Generalized beta convolution model of the true intensity for the Illumina BeadArrays

Rohmatul Fajriyah

Institute of Statistics, TU Graz, Austria

Dept. of Statistics, Univ Islam Indonesia, Jogjakarta, Indonesia

fajriyah@student.tugraz.at

Abstract

Microarray data come from many steps of production and have been known to contain noise. The pre-processing is implemented to reduce the noise, where the background is corrected. Prior to further analysis, many Illumina BeadArrays users had applied the convolution model, a model which had been adapted from when it was first developed on the Affymetrix platform, to adjust the intensity value: corrected background intensity value.

Several models based on different underlying distributions and or parameters estimation methods have been proposed and applied. For instance : the exponential-gamma, the normal-gamma and the exponential-normal convolutions with a maximum likelihood estimation, non-parametric, Bayesian and moment methods of the parameters estimation, including two recent exponential-lognormal and gamma-lognormal convolutions.

In this paper, we propose models and derive the background corrected true intensity value based on the generalized betas and the generalized beta-normal convolutions as a generalization of the existing models.

Key Words: background correction, additive error, generalized beta distribution fam-

ily, Illumina BeadArrays and convolution model.

1 Introduction

It has become common knowledge that data from microarray experiments will contain some non-biological noise. Therefore, the data needs to be adjusted. In this case, implementing the pre-processing will adjust (Huber et al. [1–3]) or correct the background intensity value.

There are several steps in pre-processing where one of the steps is the background correction. In the background correction, the noise can be modelled as additive or multiplicative (*See*, Huber et al. [1, 2], Bolstad et al. [4] and Irizarry et al. [5–7], Li and Wong [8], Silver et al. [9] and Wu et al. [10]).

In the robust multi-array average (RMA), Irizarry et al. [5–7] have modeled the noise as an additive, to adjust the intensity value. Although the RMA was developed for the Affymetrix platform initially, it was also been used for the data from the Illumina platform.

Currently, there are some models to correct the intensity value of the Illumina platform available, for instance : the model-based background correction method (MBCB) from Ding et al. [11] and Xie et al. [12], the exponential-gamma from Chen et al. [13], the gamma-normal from Plancade et al. [14] and the exponential(gamma)-lognormal from Fajriyah [15].

Posekany's et al. study [16] show us that by using the Affymetrix and Invitrogen platforms the noise in microarray data is not Gaussian but far more heavy-tailed. On the other hand, Chen et al. [13] show that the noise distribution in the Illumina platform is usually skewed in different degrees.

Therefore, while the intensity values are widely accepted as a skewed distribution, the noise distribution could possibly be symmetrical or skewed. Note that in this dissertation, noise and intensity mean the negative control probes and the observed probes intensity values respectively.

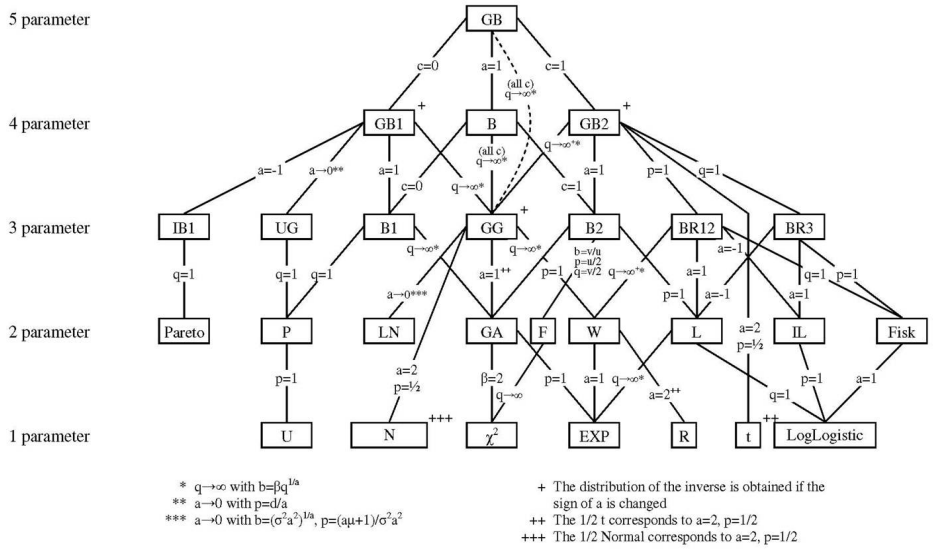

Figure 1.1: Distribution tree, [17]

McDonald and Xu [17] have introduced a distribution tree of generalized beta distributions, which is used to model the income distribution. It is similar in nature to the microarray data where the random variable is a non-negative value. This distribution tree helps us to understand the relationship among the available distributions. Moreover, quite recently, Leemis and McQueston [18] have explained the relationships among the univariate distributions in statistics. See the distribution tree from McDonald and Xu [17] in Figure 1.1.

This paper aims to present the true intensity value, the corrected background intensity, where the noise is a symmetric and skewed distribution. If the noise is a skewed distribution, the underlying distributions of the proposed convolution model are the generalized beta distributions, a generalized model of the existing ones. If the noise is a symmetrically distributed, the proposed model is a generalized beta-normal convolution, which is a generalized model of the Plancade et al. model [14].

In general, the background correction is applied toward each array, where in each array there are probes (perfect match and mismatch probes), probesets and genes (terminology for the Affymetrix platform) or bead and bead-type level probes (ter-

minology for the Illumina platform).

The current publicly available benchmarking data set for the Illumina platform is the raw data from the bead studio, which is the average of the bead-type level probes, not background corrected and of unnormalized intensity. Therefore, the background correction in this paper is applied to the gene (bead-type level probes) intensity in each array.

Suppose we have J arrays and for each array there are I regular genes and M negative control genes. Throughout the paper, the convolution model is applied for each array j and represented as follows:

$$P_i = S_i + B_i \tag{1}$$

where P_i, S_i, and B_i are the regular (observed) true/corrected background and noise intensity values respectively of the i^{th} gene, $i = 1, ..., I$. For a negative control gene w at array j, $w = 1, 2, ..., W$, the observed intensity, denoted by P_{0w} is assumed to be $P_{0w} = B_{0w}$, where B_{0w} is the noise intensity. P_i and P_{0w} are assumed to be independent.

This paper is organized as follows: Section 2 reviews previous work related to the background correction for the Illumina BeadArrays, Section 3 explains the results of our investigation and Section 4 provides discussion and remarks.

2 Previous work

2.1 Basic concepts

Definition 2.1. *Suppose X is a random variable of generalized beta distribution. McDonald and Xu [17] define the probability function of the generalized beta distribution*

as follows

$$Y \sim GB(y; a, c, d, u, v) = \frac{|a|\, y^{au-1} \left(1 - (1-c)\left(\frac{y}{d}\right)^a\right)^{v-1}}{d^{au} B(u,v) \left(1 + c\left(\frac{y}{d}\right)^a\right)^{u+v}}, 0 < y^a < \frac{d^a}{1-c}, \quad (2)$$

and zero otherwise, with $B(u,v)$ *is the beta function,*$0 \leq c \leq 1$*,* a, d, u *and* v *positive.*

Definition 2.2. *Let* X *and* Y *be two continuous random variables with density functions* $f_1(x)$ *and* $f_2(y)$ *respectively. Assume that both* $f_1(x)$ *and* $f_2(y)$ *are defined for all real numbers. Then the* convolution $f_1 * f_2$ *of* f_1 *and* f_2 *is the function given by*

$$\begin{aligned}(f_1 * f_2)(z) &= \int_{-\infty}^{+\infty} f_1(z-y) f_2(y) dy \\ &= \int_{-\infty}^{+\infty} f_2(z-x) f_1(x) dx \end{aligned} \quad (3)$$

Theorem 2.1. *Let* X *and* Y *be two independent random variables with density functions* $f_X(x)$ *and* $f_Y(y)$ *respectively defined for all* x*. Then the sum* $Z = X + Y$ *is a random variable with a density function of* $f_Z(z)$*, where* f_Z *is the convolution of* f_X *and* f_Y*.*

2.2 Background correction by RMA

In the RMA model ([4], and [5–7]), it is assumed that the intensity values are affected by the noise of the chip. The RMA model is as in the Equation (1), where $P_i = PM_i$ is the observed probe level intensity of perfect match probes of the i^{th} gene, S_i is the true intensity of the i^{th} gene, with $S_i \sim f_1(s_i; \theta_j) = \text{Exp}(\theta_j), \theta_j > 0$, and B_i is the background noise of the i^{th} gene with $B_i \sim f_2(b_i; \mu_j, \sigma_j^2) = \mathcal{N}\left(\mu_j, \sigma_j^2\right), \mu_j \in \mathbb{R}, \sigma_j^2, b_i > 0$.

