Journal of Data Science **8**(2010), 505-519

A Multivariate Method for Normalization in Affymetrix Oligonucleotide Microarray Experiments

Zhide Fang[1], Xiaohu Li[2] and Lizhe Xu[3]
[1]*Louisiana State University,* [2]*University of New Orleans*
and [3]*APHIS USDA PIADC*

Abstract: Affymetrix high-density oligonucleotide microarray makes it possible to simultaneously measure, and thus compare the expression profiles of hundreds of thousands of genes in living cells. Genes differentially expressed in different conditions are very important to both basic and medical research. However, before detecting these differentially expressed genes from a vast number of candidates, it is necessary to normalize the microarray data due to the significant variation caused by non-biological factors. During the last few years, normalization methods based on probe level or probeset level intensities were proposed in the literature. These methods were motivated by different purposes. In this paper, we propose a multivariate normalization method, based on partial least squares regression, aiming to equalize the central tendency, reduce and equalize the variation of the probe level intensities in any probeset across the replicated arrays. By so doing, we hope that one can precisely estimate the gene expression indexes.

Key words: Affymetrix GeneChip, normalization, oligonucleotide, partial least squares regression.

1. Introduction

The microarray technology makes it possible for scientists to simultaneously examine the expression profiles of a huge number of genes in living cells. It has found wide application in many areas of biomedical research, such as, gene expression, linkage analysis, Single Nucleotide Polymorphism (SNP) analysis, pathway analysis, disease diagnosis, drug discovery and evaluation, and toxicological research.

Briefly speaking, microarrays are small, solid supports onto which the selected sequences (so-called probes) of thousands of genes are immobilized at fixed locations. These immobilized probes are then employed to lock down or catch their target genes in samples based on nucleotides base-pairing rule – the hybridization in molecular biology. There are many kinds of microarrays in the market, differenced by probes attached: DNA fragments or oligonucleotides; by production methods: in situ synthesis or deposit by high precision robots; by applications: gene expression array, SNP array etc.

The single-channel oligonucleotide gene expression microarray produced by Affymetrix (GeneChip®) adopts 25 mer probes synthesized with photolithography techniques. The probes in an Affymetric GeneChip are presented in pair: the *perfect match probe* (PM) whose sequence exactly matches its target-gene sequence and the *mismatch probe* (MM) obtained by exchanging the middle (13^{th}) base of the PM with its Watson-Crick complement. There are multiple probe pairs for a given transcript and the number of these probe pairs, which together form a probeset, is fixed in the array design stage. For example, a HG-U95 array has 16 pairs per probeset while a HG-U133 array has 11 pairs. The gene expression index, employed in a later stage to identify the significantly differentially expressed genes, is obtained by summarizing the signal intensity values of probe pairs in a probeset. The summarizing method used by Affymetrix is the weighted average of the signal intensity differences ($PM - MM$) of probe pairs in a probeset. Since then, many statistical methods have been proposed. See, for example, Li and Wong (2001b), Affymetrix[1] (2002), Irizarry *et al.* (2003), Chen *et al.* (2006) and others for detailed discussion of different procedures.

In biological study, replication under the same condition (tested or control) is necessary in order to draw any meaningful conclusion. This will certainly introduce variability (or noise) to the feature intensities in microarray experiments. There are two types of variability in microarray data. One is the biological variability, indicated by the gene intensity difference under different conditions. It is the first goal of a microarray experiment to reveal this biological change. Another is the noise or system variability caused by non-biological factors such as the physical difference among arrays, the random errors in RNA sample preparation processes, other variations from hybridization and image processing steps, etc. (see, for example, Hartemink, 2001). With these system noises complicating the biological differences, down-stream analyses on microarray data will definitely increase the false positive rate or the false negative rate. Thus, a proper normalization procedure is appealed before any down-stream analysis to detect significantly differential genes is carried on. A successful procedure should be able to remove the system noises while retaining the existing biological difference.

