社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－BIO－10 (1)
2007／9／13

マイクロアレイにおける二遺伝子間の相関係数の非計量多次元尺度構成法を用いたロブストな計量

田口善弘 [1,2]

[1] 中央大学理工学部物理学科,[2] 中央大学理工学研究所

マイクロアレイにおいて、遺伝子発現プロファイル間の相関を求めるのは極めて重要である。しかし、マイクロアレイにはノイズがつき物である。したがって、個々の発現プロファイルに含まれるノイズをいかに低減するかが大切になる。通常、個々のペアを考慮してノイズ低減が行われ、二個以上の遺伝子発現プロファイルの相互作用は考慮されない。本論文では、非計量多次元尺度構成法を用いれば、遺伝子発現のモデルデータと実際の細胞分裂周期実験のマイクロアレイにおいて、ノイズ低減が可能であることを示す。

A robust measure of correlation between two genes on a microarray using non-metric multidimensional scaling

Y-h. Taguchi

Dept. Phys., Chuo. Univ., tag@granular.com

It is a very important task to estimate correlation between gene expression profiles in microarray. However, it is unavoidable for them to include many noises. Thus, it is a key to reduce noises included into individual gene expression profile. Usually, this procedure is based upon only each pair of gene expression profiles and mutual relationships among more than two gene expression profiles are ignored. In this paper, we have demonstrated that non-metric multidimensional scaling method can reduce both artificial noises in model gene expression profiles and natural noises in cell division cycle microarray experiments.

1 Introduction

DNA microarrays are important technique to validate the amount of transcription of individual gene. Especially, it is remarkable that it can measure all of genes at once. On the other hand, it includes many noises mainly due to technical reasons. Since there are also biological variations of samples, resulting measurement is unavoidably full of noises. In addition to this, because of high cost of usage of microarray chips, it is hopeless to get many replications enough to suppress noises by averaging. As a result, it is needed to reduce noises within microarray measurement.

Especially, the noise-free estimation of correlation coefficients is very important. Recently, Hardin et al[2] proposed to employ biweight correlation and reported that it can achieve better performance than other measures for artificial/real data set. In this paper, we propose to use non-metric multidimensional scaling method[1] (nMDS) to reduce noises which cannot be reduced by biweight correlation. Biweight correlation tries to minimize "error" within a pair of gene expression profiles. However, each pair is not independent from each other, since there are $N(N-1)/2$ pairs if there are N gene expression profiles. Thus, noise reduction is possible also by considering consistency between correlation coefficients. For example, for three gene expression profiles i, j and k, if both of correlation

coefficients of pairs (i,j) and (i,k) are positive, correlation coefficients of the pair (j,k) should be positive. Pairwise noise reduction like biweight correlation cannot consider this information. nMDS is the method to get configuration where distances between objects has the same rank order as those of pre-defined dissimilarity between objects. Thus, we can reduce noise by using low dimensional configuration obtained by nMDS to compute correlation coefficients since noise will be excluded as relationships which cannot be embedded into low dimensions.

2 Materials and Methods

The data used in this paper are artificial one and real one. As an artificial data, we have used simulated sinusoidal time sequential data,

$$s_i(t) = \sin(2\pi t/T + \delta_i) \quad (1)$$

$$n_i(t) = \epsilon_{ti} \quad (2)$$

$$x_i(t) = s_i(t) + n_i(t) \quad (3)$$

where $t = 1, \ldots, T$ and $i = 1, \ldots, N$. $\delta_i \in (0, 2\pi]$ is random phase and $\epsilon_{ti} \in \mathrm{N}(0, \sigma)$. We regard these as gene expression profile of ith gene at time t. Thus, genes are regarded to have perioicity. δ_i is unique phase for each gene and ϵ_{ti} is noise at time t for gene i. $s_i(t)$ is signal with strict periodicity, $n_i(t)$ is noise added to signal, and we observe x_i. Since $\langle s_i^2(t)\rangle_t = 1/2$ and $\langle n_i^2(t)\rangle_t = \sigma^2$, where

$$\langle \cdots \rangle_t \equiv \frac{\sum_t \cdots}{T},$$

noise dominates signal if $\sigma > 1/\sqrt{2}$. As a real data, we have employed Spellman's data set[4]. We have selected top most up regulated 1000 genes by using variance from all of α factor-based synchronization data set.

