With Michael Newton, I plan to work on an aspect of the NIH-funded SUCCEED (Study to Understand Cervical Cancer Early Endpoints and Determinants) project. Prof. Newton's biological collaborators on this project are Paul Ahlquist and Paul Lambert. Below I describe (1) biological background for the project, (2) our proposed mathematical model based on this biological background, and (3) our next steps in the project.

1. Biological background.

The aspect of the project that I plan to work on is as follows. Prof. Newton and his collaborators have whole genome expression profiles for $n = 128$ tissue samples, divided into four pathologic groups: putatively normal samples, early stage lesions (cervical intraepithelial neoplasia [CIN] 1 and 2), later stage lesions (CIN 3) and frank cancer. Roughly an equal number of tissue samples are in each group. Each tissue sample was measured by an Affymetrix whole genome microarray, which aims to measure the expression of essentially all genes in the genome (it contains about 54,000 probe sets).

The SUCCEED members have already carried out several analyses of these data, including analyses that aim to identify genes showing various patterns of differential expression among the four pathological groupings. However, an alternative analysis can potentially be useful in helping us understand these preliminary findings.

Motivation for the alternative analysis is given by the following biological and technical facts. Each expression profile is measured from a collection of around 1000 cells - so it represents a mixture of theoretical profiles. The stages of cancer progression of cervical tissue are characterized in part by changes in the proportion of cells of particular types. E.g., normal tissue is organized in layers with more well-differentiated cells at the surface and with less differentiated, but more actively dividing cells further inside the tissue. Neoplastic lesions shift the balance of types, at least partly by having relatively more of the less differentiated types and having fewer of the well-differentiated types. Note that the different types will have different gene-expression profiles.

2. Mathematical model.

The above motivates the following mathematical model. We have gene-expression profiles for $n = 128$ tissue samples. Each gene-expression profile $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,G})$ consists of the (raw) expression levels for $G \approx 54000$ genes. We also know the stage s_i of the cancer each tissue sample was taken from. We think of these $x_{i,g}$ and s_i as realizations of random variables $X_{i,g}$ and S_i.

Since each tissue sample is in reality a mixture of T ($T > 1$ but not too large) different cell types, we model each expression profile X_i as a mixture of type-conditional profiles. In particular, we represent "the type of a particular cell from tissue sample i being of some type" by a (hidden) random variable T_i, so each T_i takes values in $\{1, \ldots, T\}$, and $P(T_i = t)$ is the fraction of cells of type t in the tissue sample i.

The tissue samples' stages s_i are fixed, just based on what tissues the experimenters chose to look at, and there does not seem to be much to be gained by modeling them. For this reason, we are

interested in the conditional joint distribution of the $X_{i,g}$ given that $S_i = s_i$ for $i = 1, \ldots, n$. In other words, we want to specify some parametric form (and ultimately learn the parameters) for

$$P(\cap_{g=1}^{G} \cap_{i=1}^{n} \{X_{i,g} \in A_{i,g}\} | \cap_{i=1}^{n} \{S_i = s_i\})$$

We specify the parametric form in several steps.

Step 1. First, we assume that the expression levels are independent by genes, i.e., $X_{i,g}$ and $X_{i',g'}$ are independent for any $g \neq g'$. (But we do not assume that the profiles are independent by tissue sample, i.e., for a fixed gene g, $X_{i,g}$ and $X_{i',g}$ are not necessarily independent.) Thus:

$$\begin{aligned} &P(\cap_{g=1}^{G} \cap_{i=1}^{n} \{X_{i,g} \in A_{i,g}\} | \cap_{i=1}^{n} \{S_i = s_i\}) \\ &= \prod_{g=1}^{G} P(\cap_{i=1}^{n} \{X_{i,g} \in A_{i,g}\} | \cap_{i=1}^{n} \{S_i = s_i\}) \end{aligned}$$

