Statistical methods for microarray data

Lorenz Wernisch

July 21, 2001

Contents

1 Introduction

Microarray analysis is particularly interesting from a statistical point of view since already a wide range of methods has been applied in the interpretation of the data and there is doubtless more to come. The purpose of this introduction is to collect analysis methods proposed in several papers, provide the details that are often missing, use a uniform notation, indicate in which direction results might be generalized or applied, and thus make the material more accessible to the non-statistician. Essentially this is a cleaned version of the notes I made when reading these papers. As a short glance at this review shows though, the reader is expected to be fairly familiar with mathematical concepts such differentiation and integration, manipulation of equations, some vector and matrix algebra, concepts from probability theory such as Bayes theorem or the mathematical properties of the Normal probability density function.

2 Modeling gene expression by hierarchical models

We review here an approach suggested by Newton et al [2]. Suppose we have two variables R and G (for example, the intensities in the two channels of a microarray experiment) with $R \sim \text{Gamma}(a, \theta_r)$ and $G \sim \text{Gamma}(a, \theta_g)$ and both $\theta_r, \theta_g \sim \text{Gamma}(a_0, \nu)$. The ratio of the expectation of R and G is $\rho = E(R)/E(G) = \theta_g/\theta_r$ (equation 15). Given realizations r and g of R and G (as well as the parameters a, a_0, and ν), the probability distribution of ρ is simply the distribution of the quotient of a $\text{Gamma}(a + a_0, g + \nu)$ by

a Gamma$(a+a_0, r+\nu)$ distribution (equation 32). From 28 we get

$$\begin{aligned} p(\rho \mid r,g) &= \frac{\Gamma(2a+2a_0)}{\Gamma(a+a_0)^2}\left(\frac{r+\nu}{g+\nu}\right)^{a+a_0}\rho^{a+a_0-1}\left(\rho+\frac{r+\nu}{g+\nu}\right)^{-2(a+a_0)} \\ &\propto \rho^{-(a+a_0+1)}\left(\frac{1}{\rho}+\frac{g+\nu}{r+\nu}\right)^{-2(a+a_0)} \end{aligned} \tag{1}$$

The *mode* is easy to calculate from $\partial p(\rho \mid r,g)/\partial\rho = 0$

$$\rho_{\text{mode}} = \frac{(r+\nu)}{(g+\nu)}\,\frac{(a+a_0-1)}{(a+a_0+1)} \tag{2}$$

From equations 27, 15, and 16 the mean of ρ is

$$\rho_{\text{mean}} = \frac{(r+\nu)}{(g+\nu)}\,\frac{(a+a_0)}{(a+a_0-1)} \tag{3}$$

which is very close to the mode for $a+a_0$ large enough. Hence a good and simple estimate of ρ might actually be $\hat{\rho} = (r+\nu)/(g+\nu)$, where ν plays the role of a pseudocount familiar from Bayesian contexts.

A way to obtain values for ν and for the two other parameters a and a_0 is by maximum likelihood estimation. The density of a variable in a hierarchical model is shown in equation 31. The joint distribution for R and G then is the product of the two pdfs

$$p_A(r,g \mid \nu,a,a_0) = \left(\frac{\Gamma(a+a_0)}{\Gamma(a)\Gamma(a_0)}\right)^2 \frac{(rg)^{a-1}\nu^{2a_0}}{((r+\nu)(g+\nu))^{a+a_0}} \tag{4}$$

If r_k and g_k are the intensities in the two channels for gene k the log-likelihood function in ν, a, a_0 to maximize is $\sum_k \log(p_A(r_k,g_k \mid \nu,a,a_0))$. Newton et al [2] call this the Gamma-Gamma hierarchical model. The whole procedure thus consists in firstly estimating ν by maximum likelihood (values of about 0.5 the average intensities are reported), and secondly, to estimate $\rho_k = E(R_k)/E(G_k)$ by $(r_k+\nu)/(g_k+\nu)$. This shrinkage factor corrects the increasing uncertainty in small intensities.