Assuming independence, the joint density of the two-dimensional random variables (S_i, B_i) is

$$f_{S_i,B_i}(s_i, b_i; \mu_j, \sigma_j^2, \theta_j) = \theta_j e^{-s_i\theta_j} f_2\left(b_i; \mu_j, \sigma_j^2\right), s_i > 0\,. \tag{4}$$

Furthermore, the transformation formula for two-dimensional densities gives the joint density of S_i and P_i is

$$\begin{aligned} &f_{S_i,P_i}\left(s_i, p_i; \mu_j, \sigma_j^2, \theta_j\right) \\ &= \theta_j e^{\left(\frac{\theta_j^2\sigma_j^2}{2} - (p_i-\mu_j)\theta_j\right)} f_2\left(s_i; p_i - \mu_j - \sigma_j^2\theta, \sigma_j^2\right), 0 < s_i < p_i \end{aligned} \tag{5}$$

From equation (5) we get the marginal density of P_i and the conditional density of S_i given P_i in equations (6) and (7) below, respectively:

$$f_{P_i}(p_i) = \theta e^{\left\{\frac{\sigma_j^2}{2\theta_j^2} - \frac{(p_i-\mu_j)}{\theta_j}\right\}} \left(\Phi\left(\frac{\mu_{S.P}}{\sigma_j}\right) + \Phi\left(\frac{p_i - \mu_{S.P}}{\sigma_j}\right) - 1\right) \tag{6}$$

$$f_{S_i|P_i}(s_i \mid p_i) = \frac{f_2(s; \mu_{S,P}, \sigma_j^2)}{\left(\Phi\left(\frac{\mu_{S.P}}{\sigma_j}\right) + \Phi\left(\frac{p-\mu_{S.P}}{\sigma_j}\right) - 1\right)} \tag{7}$$

where $\mu_{S.P} = p_i - \mu_j - \sigma_j^2\theta_j$.

The corrected background intensity is computed by the conditional expectation

$$E(S_i \mid P_i = p_i) = \frac{1}{\left(\Phi\left(\frac{\mu_{S.P}}{\sigma_j}\right) + \Phi\left(\frac{p_i-\mu_{S.P}}{\sigma_j}\right) - 1\right)} \int_0^{p_i} f_2\left(s_i; \mu_{S.P}, \sigma_j^2\right) ds \tag{8}$$

The substitution $s_i = \mu_{S,P} + \sigma_j t_i$, yields the corrected background intensity in the

Equation (8) equal to

$$\mu_{S.P} + \sigma_j \frac{\phi\left(\frac{\mu_{S.P}}{\sigma_j}\right) - \phi\left(\frac{p_i - \mu_{S.P}}{\sigma_j}\right)}{\Phi\left(\frac{\mu_{S.P}}{\sigma_j}\right) + \Phi\left(\frac{p_i - \mu_{S.P}}{\sigma_j}\right) - 1} \tag{9}$$

2.3 Exponential-normal MBCB

Xie et al. [12] use the same underlying distributions as the RMA for the background correction. The differences between the MBCB and the RMA ([4], and [5–7]) are

1. Xie et al. [12] take the infinite value for the upper bound of the integral to compute the marginal density function and the conditional expectation of the true intensity value. On the other hand, the RMA puts p as the upper bound of the integral.

   The corrected background intensity of this model is

   $$\mu_{S.P} + \sigma_j \frac{\phi\left(\frac{\mu_{S.P}}{\sigma_j}\right)}{\Phi\left(\frac{\mu_{S.P}}{\sigma_j}\right)} \tag{10}$$

2. Under the convolution model (1), where the true intensity value is assumed exponentially distributed and the noise is normally distributed, we then need to estimate the parameters θ_j, μ_j, and σ_j^2. Xie et al. [12] offer three parameters estimation methods: the non-parametric, maximum likelihood and Bayesian. On the other hand, the RMA applies the *ad-hoc* method.

Ding et al. [11] use the exponential-normal convolution model to correct the background of the Illumina platform by using a Markov chain Monte Carlo simulation.

2.4 Gamma-normal convolution

Plancade et al. [14] introduced gamma-normal convolution to model the background correction of the Illumina BeadArrays. The model is based on the RMA background correction of Affymetrix GeneChips. Plancade et al. [14] assume that the true intensity value is gamma distributed and the noise is normally distributed.

Under the model background correction in (1), f_{P_i} is the convolution product of f_{S_i} and f_{B_i}. The true intensity S_i is computed by the conditional expectation of S_i given $P_i = p_i$:

$$E\left(S_i \mid P_i = p_i\right) = \tilde{S}_i\left(p_i\right) = \frac{\int s_i f^{\text{gam}}_{\alpha_j,\theta_j}(s) f^{\text{norm}}_{\mu_j,\sigma_j}\left(p_i - s_i\right) ds}{\int f^{\text{gam}}_{\alpha_j,\theta_j}(s) f^{\text{norm}}_{\mu_j,\sigma_j}(p_i - s_i) ds} \tag{11}$$

where $f^{\text{gam}}_{\alpha_j,\theta_j}\left(x_i;\alpha_j,\theta_j\right) = \frac{\theta_j^{\alpha_j} x_i^{\alpha_j - 1} e^{-\theta_j x_i}}{\Gamma\left(\alpha_j\right)}, \quad \alpha_j, \theta_j, x_i > 0$ is the gamma density.

When S_i is gamma distributed and B_i is normally distributed, then the equation (11) does not have analytic expression as it does in Equations (9) and (10). Therefore, Plancade et al. [14] implemented the Fast Fourier Transform to estimate the parameters and to correct the background. For the background correction with Fast Fourier Transform, Equation (11) is rewritten as

$$\widetilde{S}_i(p_i \mid \Theta) = \frac{\alpha_j \theta_j \int f^{\text{gam}}_{\alpha_j+1,\theta_j}\left(s_i\right) f^{\text{norm}}_{\mu_j,\sigma_j}\left(p_i - s_i\right) ds_i}{\int f^{\text{gam}}_{\alpha_j,\theta_j}\left(s_i\right) f^{\text{norm}}_{\mu_j,\sigma_j}\left(p_i - s_i\right) ds}, \tag{12}$$

where $\Theta = \left(\mu_j, \sigma_j, \alpha_j, \theta_j\right)$, and $s_i f^{\text{gam}}_{\alpha_j,\theta_j}(s_i) = \alpha_j \theta_j f^{\text{gam}}_{\alpha_j+1,\theta_j}\left(s_i\right)$ is valid for every $s_i > 0$.

2.5 Exponential-gamma convolution

Chen et al. [13] proposed in favor of the distribution of the true intensity and its noise, under the convolution model of Equation (2), the exponential and gamma distributions respectively. Therefore, $S_i \sim f_1(s_i;\theta_j) = \text{Exp}\left(\theta_j\right)$, and $B \sim f\left(b_i;\alpha_j,\beta_j\right) = \text{GAM}\left(\alpha_j,\beta_j\right)$, where $s_i, b_i, \theta_j, \alpha_j, \beta_j > 0$.

The corrected background intensity for the proposed model ([13]) is :

$$p_i - \frac{\int_0^{p_i} b_i^{\alpha_j} e^{-\left(\frac{1}{\beta_j}-\theta_j\right)b_i} db_i}{\int_0^{p_i} b_i^{\alpha_j-1} e^{-\left(\frac{1}{\beta_j}-\theta_j\right)b_i} db_i}. \tag{13}$$

2.6 Exponential-lognormal convolution, [15]

Under model (1), when the true intensity S_i is assumed to be exponentially distributed $S_i \sim f_1\left(s_i;\theta_j\right) = \theta_j e^{-\theta_j s_i}, \theta_j, s_i > 0$, and the background noise B is assumed to be lognormally distributed, $B_i \sim f_2(b_i;\mu_j,\sigma_j^2) = \frac{e^{-\frac{\left(\ln b_i-\mu_j\right)^2}{2\sigma_j^2}}}{b_i\sigma_j\sqrt{2\pi}}, \mu_j \in \mathbb{R}, \sigma_j^2, b_i > 0$, the corrected background intensity is

$$p_i - \frac{e^{\mu_j+\frac{\sigma_j^2}{2}} C_{2,j}}{C_{1,j}} \tag{14}$$

where
$C_{2,j} = \sum_{k=0}^{\infty} \frac{\theta_j^k}{k!} e^{k\left(\mu_j+\frac{k+2}{2}\sigma_j^2\right)} \Phi\left(\frac{\ln p_i - \left(\mu_j+(k+1)\sigma_j^2\right)}{\sigma_j}\right)$, and
$C_{1,j} = \sum_{k=0}^{\infty} \frac{\theta_j^k}{k!} e^{k\left(\mu_j+\frac{k}{2}\sigma_j^2\right)} \Phi\left(\frac{\ln p_i - \left(\mu_j+k\sigma_j^2\right)}{\sigma_j}\right)$

