Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v9n1p58

Estimation of correlation between various types of pixel intensities in a single spot
By Kalesh M. Karun et al.

Published: 26 April 2016

Electronic Journal of Applied Statistical Analysis
Vol. 09, Issue 01, April 2016, 58-69
DOI: 10.1285/i20705948v9n1p58

Estimation of correlation between various types of pixel intensities in a single spot

Kalesh M. Karun[a], Binu V. S.*[b], N. Sreekumaran Nair[c], K. Manjunatha Prasad[d], Kala M. Karun[e], Keerthana Prasad[f], and K. M. Girisha[g]

[a] *PhD Scholar, Department of Statistics, Manipal University, Manipal, India, Pin: 576104*
[b] *Associate Professor, Department of Statistics, Manipal University, Manipal, India, Pin: 576104*
[c] *Professor and Head, Department of Statistics, Manipal University, Manipal, India, Pin: 576104*
[d] *Professor, Department of Statistics, Manipal University,Manipal, India, Pin: 576104*
[e] *Assistant Professor, Department of Information and Technology, Adi Shankara Institute of Engineering and Technology,Kerala University, India, Pin: 683574*
[f] *Associate Professor, School of Information Sciences, Manipal University, Manipal, India, Pin: 576104*
[g] *Professor and Head, Department of Medical Genetics, Kasturba Medical College, Manipal University, Manipal, India, Pin: 576104*

Published: 26 April 2016

In complementary DNA (cDNA) microarray experiments, the measurement of interest is signal intensity ratio of spots. Each spot have four types of pixel intensities namely red foreground, green foreground, red background and green background. The uncertainty associated with the intensity ratio of a spot depends on the correlations between intensities of these pixels. In this article, we propose a method to estimate correlations between various types of pixel intensities within a spot using a modified form of Moran's I spatial autocorrelation. We estimate these correlations for eight selected spots from image files downloaded from the Gene Expression Omnibus (GEO) database. These estimated correlations are used for finding the uncertainty associated with each of the selected eight spots using the theory of error propagation.

*Corresponding author: binu.vs@manipal.edu

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

We observed that the estimated uncertainty associated with intensity ratio of a spot is less if we consider the correlation between various pixel intensities compared to assuming zero correlation.

keywords: Microarray, cDNA microarray, Pixel intensity ratio, spatial autocorrelation, Moran's I, Uncertainty.

1 Introduction

Microarray technologies are the most successful methodologies for high-throughput and large-scale genomic analyses. The complementary DNA (cDNA) microarray technologies initially were designed to measure the expression levels of RNA transcripts derived from thousands of genes within a genome in a single experiment at a time. However, cDNA microarrays in biomedical research are not limited to detection of gene expression level. This technology helps to relate physiological cell states to gene expression patterns for studying tumors, diseases progression, cellular response to stimuli, and drug target identification. cDNA microarrays are also used to detect single nucleotide polymorphisms (SNPs), alterations in gene copy-number, aberrations in methylation patterns, alternative RNA splicing, and pathogen detection etc. Thus microarry technology plays significant role in genomic analyses (Trevino et al., 2007).

A cDNA microarray is a glass slide or nylon chip of size that varies from 0.5 x 0.5 cm to 2.5 x 7.5 cm on to which thousands of probes (portions of single stranded DNA) are fixed using robotic spotting. Each spot contains several copies of known DNA sequence which is reverse complement to a target RNA sequence (Samavi et al., 2006). In a typical cDNA microarray experiment, the first step is to extract the cellular messenger RNA from diseased and normal cells and then reverse transcribed into cDNA and labelled with two different fluorescent dyes, one with fluoresce green (Cy3) and the other with fluoresce red (Cy5). These samples are then purified, mixed together and simultaneously hybridized to microarray chip. Later, the microarrays are washed, dried and passed through two laser beams. A photomultiplier tube (PMT) is used to capture the fluorescent lights emitted from these two lasers and an analogue-to-digital converter (ADC) converts the intensity of the red and green light from each spot into digital signal. The idea behind microarrays is to compute unique signal for each gene that is directly proportional to the quantity of mRNA that was hybridized on the chip (Baldi and Hatfield, 2002). In cDNA microarray experiments, the measurement of interest is average of red to green pixel intensity ratio at each spot that gives the expression level of a particular gene in a diseased sample compared to normal. For a spot, let $\bar{R}_f$,$\bar{R}_b$, $\bar{G}_f$ and $\bar{G}_b$ denote the average foreground red pixel intensity, average background red pixel intensity, average foreground green pixel intensity and average background green pixel intensity respectively. Then the background corrected intensity ratio for a spot is given by (Causton et al., 2009).