We have applied nMDS to these two data sets. In this study, negative signed correlation coefficient between gene expression profiles are used as dissimilarities.

3 Results

The purpose of noise reduction is to recover correlation coefficients ρ_{ij}^s between $s_i(t)$s from the correlation coefficients ρ_{ij}^x between $x_i(t)$s,

$$\rho_{ij}^y \equiv \frac{\left\langle \left(y_i(t) - \langle y_i(t)\rangle_t\right)\left(y_j(t) - \langle y_j(t)\rangle_t\right)\right\rangle_t}{\sqrt{\left\langle \left(y_i(t) - \langle y_i(t)\rangle_t\right)^2\right\rangle_t \left\langle \left(y_j(t) - \langle y_j(t)\rangle_t\right)^2\right\rangle_t}},$$

where $y = s$ or x.

In order to see how well each methods recover correlation coefficients for noise-free expression profile $s_i(t)$ from those with noises $x_i(t)$, we have computed correlation coefficients between distance d_{ij}^s obtained from ρ_{ij}^s and distances d_{ij} obtained by several methods,

$$\rho\left(d_{ij}^s, d_{ij}\right) \equiv \frac{\left\langle \left(d_{ij}^s - \left\langle d_{ij}^s \right\rangle\right)\left(d_{ij} - \langle d_{ij}\rangle\right)\right\rangle}{\sqrt{\left\langle \left(d_{ij}^s - \left\langle d_{ij}^s \right\rangle\right)^2\right\rangle \left\langle \left(d_{ij} - \langle d_{ij}\rangle\right)^2\right\rangle}},$$

			Two dimensional embeddings by		
d_{ij}	$\sqrt{2(1-\rho^x_{ij})}$	biweight	nMDS	PCA	normalized PCA
$\rho(d^s_{ij}, d_{ij})$	0.653 ± 0.008	0.645 ± 0.009	0.726 ± 0.006	0.628 ± 0.009	0.704 ± 0.008
$\rho(d^x_{ij}, d_{ij})$	1.0	0.975 ± 0.002	0.805 ± 0.006	0.698 ± 0.010	0.790 ± 0.007

Table 1: Correlation coefficients $\rho(d^s_{ij}, d_{ij})$ between d^s_{ij} and d_{ij} and correlation coefficients $\rho(d^x_{ij}, d_{ij})$ between d^x_{ij} and d_{ij}. In the left most column, we have shown $\rho(d^s_{ij}, d^2_{ij})$, too. For d_{ij}, we have employed biweight ($\sqrt{2(1-\rho^b_{ij})}$), distances in two dimensional embedded space obtained by nMDS, PCA and normalized PCA.

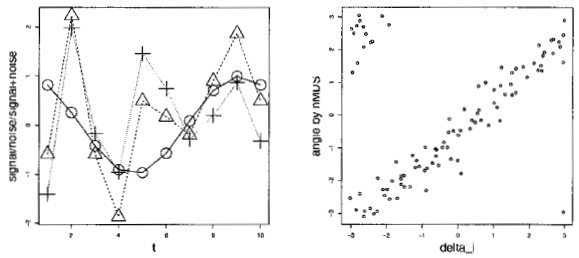
where

$$\langle \cdots \rangle \equiv \frac{\sum_{i,j} \cdots}{N(N-1)/2}$$
$$d^s_{ij} = \sqrt{2(1-\rho^s_{ij})}$$

and N is number of objects (gene expression profiles). For biweight correlation ρ^b_{ij} between gene expression profiles i and j, distance is defined as $\sqrt{2(1-\rho^b_{ij})}$. Other candidate for d_{ij} are those by nMDS, principal components analysis (PCA) and normalized PCA.