A different model, termed Gamma-Gamma-Bernoulli model, assumes that for a subset of genes the true average $E(R)$ actually equals $E(G)$. In this case the random scale parameters θ_r and θ_g are equal to a common $\theta \sim \text{Gamma}(a_0,\nu)$. An additional Bernoulli random variable Z_k indicates whether R and G are distributed with $\theta_r \neq \theta_g$ ($Z_k = 1$ with probability p) or not ($Z_k = 0$ with probability $1-p$). The null case (same distribution) is described by the pdf (see 34)

$$p_0(r,g) = \frac{\Gamma(2a+a_0)}{\Gamma(a)^2\Gamma(a_0)}\,\frac{(rg)^{a-1}\nu^{a_0}}{(r+g+\nu)^{2a+a_0}} \tag{5}$$

Estimates for the parameter a, the hyperparameters a_0, ν as well as the probability p that gene k is differentially expressed can be calculated by the EM algorithm (see section 3.2). Note that $\theta = (a, a_0, \nu, p)$ and (compare with 38)

$$\begin{aligned} F(\bar{\theta},\theta) + H(\bar{\theta}) &= \sum_k \sum_{j=0,1} p(Z_k = j \mid r_k, g_k, \bar{\theta}) \log\left(p(r_k,g_k \mid Z_k = j,\theta)p(Z_k = j \mid \theta)\right) \\ &= \sum_k \hat{z}_k \log p_A(r_k,g_k) + (1-\hat{z})_k \log p_0(r_k,g_k) + \hat{z}_k \log p + (1-\hat{z}_k)\log(1-p) \end{aligned} \tag{6}$$

p needs to be constrained to $0 \leq p \leq 1$ in the maximization of this function in the M-step. The posterior probability $p(Z_k = 1 \mid r_k, g_k, \theta)$ is (see 40)

$$\hat{z}_k = p(Z_k = 1 \mid r_k, g_k, p) = \frac{p\, p_A(r_k, g_k)}{p\, p_A(r_k, g_k) + (1-p) p_0(r_k, g_k)} \tag{7}$$

The goal is to calculate the odds of differential expression ($z_k = 1$) over nondifferential expression ($z_k = 0$)

$$\text{odds} = \frac{P(z_k = 1 \mid r, g)}{P(z_k = 0 \mid r, g)}$$

where

$$P(Z_k = 1 \mid r, g) = \int_0^1 P(Z_k = 1 \mid p, r, g) P(p \mid r, g)\, dp$$

Note that this comprises over- as well as underexpression. We use the point estimate $\hat{p}$ from the above EM algorithm to estimate this integral and obtain for the odds

$$\text{odds} \approx \frac{p(Z_k = 1 \mid r_k, g_k, \hat{p})}{p(Z_k = 0 \mid r_k, g_k, \hat{p})} = \frac{p_A(r_k, g_k)}{p_0(r_k, g_k)} \frac{\hat{p}}{1-\hat{p}} \tag{8}$$

Suppose we have several measurements $r^{(1)}, \ldots, r^{(n)}$ of R and $g^{(1)}, \ldots, g^{(n)}$ of G. Assuming that $R^{(i)} \sim \text{Gamma}(a, \theta_r)$ and $G^{(i)} \sim \text{Gamma}(a, \theta_g)$ for all i and using equation 35 the joint pdf for these measurements is

$$p_A(r^{(1)}, \ldots, r^{(n)}, g^{(1)}, \ldots, g^{(n)}) = \left(\frac{\Gamma(na + a_0)}{\Gamma(a)^n \Gamma(a_0)} \right)^2 \frac{(\prod r^{(i)} g^{(i)})^{a-1} \nu^{2a_0}}{(\sum r^{(i)} + \nu)^{na+a_0} (\sum g^{(i)} + \nu)^{na+a_0}} \tag{9}$$

The null hypothesis $R, G \sim \text{Gamma}(a, \theta)$ is described by the pdf (again by equation 35)

$$p_0(r^{(1)}, \ldots, r^{(n)}, g^{(1)}, \ldots, g^{(n)}) = \frac{\Gamma(2na + a_0)}{\Gamma(a)^{2n} \Gamma(a_0)} \frac{\left(\prod r^{(i)} g^{(i)}\right)^{a-1} \nu^{a_0}}{\left(\sum r^{(i)} + \sum g^{(i)} + \nu\right)^{2na+a_0}} \tag{10}$$

Equations 6, 7, and 8 can be applied as before. This procedure allows to take into account replications in hybridization.