$$Y = \frac{\bar{R}_f - \bar{R}_b}{\bar{G}_f - \bar{G}_b}$$

The estimated value of any measurement may deviate from the true value and this deviation is called error or uncertainty associated with that estimate. In cDNA microarray experiments, the measurement of interest is gene expression level which is measured as the background corrected intensity ratio and this estimate of signal intensity measurement may have errors or uncertainty. This uncertainty could be due to the contribution of errors associated with several components in the experiment like probe and chip used, hybridization, dye and scanner used, pin geometry in a slide etc. Usually the estimate of uncertainty for intensity ratio of each spot is reported by standard deviation by assuming the pixel intensities as independent (Binu et al., 2012).

Using the method of error propagation we can derive the uncertainty associated with background corrected intensity ratio as a function of uncertainty in individual pixel intensities. Due to the presence of chip artefacts, there will be correlation between various pixel intensities between red foreground (R_f) and red background (R_b) ; between green foreground (G_f) and green background (G_b) ; between red foreground (R_f) and green background (G_b) ; between green foreground (G_f) and red background (R_b) ; between red foreground (R_f) and green foreground (G_f) ; between red background (R_b) and green background (G_f) within a spot. Binu et al. (2012) demonstrated that the uncertainty in the estimate of intensity ratio depends on the correlation between various pixel intensities. The correlation between intensities of green and red foreground pixels in a spot can be calculated using Pearson correlation coefficient. Similarly, we can estimate the correlation between intensities of red and green background pixels using Pearson correlation coefficient. This is because we have paired values of intensities for red and green foreground pixels as each of the pixels in the foreground and background of spot have both red and green intensity values. However, it is not straightforward to estimate the remaining four correlations (between red foreground and red background; between green foreground and green background; between red foreground and green background; between green foreground and red background) by means of Pearson correlation coefficient as position wise pairing is not possible for these four types of pixels in a spot. Hence we have to look for alternative methods to estimate these correlations.

In this article, we propose a method to estimate these correlations using an expression that is similar to spatial autocorrelation measure proposed by Moran (1950). We estimate the correlation between various pixels intensities using a modified Moran's I autocorrelation. With these correlations we estimate the errors associated with background corrected intensity ratio of spots, using the theory of error propagation.

2 Materials and methods

2.1 Estimation of correlation between various pixel intensities in a spot.

The correlations between intensities of green and red foreground pixels as well as background pixels were estimated using Pearson correlation coefficient. The remaining four correlations (between Red foreground and Red background; between Green foreground and Green background; between Red foreground and Green background; between Green

foreground and Red background) were estimated using a modified form of Moran's I spatial autocorrelation measure. Moran's autocorrelation coefficient (often denoted as I) is an extension of Pearson product-moment correlation coefficient to a univariate series (Cliff and Ord, 1975), which is successfully applied in 2D image analysis (Chen et al., 2003). It is based on cross-products of the deviations from the mean and is calculated as,

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{1}$$

where $\bar{x}$ is the mean of the variable x , n is the number of observations for the variable x, w_{ij} is the distance-based weight given to each pair *(i, j)*, and is given by $\frac{1}{\sqrt{(p-k)^2+(q-l)^2}}$ where *(p,q)* is the co-ordinate position of i^{th} observation and *(k,l)* is the co-ordinate position of j^{th} observation and S_0 is the sum of all w_{ij}'s:

$$S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}.$$

We can rewrite the equation (1) as,

$$I = \frac{n}{S_0} \frac{X'WX}{X'X} \tag{2}$$

where,

$$X = \begin{pmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ . \\ . \\ . \\ x_n - \bar{x} \end{pmatrix}$$

and

$$W = \begin{pmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & \dots & w_{2n} \\ \vdots & & \ddots & \vdots \\ w_{n1} & w_{n2} & \dots & w_{nn} \end{pmatrix}$$

In the present study, we are interested to estimate the correlation of intensities between foreground and background pixels that are placed in two different locations within a spot. We can consider the intensities of n_1 foreground pixels as one set of observations

and the intensities of n_2 background pixels as another set of observations in a spot. The Moran's I given in equation (1) is defined only for a set of n observations and giving weight for each pairs. In the current situation we are interested in how foreground pixels are correlated with background pixels; therefore we modified the weight matrix in such a way that each foreground- background pixel pair will get a weight based on the distance between the pixels and all other pairs will get a weight of zero.4. Even though there are different distance functions, Euclidean distance is most often used distance function in spatial analysis (Shahid et al., 2009). So inverse of the Euclidean distance is taken as weight in this work. The correlations obtained using equation (2) for a spot was used for estimating the uncertainty associated with background corrected intensity ratio of that spot.

