is a linear relationship between i and j does not depend on identifying that one expression level changes in response to the other, but rather that the overall correlation is independent of which is which. This changes very little, but simplifies the determination of edge direction that we described in detail for the non-restricted model; it only necessitates the choice of a single threshold u_1', and eliminates some of the judgements necessary when examining both streams.

In sample studies, the performance of the RHM was very similar to that of the total hierarchical model, although the HM appeared to maintain a slightly higher accuracy when both models were applied to already-known systems.

The Nonhierarchical Model

The Nonhierarchical Model is an even simpler version of the hierarchical model, using a regular Dependency Network without any hierarchical modifications. Like the HM, it also assumes a linear relationship between the activity levels of proteins, but now it uses complete pooling instead of partial pooling, i.e. it assumes that under all conditions, the linear regression coefficients are identical. This means that it cannot be used to determine the causal relationship of which proteins regulate other proteins. From this structure, we derive equation 7 from equation 2, replacing each $\alpha_{ij}^{(k)}$ with α_{ij}:

$$x_{ink} = \alpha_{i0} + \sum_{j \neq i} \alpha_{ij}(x_{jnk}) + \sum_{j \neq i} \alpha_{ij}(\epsilon_{ink}^{M}) + \epsilon_{ink}^{I} + \epsilon_{ink}^{M} \tag{7}$$

Luo and Zhao use the same logic behind the priors for α_{ij} in this model as the logic behind the priors for α_{ijk} in the HM as described earlier. Therefore equation 8 looks very similar to equation 3:

$$\alpha_{ij} \sim (1 - w_{i,j})\delta_0(\alpha_{i,j}) + w_{i,j} N(\alpha_{ij} | a, \tau^2) \tag{8}$$

The only differences are that α_{ij} appears instead of $\alpha_{ij}^{(k)}$ and w_{ij} appears instead of $w_{ij}^{(k)}$—because all α_{ij}^{k} were assumed to be identical in this model—and the normal distribution is defined by a mean of α_{ij} depenedent on a and a variance of τ^2. Both a and τ^2 are manipulated to again gauge the sensitivity of the inferred results to the chosen values. Therefore, from this model we can only tell if proteins i and j are associated in any way, and we cannot determine the regulatory direction between any two proteins as was possible with the HM. The reason for this is explained in the next section.

MCMC

After defining these three models, Luo and Zhao use MCMC algorithms to sample from the posterior distributions in order to infer the desired network of proteins. For the HM and RHM, they use the algorithms to sample from the posterior distributions of $w_{ij}^{(k)}$ and and w_{ij}. The posterior distribution of w_{ij} establishes that proteins i and j are related (i.e., that there exists an edge between vertices i and j in the graph) because it reflects the connectedness under all k conditions. The posterior distribution of $w_{ij}^{(k)}$ is used to determine the regulatory direction between any two proteins (i.e. the direction of the edge between vertices i and j) based on how its value changes between experimental conditions. For the NHM, Luo and Zhao are only able to use the posterior distribution of w_{ij} which is why the regulatory direction cannot be determined as mentioned earlier.

As we saw in class, Monte Carlo Markov Chain methods use samples generated by a Markov Chain in which the stationary distribution matches the desired distribution. Although the algorithm used is never mentioned in the paper, the full conditional probabilities of all variables can be determined from the model, and the calculations to this end can be found in Luo and Zhao's supplementary material. Given the full conditional probabilities of all variables, it would be appropriate to try an MCMC algorithm implementing Gibbs Sampling; we conjecture that this is what Luo and Zhao may have done [10]. For example, Gibbs Sampling would be very appropriate for the distributions of w_{ij}, α_{ij}, and σ_{ij}^{α}, which have conditional distributions that are standard functions: Beta, Normal, and Inverse Gamma respectively [equations (7), (8), and (9) in Supplementary Material **S2**]. Software exists to make this process easy, such as the widely-used JAGS ("Just Another Gibbs Sampler") software. However, we do not recognize the distributions given by equations (11) and (12) in **S2**; if they are nonstandard distributions, it may be necessary to default to Metropolis-Hastings instead.

Applications and Results

In the case study portion of the paper, Zhao and Luo empirically tested their hierarchical model, as well as the RHM, the NHM, and the results of an older, previously-published method, against data from the well-understood system of Mitogen-Activated Protein Kinase signal transduction. As the case study was primarily a proof-of-concept run of the algorithm with little new content, performed on data we do not have access to, we will not discuss this portion of the paper in detail; however, it is worth noting that the hierarchical model performed approximately as well as the restricted hierarchical model, and outperformed by far the nonhierarchical model (using the Hamming distance between the generated model and the known pathway as a metric for the performance of a model).