Affymetrix's normalization method rescales the probeset expression measures by a constant factor such that these measures between two or more arrays are linearly related through the origin. Following this, other methods are proposed

[1] Affymetrix Whitepaper. Available: http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf. Accessed 6 March 2009.

in the literature. These include quantile normalization in Bolstad *et al.* (2003), piecewisely linear run median method of dChip in Li and Wong (2001a), nonlinear normalization methods in Schadt *et al.* (2001), Workman *et al.* (2002), Faller *et al.* (2003), variance-stabilizing normalization (*VSN*) in Huber *et al.* (2002) and many others. These normalization methods are proposed for different purposes. For example, the quantile normalization intends to make the empirical distribution of intensities are the same across arrays, and the *VSN* method tranforms the data such that the variance of measured intensities is independent of the mean.

In this paper, we propose a multivariate normalization method, simultaneously Partial Least Squares (*sPLS*) regression method. It is based on the Partial Least Squares (*PLS*) regression introduced by Wold (1975). This is a probe level normalization method. Detailed description of the method is given in §3. By this method, we intend to equalize the central tendency, to reduce the variation of the probe level intensities in a probeset, and to equalize these variations across the replicated arrays. Thus, this normalization can make the estimation of gene expression indexes more precisely.

We will use the data sets produced by Affymetrix GeneChip Human Genome U95 (HG-U95) arrays and HG-U133 plus 2.0 arrays to evaluate our method. During the last few years, the human genome U95A and U133 GeneChips were reference standard for gene expression studies. Gene expression data from both experimental and clinical studies are voluminous and freely available from several public repositories. This availability enable scientists to access and compare their research results to published studies. Meanwhile, The U133 is still widely used even though new generation of arrays, the Gene ST 1.0 and Exon ST 1.0, are available in the market. The major advantage of the U95A and U133 is the wealth of data analysis tools available, and tools for QC are much better. Moreover, the perfect match/mismatch probe sets design of these expression arrays offer alternative way to examine gene expression as well as to detect genetic variation within and between species, a case in point for single feature polymorphism, first done in yeast to examine wild strain types.

2. Data

To evaluate our normalization method, we use the public Dilution/Mixture data. The HG-U95 arrays are employed to produce these data. Each probe set in the arrays of these two data sets contains 16 probe pairs. Thus, in each array, we have 16 pairs of probe level intensities (*PM* and *MM*) for each transcript. The normalization method can also normalize data produced by arrays in which the number of probe pairs is different from 16. As an example, we apply the method to data sets (downloaded from website, http://www.affymetrix.com/index.affy) produced from HG-U133 plus 2.0 arrays in which each probeset has 11 pairs of probe level intensities for each transcript, at the end of the paper.

In Dilution/Mixture study, the cRNA masses for sample A (liver cRNA) and for sample B (central nervous system cRNA) are $20,\ 10,\ 7.5,\ 5,\ 2.5, 1.25\, ug/(200ul)$, separately. The mixtures of these two samples are run for three combinations: $7.5 : 2.5,\ 5 : 5,\ 2.5 : 7.5$. There are 5 replicates for each dilution and mixture. Thus, the Dilution/Mixture data set contains 75 arrays. The experiments use 5 different Affymetrix scanners for each dilution and mixture.

The sample data set from Affymetrix HG-U133 plus 2.0 arrays contains 53 arrays, of which there are 33 arrays for 11 tissues (each tissue has three replicated arrays) including breast, cerebellum, heart, kidney, liver, muscle, pancreas, prostate, spleen, testes, thyroid, and there are 20 arrays for 4 mixtures (each mixture has five replicated arrays) including testes-cerebellum with ratio 1:1, testes-cerebellum with ratio 1:2, heart-testes-cerebellum with ratio 1:1:1 and heart-testes-cerebellum with ratio 2:1:3.

3. Normalization Method

Assume that there are p genes and n probe pairs for each gene in one array. Thus, one array will produce a $(n \times p)$ data matrix X of probe pair level intensities. By the design of oligonucleotide microarray, we assume that the intensities $\{X_{ig}\}_{i=1}^{n}$, for any gene g, are independent. Thus the n row vectors of X can be viewed as p-dimensional multivariate observations.