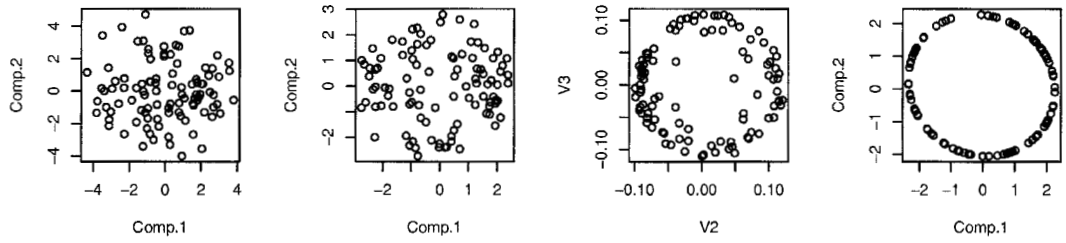
First of all, we have generated 5 (for nMDS) and 12 (for others) sets of $x_i(t)$s, and calculated mean $\rho(d^s_{ij}, d_{ij})$ within each set. The standard errors are computed from 5 or 12 mean values. In Table 1, we have shown the results for $N = 100, T = 10, \sigma = 1.0$. For comparison, in the first column of Table 1, we have shown the correlation coefficients between Pearson's correlation coefficient $\rho(d^s_{ij}, d^x_{ij}) \simeq 0.653$ between d^s_{ij} and d^x_{ij}. The purpose is to get larger $\rho(d^s_{ij}, d_{ij})$ by using other methods. Clearly biweight correlation cannot get larger $\rho(d^s_{ij}, d_{ij})$ than $\rho(d^s_{ij}, d^x_{ij})$. Biweight correlation is closer to d^x_{ij} than d^s_{ij} (see the lowest raw in Table 1 named as $\rho(d^x_{ij}, d_{ij})$). On the other hand, two dimensional embedding by nMDS has larger $\rho(d^s_{ij}, d_{ij})$ than $\rho(d^s_{ij}, d^x_{ij})$. If we apply t-test between a set of 12 mean values $\rho(d^s_{ij}, d^b_{ij})$ by biweight correlation and a set of 5 mean values $\rho(d^s_{ij}, d_{ij})$ by nMDS, P-value is 2.4×10^{-6}. Normalized PCA, which is equivalent to metric classical MDS with employing $d_{ij} = \sqrt{2(1-\rho_{ij})}$ as dissimilarities in this case but requires less computational resources, is known to be as good as nMDS[5]. However, in this case, nMDS is significantly better than normalized PCA, since P-value by t-test between these two set is 0.04. Thus, non metricity is important in this case.

One may think that it is not fair since we have used the information that signal is two dimensional (the degree of freedom is two). In order to demonstrate that solely the information of dimension of signal does not help to extract noises, we have also done simple PCA. Clearly, two dimensional embedding of ordinary PCA cannot extract noise at all (see Table 1). $\rho(d^s_{ij}, d_{ij})$ by PCA is even smaller than $\rho(d^s_{ij}, d^x_{ij})$. This is reasonable since $\langle s^2_i(t) \rangle_t \simeq 0.5$ while $\langle n^2_i(t) \rangle_t \simeq 1.0$ for $\sigma = 1.0$. This means, noise is twice larger than signals. This is demonstrated in Fig. 1 (a). Clearly noise $n_i(t)$ dominates signal $s_i(t)$. Simple selection by PCA of two dimensions which have more variance than remaining dimensions cannot reduce noises at all. If we see configuration of two dimensional embedding, this tendency becomes clearer. In Figs. 2, we have compared two dimensional



(a) (b)

Figure 1: (a) Typical example of model gene expression profile. ○: signal $s_i(t)$, +: noise $n_i(t)$, △: signal + noise $x_i(t)$. $T = 10, \sigma = 1.0$. (b) Typical comparison between δ_i and polar angle estimated from nMDS embedding (Fig. 2 (c)).

