It is important to note that this definition is somewhat broader than is often used in the wider community. Many times only methods dealing with the first problem have been referred to as background correction methods.

Unlike other array systems, such as cDNA microarrays, where pixels surrounding a spot can be used to compute the background adjustment, the probe intensities themselves must be used to determine any adjustment for Affymetrix Genechips. This is because probe locations are very densely spaced on the array.

2.2 Background Correction / Signal Adjustment Methods

2.2.1 RMA Convolution Model

The RMA convolution model background correction method is motivated by looking at the distribution of probe intensities. Figure 2.1 shows the probe intensity distribution for a group of typical arrays. We model the observed intensity as the sum of a signal and a background component. In particular, our model is that we observe $S = X + Y$, where X is signal and Y is background. Assume that X is distributed $\exp(\alpha)$ and that Y is distributed $\mathrm{N}\left(\mu, \sigma^2\right)$, with X and Y independent. Furthermore, assume that $Y \geq 0$ to avoid producing negative values. Thus, Y is normally distributed with truncation at 0. This model is motivated by the observed probe densities in Figure 2.1. Under this model the background corrected probe intensities will be given by $E(X|S = s)$. A formula for this quantity is derived below.

We define $\Phi(z)$ and $\phi(z)$ as the standard normal distribution function and density function respectively. More specifically

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}w^2\right) dw$$

and

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right).$$

Remembering that we observe only $S = X + Y$, under the conditions of this model, the density of the joint distribution of X and Y is given by

$$f_{X,Y}(x,y) = \alpha \exp(-\alpha x) \frac{1}{\sigma} \phi\left(\frac{y-\mu}{\sigma}\right) \text{ when } y > 0, x > 0$$

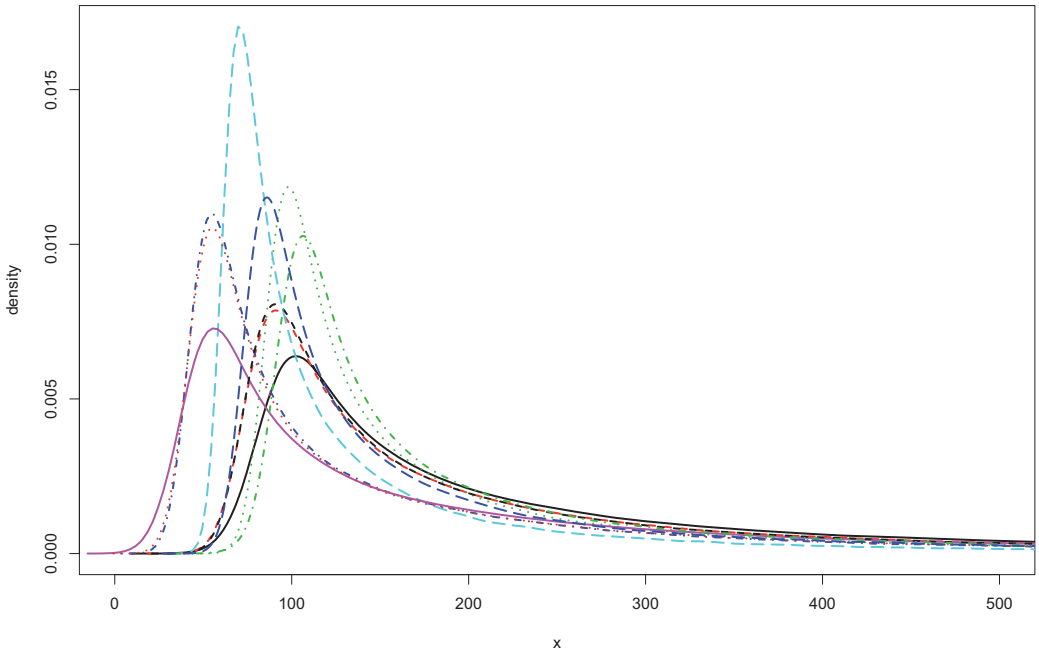

Figure 2.1: Smoothed histograms of the probe intensities for a number of arrays from the HGU95A spike-in dataset.