3 Statistical background

3.1 Distributions

Gamma distribution and Gamma function. The density function of the gamma distribution Gamma(x, a, λ) with shape parameter a and scale parameter λ is function

$$f(x) = \frac{\lambda^a}{\Gamma(a)} x^{a-1} e^{-\lambda x} \tag{11}$$

The Gamma function $\Gamma(a)$ is the normalizing factor defined as

$$\Gamma(a) = \int_0^\infty u^{a-1} e^{-u}\, du = \int_0^\infty (\lambda x)^{a-1} e^{-\lambda x}\, \lambda\, dx = \int_0^\infty \lambda^a x^{a-1} e^{-\lambda x}\, dx, \quad a > 0 \tag{12}$$

where we substituted $u = \lambda x$. Written slightly differently this equation is turned into a useful formula for evaluating integrals involving powers of x and the exponential function

$$= \int_0^\infty x^{a-1} e^{-\lambda x}\, dx = \frac{\Gamma(a)}{\lambda^a}, \quad a > 0 \tag{13}$$

We will use this formula frequently.

By partial integration the following useful relation for the gamma function is easily shown

$$\Gamma(a+1) = (a)\Gamma(a) \tag{14}$$

Starting with an integer n repeated application of it leads to $\Gamma(n) = (n-1)!$ since $\Gamma(1) = 1$.

The gamma distribution is suitable for modeling a range of nonnegative random variables. Its easy to derive further results for this distribution using the above equations 11, 13, and 14. As an example, the mean or expectation value of an gamma distributed random variable X is calculated as

$$E(X) = \int_0^\infty x \frac{\lambda^a}{\Gamma(a)} x^{a-1} e^{-\lambda x}\, dx = \frac{1}{\lambda} \frac{1}{\Gamma(a)} \int_0^\infty \lambda^{a+1} x^a e^{-\lambda x}\, dx = \frac{\Gamma(a+1)}{\lambda\, \Gamma(a)} = \frac{a}{\lambda} \tag{15}$$

Similarly the expectation of $1/X$ is

$$E\left(\frac{1}{X}\right) = \int_0^\infty \frac{1}{x} \frac{\lambda^a}{\Gamma(a)} x^{a-1} e^{-\lambda x}\, dx = \frac{\lambda}{\Gamma(a)} \int_0^\infty \lambda^{a-1} x^{a-2} e^{-\lambda x}\, dx = \frac{\lambda\, \Gamma(a-1)}{\Gamma(a)} = \frac{\lambda}{a-1} \tag{16}$$

Considering a mixture of gamma distributions is sometimes appropriate when a sample distribution looks like gamma distributed (for example, positive and strongly right skewed) but a simple gamma distribution does not fit. In this case a better fit may be achieved with

$$f(x) = \sum_i p_i \frac{\lambda_i^{a_i}}{\Gamma(a_i)} x^{a_i - 1} e^{-\lambda_i x} \tag{17}$$

where $\sum_i p_i = 1$. This is written as $X \sim \text{Gamma}(\mathbf{p}, \mathbf{a}, \lambda)$ with boldface types indicating vectors. Results for the gamma distribution are often easily extended to mixtures which are simply linear combinations of gamma functions. For example, the mean is

$$E(X) = \int_0^\infty x \sum_i p_i \frac{\lambda_i^{a_i}}{\Gamma(a_i)} x^{a_i - 1} e^{-\lambda_i x}\, dx = \sum_i p_i \frac{a_i}{\lambda_i} \tag{18}$$

Combining distributions. The combined pdf of two independent variables X and Y with pdfs $f_x(x)$, $f_y(y)$ is simply $f(x, y) = f_x(x) f_y(y)$. A change of variables from x, y to u, v with $x = X(u, v)$ and $y = Y(u, v)$ results in the following combined density function

$$f(u, v) = f_x(X(u, v)) f_y(Y(u, v))\, |J|, \quad \text{where } J = \begin{vmatrix} \frac{\partial X}{\partial u} & \frac{\partial X}{\partial v} \\ \frac{\partial Y}{\partial u} & \frac{\partial Y}{\partial v} \end{vmatrix} \tag{19}$$

J is the determinant of the Jacobi matrix of partial differentials.

As an example, let us find the distribution function of the quotient of two random variables X and Y. With $u = x/y$ and $v = y$ we have $X(u, v) = uv$ and $Y(u, v) = v$, so

$$J = \begin{vmatrix} v & u \\ 0 & 1 \end{vmatrix} = v \tag{20}$$

The joint distribution function for u, v is $f(u, v) = f_x(uv) f_y(v) v$ and we get the distribution function $f(u)$ of $u = x/y$ by integrating v out