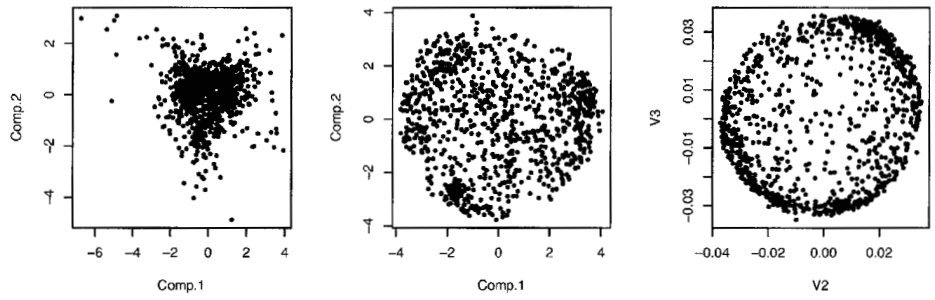


(a) (b) (c) (d)

Figure 2: Configuration of embedding of $x_i(t)$ with (a) PCA, (b) normalized PCA, (c) nMDS. For comparison, PCA results for $s_i(t)$ is also shown in Fig. (d).

embedding by PCA, normalized PCA, nMDS of $x_i(t)$ with each other. For comparison, that of PCA for signal $s_i(t)$ is also shown. It is clear that nMDS outperforms PCA and normalized PCA embedding. Especially, δ_i is reproduced very precisely. δ_is correspond to polar angles in Figs. 2. nMDS (and normalized PCA, not shown here) can reproduce δ_i with the accuracy of $24^\circ \pm 2^\circ$ (Fig. 1(b)). Thus, main deviation between d_{ij} and d^s_{ij} originates from deviation from a ring.

One may think that this example is too artificial and biased to be favorable to nMDS. In order to see if there are real examples for which nMDS is useful, following Hardin et al[2], we have tried to apply nMDS to gene expression profiles where the periodic nature is expected. We have selected top 1000 genes having larger variance in α factor-based synchronization experiments by Spellman. In Fig. 3(c), we have shown the results of two dimensional embedding by nMDS. We can see some circular arrangement which can never be extracted by other methods (see Figs. 3 (a) and (b)). Here we have employed negative signed correlation coefficients between gene expression profiles as dissimilarity



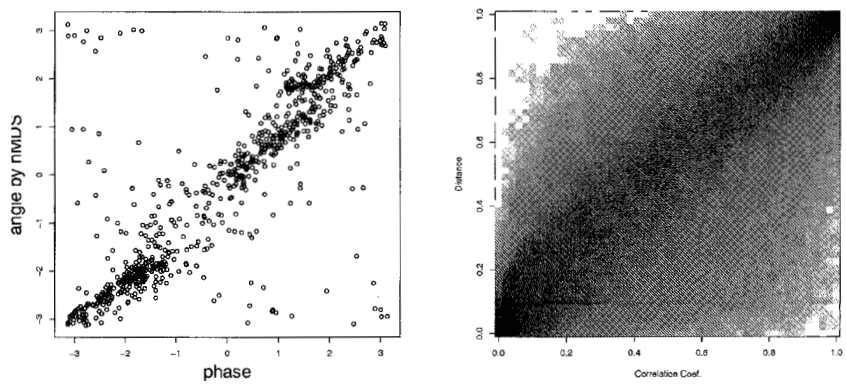
(a) (b) (c)

Figure 3: 2D embedding by (a) PCA (b) normalized PCA and (c) nMDS for Spellman's data set. Top most 1000 genes with regard to variance are consider for α factor-based synchronization experiments.