Then, we get the joint distribution of X and S from

$$f_{X,S}(x,s) = f_{X,Y}(x,s-x)\,|J|$$

where J is the Jacobian of the transformation. Now, $|J| = 1$ and so the joint distribution of X and S is

$$f_{X,S}(x,s) = \alpha\exp(-\alpha x)\frac{1}{\sigma}\phi\left(\frac{s-x-\mu}{\sigma}\right) = \alpha\exp(-\alpha x)\frac{1}{\sigma}\phi\left(\frac{x-s+\mu}{\sigma}\right)$$

The conditional distribution of X given S is

$$f_{X|S}(x|s) = \frac{f_{X,S}(x,s)}{\int_0^s f_{X,S}(x,s)\,dx}$$

where the denominator (the marginal pdf of S) is

$$\int_0^s \alpha\exp(-\alpha x)\frac{1}{\sigma}\phi\left(\frac{x-s+\mu}{\sigma}\right)dx$$

Let $w = \frac{x-s+\mu}{\sigma}$ so that $\sigma dw = dx$ and $x = \sigma w + s - \mu$. Making the substitution, the integral becomes

$$\begin{aligned}
&\int_{\frac{-s+\mu}{\sigma}}^{\frac{\mu}{\sigma}} \alpha\exp(-\alpha(\sigma w+s-\mu))\phi(w)\,dw \\
&= \alpha\exp(-\alpha(s-\mu))\int_{\frac{-s+\mu}{\sigma}}^{\frac{\mu}{\sigma}} \exp(-\alpha\sigma w)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}w^2\right)dw
\end{aligned}$$

Now, we consider the integral on the right hand side

$$\begin{aligned}
&\int_{\frac{-s+\mu}{\sigma}}^{\frac{\mu}{\sigma}} \frac{1}{\sqrt{2\pi}}\exp(-\alpha\sigma w)\exp(-\frac{1}{2}w^2)dw \\
&= \int_{\frac{-s+\mu}{\sigma}}^{\frac{\mu}{\sigma}} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(w^2+2\alpha\sigma w\right)\right)dw \\
&= \exp(\frac{1}{2}\alpha^2\sigma^2)\int_{\frac{-s+\mu}{\sigma}}^{\frac{\mu}{\sigma}} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(w^2+2\alpha\sigma w+\alpha^2\sigma^2\right)\right)dw \\
&= \exp(\frac{1}{2}\alpha^2\sigma^2)\int_{\frac{-s+\mu}{\sigma}}^{\frac{\mu}{\sigma}} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(w+\sigma\alpha)^2\right)dw
\end{aligned}$$

Let $z = w + \sigma\alpha$ and then the integral becomes

$$\int_{\frac{-s+\mu}{\sigma}+\alpha\sigma}^{\frac{\mu}{\sigma}+\alpha\sigma} \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}z^2)dz = \Phi\left(\frac{s-\mu-\alpha\sigma^2}{\sigma}\right)+\Phi\left(\frac{\mu+\alpha\sigma^2}{\sigma}\right)-1$$

and the denominator is

$$\alpha\exp\left(\frac{1}{2}\alpha^2\sigma^2-\alpha(s-\mu)\right)\left[\Phi\left(\frac{s-\mu-\alpha\sigma^2}{\sigma}\right)+\Phi\left(\frac{\mu+\alpha\sigma^2}{\sigma}\right)-1\right]$$

thus,

$$
\begin{aligned}
f_{X|S}(x|s) &= \frac{f_{X,S}(x,s)}{\int_0^s f_{X,S}(x,s)\,dx} \\
&= \frac{\alpha \exp(-\alpha x)\,\frac{1}{\sigma}\phi\left(\frac{x-s+\mu}{\sigma}\right)}{\alpha \exp(\frac{1}{2}\alpha^2\sigma^2 - \alpha(s-\mu))\left[\Phi\left(\frac{s-\mu-\alpha\sigma^2}{\sigma}\right)+\Phi\left(\frac{\mu+\alpha\sigma^2}{\sigma}\right)-1\right]} \\
&= \frac{\exp\left(-\alpha x+\alpha(s-\mu)-\frac{1}{2}\alpha^2\sigma^2\right)\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-s+\mu)^2\right)}{\left[\Phi\left(\frac{s-\mu-\alpha\sigma^2}{\sigma}\right)+\Phi\left(\frac{\mu+\alpha\sigma^2}{\sigma}\right)-1\right]} \\
&= \frac{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}\left(x^2-2x(s-\mu)+(s-\mu)^2+2\sigma^2\alpha x-2\sigma^2\alpha(s-\mu)+\alpha^2\sigma^4\right)\right)}{\left[\Phi\left(\frac{s-\mu-\alpha\sigma^2}{\sigma}\right)+\Phi\left(\frac{\mu+\alpha\sigma^2}{\sigma}\right)-1\right]} \\
&= \frac{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}\left(x^2-2x(s-\mu-\alpha\sigma^2)+(s-\mu)^2-2(s-\mu)\sigma^2\alpha+\alpha^2\sigma^4\right)\right)}{\left[\Phi\left(\frac{s-\mu-\alpha\sigma^2}{\sigma}\right)+\Phi\left(\frac{\mu+\alpha\sigma^2}{\sigma}\right)-1\right]} \\
&= \frac{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}\left(x-\left(s-\mu-\alpha\sigma^2\right)\right)^2\right)}{\left[\Phi\left(\frac{s-\mu-\alpha\sigma^2}{\sigma}\right)+\Phi\left(\frac{\mu+\alpha\sigma^2}{\sigma}\right)-1\right]}
\end{aligned}
$$