Step 2. We decompose the profiles into mixtures based on type. I.e., for fixed g, we stipulate that

$$\begin{aligned} &P(\cap_{i=1}^{n} \{X_{i,g} \in A_{i,g}\} | \cap_{i=1}^{n} \{S_i = s_i\}) \\ &= \sum_{(t_1,\ldots,t_n)} P(\cap_{i=1}^{n} \{T_i = t_i\} | \cap_{i=1}^{n} \{S_i = s_i\}) P(\cap_{i=1}^{n} \{X_{i,g} \in A_{i,g}\} | \cap_{i=1}^{n} \{S_i = s_i\}, \cap_{i=1}^{n} \{T_i = t_i\}) \\ &= \sum_{(t_1,\ldots,t_n)} P(\cap_{i=1}^{n} \{T_i = t_i\} | \cap_{i=1}^{n} \{S_i = s_i\}) P(\cap_{i=1}^{n} \{X_{i,g} \in A_{i,g}\} | \cap_{i=1}^{n} \{T_i = t_i\}) \end{aligned}$$

where the sums over $(t_1, \ldots, t_n)$ are over all combinations of t_i in $\{1, \ldots, T\}$. (If T is large, there will be a lot of terms here, but we have already required that T is not too large.)

Step 3. We define the mixing proportions. We assume that (i) T_i are conditionally independent given the S_i, and (ii) $P(T_i = t_i | \cap_{i=1}^{n} \{S_i = s_i\}) = P(T_i = t_i | S_i = s_i)$. These are reasonable assumptions, because biologically we think that the proportion of cells of various types in a tissue depends (only or at least primarily) on the tissue's type. So:

$$P(\cap_{i=1}^{n} \{T_i = t_i\} | \cap_{i=1}^{n} \{S_i = s_i\}) = \prod_{i=1}^{n} P(T_i = t_i | S_i = s_i) = \prod_{i=1}^{n} p_{s_i,t_i}$$

where $(p_{1,t}, \ldots, p_{4,t})_{t=1}^{T}$ are parameters that we want to find.

Step 4. We stipulate that the gene expression levels (for a fixed gene g) for tissue samples i and i' are conditionally independent given that $T_i \neq T_{i'}$. (This assumption follows [K], Section 3.) This assumption is reasonable, since if a cell from tissue i is of a different type than a cell from tissue i', we shouldn't think that their expression levels of gene g are related. (In contrast, if the cells are of the same type, then their levels would be related.)

In order to express the assumption precisely, we first define some index sets:

$$\mathscr{I}_t = \mathscr{I}_t(t_1, \ldots, t_n) = \{i : t_i = t\}$$

Note that the $\mathscr{I}_t$ are disjoint sets. Also note that the $\mathscr{I}_t$ depend on $(t_1,\ldots,t_n)$, though we will not always write this explicitly below, in order to save space.

Using this notation, we express the stipulation stated above:

$$\begin{aligned}
&P(\cap_{i=1}^n\{X_{i,g}\in A_{i,g}\}\,|\,\cap_{i=1}^n\{T_i=t_i\})\\
&=\prod_{t=1}^T P(\cap_{i\in\mathscr{I}_t(t_1,\ldots,t_n)}\{X_i\in A_i\}\,|\,\cap_{i=1}^n\{T_i=t_i\})\\
&=\prod_{t=1}^T P(\cap_{i\in\mathscr{I}_t(t_1,\ldots,t_n)}\{X_i\in A_i\}\,|\,\cap_{i\in\mathscr{I}_t(t_1,\ldots,t_n)}\{T_i=t\})
\end{aligned}$$