Discussion

Connections

It is no surprise that many key points in the paper by Luo and Zhao relate closely to what we have learned in class. We were first introduced to the concept of prior and posterior distributions in lecture when we were discussing Hidden Markov Chains, and also spent substantial time discussing MCMC methods, particularly the Metropolis / Metropolis-Hastings Algorithm and Gibbs Sampling. Additionally, not all DGE networks are created using Bayesian hierarchical models; other similar leading models include Markov Random Fields and Bayesian Networks. In class, we were introduced to MRFs as structures that can simultaneously have Markov Chain properties but also be an undirected graph; Bayesian Networks, too, relate to concepts in stochastic processes as they are grounded in the often-discussed Markov property.

In the remainder of this section, we aim to clarify the similarities and differences between these three methods—Bayesian Networks (BNs), Dependency Networks (DNs), and Markov Random Fields (MRFs)—in more detail. While only one of these methods is a part of the paper, they are all commonly used in bioinformatics to achieve the goal of creating an accurate DGE network, and each has its own strengths and weaknesses.

Bayesian Networks

Bayesian networks are one of the most common tools used to analyze gene expression networks. A Bayesian network consists of a directed, acyclic graph $G = (V, E)$ where $V = variables$, together with a conditional distribution for each variable $v \in V$. The characteristic feature of this graph is the requirement of Markov-Chain-type dependence between nodes: each variable is conditionally independent of all vertices other than its parent nodes. In other words,

$$\text{(value of node} \mid \text{parents)} = \text{(value of node} \mid \text{parents, all other variables)}$$

Together, the distributions and graph represent the joint probability $\mathbb{P}(v_1, v_2, ..., v_n)$ $\forall$ $v_i \in V$. From here, techniques such as flow cytometry are often used to determine a notion of causation among the vertices.

Dependency Networks

Dependency Networks were introduced to the field as an improvement on Bayesian Networks, which, while useful, can often be misleading. In a Bayesian Network, it can be easy to misinterpret the claim of causality between nodes, because the model cannot account for hidden factors that also might be present in the system but were not included in the model [3]. However, just stating that the edges in a Bayesian Network show a correlation, not a causation, is too weak to be general; in that case, we would need to add more edges between any two edges that are not completely independent of one another, even if the causal link between them is less direct.

Completing this revised network creates, unsurprisingly, a Dependency Network just as we defined it in the beginning of the paper. It is only from this full Dependency Network that we then consider additional sources of analysis in order to bring a more rigorous definition of causation back into the graph. []

Markov Random Fields for DGE Analysis

In contrast to Bayesian networks, MRF models make use of undirected graphs and these are allowed to have cycles. Like the Bayesian Network though, it still makes use of the Markov property. There are certain situations that only an MRF model can capture, and others that only the Bayesian Network can capture. MRF models succeed over Bayesian Networks when there is a cyclic dependency among the vertices of the graph. Obviously Bayesian Networks fail here because they must be acyclic. On the other hand, Bayesian Networks can depict induced dependecy which a MRF cannot because it is undirected. An induced dependency occurs when a child node has multiple parents and information is able to flow in between the two parents nodes. [11]

Conclusions

From this paper, readers can conclude that the hierarchical model is a reliable model for creating simple DGE networks—it gave better results than the nonhierarchical model, and performed well even in its restricted version. With the Hierarchical model, it was possible to create a regulatory network of the proteins, depicting not only associations between gene expression, but also inferring the direction of causation within the network—a result for which it was not possible to use the nonhierarchical model. By allowing information to be borrowed across experiments while still allowing for differences between individual variables, the Bayesian Hierarchical model lead to more accurate inferences on the relationships between genes than previously known.

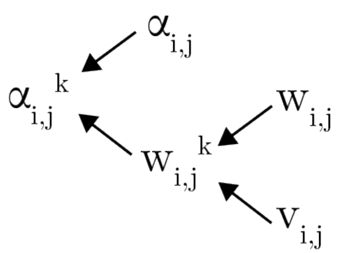
In the coming months, we intend to continue our investigation of the hierarchical model, as well as other algorithms for determining protein expression networks, in Professor Zhao's laboratory group, using many of the same methods we have discussed. Using what we have learned from this paper about the intersection of stochastic processes and Bayesian statistics, we hope to be able to determine more about the structure of regulatory gene networks and perhaps even improve upon the accuracy seen in Luo and Zhao's paper.