Let $a = s-\mu-\sigma^2\alpha$ and $b=\sigma$

Therefore, the conditional distribution of x given S is

$$f(x|s) = \frac{\frac{1}{b}\phi\left(\frac{x-a}{b}\right)}{\left[\Phi\left(\frac{a}{b}\right)+\Phi\left(\frac{s-a}{b}\right)-1\right]}$$

and so

$$E(x|s) = \frac{1}{\Phi\left(\frac{a}{b}\right)+\Phi\left(\frac{s-a}{b}\right)-1}\int_0^s \frac{x}{b}\phi\left(\frac{x-a}{b}\right)dx$$

Let $z=\frac{x-a}{b}$ so $dz=\frac{dx}{b}$. Thus

$$
\begin{aligned}
\int_0^s \frac{x}{b}\phi\left(\frac{x-a}{b}\right)dx &= \int_{-a/b}^{\frac{s-a}{b}}(bz+a)\,\phi(z)dz \\
&= a\int_{-a/b}^{\frac{s-a}{b}}\phi(z)dz + b\int_{-a/b}^{\frac{s-a}{b}} z\phi(z)dz \\
&= a\left[\Phi\left(\frac{s-a}{b}\right)+\Phi\left(\frac{a}{b}\right)-1\right]+b\left[\phi\left(\frac{a}{b}\right)-\phi\left(\frac{s-a}{b}\right)\right]
\end{aligned}
$$

and so

$$E(X|S=s) = a + b\frac{\phi\left(\frac{a}{b}\right)-\phi\left(\frac{s-a}{b}\right)}{\Phi\left(\frac{a}{b}\right)+\Phi\left(\frac{s-a}{b}\right)-1}$$

In most Affymetrix micorarray applications $\phi\left(\frac{s-a}{b}\right)$ is neglible and $\Phi\left(\frac{s-a}{b}\right)$ is close to one. So in practice, it is only necessary to compute the first term in the numerator and the first term in the denominator to make the adjustment.

It is somewhat troublesome to estimate the parameters μ, σ and α. Some approaches are either painfully slow (the EM algorithm) or numerically unstable (Newton methods). An ad-hoc approach is used to estimate the parameters. First, a non-parametric density estimate of the observed probe intensities on an array is taken, the mode of which is used as the estimate of μ. Then, the variabilty of the lower tail about μ is used for σ and an exponential is fitted to the right tail to estimate α.

In this thesis, we have elected to only adjust PM probe intensities because we focus on expression measures which use only PM probes, but in principle we could adjust MM probe intensities using this method, either separately or together with the PM probe intensities.

2.2.2 Methods Proposed by Affymetrix

There are two separate adjustment steps that have been proposed by Affymetrix (2002). For our analysis, they are considered both separately and in the sequence in which they are used in the MAS 5.0 software (Affymetrix, 2001a), which is the location specific correction followed by the ideal mismatch adjustment. It should be noted that we created our own implementations of these methods based upon the available documentation and there may be some slight differences from the Affymetrix software.

Location Specific Correction

The goal of this step is to remove overall background noise. Each array is divided into a set of regions, then a background value for that is grid estimated. Then each probe intensity is adjusted based upon a weighted average of each of the background values. The weights are dependent on the distance from the centroid of each of the grids. In particular, the weights are

$$w_k(x,y) = \frac{1}{d_k^2(x,y) + \text{smooth}}$$

where $d_k(x,y)$ is the euclidean distance from location x,y to the centroid of region k. The default value for smooth is 100. Special care is taken to avoid negative values or other numerical problems