Suppose we have two $(n \times p)$ data matrices X, Y of intensities generated from two arrays. When two experiments are technically replicated, with the same sample being applied on more than one array, the multivariate gene expression vectors should be identical except random errors if there are no systematic errors in the experiments. When the target samples in two arrays are from two conditions (for example, tumor or normal), the vectors should be highly correlated if no experimental errors because of the assumption that the percentage of differentially expressed genes is small. Thus, the intensities in Y can be predicted from those in X through statistical regression models and vice verse. We may apply prediction techniques to reduce the systematic errors. However, the facts that the genes in an array are correlated and that $p >> n$ make it impossible to apply the traditional multivariate regression for prediction purpose. Partial least squares (*PLS*) is especially useful in this situation-building prediction equations when there are large number of explanatory variables and small number of sample data – because it intends to search a set of factors or latent variables that explain as much as possible of the covariance between X variables and Y variables, and then to construct prediction equations linking the factors to the Y variables (see Garthwaite, 1994 or Abdi, 2003 for details). In other words, factors are mutually independent (orthogonal) linear combinations of X variables, and are obtained in order to provide maximum correlation with Y variables. Since its first appearance (Wold, 1975), *PLS* regression has found its application in analysis of experiments in many areas. These, for example, include the most recent application in Electroencephalography (Alm *et al.*, 2009) and the earlier application in Analytical chemistry (Delaguardia *et al.*, 1996), in Ecology (Eriksson *et al.*, 1995), in Agriculture and food science (Kays *et al.*, 1996) and many others. A recent review about *PLS* can be found in Rosipal and Krämer (2006) and the references therein.

Specifically, *PLS* obtains factors by iteration. A pair of vectors $\alpha = X\mathbf{u}$ and $\beta = Y\mathbf{v}$ are found such that $||\mathbf{u}|| = ||\alpha|| = 1$ and the cross-product $b = \alpha^T\beta$ is maximum. Then, α is the first factor we are searching. With $\mathbf{q} = X^T\alpha$ and X, Y being replaced by residuals: $X - \alpha\mathbf{q}^T$, $Y - b\alpha\mathbf{v}^T$, repeat the above maximizing process until a pre-chosen number of factors are obtained. Denote Q, V as matrices consisting of vectors $\mathbf{q}$, $\mathbf{v}$, respectively, and B as a diagonal matrix consisting diagonal elements b. Then, *PLS* regression predicts Y by $\hat{Y} = X(Q^T)^+BV^T$, where $(Q^T)^+$ is the Moore-Penrose inverse of Q^T. See, for example, Abdi (2003) for details.

An algorithm to search the factors is available in the literature (Hoskuldsson, 1988). It is as follows.

Step 1. Initialize the vector β.

Step 2. Let $\mathbf{u} = X^T\beta\dfrac{1}{\text{norm}(X^T\beta)}$.

Step 3. Let $\alpha = X\mathbf{u}\frac{1}{\text{norm}(X\mathbf{u})}$.

Step 4. Let $\mathbf{v} = Y^T\alpha\frac{1}{\text{norm}(Y^T\alpha)}$.

Step 5. Let $\beta = Y\mathbf{v}$.

Step 6. Repeat Steps 1 to 5 for the next factor with X, Y replacing by residuals: $X - \alpha\mathbf{q}^T$, $Y - b\alpha\mathbf{v}^T$ respectively if there is convergence; otherwise return to Step 2.

Note that the vector $\mathbf{q}$ and the scalar b are calculated as in previous paragraph. We claim that there is convergence whenever $\text{norm}(\alpha_{new} - \alpha)/\text{norm}(\alpha) < 10^{-6}$.

In an oligonucleotide microarray experiment, there usually are $N > 2$ replicates: $X_1, \ldots, X_N$. The *sPLS* normalization method to probe level data includes two steps. First, we obtain the median of N chips. Then, by using the median as the dependent variables Y and X_j as the independent variables, the *PLS* prediction $\hat{Y}$ is the normalized probe pair level data of j^{th} chip, $j = 1, 2, \ldots, N$.