$$f(u) = \int f(u, v)\, dv = \int f_x(uv) f_y(v)\, v\, dv \tag{21}$$

Calculating the expectation from this distribution is straightforward

$$E(u) = \int f(u)\, du = \int \int f_x(uv) f_y(v)\, uv\, dv\, du \tag{22}$$

Similarly, for the product $u = xy$, $v = y$, or $X(u, v) = u/v$, $Y(u, v) = v$

$$J = \begin{vmatrix} 1/v & -u/v^2 \\ 0 & 1 \end{vmatrix} = \frac{1}{v} \tag{23}$$

and the distribution function of $u = xy$ is

$$f(u) = \int f(u, v)\, dv = \int f_x(u/v) f_y(v)\, \frac{1}{v}\, dv \tag{24}$$

For microarray analysis the joint distribution of $u = xy$ and $v = x/y$ is of interest (if x and y are the expression levels in two channels, u is the differential expression and v is total expression). Solving for x and y gives $x = \sqrt{uv}$, $y = \sqrt{u/v}$ and

$$J = \begin{vmatrix} \frac{\sqrt{v}}{2\sqrt{u}} & \frac{\sqrt{u}}{2\sqrt{v}} \\ \frac{1}{\sqrt{v}\, 2\sqrt{u}} & -\frac{\sqrt{u}}{2v^{3/2}} \end{vmatrix} = -\frac{1}{4v} - \frac{1}{4v} = -\frac{1}{2v} \tag{25}$$

The pdf for $u = xy$ and $v = x/y$ is

$$f(u,v) = f_x\left(\sqrt{uv}\right) f_y\left(\sqrt{u/v}\right) \frac{1}{2v} \tag{26}$$

If X and Y are two *independent* random variables then the expectation $E(g(X)h(Y))$ of the product of any functions of X and Y is $E(g(X)h(Y)) = E(g(x))\,E(h(y))$, the product of the expectations for each variable. Hence

$$E(XY) = E(X)E(Y) \text{ and } E(X/Y) = E(X)E(1/Y) \tag{27}$$

Combining gamma distributions. The application of the above equations to gamma distributions is straightforward. First, let us calculate the distribution of the quotient of two gamma distributed variables. Let random variables X and Y be distributed as $X \sim \text{Gamma}(a,\lambda), Y \sim \text{Gamma}(b,\theta)$, then $u = x/y$ has pdf (using equations 21 and 13)

$$\begin{aligned} f(u) &= \int f_x(uy) f_y(y)\, y\, dy = \int_0^\infty \frac{\lambda^a}{\Gamma(a)}(uy)^{a-1}e^{-\lambda uy}\frac{\theta^b}{\Gamma(b)}y^{b-1}e^{-\theta y}\, y\, dy \\ &= \frac{\lambda^a\theta^b}{\Gamma(a)\Gamma(b)}u^{a-1}\int_0^\infty y^{a+b-1}e^{-y(\lambda u+\theta)}\, dy = \frac{\lambda^a\theta^b}{\Gamma(a)\Gamma(b)}u^{a-1}\frac{\Gamma(a+b)}{(\lambda u+\theta)^{a+b}} \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{(u\lambda/\theta)^{a-1}}{(1+u\lambda/\theta)^{a+b}}\frac{\lambda}{\theta} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\left(\frac{\theta}{\lambda}\right)^b\frac{u^{a-1}}{(u+\theta/\lambda)^{a+b}} \end{aligned} \tag{28}$$

Similarly, for mixture distributions $X \sim \text{Gamma}(\mathbf{p},\mathbf{a},\lambda)$ and $Y \sim \text{Gamma}(\mathbf{q},\mathbf{b},\theta)$

$$f(u) = \sum_{ij} p_i q_j \frac{\Gamma(a_i+b_j)}{\Gamma(a_i)\Gamma(b_j)}\left(\frac{\theta_j}{\lambda_i}\right)^{b_j}\frac{u^{a_i-1}}{(u+\theta_j/\lambda_i)^{a_i+b_j}} \tag{29}$$

Distributions may also be combined in the sense that the parameters of one are random variables following a different distribution. Such combinations are called *hierarchical models*. This is particularly easy for Gamma distributions. Suppose a random variable X is $\text{Gamma}(a,\theta)$ distributed but with a scale parameter θ itself random with $\theta \sim \text{Gamma}(a_0,\nu)$. The joint probability for X and θ is

$$\begin{aligned} p(x,\theta) &= p(x \mid \theta)p(\theta) = \frac{\theta^a}{\Gamma(a)}x^{a-1}e^{-\theta x}\,\frac{\nu^{a_0}}{\Gamma(a_0)}\theta^{a_0-1}e^{-\nu\theta} \\ &= \frac{\nu^{a_0}}{\Gamma(a)\Gamma(a_0)}x^{a-1}\theta^{a+a_0-1}e^{-\theta(x+\nu)} \end{aligned} \tag{30}$$