2.7 Gamma-lognormal convolution, [15]

Under model (1), when the true intensity S_i is assumed to be gamma distributed $S_i \sim f_1(s_i;\alpha_j,\beta_j) = \frac{s_i^{\alpha_j-1} e^{-\frac{s_i}{\beta_j}}}{\beta_j^{\alpha_j}\Gamma\left(\alpha_j\right)}, \alpha_j, \beta_j, s_i > 0$, and the background noise B_i is assumed to be lognormally distributed, $B_i \sim f_2(b_i;\mu_j,\sigma_j^2) = \frac{e^{-\frac{\left(\ln b_i-\mu_j\right)^2}{2\sigma_j^2}}}{b_i\sigma_j\sqrt{2\pi}}, \mu_j \in \mathbb{R}, \sigma_j^2, b_i > 0$, the corrected background intensity is

$$\frac{p_i C_{4,j}}{C_{3,j}} \tag{15}$$

where

$$C_{4,j} = \sum_{k=0}^{\infty} \sum_{n=0}^{\infty} \frac{(-1)^k \binom{\alpha_j}{k} e^{(k+n)\left(\mu_j+(k+n)\frac{\sigma_j^2}{2}\right)} \Phi\left(\frac{\ln p_i - \left(\mu_j+(k+n)\sigma_j^2\right)}{\sigma_j}\right)}{p_i^k \beta_j^n n!}, \text{ and}$$

$$C_{3,j} = \sum_{k=0}^{\infty} \sum_{n=0}^{\infty} \frac{(-1)^k \binom{\alpha_j - 1}{k} e^{(k+n)\left(\mu+(k+n)\frac{\sigma_j^2}{2}\right)} \Phi\left(\frac{\ln p_i - (\mu_j+(k+n)\sigma_j^2)}{\sigma_j}\right)}{p_i^k \beta_j^n n!}$$

The exponential-lognormal and gamma-lognormal models, [15] implements three methods for the parameters estimation: Maximum likelihood estimation (MLE), method of moments, and plug-in.

3 Results

In the subsequent sections, we will explain the generalized beta convolution model and its corrected background intensity value.

3.1 Generalized beta distribution convolution

3.1.1 The joint density function

Under the convolution model of Equation (1), where P_i is the observed intensity of regular probes of the i^{th} gene, S_i is the true intensity of the i^{th} gene, with

$$\begin{aligned} S_i &\sim f_1\left(s_i; a_{1,j}, c_{1,j}, d_{1,j}, u_{1,j}, v_{1,j}\right) \\ &= \frac{\left|a_{1,j}\right| s_i^{a_{1,j}u_{1,j}-1} \left(1-\left(1-c_{1,j}\right)\left(\frac{s_i}{d_{1,j}}\right)^{a_{1,j}}\right)^{v_{1,j}-1}}{d_1^{a_{1,j}u_{1,j}} B\left(u_{1,j}, v_{1,j}\right)\left(1+c_{1,j}\left(\frac{s_i}{d_{1,j}}\right)^{a_{1,j}}\right)^{u_{1,j}+v_{1,j}}}, \qquad (16) \\ 0 \le c_{1,j} \le 1, &a_{1,j}, d_{1,j}, u_{1,j} \text{ and } v_{1,j} \text{ positive}, s_i > 0 \end{aligned}$$

and B_i is the background noise with

$$\begin{aligned} B_i &\sim f_2\left(b_i; a_{2,j}, c_{2,j}, d_{2,j}, u_{2,j}, v_{2,j}\right) \\ &\frac{\left|a_{2,j}\right| b_i^{a_{2,j}u_{2,j}-1}\left(1-\left(1-c_{2,j}\right)\left(\frac{b_i}{d_{2,j}}\right)^{a_{2,j}}\right)^{v_{2,j}-1}}{d_2^{a_{2,j}u_{2,j}} B\left(u_{2,j}, v_{2,j}\right)\left(1+c_{2,j}\left(\frac{b_i}{d_{2,j}}\right)^{a_{2,j}}\right)^{u_{2,j}+v_{2,j}}}, \\ &0 \leq c_{2,j} \leq 1, a_{2,j}, d_{2,j}, u_{2,j} \text{ and } v_{2,j} \text{ positive}, b_i > 0 \end{aligned} \tag{17}$$

The joint density function of S_i and B_i is :

$$\begin{aligned} f_{S_i,B_i}\left(s_i, b_i\right) = &\frac{\left|a_1\right| s_i^{a_{1,j}u_{1,j}-1}\left(1-\left(1-c_{1,j}\right)\left(\frac{s_i}{d_1}\right)^{a_{1,j}}\right)^{v_{1,j}-1}}{d_1^{a_{1,j}u_{1,j}} B\left(u_{1,j}, v_{1,j}\right)\left(1+c_{1,j}\left(\frac{s_i}{d_{1,j}}\right)^{a_{1,j}}\right)^{u_{1,j}+v_{1,j}}} \times \\ &\frac{\left|a_{2,j}\right| b_i^{a_{2,j}u_{2,j}-1}\left(1-\left(1-c_{2,j}\right)\left(\frac{b_i}{d_{2,j}}\right)^{a_{2,j}}\right)^{v_{2,j}-1}}{d_2^{a_{2,j}u_{2,j}} B\left(u_{2,j}, v_{2,j}\right)\left(1+c_{2,j}\left(\frac{b_i}{d_{2,j}}\right)^{a_{2,j}}\right)^{u_{2,j}+v_{2,j}}} \end{aligned} \tag{18}$$

The joint density function of S_i and P_i is

$$\begin{aligned} f_{S_i,P_i}\left(s_i, p_i\right) = &\frac{\left|a_1\right| s_i^{a_{1,j}u_{1,j}-1}\left(1-\left(1-c_{1,j}\right)\left(\frac{s_i}{d_{1,j}}\right)^{a_{1,j}}\right)^{v_{1,j}-1}}{d_1^{a_{1,j}u_{1,j}} B\left(u_1, v_{1,j}\right)\left(1+c_{1,j}\left(\frac{s_i}{d_{1,j}}\right)^{a_{1,j}}\right)^{u_{1,j}+v_{1,j}}} \times \\ &\frac{\left|a_{2,j}\right| \left(p_i-s_i\right)^{a_{2,j}u_{2,j}-1}\left(1-\left(1-c_{2,j}\right)\left(\frac{(p_i-s_i)}{d_{2,j}}\right)^{a_{2,j}}\right)^{v_{2,j}-1}}{d_2^{a_{2,j}u_{2,j}} B\left(u_{2,j}, v_{2,j}\right)\left(1+c_{2,j}\left(\frac{(p_i-s_i)}{d_{2,j}}\right)^{a_{2,j}}\right)^{u_{2,j}+v_{2,j}}} \end{aligned} \tag{19}$$

3.1.2 The marginal density function

The marginal density function of P_i is

$$
\begin{aligned}
f_{P_i}(p_i) &= \int_0^{p_i} f_{S_i,P_i}(s_i,p_i)\,ds_i \\
&= K\sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\sum_{r=0}^{\infty}\left\{\frac{(-1)^{l+m+n+r}\left(1-c_{1,j}\right)^{l}\left(1-c_{2,j}\right)^{m}c_{1,j}^{n}c_{2,j}^{r}}{d_{1,j}^{a_{1,j}(l+n)}d_{2,j}^{a_{2,j}(m+r)}}\times\right. \\
&\binom{v_{1,j}-1}{l}\binom{v_{2,j}-1}{m}\binom{u_1+v_1+n-1}{n}\binom{u_{2,j}+v_{2,j}+r-1}{r}\times \\
&\left.\int_0^{p_i} s_i^{a_{1,j}(u_{1,j}+l+n)-1}\left(p_i-s_i\right)^{a_{2,j}\left(u_{2,j}+m+r\right)-1}ds_i\right\} \qquad (20)
\end{aligned}
$$

Let $\frac{s_i}{p_i}=z_i$, then the equation (20) becomes

$$
K_1 p_i^{a_{1,j}u_1+a_{2,j}u_{2,j}-1}C_{5,j} \qquad (21)
$$

where

$$
K_1=\frac{\mid a_{1,j}\mid\mid a_{2,j}\mid}{d_1^{a_{1,j}u_{1,j}}d_2^{a_{2,j}u_{2,j}}B(u_{1,j},v_{1,j})B(u_{2,j},v_{2,j})},
$$

and

$$
\begin{aligned}
C_{5,j}=\sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\sum_{r=0}^{\infty}&\left\{\frac{(-1)^{l+m+n+r}\left(1-c_{1,j}\right)^{l}\left(1-c_{2,j}\right)^{m}c_{1,j}^{n}c_{2,j}^{r}\binom{v_{1,j}-1}{l}\binom{v_{2,j}-1}{m}}{d_{1,j}^{a_{1,j}(l+n)}d_{2,j}^{a_{2,j}(m+r)}}\times\right. \\
&\binom{u_{1,j}+v_{1,j}+n-1}{n}\binom{u_{2,j}+v_{2,j}+r-1}{r}p_i^{a_{1,j}(l+n)+a_{2,j}(m+r)}\times \\
&\left.\mathrm{B}\left(a_{1,j}\left(u_{1,j}+l+n\right)-1,a_{2,j}\left(u_{2,j}+m+r\right)-1\right)\right\}
\end{aligned}
$$