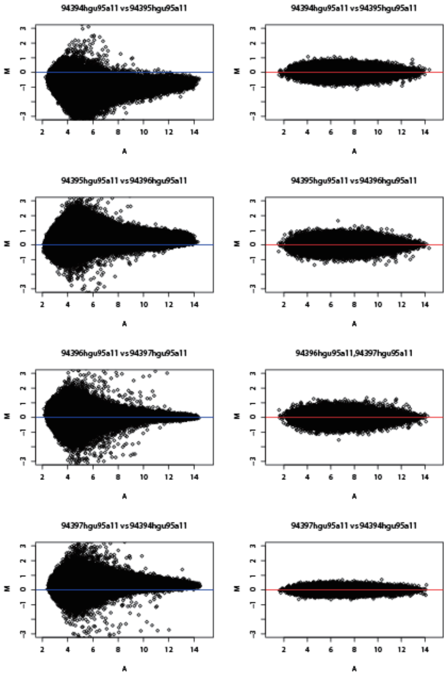

Figure 1: MA plots for *PM* values at probe level: the left panel includes plots before normalization and the right panel includes corresponding plots after *sPLS* normalization.

Note that in this paper cross-validation procedure is applied to determine the number of chosen factors in every *PLS* regression. For the data sets produced by HG-U95 arrays, since there are $n = 16$ probe pairs in each chip, we split the data into four groups, each being a $(4 \times p)$ matrix. For every k from 2 to 16, one group is omitted, and the Y, X_j from other three groups are used to search k factors. The resulted *PLS* prediction equation is used to predict Y values for omitted group. Then the norm of the difference between the omitted group's Y values and predicted values are calculated. Repeat the process until each group has been omitted once and add all four norms to get the total error: $E(k)$. If $E(k) < E(k+1)$, then we choose k as the number of factors in *PLS* regression (see Garthwaite, 1994). We usually have $k = 12$ for HG-U95 data.

4. Results and Discussions

As an illustration, we apply in this section the *sPLS* normalization method proposed in the previous section as well as some other normalization methods in the literature to the data sets described in §2. It should be remarked here that the log transformation with base 2 is applied to raw probe level intensities before normalization. All calculations were performed in R (http://www.r-project.org). Codes are available upon request.

4.1 Probe level assessment of normalization

The M vs A plots are used to evaluate the performace of the proposed normalization method, where $M = \log_2(x) - \log_2(y)$ and $A = \frac{1}{2}[\log_2(x) + \log_2(y)]$ for two intensities, x, y, from two arrays. As an example, we present in Figure 1 four pairwise M vs A plots for liver dilution series data (PM intensities of 94394hgu95a11, 94395hgu95a11, 94396hgu95a11 and 94397hgu95a11, at concentration 10 (ug/(200 ul)). We have the following observations from the plots. (1) The plots in the left panel, which are based upon the data before normalization, clearly show significant deviations from the horizontal line, $M = 0$. This suggests the necessity of proper normalization. (2) In the corresponding plots in the right panel, which are based upon the data after *sPLS* normalization method, the point clouds are centered tightly around the horizontal line $M = 0$. These imply that *sPLS* normorlization method successfully removes the system noise arised due to replication. Other six pairwise M vs A plots for liver dilution series data at concentration 10 (ug/(200 ul)) are similar and are not presented here.

Next we compare the performance of our normalization method with those of Affymetrix scale normalization, the quantile normalization and the *VSN* method. Subject to limitations of space, we only choose these three methods for comparison because to our knowledge they, similar to our method, are methods based on the probe level intensities and the *VSN* method intends to stabilize the variances of measured intensities.