and embedded them into two dimensional space. The biological significance of circular arrangement obtained by nMDS in this data set has already been reported partially[3]. In Fig. 4(a), we have compared phase obtained by sinusoidal fittings $\langle x_i(t)\sin(2\pi t/T)\rangle_t$ and $\langle x_i(t)\cos(2\pi t/T)\rangle_t$ and polar angle estimated by nMDS embedding (Fig. 3(c)). It is clear that nMDS successfully reproduce phases. While for sinusoidal fittings we need the information of period ($T = 66$min.[4]), nMDS does not require this information. Thus, nMDS is more unsupervised. The fact that nMDS can reproduce phases in unsupervised manner has already reported for the fission yeast[6]. One may think that Fourier transformation also can extract the correct period, thus we can get phase by using period with the largest Fourier transformation spectrum. However, period extracted by Fourier analysis must be T'/N, where T' is the length of time sequence and $N = 1, 2, \ldots, T'/2$. Generally, there are no reasons that T'/T is integer, thus nMDS is more suitable method to extract phases at least when T'/T is not integer.

Of course, the simple appearance of ring in the embedded space by nMDS does not always mean biologically significant structure. For example, it is pointed out that a set of random walks analyzed by nMDS exhibits ring, too[7]. However, if the ring is outcome of random walk, period T must be equal to T'. In this case, $T = 66$min. while $T' = 119$. Thus the appearance of ring cannot be explained as being accidental. Moreover, the fact that phase obtained by assuming $T = 66$ min. agree with phase obtained by nMDS shows that the ring has surely biological base, since T is really the period of cell division cycle.

Then in order to compare distance obtained by nMDS and that by correlation coefficients, we have shown the comparison between ascending rank order of distance d_{ij} between gene expression profiles in the embedded space by nMDS and the distance $\sqrt{2(1-\rho_{ij})}$ by correlation coefficient ρ_{ij} between gene expression profiles i and j (Fig. 4(b)). Although there are globally well correlation between two rank orders, there are some pairs of gene expression profiles for which two ranks differ. In Fig. 5, we have shown scatter plot of pairs of gene expression profiles which are located at upper left corners in Fig. 4. This



(a) (b)

Figure 4: (a) Comparison between phase by sinusoidal fitting (horizontal axis) and polar angle (vertical axis) estimated by nMDS embedding (Fig. 3(c)). Vertical axis is shifted such that points are located at diagonal region. (b) Scatter plot of rank order of all pairwise distances of the 1000 most variable genes in the yeast data. Horizontal axis is rank order of $\sqrt{2(1-\rho_{ij})}$ where ρ_{ij} is the correlation coefficient between ith and jth gene expression profile. Vertical axis is the rank order of distance d_{ij} obtained by nMDS (Fig. 3(c)). The blackest squares represent 6800 pairs of genes. The lightest squares represents two pair of genes. Grey scale is proportional to log transformed number of pairs. Both horizontal and vertical axes are normalized such that the range spans from 0 to 1 and takes the smallest values of distances at bottom left corner. This means rank order is ascending.

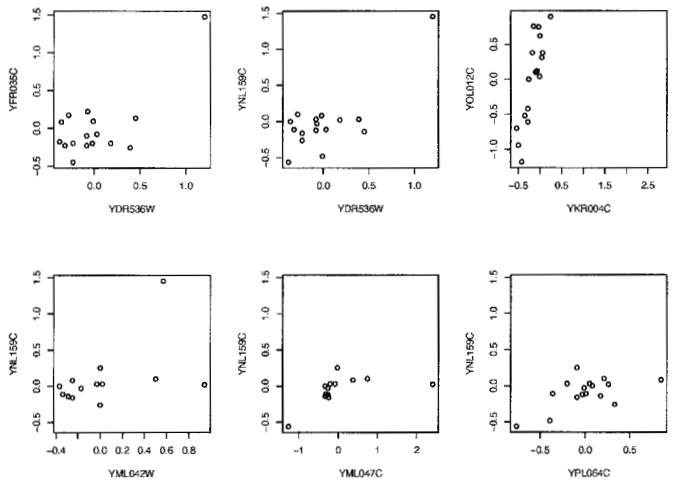


Figure 5: Each points represent gene expression profiles for specified genes. d_{ij} by nMDS shows larger values and distance $\sqrt{2(1-\rho_{ij})}$ computed from Pearson's correlation coefficients ρ_{ij} takes smaller values for these group of 6 pairs.