3.1.3 The conditional density function

The conditional density function of S_i where it is known that $P_i = p_i$ is

$$\begin{aligned} f_{S_i|P_i}\left(s_i \mid p_i\right) &= \frac{f_{S_i,P_i}\left(s_i, p_i\right)}{f_{P_i}\left(p_i\right)} \\ &= \frac{s_i^{a_{1,j}u_{1,j}-1}\left(1-\left(1-c_{1,j}\right)\left(\frac{s_i}{d_{1,j}}\right)^{a_{1,j}}\right)^{v_{1,j}-1}\left(p_i-s_i\right)^{a_{2,j}u_{2,j}-1}}{p_i^{a_{1,j}u_{1,j}+a_{2,j}u_{2,j}-1}C_{5,j}\left(1+c_{1,j}\left(\frac{s_i}{d_{1,j}}\right)^{a_{1,j}}\right)^{u_{1,j}+v_{1,j}}} \times \\ &\quad \frac{\left(1-\left(1-c_{2,j}\right)\left(\frac{(p_i-s_i)}{d_{2,j}}\right)^{a_{2,j}}\right)^{v_{2,j}-1}}{\left(1+c_{2,j}\left(\frac{(p_i-s_i)}{d_{2,j}}\right)^{a_{2,j}}\right)^{u_{2,j}+v_{2,j}}} \end{aligned} \tag{22}$$

3.1.4 The corrected background intensity

The corrected background intensity under this generalized beta convolution is

$$p_i \frac{C_{6,j}}{C_{5,j}} \tag{23}$$

where

$$\begin{aligned} C_{6,j} = &\sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\sum_{r=0}^{\infty}\left\{\frac{(-1)^{l+m+n+r}\left(1-c_{1,j}\right)^{l}\left(1-c_{2,j}\right)^{m}c_{1,j}^{n}c_{2,j}^{r}\binom{v_{1,j}-1}{l}\binom{v_{2,j}-1}{m}}{d_{1,j}^{a_{1,j}(l+n)}d_{2,j}^{a_{2,j}(m+r)}} \times \right. \\ &\binom{u_{1,j}+v_{1,j}+n-1}{n}\binom{u_{2,j}+v_{2,j}+r-1}{r}p_i^{a_{1,j}(l+n)+a_{2,j}(m+r)} \times \\ &\left.\mathrm{B}\left(a_{1,j}\left(u_{1,j}+l+n\right), a_{2,j}\left(u_{2,j}+m+r\right)-1\right)\right\} \end{aligned}$$

3.1.5 The likelihood function

The likelihood function ($\mathbf{L}$) to estimate $a_{1,j}, c_{1,j}, d_{1,j}, u_{1,j}, v_{1,j}, a_{2,j}, c_{2,j}, d_{2,j}, u_{2,j}$ and $v_{2,j}$ is

$$
\begin{aligned}
&= \prod_{i=1}^{I} \frac{\left|a_{1,j}\right|\left|a_{2,j}\right| p_i^{a_{1,j}u_1 + a_{2,j}u_{2,j} - 1} C_{5,j}}{d_2^{a_{1,j}u_{1,j}} d_{2,j}^{a_{2,j}u_{2,j}} \mathrm{B}\left(u_1, v_{1,j}\right) B\left(u_{2,j}, v_{2,j}\right)} \times \\
&\prod_{w=1}^{W} \frac{\left|a_{2,j}\right| b_{0w}^{a_{2,j}u_{2,j}-1} \left(1 - \left(1 - c_{2,j}\right)\left(\frac{b_{0w}}{d_{2,j}}\right)^{a_{2,j}}\right)^{v_{2,j}-1}}{d_2^{a_{2,j}u_{2,j}} B\left(u_{2,j}, v_{2,j}\right) \left(1 + c_{2,j}\left(\frac{b_{0w}}{d_{2,j}}\right)^{a_{2,j}}\right)^{u_{2,j}+v_{2,j}}}
\end{aligned} \tag{24}
$$

The log-likelihood function l is

$$
\begin{aligned}
&= \sum_{i=1}^{I} \Bigg\{ \ln\left(\left|a_{1,j}\right|\right) + \ln\left(\mid a_{2,j} \mid\right) + \left(a_{1,j}u_{1,j} + a_{2,j}u_{2,j} - 1\right)\ln\left(p_i\right) \\
&+ \ln\left(C_{5,j}\right) - \left(a_{1,j}u_{1,j}\right)\ln\left(d_{1,j}\right) - \left(a_{2,j}u_{2,j}\right)\ln\left(d_{2,j}\right) - \ln\left(\mathrm{B}\left(u_{1,j}, v_{1,j}\right)\right) \\
&- \ln\left(\mathrm{B}\left(u_{2,j}, v_{2,j}\right)\right) \Bigg\} + \sum_{w=1}^{W} \Bigg\{ \ln\left(\mid a_{2,j} \mid\right) + \left(a_{2,j}u_{2,j} - 1\right)\ln\left(b_{0w}\right) \\
&+ \left(v_{2,j} - 1\right)\ln\left(\left(1 - \left(1 - c_{2,j}\right)\left(\frac{b_{0w}}{d_{2,j}}\right)^{a_{2,j}}\right)\right) - \left(a_{2,j}u_{2,j}\right)\ln\left(d_{2,j}\right) \\
&- \ln\left(\mathrm{B}\left(u_{2,j}, v_{2,j}\right)\right) - \left(u_{2,j} + v_{2,j}\right)\ln\left(\left(1 + c_{2,j}\left(\frac{b_{0w}}{d_{2,j}}\right)^{a_{2,j}}\right)\right) \Bigg\}
\end{aligned} \tag{25}
$$

The likelihood equations are as follows

$$
\begin{aligned}
\frac{\partial l}{\partial a_{2,j}} = \sum_{w=1}^{W} \Bigg(&\frac{1}{\left|a_{2,j}\right|} + u_{2,j}\ln\left(b_{0w}\right) + \left(v_{2,j} - 1\right)\frac{-\left(1 - c_{2,j}\right)\ln\left(\frac{b_{0w}}{d_{2,j}}\right)\left(\frac{b_{0w}}{d_{2,j}}\right)^{a_{2,j}}}{\left(1 - \left(1 - c_{2,j}\right)\left(\frac{b_{0w}}{d_{2,j}}\right)^{a_{2,j}}\right)} \\
&- u_{2,j}\ln\left(d_{2,j}\right) - \left(u_{2,j} + v_{2,j}\right)\frac{c_{2,j}\ln\left(\frac{b_{0w}}{d_{2,j}}\right)\left(\frac{b_{0w}}{d_{2,j}}\right)^{a_{2,j}}}{1 + c_{2,j}\left(\frac{b_{0w}}{d_{2,j}}\right)^{a_{2,j}}} \Bigg) = 0
\end{aligned} \tag{26}
$$

$$\frac{\partial l}{\partial c_{2,j}} = \sum_{w=1}^{W} \left(\left(v_{2,j} - 1\right) \frac{\left(\frac{b_{0w}}{d_{2,j}}\right)^{a_{2,j}}}{\left(1 - \left(1 - c_{2,j}\right)\left(\frac{b_{0w}}{d_{2,j}}\right)^{a_{2,j}}\right)} - \left(u_{2,j} + v_{2,j}\right) \frac{\left(\frac{b_{0w}}{d_{2,j}}\right)^{a_{2,j}}}{\left(1 + c_{2,j}\left(\frac{b_{0w}}{d_{2,j}}\right)^{a_{2,j}}\right)} \right) = 0 \tag{27}$$

$$\frac{\partial l}{\partial d_{2,j}} = \sum_{w=1}^{W} \left(\left(v_{2,j} - 1\right) \frac{-\left(1 - c_{2,j}\right) b_{0w}^{a_{2,j}} \left(-a_{2,j}\right) d_{2,j}^{-\left(a_{2,j}+1\right)}}{\left(1 - \left(1 - c_{2,j}\right)\left(\frac{b_{0w}}{d_{2,j}}\right)^{a_{2,j}}\right)} - \frac{a_{2,j} u_{2,j}}{d_{2,j}} - \left(u_{2,j} + v_{2,j}\right) \frac{c_{2,j} b_{0w}^{a_{2,j}} \left(-a_{2,j}\right) d_{2,j}^{-\left(a_{2,j}+1\right)}}{\left(1 + c_{2,j}\left(\frac{b_{0w}}{d_{2,j}}\right)^{a_{2,j}}\right)} \right) = 0 \tag{28}$$