The minimum and maximum of Moran's I ranges from -1 to 1. If the weight matrix W is a symmetric matrices then the extreme values of the quadratic form $X'WX/X'X$ for all X in R are simply the smallest and largest eigenvalue of W (Jong et al., 1984; Bellman et al., 1970). If the weight matrix W is not symmetric, the extreme values of the quadratic form can be found by noting that $X'WX = X'SX$, where $S = (W + W')/2$, so the extreme values of the quadratic form is the smallest and largest eigenvalues of S.

We have taken 21 by 21 pixel grid size for each selected spot because 21 by 21 is considered as boundary of the spot grid. If we take more than 21 by 21 it will lead to overlap between the pixels of nearby spots. We calculated the range of Moran's I based on the weight matrix which is not symmetrical and we found that Moran's I ranges from -1.85 to 1.85.

3 Estimation of the uncertainty associated with intensity ratio of a single spot

It was observed that the uncertainty associated with the final intensity measurement depends on the correlation between intensities of various pixels within a spot. Using the calculated correlations we estimated the uncertainty associated with the intensity ratios of the selected eight spots by theory of error propagation. The uncertainty associated with the final intensity ratio is estimated using the Theory of error propagation by

$$\sigma_y^2 = J_x C_x J_x' \tag{3}$$

where, J_x is called Jacobian (or sensitivity) matrix, and C_x is called variance-covariance matrix Binu et al. (2012).

The Jacobian and the covariance matrix in this case are given by

$$\begin{aligned} J_x &= \begin{pmatrix} \frac{\partial_y}{\partial_{\bar{R}_f}} & \frac{\partial_y}{\partial_{\bar{R}_b}} & \frac{\partial_y}{\partial_{\bar{G}_f}} & \frac{\partial_y}{\partial_{\bar{G}_b}} \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{\bar{G}_f - \bar{G}_b} & \frac{-1}{\bar{G}_f - \bar{G}_b} & \frac{-\bar{R}_f - \bar{R}_b}{(\bar{G}_f - \bar{G}_b)^2} & \frac{\bar{R}_f - \bar{R}_b}{(\bar{G}_f - \bar{G}_b)^2} \end{pmatrix} \end{aligned}$$

$$C_x = \begin{pmatrix} \sigma^2_{R_f} & \sigma_{R_fR_b} & \sigma_{R_fG_f} & \sigma_{R_fG_b} \\ \sigma_{R_bR_f} & \sigma^2_{R_b} & \sigma_{R_bG_f} & \sigma_{R_bG_b} \\ \sigma_{G_fR_f} & \sigma_{G_fR_b} & \sigma^2_{G_f} & \sigma_{G_fG_b} \\ \sigma_{G_bR_f} & \sigma_{G_bR_b} & \sigma_{G_bG_f} & \sigma^2_{G_b} \end{pmatrix}$$

where $\sigma^2_{R_f}$,$\sigma^2_{R_b}$,$\sigma^2_{G_f}$ and $\sigma^2_{G_b}$ are the variances of intensities of red foreground pixels, red background pixels, green foreground pixels and green background pixels respectively. The off-diagonal elements are the covariance between the intensities of all possible pairs of the above four types of pixels in the spot. Covariances were obtained by multiplying the correlations between corresponding pairs of pixels obtained with their standard deviations. Using equation (3), we have

$$\sigma^2_Y = \frac{\sigma^2_{R_f} + \sigma^2_{R_b}}{(G_f - G_b)^2} + \frac{(R_f - R_b)^2(\sigma^2_{G_f} + \sigma^2_{G_b})}{(G_f - G_b)^4} - 2\frac{\sigma_{R_fR_b}}{(G_f - G_b)^2} - 2\frac{(R_f - R_b)^2(\sigma_{G_fG_b})}{(G_f - G_b)^4} -$$

$$2\frac{(R_f - R_b)}{(G_f - G_b)^3}[\sigma_{R_fG_f} - \sigma_{R_fG_b} - \sigma_{R_bG_f} + \sigma_{R_bG_b}] \tag{4}$$