The pdf of X is obtained by integrating the joint probability over θ (using 13)

$$p(x) = \int_0^\infty p(x,\theta)\, d\theta = \frac{\Gamma(a+a_0)}{\Gamma(a)\Gamma(a_0)}\frac{x^{a-1}\nu^{a_0}}{(x+\nu)^{a+a_0}} \tag{31}$$

Applying Bayes formula gives the posterior probability of θ assuming data x and parameters a, $a0$, and ν.

$$p(\theta \mid x) = \frac{p(x \mid \theta)p(\theta)}{p(x)} = \frac{p(x,\theta)}{p(x)} = \frac{(x+\nu)^{a+a_0}}{\Gamma(a+a_0)}\theta^{a+a_0-1}e^{-\theta(x+\nu)} \tag{32}$$

Which is a gamma distribution Gamma$(a + a_0, x + \nu)$ again.

If a second variable Y is also Gamma(a, θ) distributed the marginal joint pdf can be calculated in a fashion similar to 30

$$\begin{aligned} p(x, y, \theta) &= p(x \mid \theta)\, p(y \mid \theta)\, p(\theta) = \frac{\theta^a}{\Gamma(a)} x^{a-1} e^{-\theta x} \frac{\theta^a}{\Gamma(a)} y^{a-1} e^{-\theta y} \frac{\nu^{a_0}}{\Gamma(a_0)} \theta^{a_0 - 1} e^{-\nu\theta} \\ &= \frac{\nu^{a_0}}{\Gamma(a)^2 \Gamma(a_0)} (xy)^{a-1} \theta^{2a + a_0 - 1} e^{-\theta(x+y+\nu)} \end{aligned} \tag{33}$$

As usual, the joint pdf of X and Y is obtained by integrating over θ (using 13)

$$p(x, y) = \int_0^\infty p(x, y, \theta)\, d\theta = \frac{\Gamma(2a + a_0)}{\Gamma(a)^2 \Gamma(a_0)} \frac{(xy)^{a-1} \nu^{a_0}}{(x + y + \nu)^{2a + a_0}} \tag{34}$$

Generalizing to m variables $X_1, \ldots, X_n \sim$ Gamma(a, θ) with $\theta \sim$ Gamma(a_0, ν)

$$p(x_1, \ldots, x_n) = \int_0^\infty p(x_1, \ldots, x_n, \theta)\, d\theta = \frac{\Gamma(na + a_0)}{\Gamma(a)^n \Gamma(a_0)} \frac{(\prod x_i)^{a-1} \nu^{a_0}}{(\sum x_i + \nu)^{na + a_0}} \tag{35}$$

3.2 Algorithms

Expectation-Maximization. One application of the EM algorithm is the classification of random variables $X_i, i = 1, \ldots, n$, where random variables $Y_i, i = 1, \ldots, n$ determine to which class $y_i = 1, \ldots, m$ each X_i belongs. The distribution $p(x_i \mid y_i)$ of X_i depends on its class y_i. The details of these distributions as well as the (prior) probability of membership in class are determined by model parameters θ. Known are only the realizations x_i of X_i but not the class (and hence distribution) of each X_i. The task is to find a maximum likelihood estimate of the parameters θ and the probability $p(y_i \mid x_i, \theta)$ that X_i belongs to class y_i under model θ.

From $p(x) = p(x, z)/p(z \mid x)$ applied to the above probabilities, taking logarithm and multiplying by $p(y_i \mid x_i, \bar\theta)$ gives

$$p(y_i \mid x_i, \bar\theta) \log p(x_i \mid \theta) = p(y_i \mid x_i, \bar\theta) \log p(x_i, y_i \mid \theta) - p(y_i \mid x_i, \bar\theta) \log p(y_i \mid x_i, \theta) \tag{36}$$

Summing over $y_i = 1, \ldots, m$ and $i = 1, \ldots, n$ (and using $\sum_{y_i} p(y_i \mid x_i, \theta) = 1$) we have for the loglikelihood of θ

$$\begin{aligned} L(\theta) &= \sum_i \log p(x_i \mid \theta) \\ &= \sum_i \sum_{y_i} p(y_i \mid x_i, \bar\theta) \log p(x_i, y_i \mid \theta) - \sum_i \sum_{y_i} p(y_i \mid x_i, \bar\theta) \log p(y_i \mid x_i, \theta) \end{aligned} \tag{37}$$