$$\frac{\partial l}{\partial u_{2,j}} = \sum_{w=1}^{W} \left(a_{2,j} \ln\left(b_{0w}\right) + a_{2,j} \ln\left(d_{2,j}\right) - \frac{\frac{\partial \mathrm{B}\left(u_{2,j}, v_{2,j}\right)}{\partial u_{2,j}}}{\mathrm{B}\left(u_{2,j}, v_{2,j}\right)} - \ln\left(1 + c_{2,j}\left(\frac{b_{0w}}{d_{2,j}}\right)^{a_{2,j}}\right) \right) = 0 \tag{29}$$

$$\frac{\partial l}{\partial v_{2,j}} = \sum_{w=1}^{W} \left(\ln\left(1 - \left(1 - c_{2,j}\right)\left(\frac{b_{0w}}{d_{2,j}}\right)^{a_{2,j}}\right) - \frac{\frac{\partial \mathrm{B}\left(u_{2,j}, v_{2,j}\right)}{\partial v_{2,j}}}{\mathrm{B}\left(u_{2,j}, v_{2,j}\right)} - \ln\left(1 + c_{2,j}\left(\frac{b_{0w}}{d_{2,j}}\right)^{a_{2,j}}\right) \right) = 0 \tag{30}$$

$$\frac{\partial l}{\partial a_{1,j}} = \sum_{i=1}^{I} \left(\frac{1}{|a_{1,j}|} + u_{1,j} \ln\left(p_i\right) + \frac{\frac{\partial C_{5,j}}{\partial a_{1,j}}}{C_{5,j}} - u_{1,j} \ln\left(d_{1,j}\right) \right) = 0 \tag{31}$$

$$\frac{\partial l}{\partial c_{1,j}} = \sum_{i=1}^{I} \left(\frac{\frac{\partial C_{5,j}}{\partial c_{1,j}}}{C_{5,j}} \right) = 0 \tag{32}$$

$$\frac{\partial l}{\partial d_{1,j}} = \sum_{i=1}^{I} \left(\frac{\frac{\partial C_{5,j}}{\partial d_{1,j}}}{C_{5,j}} - \frac{a_{1,j} u_{1,j}}{d_{1,j}} \right) = 0 \tag{33}$$

$$\frac{\partial l}{\partial u_{1,j}} = \sum_{i=1}^{I} \left(a_{1,j} \ln\left(p_i\right) + \frac{\frac{\partial C_{5,j}}{\partial u_{1,j}}}{C_{5,j}} - a_{1,j} \ln\left(d_{1,j}\right) - \frac{\frac{\partial \mathrm{B}\left(u_{1,j}, v_{1,j}\right)}{\partial u_{1,j}}}{\mathrm{B}\left(u_{1,j}, v_{1,j}\right)} \right) = 0 \tag{34}$$

$$\frac{\partial l}{\partial v_{1,j}} = \sum_{i=1}^{I} \left(\frac{\frac{\partial C_{5,j}}{\partial v_{1,j}}}{C_{5,j}} - \frac{\frac{\partial \mathrm{B}\left(u_{1,j}, v_{1,j}\right)}{\partial v_{1,j}}}{\mathrm{B}\left(u_{1,j}, v_{1,j}\right)} \right) = 0 \tag{35}$$

where

$$\begin{aligned} \frac{\partial \mathrm{B}\left(u_{2,j}, v_{2,j}\right)}{\partial u_{2,j}} &= \Gamma\left(v_{2,j}\right) \frac{\Gamma\left(u_{2,j}\right) \left(-\gamma + \sum_{k=1}^{u_{2,j}-1} \frac{1}{k} \right) - \Gamma\ \left(u_{2,j}\right) \left(-\gamma + \sum_{k=1}^{u_{2,j}+v_{2,j}-1} \frac{1}{k} \right)}{\Gamma\left(u_{2,j} + v_{2,j}\right)} \\ &= \mathrm{B}\left(u_{2,j}, v_{2,j}\right) \left(\sum_{k=1}^{u_{2,j}-1} \frac{1}{k} - \sum_{k=1}^{u_{2,j}+v_{2,j}-1} \frac{1}{k} \right) \end{aligned} \tag{36}$$

$$\frac{\partial \mathrm{B}\left(u_{2,j},v_{2,j}\right)}{\partial v_{2,j}} = \Gamma\left(u_{2,j}\right)\frac{\Gamma\left(v_{2,j}\right)\left(-\gamma+\sum_{k=1}^{v_{2,j}-1}\frac{1}{k}\right)-\Gamma\ \left(v_{2,j}\right)\left(-\gamma+\sum_{k=1}^{u_{2,j}+v_{2,j}-1}\frac{1}{k}\right)}{\Gamma\left(u_{2,j}+v_{2,j}\right)}$$
$$= \mathrm{B}\left(u_{2,j},v_{2,j}\right)\left(\sum_{k=1}^{v_{2,j}-1}\frac{1}{k}-\sum_{k=1}^{u_{2,j}+v_{2,j}-1}\frac{1}{k}\right) \tag{37}$$

and suppose $C_{5,j}$ is written as $\sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\sum_{r=0}^{\infty}C_{5lmnr}$ then

$$\frac{\partial C_{5,j}}{\partial a_{1,j}} = \sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\sum_{r=0}^{\infty}\left[C_{5lmnr}\left(\left(l+n\right)\ln\left(\frac{p_i}{d_{1,j}}\right)+\left(u_{1,j}+l+n\right)\times\right.\right.$$
$$\left.\left.\left(\sum_{k=1}^{a_{1,j}\left(u_{1,j}+l+n\right)-1}\frac{1}{k}-\sum_{k=1}^{a_{1,j}\left(u_{1,j}+l+n\right)-a_{2,j}\left(v_{2,j}+m+r\right)-2}\frac{1}{k}\right)\right)\right] \tag{38}$$

$$\frac{\partial C_{5,j}}{\partial c_{1,j}} = \sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\sum_{r=0}^{\infty}\left[C_{5lmnr}\left(\frac{\left(1-c_{1,j}\right)n-lc_{1,j}}{c_{1,j}\left(1-c_{1,j}\right)}\right)\right] \tag{39}$$

$$\frac{\partial C_{5,j}}{\partial d_{1,j}} = \sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\sum_{r=0}^{\infty}\left[C_{5lmnr}\left(\frac{-a_{1,j}\left(l+n\right)}{d_{1,j}}\right)\right] \tag{40}$$

$$\frac{\partial C_{5,j}}{\partial u_{1,j}} = \sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\sum_{r=0}^{\infty}\left[C_{5lmnr}\left(\left(\sum_{k=1}^{u_{1,j}+v_{1,j}+n-1}\frac{1}{k}-\sum_{k=1}^{u_{1,j}+v_{1,j}-1}\frac{1}{k}\right)+\right.\right.$$
$$\left.\left.a_{1,j}\left(\sum_{k=1}^{a_{1,j}\left(u_{1,j}+l+n\right)-1}\frac{1}{k}-\sum_{k=1}^{a_{1,j}\left(u_{1,j}+l+n\right)+a_{2,j}\left(v_{2,j}+m+r\right)-2}\frac{1}{k}\right)\right)\right] \tag{41}$$

$$\frac{\partial C_{5,j}}{\partial v_{1,j}} = \sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\sum_{r=0}^{\infty}\left[C_{5lmnr}\left(\left(\sum_{k=1}^{v_{1,j}-1}\frac{1}{k} - \sum_{k=1}^{v_{1,j}-l-1}\frac{1}{k}\right) + \left(\sum_{k=1}^{u_{1,j}+v_{1,j}+n-1}\frac{1}{k} - \sum_{k=1}^{u_{1,j}+v_{1,j}-1}\frac{1}{k}\right)\right)\right] \tag{42}$$

$$\begin{aligned}\frac{\partial \mathrm{B}\left(u_{1,j}, v_{1,j}\right)}{\partial u_{1,j}} &= \Gamma\left(v_{1,j}\right)\frac{\Gamma\left(u_{1,j}\right)\left(-\gamma + \sum_{k=1}^{u_{1,j}-1}\frac{1}{k}\right) - \Gamma\ \left(u_{1,j}\right)\left(-\gamma + \sum_{k=1}^{u_{1,j}+v_{1,j}-1}\frac{1}{k}\right)}{\Gamma\left(u_{1,j}+v_{1,j}\right)} \\ &= \mathrm{B}\left(u_{1,j}, v_{1,j}\right)\left(\sum_{k=1}^{u_{1,j}-1}\frac{1}{k} - \sum_{k=1}^{u_{1,j}+v_{1,j}-1}\frac{1}{k}\right)\end{aligned} \tag{43}$$

$$\begin{aligned}\frac{\partial \mathrm{B}\left(u_{1,j}, v_{1,j}\right)}{\partial v_{1,j}} &= \Gamma\left(u_{1,j}\right)\frac{\Gamma\left(v_{1,j}\right)\left(-\gamma + \sum_{k=1}^{v_{1,j}-1}\frac{1}{k}\right) - \Gamma\ \left(v_{1,j}\right)\left(-\gamma + \sum_{k=1}^{u_{1,j}+v_{1,j}-1}\frac{1}{k}\right)}{\Gamma\left(u_{1,j}+v_{1,j}\right)} \\ &= \mathrm{B}\left(u_{1,j}, v_{1,j}\right)\left(\sum_{k=1}^{v_{1,j}-1}\frac{1}{k} - \sum_{k=1}^{u_{1,j}+v_{1,j}-1}\frac{1}{k}\right)\end{aligned} \tag{44}$$

and γ is the Euler-Mascheroni constant.

