Comparison of the risk ratio of background correction models for Illumina BeadArrays

Rohmatul Fajriyah
fajriyah@student.tugraz.at

Introduction

Background correction (BC) plays an important role in microarray data processing, since some steps in producing microarray data contribute the noise. Irizarry et al. [3] introduced the convolution model for the Affymetrix platform single chanel and since then it also has been implemented to other platforms, such as two-colours/chanels and beadArrays, where for each platform the adaptive models has been developed.
We have developed the exponential and gamma -lognormal convolution models (Fajriyah, 2013) and compared their performances to the existing models. Here, we study the mean absolute deviation of the BC of all existing convolution models and compare their risk ratio.

Conclusion

Based on the simulation study, we conclude that:

1. The Exponential-Normal models, except ENrma, show that their risk ratio always lower than 1.
2. The Gamma-Normal model shows unpredictable behaviour in regard of measuring the risk ratio. There is a situation where this model could not be implemented, because its produce *NaN* value. This behaviour is similar to the previous result of Fajriyah [2].
3. The **proposed models** provide the best performance by showing a consistency moderate risk ratio, particulalry the ELNnp and ELNp models where their risk ratio is lower than 1.
4. In general the choice of background correction method is at user's hand, by compromising between, low risk ratio, low bias in parametrization, and low bias in background correction.

References

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

[2] Rohmatul Fajriyah, A study of convolution models for background correction of BeadArrays, submitted to *Austrian Journal of Statistics*, presented at Doctoral School of Mathematics Seminar, May 6, 2013.

[3] Rafael A Irizarry, Benjamin M Bolstad, Francois Collin, Leslie M Cope, Bridget Hobbs, and Terence P Speed, 2003. Summaries of Affymetrix GeneChip probe level data, *Nucleic Acids Research*, 31(4), e15.

[4] Sandra Plancade, Yves Rozenholc, and Eiliv Lund, 2012. Generalization of the normal-exponential model: exploration of a more accurate parameterisation for the signal distribution on Illumina BeadArrays, *arXiv.org*, arXiv:1112.4180v2.

[5] Yang Xie, Xinlei Wang and Michael Story, 2009. Statistical methods of background correction for Illumina BeadArray data, *Bioinformatics*, 25(6):751-757.

Acknowledgements

I would like to thank to **Prof. Istvan Berkes** and Austrian Science Fund (FWF), Project P24302, Institute of Statistics, TU Graz, Austria.

Measuring the Risk Ratio

The excess risk ratio, of using particular model i where the true model j is known, is defined as follows:

$$R(i) = \frac{MAD(\hat{S}_i)}{MAD(\hat{S}_j)} \tag{1}$$

where

$$MAD(\hat{S}) = \tfrac{1}{N}\sum_{l=1}^{N}\Big(\frac{1}{n_r}\sum_{k=1}^{n_r} \mid \hat{S}(X_k^l \mid \hat{\Theta}_l) - S_k^l \mid \Big)$$

X is the intensity value of regular probes, $\hat{\Theta}_l$ is the estimated parameters from the simulated array l, and S is the true intensity value.

We compare the excess risk ratio of the existing models through the simulation based on the Illumina benchmarking data set. The simulation is conducted as follows:

1. Select a particular model as the 'true' model.
2. Generate the samples: regular and negative control, based on the parameters of this 'true' model, for each model
3. Estimate the parameters and the true intensity value for each model
4. Compute the MAD value
5. Compute the risk ratio

The sample size of the regular probes is 25000(n_r), the negative control probes is 1000 and the simulation is done $N = 100$ times .

Convolution models for Background Correction

In general the convolution model is

$$P = S + B \tag{2}$$

where **P** is probe intensity, **S** is the true intensity and **B** is the noise.

The models, that we studied, are

ENrma Exponential-Normal convolution model, where the parameter estimation method is *ad-hoc* (**the pioneer**, Irizarry et al. [3]).

ENmbcb Exponential-Normal convolution model, where parameter estimation methods are non-parametric, MLE and bayesian (Xie et al. [5]).

EG Exponential-Gamma convolution model, where parameter estimation method is MLE (Chen et al. [1]).

GN Gamma-Normal convolution model, where parameter estimation method is MLE (Plancade et al. [4]).

E/G LN Exponential/Gamma-Lognormal convolution models, where parameter estimation methods are nonparametric, MLE and plug-in (**proposed methods**, Fajriyah, R. [2]).

Results

The simulation study results are presented at Table 1 below.

True models| Method	ENrma	ENnp	ENmle	ENmcmc	GN	ELNnp	ELNmle	ELNp	GLNnp	GLNmle
ENrma	1.000	0.638	0.638	0.638	NaN	0.824	46.497	0.961	2.796	1.434
ENnp	2.871	1.000	1.000	1.000	1.000	1.003	42.762	1.040	1.091	1.098
ENmle	3.199	1.0005	1.000	1.000	1.000	1.000	47.832	1.015	1.020	1.060
ENbayes	3.204	1.000	1.000	1.000	1.000	1.001	47.347	1.015	1.019	1.059
GN	1.836	1.232	1.903	1.903	1.000	1.237	56.436	1.296	3.294	1.819
ELNnp	2.809	0.998	0.997	0.998	0.998	1.000	42.931	1.034	1.086	1.092
ELNmle	0.071	0.022	0.0228	0.022	0.022	0.022	1.000	0.022	0.022	0.023
ELNnp	4.484	1.000	1.000	1.000	1.000	1.000	38.481	1.000	1.000	1.002
GLNnp	0.614	0.410	0.590	0.590	0.341	0.413	16.082	0.438	1.000	1.357
GLNmle	3.525	0.922	0.943	0.943	0.923	0.922	41.740	0.927	0.938	1.000

Table 1: Comparison of the risk ratio for each model

From the Table 1 above, we can see that *GN* model produce NaN value when it is assumed that the true model is ENrma. This result is quite unexpected, since the parameters are based on the benchmarking data set. By excluding this value from the computation and we compute the average of the risk ratio for each model, in fact the GN model produces the lowest risk ratio. It is followed by the ENnp, ELNnp, ELNp, ENmle, ENbayes, GLNmle, GLNnp, and ENrma.