The proposed modified Moran's I were obtained for selected eight spots from the image file which is publically available in Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM16390). These are the image files from dual channel microarray experiment of Atlantic Salmon Head Kidney Study. This study is to identify the differentialy expressed genes in two situations such as non-infected and infected Salmo salar head kidney. Non-infected Salmo salar head kidney is labeled with Cy3 and Piscirickettsia salmonis-infected Salmo salar head kidney with Cy5. The image taken from the database contains 24 blocks and each block contains 196 spots, i.e. 1414 rows and columns of spots. We have taken 8 spots from the first block for the current study. We used GenePix Pro7 software (Molecular Devices, Inc, USA) for pixel intensity extraction from the image files and we also used R (version 3.0.1) for fixed circle segmentation, estimation of correlations and uncertainty for each spot. For each of the selected 8 spots we estimated the uncertainty associated with background corrected intensity ratio using equation(4). We also estimated the uncertainty associated with intensity ratios by assuming zero correlation between the various type of pixels of the spots.

4 Results

We have estimated correlations between intensities of various types of pixels for selected eight spots. The correlation of pixel intensities between foreground and background of the spot is estimated using the Moran's I index. Correlation between red and green pixel intensities of foreground as well as background is estimated using Pearson correlation coefficient. The correlations estimated using Moran's I and Pearson coefficient

are given in the Table 1. Using these correlations of pixel intensities, we estimated the uncertainty associated with the intensity ratios of the selected eight spots by theory of error propagation. We also estimated the uncertainty associated with the intensity ratio of the eight spots by assuming zero correlation between intensities of various pixels of a spot. Table 2 gives the values of estimated uncertainty associated with the selected eight spots. For each of the 8 spots used in our study, we observed that the uncertainty estimated using the estimated correlations is less compared to the uncertainty estimated by assuming zero correlations between the pixel intensities.

Table 1: Estimated correlation values for selected eight spots

	Modified Moran's I value between				Pearson r value between	
Spot no	R_f and R_b	G_f and G_b	R_f and G_b	G_f and R_b	R_f and G_f	R_b and G_b
S3	-0.3313	-0.3316	-0.4334	-0.4203	0.8294	0.7399
S4	-0.3590	-0.3036	-0.3965	-0.4329	0.8231	0.7058
S7	-0.3956	-0.5492	-0.6035	-0.4033	0.6127	0.5667
S8	-0.4635	-0.5692	-0.6919	-0.3913	0.6653	0.6604
S11	-0.5033	-0.4687	-0.6072	-0.4960	0.5496	0.5394
S12	-0.6400	-0.7201	-0.7265	-0.6442	0.1843	0.3794
S19	-0.6556	-0.6555	-0.6207	-0.7144	0.6087	0.5629
S20	-0.5104	-0.4932	-0.4674	-0.6515	0.7695	0.6810

Table 2: Estimated intensity ratio and uncertainty associated with background corrected intensity ratio for selected eight spots

		Uncertainty	
Spot no	Intensity Ratio(IR)	with correlation	without correlation
S3	0.9989	0.1767	1.4119
S4	1.0424	0.2206	1.5662
S7	1.4006	0.8416	2.1763
S8	1.5344	0.5863	1.7540
S11	1.3794	0.4949	1.2389
S12	1.3032	0.4876	0.6154
S19	0.7955	0.1087	0.2805
S20	0.7572	0.0869	0.4215

We computed the Spearman's rank correlation between the six estimated correlations given in the Table 1 with the estimated uncertainty given in column 3 of Table 2. The estimated Spearman's rank correlations are given in Table 3. The values of estimated correlation between the G_f and R_b is observed to have positive correlation with estimated uncertainty and the remaining five correlations are observed to be weak.