The entropy of the hidden variables is

$$H(\bar\theta) = \sum_i \sum_{y_i} p(y_i \mid x_i, \bar\theta) \log p(y_i \mid x_i, \bar\theta)$$

We define

$$\begin{aligned} F(\bar\theta, \theta) &= \sum_i \sum_j p(y_i \mid x_i, \bar\theta) \log p(x_i, y_i \mid \theta) - H(\bar\theta) \\ &= \sum_i \sum_j p(y_i \mid x_i, \bar\theta) \log \left(p(x_i \mid y_i, \theta) p(y_i \mid \theta) \right) - H(\bar\theta) \end{aligned} \tag{38}$$

and the Kullback-Leibler divergence (or relative entropy)

$$D(\bar{\theta}, \theta) = H(\bar{\theta}) - \sum_i \sum_{y_i} p(y_i \mid x_i, \bar{\theta}) \log p(y_i \mid x_i, \theta)$$

By subtracting and adding the entropy $H(\bar{\theta})$ and using the above notation the loglikehood might be decomposed into

$$L(\theta) = F(\bar{\theta}, \theta) + D(\bar{\theta}, \theta) \tag{39}$$

Since the relative entropy $D(\bar{\theta}, \theta)$ is nonnegative for all θ and $D(\bar{\theta}, \theta) = 0$ for $\theta = \bar{\theta}$, we have that $F(\bar{\theta}, \theta)$, as a function of θ, is always smaller than $L(\theta)$ for any θ and that $L(\bar{\theta}) = F(\bar{\theta}, \bar{\theta})$. Together this implies that if $F(\bar{\theta}, \theta) > F(\bar{\theta}, \bar{\theta})$ then $L(\theta) > L(\bar{\theta})$. The relative entropy $D(\bar{\theta}, \theta)$, as a function of θ, reaches a minimum in $\theta = \bar{\theta}$, hence its derivative is also zero in $\bar{\theta}$. This means the gradient of $L(\theta)$ and $F(\bar{\theta}, \theta)$ is the same in $\bar{\theta}$, which means that if $\bar{\theta}$ is not an extremum of L so it cannot be one of $F(\bar{\theta}, \theta)$ either (of, if $L(\theta)$ can be increased so can $F(\bar{\theta}, \theta)$).

These considerations provide the rationale for the following algorithm. Start with an arbitrary distribution $\bar{\theta} = \theta_0$ and iterate an E- and an M-step improving $L(\theta)$ until convergence. The *E-step* consists in computing the posterior probability $p(y_i \mid x_i, \bar{\theta})$ based on the current parameters $\bar{\theta}$ by Bayes formula

$$p(y_{i0} \mid x_i, \bar{\theta}) = \frac{p(x_i \mid y_{i0}, \bar{\theta}) p(y_{i0} \mid \bar{\theta})}{\sum_{y_i} p(x_i \mid y_i, \bar{\theta}) p(y_i \mid \bar{\theta})} \tag{40}$$

This is the probability that x_i belongs to class y_{i0} taking into account what probability this class y_{i0} would give x_i and what overall prior probability any point has of being a member this class; these probabilities are all given in $\bar{\theta}$. In the *M-step* $F(\bar{\theta}, \theta)$ as a function of θ is maximized; the probabilities $p(y_i \mid x_i, \bar{\theta})$ calculated in the previous E-step are kept fixed in this process. $\bar{\theta}$ is now set to θ. These two steps are repeated until the improvements in $L(\theta)$ are small or $\bar{\theta}$ is similar to θ. Note that this algorithm might get stuck in local maxima (or, theoretically, in saddle points).

The major advantage of the EM algorithm is that the classification probabilities $p(y_{i0} \mid x_i, \bar{\theta})$ don't need to be optimized together with the other parameters θ. The focus is alternatingly on θ and on the classification probabilities. For a discussion of the EM algorithm and variations relating $F(\theta, \theta)$ to the concept of the free energy see [1].

References

[1] R.M. Neal and G.E Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In M.I. Jordan, editor, *Learning in Graphical Models*, 1998. http://www.gatsby.ucl.ac.uk/Hinton/absps/emk.ps.

[2] M. A. Newton, C. M. Kendziorski, C. S. Richmond, F. R. Blattner, and K. W. Tsui. On Differential Variability of Expression Ratios: Improving Statistical Inference about Gene Expression Changes from Microarray Data. *J. Comput. Biol.*, 8:37–52, 2001.