3.2 Generalized beta-normal convolution

Although Figure 1.1 covers normal distribution, we can not derive the formula of the true intensity value when the noise is normal, from Equation (1). The normal distribution in Figure 1.1 is the normal distribution with one parameter. Therefore, in this section, we derive the formula to compute the corrected background intensity when the noise is symmetrically distributed, a normal distribution.

3.2.1 The joint density function

Under the convolution model in Equation (1), where P_i is the observed intensity of the regular i^{th} gene, S_i is the true intensity of the i^{th} gene, with

$$\begin{aligned} S_i \sim & f_1\left(s_i; a_j, c_j, d_j, u_j, v_j\right) \\ = & \frac{\left|a_j\right| s_i^{a_j u_j - 1}\left(1-\left(1-c_j\right)\left(\frac{s_i}{d_j}\right)^{a_j}\right)^{v_j-1}}{d^{a_j u_j}\mathrm{B}\left(u_j, v_j\right)\left(1+c_j\left(\frac{s_i}{d_j}\right)^{a_j}\right)^{u_j+v_j}}, \\ & 0 \leq c_j \leq 1; a_j, d_j, u_j, v_j, s_i > 0 \end{aligned} \tag{45}$$

and B is the background noise with

$$B_i \sim f_2\left(b_i; \mu_j, \sigma_j^2\right) = \frac{e^{-\frac{1}{2\sigma_j^2}\left(b_i-\mu_j\right)^2}}{\sqrt{2\pi}\sigma_j}, \mu_j \in \mathbb{R}, \sigma_j^2 > 0, b_i > 0 \tag{46}$$

The joint density function of S_i and B_i is

$$f_{S_i,B_i}\left(s_i, b_i\right) = \frac{\left|a_j\right| s_i^{a_j u_j - 1}\left(1-\left(1-c_j\right)\left(\frac{s_i}{d_j}\right)^{a_j}\right)^{v_j-1}}{d_j^{a_j u_j}\mathrm{B}\left(u_j, v_j\right)\left(1+c_j\left(\frac{s_i}{d_j}\right)^{a_j}\right)^{u_j+v_j}} \frac{e^{-\frac{1}{2\sigma_j^2}\left(b_i-\mu_j\right)^2}}{\sqrt{2\pi}\sigma_j} \tag{47}$$

The joint density function of S_i and P_i is

$$f_{S_i,P_i}\left(s_i, p_i\right) = \frac{\left|a_j\right| s_i^{a_j u_j - 1}\left(1-\left(1-c_j\right)\left(\frac{s_i}{d_j}\right)^{a_j}\right)^{v_j-1}}{d_j^{a_j u_j}\mathrm{B}\left(u_j, v_j\right)\left(1+c_j\left(\frac{s_i}{d_j}\right)^{a_j}\right)^{u_j+v_j}} \frac{e^{-\frac{\left(p_i-s_i-\mu_j\right)^2}{2\sigma_j^2}}}{\sqrt{2\pi}\sigma_j} \tag{48}$$

3.2.2 The marginal density function

The marginal density function of P_i is

$$f_{P_i}(p_i) = \frac{|a_j|}{d_j^{a_j u_j} \mathrm{B}(u,v)\sqrt{2\pi}\sigma_j} \sum_{l=0}^{\infty}\sum_{m=0}^{\infty} \left\{ \frac{(-1)^{l+m}\left(1-c_j\right)^l c_j^m \binom{v_j-1}{l}}{d^{a_j(l+m)}} \times \binom{u_j+v_j+m-1}{m} \int_0^{p_i} s_i^{a_j(u_j+l+m)-1} e^{-\frac{(s_i-p_i-\mu_j)^2}{2\sigma_j^2}} ds_i \right\} \tag{49}$$

Let $\frac{\left(s_i-\left(p_i-\mu_j\right)\right)}{\sigma_j} = z_i$, and the equation (48) becomes

$$= \frac{|a_j|}{d_j^{a_j u_j} \mathrm{B}\left(u_j,v_j\right)\sqrt{2\pi}} \sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty} \left\{ \frac{(-1)^{l+m}\left(1-c_j\right)^l c_j^m \binom{v_j-1}{l}}{d_j^{a_j(l+m)}\left(p_i-\mu_j\right)^n} \times \binom{u_j+v_j+m-1}{m} \binom{a_j\left(u_j+l+m\right)-1}{n} \left(p_i-\mu_j\right)^{a_j\left(u_j+l+m\right)-1} \sigma_j^n \times \int_{-\frac{\left(p_i-\mu_j\right)}{\sigma_j}}^{\frac{\mu_j}{\sigma_j}} z_i^n e^{-\frac{z_i^2}{2}} dz_i \right\} \tag{50}$$

Let $\frac{z_i^2}{2} = x_i$, the equation (50) becomes

$$K_2 C_{7,j} \tag{51}$$

where

$$K_2 = \frac{\left|a_j\right| p_i^{a_j u_j - 1}}{2\sqrt{\pi} d_j^{a_j u_j} B\left(u_j, v_j\right)}, \text{ and}$$

$$C_{7,j} = \sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\left\{\frac{(-1)^{l+m}\left(1-c_j\right)^l c_j^m \binom{v_j-1}{l}\binom{u_j+v_j+m-1}{m}}{d_j^{a_j(l+m)}\left(p_i-\mu_j\right)^n}\times\right.$$

$$\binom{a_j\left(u_j+l+m\right)-1}{n}\left(p_i-\mu_j\right)^{a_j(l+m)}\sigma_j^n 2^{\frac{n}{2}}\left(\gamma\left(\frac{n+1}{2},\left(\frac{\mu_j}{\sigma_j}\right)^2\right)-\right.$$

$$\left.\left.\gamma\left(\frac{n+1}{2},\left(\frac{\left(p_i-\mu_j\right)}{\sigma_j}\right)^2\right)\right)\right\}, \text{ and}$$

$\gamma\left(\blacksquare,\blacksquare\right)$ is the lower incomplete gamma function

3.2.3 The conditional density function

The conditional density function of S_i where it is known that $P_i = p_i$ is

$$f_{S_i|P_i}\left(s_i \mid p_i\right)$$

$$= \frac{\sqrt{2}p_i^{1-a_j u_j}}{C_{7,j}\sigma_j}\frac{s_i^{a_j u_j - 1}\left(1-\left(1-c_j\right)\left(\frac{s_i}{d_j}\right)^{a_j}\right)^{v_j-1} e^{-\frac{\left(p_i-s_i-\mu_j\right)^2}{2\sigma_j^2}}}{\left(1+c_j\left(\frac{s_i}{d_j}\right)^{a_j}\right)^{u_j+v_j}} \tag{52}$$

3.2.4 The corrected background intensity

The corrected background intensity under this generalized beta convolution is

$$p_i\frac{C_{8,j}}{C_{7,j}} \tag{53}$$

where

$$C_{8,j} = \sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\left\{\frac{(-1)^{l+m}\left(1-c_j\right)^{l} c_j^{m}\binom{v_j-1}{l}\binom{u_j+v_j+m-1}{m}}{d_j^{a_j(l+m)}\left(p_i-\mu_j\right)^{n}}\times\right.$$

$$\binom{a_j\left(u_j+l+m\right)}{n}\left(p_i-\mu_j\right)^{a_j(l+m)}\sigma_j^{n}2^{\frac{n}{2}}\left(\gamma\left(\frac{n+1}{2},\left(\frac{\mu_j}{\sigma_j}\right)^{2}\right)-\right.$$

$$\left.\left.\gamma\left(\frac{n+1}{2},\left(\frac{\left(p_i-\mu_j\right)}{\sigma_j}\right)^{2}\right)\right)\right\},\ \text{and}$$

$\gamma\left(\boldsymbol{.},\boldsymbol{.}\right)$ is the lower incomplete gamma function

3.2.5 The likelihood function

The likelihood function ($\mathbf{L}$) to estimate $a_j, c_j, d_j, u_j, v_j, \mu_j$ and σ_j^2 is

$$= \prod_{i=1}^{I}\frac{\left|a_j\right| p_i^{a_j u_j-1} C_{7,j}}{2\sqrt{\pi} d_j^{a_j u_j}\mathrm{B}\left(u_j,v_j\right)}\prod_{w=1}^{W}\frac{e^{-\frac{1}{2\sigma_j^2}\left(b_{0w}-\mu_j\right)^2}}{\sqrt{2\pi}\sigma_j} \tag{54}$$