Table 3: Estimated Spearman's rank correlation between various pixel intensity correlations and estimated uncertainties

	R_f and R_b	G_f and G_b	R_f and G_b	G_f and R_b	R_f and G_f	R_b and G_b
Spearman's correlation	0.29	-0.14	-0.33	0.74	-0.33	-0.33

5 Discussion and Conclusion

According to us, this is the first study that attempts to estimate the correlation between intensities of various types of pixels in a spot. In 2005, Claus et al., compared five spatial correlation structures like exponential, Gaussian, linear, rational quadratic and spherical for selected genes that accommodate errors by allowing the spatial correlation among pixels in the microarray slide (Ekstrøm Claus et al., 2005). In this study the authors considered only the immediate neighbouring pixels in the spatial correlation models. In 2010, Bergmann et al., proposed two quantities as measure of spot quality

which uses the spatial correlation between intensities of pixels in a spot. The main limitations of proposed methods are the assumption of multivariate normal distribution of intensities of pixels in the spot which may not be always true and considered only correlation between neighbouring pixels (Bergemann and Zhao, 2010). In 2012, Binu et al., studied the role of correlation between pixel intensities within a spot in cDNA microarray chip on uncertainty estimation. In that study the authors assumed equal correlation between the various types of pixel intensities of spot which is not true. This study demonstrated the importance of correlation between the intensities of pixels while estimating the uncertainty associated with each spot. The estimated uncertainty associated with the intensity ratio of each spot in a microarray chip can be used in further analysis which may change the statistical inference about the expresion level of genes. Hence it is recommended to consider correlation between intensities of various types of pixels in a spot while estimating the uncertainty associated with those spots.

In cDNA microarray experiments, the measurement of interest is gene expression level which is measured as the background corrected intensity ratio and this estimate of signal intensity measurement may have errors or uncertainty. Usually the estimate of uncertainty for intensity ratio of each spot is reported by standard deviation by assuming the pixel intensities as independent. This uncertainty could be due to the contribution of errors associated with several components in the experiment like probe and chip used, hybridization, dye and scanner used, pin geometry in a slide etc. Due to these chip artefacts, there will be correlation between various pixel intensities. The uncertainty associated with the estimated intensity ratio depends on the correlation between the intensities of various types of pixels in a spot. So we have estimated the correlation between various pixels intensities in a spot using a modified form of Moran's I spatial auto correlation index and Pearson product moment correlation coefficient. We observed that the estimated uncertainty associated with intensity ratio is less if we consider the various pixel intensities correlations compared to assuming zero correlation in a spot. Hence it is recommended to consider correlation between intensities of various types of pixels in a spot for estimating uncertainty associated with the intensity ratio. The estimated uncertainty thus obtained can be change the statistical inference about the expression level of genes in microarray experiments.

References

Baldi, P. and Hatfield, G. W. (2002). *DNA microarrays and gene expression: from experiments to data analysis and modeling*. Cambridge University Press.

Bellman, R., Bellman, R. E., Bellman, R. E., and Bellman, R. E. (1970). *Introduction to matrix analysis*, volume 960. SIAM.

Bergemann, T. L. and Zhao, L. P. (2010). Signal quality measurements for cdna microarray data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 7(2):299–308.

Binu, V. S., N S, N., Prasad K, M., and M K, K. (2012). Estimation of uncertainty

associated with intensity ratio in cdna microarray experiments. *Research & Reviews: A Journal of statistics*, 1(2):24–33.

Causton, H., Quackenbush, J., and Brazma, A. (2009). *Microarray gene expression data analysis: a beginner's guide*. John Wiley & Sons.

Chen, T.-J., Chuang, K.-S., Wu, J., Chen, S. C., Hwang, M., and Jan, M.-L. (2003). A novel image quality index using moran i statistics. *Physics in Medicine and Biology*, 48(8):N131.

Cliff, A. and Ord, J. (1975). Model building and the analysis of spatial pattern in human geography. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 297–348.

Ekstrøm Claus, T., Søren, B., Mats, R., et al. (2005). Pixel-level signal modelling with spatial correlation for two-colour microarrays. *Statistical Applications in Genetics and Molecular Biology*, 4(1):1–16.

Jong, P. d., Sprenger, C., and Veen, F. V. (1984). On extreme values of moran's i and geary's c. *Geographical Analysis*, 16(1):17–24.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, pages 17–23.

Samavi, S., Shirani, S., and Karimi, N. (2006). Real-time processing and compression of dna microarray images. *Image Processing, IEEE Transactions on*, 15(3):754–766.

Shahid, R., Bertazzon, S., Knudtson, M. L., and Ghali, W. A. (2009). Comparison of distance measures in spatial analytical modeling for health service planning. *BMC health services research*, 9(1):200.

Trevino, V., Falciani, F., and Barrera-Saldaña, H. A. (2007). Dna microarrays: a powerful genomic tool for biomedical and clinical research. *Molecular Medicine*, 13(9-10):527.