means, for these pairs, distance by Pearson correlation coefficient is smaller while that by nMDS is larger. In other words, Pearson correlation coefficient says that these two are similar while nMDS result says that they are not. For most graphs in Figs. 5, there are outliers and they essential make each of pair gene expression profile look similar. Although some of graphs lack outliers, this is due to missing values, since upper/lower limits of axes are decided by considering all observations, even if points are not plotted due to lack of one of a pair of observations. For example, for the pair of YOL012C and YKR004C (top right corner), although upper limit of horizontal axis is 2.5, there are no points plotted with this value. This means that YKR004C has gene expression as large as 2.5 although corresponding profile of YOL012C is missing. Thus, TKR004C has outliers even if it is not plotted. These outliers cause larger correlation coefficient (i.e., smaller distance). If we exclude outliers, this similarity is doubtful.Thus, distance by nMDS is more trusted than that by Person correlation coefficients. Hardin et al[2] have reported that biweight correlation has similar function, but biweight transformation $T(X)$ for X,

1. $m =$ the median absolute value of X
2. cutoff $= 6 \times m$
3. $$\begin{aligned} T(X) &= (1-(X/\text{cutoff})^2)^2 && ,\text{if abs}(X) < \text{cutoff} \\ T(X) &= 0 && ,\text{if abs}(X) \geq \text{cutoff} \end{aligned}$$

explicitly intends to exclude outliers. On the other hand, nMDS does not aim such an specific purpose, thus nMDS is regarded as having the power that automatically reduce

noises by outliers without any supervision. This shows that importance to consider mutual relationship among more than two gene expression profiles.

4 Conclusion

In this paper, we have proposed usage of nMDS for noise reduction procedure of microarray experiments. For model (simulated) gene expression profiles, nMDS outperform biweight correlation which is reported to outperform other methods[2]. nMDS turns out to have ability of noise reduction for real gene expression profile, too. Without supervision, nMDS can exclude similarity caused accidentally by outliers, although biweight which is designed for this purpose have the same ability.

5 Acknowledge

The author thanks Prof. Y. Oono and S. Rajaram for their critical readings of this manuscript. This work has been partially supported by the Grant-in-Aid for Creative Scientific Research No.19500254 of the Ministry of Education, Culture, Sports, Science and Technology (MEXT) from 2007 to 2008. We are grateful for their support.

References

[1] Y-h. Taguchi, Gene arrangement for cell division cycle microarray experiments without sinusoidal fittings, 2006-BIO-7, (2007) 173-180.

[2] J. S. Hardin, A. Mitani, L. Hicks, and B. VanKoten, A robust measure of correlation between two genes on a microarray, BMC Bioinformatics, 8, (2007) 220.

[3] Y-h. Taguchi and S. Rajaram, Noise reduction procedure for microarray experiments with non-metric multidimensional scaling method, IPSJ SIG Tech. Rep., 2005-BIO-4, (2006) 9-15.

[4] P. Spellman, G. Sherlock, M. Zhang, V. Iyer, K. Anders, M. Eisen, P. Brown, D. Botstein and B. Fitcher, Comprehensive identification of cell cycle-regulated genes of Yeast *Saccharomyces cervisiae* by microarray hybridization, Mol. Biol. Cell, 9, (1998) 3273-3297.

[5] Y-h. Taguchi and Y. Oono, Relational patterns of gene expression via non-metric multidimensional scaling analysis, Bioinformatics 21 (2005) 730-740.

[6] Y-h. Taguchi, Detecting cell cycle regulated genes of *Schizosaccharomycs pombe* by using non-metric multidimensional scaling without sinusoidal fitting, IPSJ SIG Tech. Rep., 2005-BIO-3, (2005) 59-66.

[7] S. Rajaram, private communication.