Appendix and Figures

[**A.1**] Reference for Variable Definitions

Variable	Definition
X_i	Value of the i^{th} node of graph G; the expression level of the i^{th} protein
$\tilde{x}_{ink}$	True activity level of the i^{th} protein in the n^{th} cell under the k^{th} experimental condition
x_{ink}	Measured activity level of the i^{th} protein in the n^{th} cell under the k^{th} experimental condition; $= \tilde{x}_{ink} - \epsilon^M_{ink}$
ϵ^M_{ink}	Measurement error $\sim N(0, (\sigma^M)^2)$
$\alpha^{(k)}_{ij}$	Coefficient of x term in the linear regression relating i and j under condition k
α_{ij}	Coefficient of x term in the linear regression relating i and j over all conditions
z_{ij}	Indicator function: 1 if i and j are linearly related; 0 otherwise
$w^{(k)}_{ij}$	Probability that z_{ij} is 1 under experimental condition k
w_{ij}	Probability that z_{ij} is 1 over all experimental conditions
v_{ij}	Hyperparameter for the Beta distribution to define $w^{(k)}_{ij}$
u_1	Threshold for inferring that there is an association between i and j, based on the avgerage of w_{ij} and w_{ji}

[**A.2**] Hierarchical Dependence of Variables



$A \to B$ denotes that B depends on A.

[A.3] Suppose that $\mathbb{P}(X|\theta)$ is a simple Bernoulli distribution. Then, to see that the beta distibution

$$\mathbb{P}(\theta) \sim Beta(\alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta}$$

is a useful conjugate prior, we plug both distributions into Bayes' formula:

$$\begin{aligned}\mathbb{P}(\theta|X) &\propto \mathbb{P}(X|\theta)\mathbb{P}(\theta)\\ &\propto (\theta^{X}(1-\theta)^{1-X})(\theta^{\alpha-1}(1-\theta)^{\beta-1})\\ &= \theta^{X+\alpha-1}(1-\theta)^{(1-X)+\beta-1}\end{aligned}$$

Hence this returns another Beta distribution (after some normalization) with hyperparameters $(X+\alpha-1)$ and $(1-X+\beta)$.
Thus we see that conjugate priors are useful because, when multiplied by the likelihood function, the conjugate prior gives a new distribution of the same form as itself (in this case, a Beta distribution).

[A.4] From equation 5:

$$\mathbb{P}(A|C) = \int \mathbb{P}(A,B|C)dB = \int \frac{(\mathbb{P}(A,B,C)}{\mathbb{P}(C)}dB$$

$$= \int \frac{(\mathbb{P}(A,B,C)}{\mathbb{P}(B,C)} \cdot \frac{\mathbb{P}(B,C)}{\mathbb{P}(C)}dB = \int \mathbb{P}(A|B,C) \cdot \mathbb{P}(B|C)dB$$

References

[1] J. Lee, et. al, Statistics in Biosciences., vol 10, No. 7, 1-20 (2013)
[2] R. Luo and H. Zhao, Annals App. Stat., vol. 5, No. 2A, 725745 (2011)
PAREonline.net, n.d. Web.
[3] David, Heckerman, David M. Chickering, Christopher Meek, Robert Rounthwaite, and Carl Kadie. ”Dependency Networks for Inference, Collaborative Filtering, and Data Visualization.” Journal of Machine Learning Research 1 (2000): 49-75. Web.
[4] Osborne, Jason W., and Elaine Waters. “Four Assumptions Of Multiple Regression That Researchers Should Always Test.” Practical Assessment, Research and Evaluation.
[5] Fink, Daniel. ”A Compendium of Conjugate Priors.” Environmental Statistics Group (1997): n. pag. Web.
[6] Kulis, Brian. ”Conjugate Priors.” CSE 788.04: Topics in Machine Learning. Lecture.
[7] Weisstein, Eric W. ”Beta Distribution.” From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/BetaDistribution.html
[8] ”Unit 2: Prior Distributions.” BIO249 Bayesian Methodology in Biostatistics. Massachusetts, Cambridge. Lecture.
[9] Lam, Patrick. ”MCMC Methods: Gibbs Sampling and the Metropolis-Hastings Algorithm.” Massachusetts, Cambridge. Lecture.
[10] D. Peer, Bayesian network analysis of signaling networks: A primer. Sci. STKE 2005, pl4 (2005).
[11] ”Undirected Models.” Brigham Young University CS Department. Brigham Young University, 8 June 2010. Web.