The log-likelihood function l is

$$= \sum_{i=1}^{I}\left\{\ln\left(\left|a_j\right|\right)+\left(a_j u_j-1\right)\ln\left(p_i\right)+\ln(C_{7,j})-\ln(2)-\frac{1}{2}\ln(\pi)-a_j u_j\ln\left(d_j\right)-\right.$$

$$\left.\ln\left(\mathrm{B}\left(u_j,v_j\right)\right)\right\}+\sum_{w=1}^{W}\left\{-\frac{\left(b_{0w}-\mu_j\right)^2}{2\sigma_j^2}-\frac{1}{2}\left(\ln(2)+\ln(\pi)\right)-\ln(\sigma_j)\right\} \tag{55}$$

The likelihood equations are as follows

$$\frac{\partial l}{\partial \mu_j} = \sum_{w=1}^{W}\left(\frac{\left(b_{0w}-\mu_j\right)}{\sigma_j^2}\right)=0 \tag{56}$$

$$\frac{\partial l}{\partial \sigma_j} = \sum_{w=1}^{W} \left(\frac{\left(b_{0w} - \mu_j\right)^2}{\sigma_j^3} - \frac{1}{\sigma_j} \right) = 0 \tag{57}$$

$$\frac{\partial l}{\partial a_j} = \sum_{i=1}^{I} \left(\frac{1}{|a_j|} + u_j \ln\left(p_i\right) + \frac{\frac{\partial C_{7,j}}{\partial a_j}}{C_{7,j}} - \mu_j \ln\left(d_j\right) \right) = 0 \tag{58}$$

$$\frac{\partial l}{\partial c_j} = \sum_{i=1}^{I} \left(\frac{\frac{\partial C_{7,j}}{\partial c_j}}{C_{7,j}} \right) = 0 \tag{59}$$

$$\frac{\partial l}{\partial d_j} = \sum_{i=1}^{I} \left(\frac{\frac{\partial C_{7,j}}{\partial d_j}}{C_{7,j}} - \frac{a_j \mu_j}{d_j} \right) = 0 \tag{60}$$

$$\frac{\partial l}{\partial u_j} = \sum_{i=1}^{I} \left(a_j \ln\left(p_i\right) + \frac{\frac{\partial C_{7,j}}{\partial u_j}}{C_{7,j}} - \frac{\frac{\partial \mathrm{B}\left(u_j, v_j\right)}{\partial u_j}}{\mathrm{B}\left(u_j, v_j\right)} \right) = 0 \tag{61}$$

$$\frac{\partial l}{\partial v_j} = \sum_{i=1}^{I} \left(\frac{\frac{\partial C_{7,j}}{\partial v_j}}{C_{7,j}} - \frac{\frac{\partial \mathrm{B}\left(u_j, v_j\right)}{\partial v_j}}{\mathrm{B}\left(u_j, v_j\right)} \right) = 0 \tag{62}$$

where

$$\begin{aligned} \frac{\partial \mathrm{B}\left(u_j, v_j\right)}{\partial u_j} =& \Gamma\left(v_j\right) \frac{\Gamma\left(u_j\right) \left(-\gamma + \sum_{k=1}^{u_j - 1} \frac{1}{k} \right) - \Gamma\ \left(u_j\right) \left(-\gamma + \sum_{k=1}^{u_j + v_j - 1} \frac{1}{k} \right)}{\Gamma\left(u_j + v_j\right)} \\ =& \mathrm{B}\left(u_j, v_j\right) \left(\sum_{k=1}^{u_j - 1} \frac{1}{k} - \sum_{k=1}^{u_j + v_j - 1} \frac{1}{k} \right) \end{aligned} \tag{63}$$

$$\frac{\partial \mathrm{B}\left(u_j, v_j\right)}{\partial v_j} = \Gamma\left(u_j\right) \frac{\Gamma\left(v_j\right)\left(-\gamma + \sum_{k=1}^{v_j-1} \frac{1}{k}\right) - \Gamma\ \left(v_j\right)\left(-\gamma + \sum_{k=1}^{u_j+v_j-1} \frac{1}{k}\right)}{\Gamma\left(u_j + v_j\right)}$$
$$= \mathrm{B}\left(u_j, v_j\right)\left(\sum_{k=1}^{v_j-1} \frac{1}{k} - \sum_{k=1}^{u_j+v_j-1} \frac{1}{k}\right) \tag{64}$$

and suppose $C_{7,j}$ is written as $\sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty} C_{7lmn}$ then

$$\frac{\partial C_{7,j}}{\partial a_j} = \sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\left[C_{7lmn}\left(\left(l+m\right)\ln\left(\frac{p_i-\mu_j}{d_j}\right) + \left(u_j+l+m\right)\times \left(\sum_{k=1}^{a_j\left(u_j+l+m\right)-1} \frac{1}{k} - \sum_{k=1}^{a_j\left(u_j+l+m\right)-n-1} \frac{1}{k}\right)\right)\right] \tag{65}$$

$$\frac{\partial C_{7,j}}{\partial c_j} = \sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\left[C_{7lmn}\left(\frac{\left(1-c_j\right)m - lc_j}{c_j\left(1-c_j\right)}\right)\right] \tag{66}$$

$$\frac{\partial C_{7,j}}{\partial d_j} = \sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\left[C_{7lmn}\left(-\frac{a_j\left(l+m\right)}{d_j}\right)\right] \tag{67}$$

$$\frac{\partial C_{7,j}}{\partial u_j} = \sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\left[C_{7lmn}\left(\left(\sum_{k=1}^{u_j+v_j+m-1} \frac{1}{k} - \sum_{k=1}^{u_j+v_j-1} \frac{1}{k}\right) + a_j\left(\sum_{k=1}^{a_j\left(u_j+l+m\right)-1} \frac{1}{k} - \sum_{k=1}^{a_j\left(u_j+l+m\right)-n-1} \frac{1}{k}\right)\right)\right] \tag{68}$$

$$\frac{\partial C_{7,j}}{\partial v_j} = \sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\left[C_{7lmn}\left(\left(\sum_{k=1}^{v_j-1}\frac{1}{k} - \sum_{k=1}^{v_j-l-1}\frac{1}{k}\right) + a_j\left(\sum_{k=1}^{u_j+v_j+m-1}\frac{1}{k} - \sum_{k=1}^{u_j+l+m-n-1}\frac{1}{k}\right)\right)\right] \tag{69}$$

4 Discussion and remarks

We have studied the additive models of background correction for BeadArrays and proposed the generalized model where the true intensity and the noise are assumed to be skewed distribution and where the true intensity is a skewed but the noise is symmetrics distribution. In this paper, we have shown the corrected background intensity value of the proposed models.

This proposed model is a generalization of the available convolution models as in papers [4], [5–7], [13], [15], [5–7], [14] and [12]. The generalization comes from the property of the tree-generalized beta distributions [17] and is explained in [19] and [17]. The parameters of the generalized beta distribution are a, d, c, u and v. The gamma, exponential and lognormal distributions are special cases of the generalized beta distribution.

The gamma distribution is the generalized beta distribution when $c = 1, v \to \infty, d = \beta v^{\frac{1}{a}}$ and $a = 1$; the exponential distribution is the generalized beta distribution when $c = 1, v \to \infty, d = \beta v^{\frac{1}{a}}$ and $a = 1, p = 1$; and the lognormal distribution is the generalized beta distribution when $c = 1, v \to \infty, d = \beta v^{\frac{1}{a}}$ and $\beta = (\sigma^2 a^2)^{\frac{1}{2}}, u = \frac{(a\mu+1)}{\sigma^2 a^2}$ and $a \to 0$.

There are some aspects to be considered while implementing these models:

1. parameters estimation

   In parameters estimation, there are some methods have been suggested by some researchers. Mc Donald and Xu [17] used and suggested: the method of maximum likelihood (also was used by Fajriyah [20–22]), the method of moments and the maximum product spacing estimation.

   When $c = 1$, the generalized beta distribution is a generalized beta of the second kind. Graf and Nedyalkova [23] and Graf et al. [24] have observed that the pseudo maximum likelihood (Huber [25], Freedman [26] and Pfeffermann et al. [27]), the nonlinear least squares on the quantile function (Dagum [28]) and the nonlinear fit for indicator can be implemented to estimate the parameters of the generalized beta of the second kind. The available VGAM *package* in R helps to estimate the parameters of this distribution.

   The existing convolution models use various methods:

   (a) the *ad-hoc* method which is implemented by the RMA method, more details can be found in [5–7], [29] and [12]

   (b) Markov chain Monte Carlo simulations, more details can be found in [11]

   (c) Maximum likelihood, nonparametrics and method of moments, more details can be found in [13], [15] and [12]

   (d) Plug-in method, more details can be found in [15]

   (e) Fast Fourier transform, more details can be found in [14]

   In general, we first need to provide the initial parameters to optimize the log-likelihood function in Equations (25) and (36). The initial parameters of the noise are easily provided since the benchmarking data set of the negative control probes is available publicly. The initial parameters of the true intensity can be estimated from the observed intensity data substracted by the mean (or median) of the negative control intensity.