Step 5. We plug in the Gamma-Gamma model from [K]. In particular, following [K], we stipulate that for each fixed tissue type t the measure on $\text{Borel}(\mathbb{R}^{|\mathscr{I}_t(t_1,\ldots,t_n)|})$ which is induced by

$$(\times_{i\in\mathscr{I}_t(t_1,\ldots,t_n)}A_i)\mapsto P(\cap_{i\in\mathscr{I}_t(t_1,\ldots,t_n)}\{X_i\in A_i\}\,|\,\cap_{i\in\mathscr{I}_t(t_1,\ldots,t_n)}\{T_i=t\})$$

has the following density (wrt Lebesgue):

$$f_0(x_1,\ldots,x_{|\mathscr{I}_t|};\theta_t')=\int_{\mathbb{R}}\Big(\prod_{i=1}^{|\mathscr{I}_t|}f_{obs}(x_i|\mu;\alpha_t)\Big)\pi(\mu;\theta_t')\,d\mu$$

which is parameterized by $\theta_t'=\{\alpha_t,\alpha_{0,t},\nu_t\}$. Here, f_{obs} is a gamma density and π is an inverse-gamma density, with specific functional forms given in [K], Section 4.

Step 6. Finally, putting all the pieces above together,

$$\begin{aligned}
&P(\cap_{g=1}^G\cap_{i=1}^n\{X_{i,g}\in A_{i,g}\}\,|\,\cap_{i=1}^n\{S_i=s_i\})\\
&=\prod_{g=1}^G\sum_{(t_1,\ldots,t_n)}\Big[\prod_{i=1}^n p_{s_i,t_i}\Big]\Big[\prod_{t=1}^T P(\cap_{i\in\mathscr{I}_t(t_1,\ldots,t_n)}\{X_i\in A_i\}\,|\,\cap_{i\in\mathscr{I}_t(t_1,\ldots,t_n)}\{T_i=t\})\Big]
\end{aligned}$$

and the measure induced by

$$(\times_{g=1}^G\times_{i=1}^n A_{i,g})\mapsto P(\cap_{g=1}^G\cap_{i=1}^n\{X_{i,g}\in A_{i,g}\}\,|\,\cap_{i=1}^n\{S_i=s_i\})$$

has density

$$f(x_{1,1},\ldots,x_{n,G};\theta)=\prod_{g=1}^G f_g(x_{1,g},\ldots,x_{n,g})$$

where $\theta=(\theta_t',p_{1,t},\ldots,p_{4,t})_{t=1}^T$ and

$$f_g(x_{1,g},\ldots,x_{n,g};\theta)=\sum_{(t_1,\ldots,t_n)}\Big[\prod_{i=1}^n p_{s_i,t_i}\Big]\Big[\prod_{t=1}^T f_0(\mathbf{x}_{\mathscr{I}_t(t_1,\ldots,t_n),g};\theta_t')\Big]$$

and $\mathbf{x}_{\mathscr{I}_t(t_1,\ldots,t_n),g} = (x_{i,g} : i \in \mathscr{I}_t(t_1,\ldots,t_n))$.

3. Next steps.

Our immediate next step in the project is to develop an algorithm (perhaps based on EM) to find the parameters θ that maximize the likelihood based on our given data $(x_{i,g})$. Note that the log likelihood is

$$\begin{aligned}
\ell(\theta) &= \log f(x_{1,g},...,x_{n,G};\theta) \\
&= \sum_{g=1}^{G} \log f_g(x_{1,g},\ldots,x_{n,g};\theta) \\
&= \sum_{g=1}^{G} \log \sum_{(t_1,\ldots,t_n)} \Big[\prod_{i=1}^{n} p_{s_i,t_i}\Big] \Big[\prod_{t=1}^{T} f_0(\mathbf{x}_{\mathscr{I}_t(t_1,\ldots,t_n),g};\theta_t')\Big]
\end{aligned}$$

After we have an algorithm to find the MLE, we will investigate aspects of the type-specific gene-expression profiles and also mixing proportions of different types at each stage in the MLE. We will also investigate sensitivity to the number of types T.

Depending on the results of this investigation, we will either look for a different model, or we will link in other relevant data into the above model, in order to get a clearer picture of cervical cancer progression.

References.

[K] Kendziorski, C.M., M.A. Newton, H. Lan, and M.N. Gould (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. Statistics in Medicine, 22, 3899-914.

[Y] Yuan, Ming, Ping Wang, Deepayan Sarkar, Michael Newton, and Christina Kendziorski. Parametric Empirical Bayes Methods for Microarrays. R vignette, April 13, 2011.