   Secondly, once the initial parameters are available, then they will be used to optimize the likelihood function by implementing the optimization method. There

are some packages in R which can be used to compute the parameters of the model, for example the *optim* or *optimx* package. These parameters are then used to compute the corrected background intensity based on the formula of the choosen model. Remember that the background correction is implemented for each array.

2. the corrected background intensity computation

   The corrected background intensity computation includes computations of the infinite summations: $C_{5,j}, C_{6,j}, C_{7,j}$ and $C_{8,j}$. In the author's experience (in [15]) these infinite summations are close to being constant after certain terms. As a consequence, the ratios of $\frac{C_{6,j}}{C_{5,j}}$ and $\frac{C_{8,j}}{C_{7,j}}$ are able to be computed. Therefore the difficulty in computing the summations used to compute the corrected background intensity can be eliminated. A sophisticated program written in R, C, *Python* and its paralellisation, could help to speed up the computation.

3. the benchmarking data set

   During the implementation of this generalized estimator, the Illumina users need to be aware of the availability of the Illumina Spike-in data set. Once the model is fitted into this data set, the model can then be used to adjust the intensity value.

   Apart from the benchmarking criteria for the Affymetrix GeneChips, in the author's knowledge, the benchmarking criteria for the Illumina BeadArrays have not been formalized yet. Some researchers, i.e. [13], [14], [30] and [12] have developed the criteria to assess which background correction methods perform better than the others for the Illumina BeadArrays. These criteria together with the criteria in the Affycomp *package* ([31] and [32]) can be used as the benchmarking criteria for the Illumina BeadArrays. These have been implemented by Fajriyah [15]. The method which has been used by Shi et al. [33] also can be used to assess the best performance of the background correction methods.

4. the negative control data set

It is possible that the negative control probes set data is unavailable. In this case, we can adapt the proposed model to the convolution model for background correction without the negative control probes intensities, as in the RMA model.

The application of this generalized model towards other platforms, such as the Affymetrix, is possible by considering the points above.

Acknowledgements:
This paper is part of the author's PhD dissertation written under the direction of Professor István Berkes. We would like to thank *Paulo Canas Rodrigues, PhD* for his comments. Financial support from the Austrian Science Fund (FWF), Project P24302-N18 is gratefully acknowledged. We would also like to thank the anonymous reviewers for their valuable remarks in leading to an improvement of this paper.

Conflicts of interest: None

References

[1] W. Huber, A. von Heydebreck, and M. Vingron Error models for microarray intensities Technical Report Paper 6, Bioconductor Project Working Papers, 2004.

[2] W. Huber, A. von Heydebreck, and M. Vingron An introduction to low-level analysis methods of DNA microarray data Technical Report Paper 9, Bioconductor Project Working Papers, 2005a.

[3] W. Huber, R. A. Irizarry, and R. Gentleman, Bioinformatics and Computational Biology Solutions Using R and Bioconductor; chapter Prepocessing Overview, Springer, 2005b.

[4] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed, *A Comparison of*

Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance, Bioinformatics, 2003, **19**(2), 185–193.

[5] R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed, *Summaries of Afymetrix GeneChip probe level data, Nucleic Acids Research*, 2003a, **31**(4).

[6] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed, *Exploration, Normalization and Summaries of High Density Oligonucleotide Array Probe Level Data, Biostatistics*, 2003b, **4**(2), 249–264.

[7] R. A. Irizarry, Z. Wu, and H. A. Jaffee, *Comparison of Affymetrix geneChip expression measures, Bioinformatics*, 2006, **22**(7), 789–794.

[8] C. Li and W. H. Wong, *Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection, Proceeding National Academy of Sciences*, 2001, **98**(1), 31–36.

[9] J. D. Silver, M. E. Ritchie, and G. K. Smyth, *Microarray background correction: maximum likelihood estimation for the normal-exponential convolution model, Biostatistics*, 2009, **10**, 352–363.

[10] Z. Wu, R. A. Irizarry, R. Gentleman, F. Martinez-Murillo, and F. Spencer, *A model-based background adjustment for oligonucleotide expression arrays, Journal of the American Statistical Association*, 2004, **99**(468), 909 – 917.

[11] L.-H. Ding, Y. Xie, S. Park, G. Xiao, and M. D. Story, *Enhanced identification and biological validation of differential gene expression via Illumina whole-genome expression arrays through the use of the model-based background correction methodology, Nucleic Acids Research*, 2008, **36**(10: e58).

[12] Y. Xie, X. Wang, and M. D. Story, *Statistical methods of background correction for Illumina BeadArray data, Bioinformatics*, 2009, **25**(6), 751–757.

[13] M. Chen, Y. Xie, and M. D. Story, *An Exponential-Gamma Convolution Model for Background Correction of Illumina BeadArray Data, Communication in Statistics: Theory and Methods*, 2011, **40**(17), 3055–3069.

[14] S. Plancade, Y. Rozenholc, and E. Lund, *Generalization of the normal-exponential model: exploration of a more accurate parameterisation for the signal distribution on Illumina BeadArrays, BMC Bioinformatics*, 2012, **13**(329).

[15] R. Fajriyah, *A Study of convolution models for background correction of BeadArrays, accepted paper at Austrian Journal of Statistics*, 2014.

[16] A. Posekany, K. Felsenstein, and P. Sykacek, *Biological assessment of robust noise models in microarray data analysis, Bioinformatics*, 2011, **27**(6), 807–814.

[17] J. B. McDonald and Y. J. Xu, *A generalization of the beta distribution with applications, Journal of Econometrics*, 1995, **66**, 133–152.

[18] L. M. Leemis and J. T. McQueston, *Univariate Distribution Relationships, The American Statistician*, 2008, **62**(1), 45 – 53.

[19] J. B. McDonald, *Some generalized functions for the distribution of income, Econometrica*, 1984, **52**(3).

[20] R. Fajriyah. Statistical analysis of the economic performance in Indonesia, Part I - Simplex method, 55th ISI Session Conference, 2005a.

[21] R. Fajriyah. Statistical analysis of the economic performance in Indonesia, Part II - Grad method, ICREM 2 Conference, INSPEM, University Putra Malaysia, 2005b.

[22] R. Fajriyah, *The pdf's estimation by grad method and its Gini index, Karya Asli Lorekan Matematik*, 2008, **1**(2), 021 – 027.

[23] M. Graf and D. Nedyalkova, Fitting the Generalized Beta Distribution of the Second Kind to the Empirical Income Distribution from the Aggregate Laeken Indicators, 2010.

[24] M. Graf, D. Nedyalkova, R. Münnich, J. Seger, and S. Zins Parametric Estimation of Income Distributions and Indicators of Poverty and Social Exclusion Technical Report 2.1, AMELI, 2011.

[25] P. J. Huber In *The behavior of maximum likelihood estimates under nonstandard conditions, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1: Statistics, pp. 221 – 233, Berkeley ,California, 1967. Univ. of Calif. Press.

[26] D. A. Freedman, *On the so-called "Huber sandwich estimator" and "robust standard errors", The American Statistician*, 2006, **60**, 299 – 302.

[27] D. Pfeffermann, C. J. Skinner, D. J. Holmes, H. Goldstein, and J. Rasbash, *Weighting for unequal selection probabilities in multilevel models, Journal of the Royal Statistical Society B*, 1998, **60**(Part 1), 23 – 40.

[28] C. Dagum, *A New Model of Personal Income Distribution: Specification and Estimation, Economie Appliquée*, 1977, **30**(413 - 437).

[29] M. McGee and Z. Chen, *Parameter estimation for the convolution model for background correction of affymetrix genechip data, Statistical Applications in Genetics and Molecular Biology*, 2006, **5**(24).

[30] A. Shamilov, Y. M. Kantar, and I. Usta In *On a Functional defined by means of Kullback-Leibler Measure and Its Statistical Applications, Proceedings of the 9th WSEAS International Conference on Applied Mathematics*, pp. 632–637, 2006.

[31] L. M. Cope, R. A. Irizarry, H. A. Jaffee, Z. Wu, and T. P. Speed, *A benchmark for Affymetrix GeneChip expression measures, Bioinformatics*, 2004, **20**, 323–331.

[32] affycomp: Graphics Toolbox for Assessment of Affymetrix Expression Measures. R. A. Irizarry and Z. Wu R package version 1.38.0 (with contributions from Simon Cawley) ed., 2013.

[33] W. Shi, A. Oshlack, and G. K. Smyth, *Optimizing the noise versus bias trade-off for Illumina whole genome expression Beadchips, Nucleic Acids Research*, 2010, **38**(22: e204).