We obtain in Figure 2 densities of probe intensities ($\log_2(PM)$) for two genes (left, right panel separately in the figure) randomly selected from the liver dilution data. Five curves in each plot represent five arrays 94394hgu95a11, 94395hgu95a11, 94396hgu95a11, 94397hgu95a11 and 94398hgu95a11 at concentration 10 (ug/(200 ul)). Plots in the top row are based on the data before normalization, while plots in the following rows are for the data after the scale normalization, the quantile normalization, the *VSN* normalization and the *sPLS* normalization sequentially. From these density plots we observe that (a) there have obvious discrepancies among the distributions for the replicated arrays before normalization and this once again implies the necessity of normalization. (b) All four normalization methods can remove certain levels of the distribution discrepancies among arrays. However, while both the scale method and the quantile method reduce more density discrepancies in high intensity areas than those in the low intensity areas, both *VSN* and *sPLS* methods greatly remove this discrep-

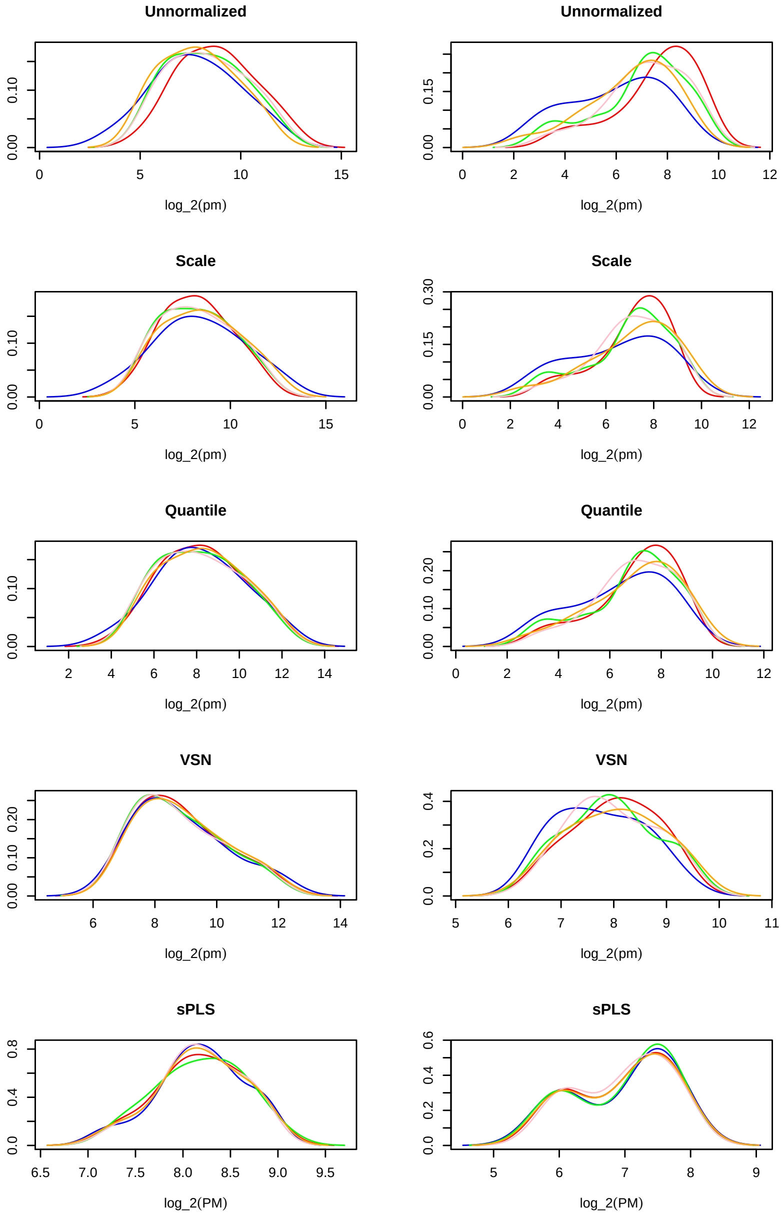

Figure 2: Densities of $\log_2(PM)$ of two randomly selected genes at probe level.

ancy in the whole intensity area, with the *sPLS* method performing best. (c) The intensity range in Figure 2 indicates that the experimental variability in the data set of probe intensities for a gene in a single array becomes much smaller after *sPLS* normalization (note that the horizontal scale ranges are approximately from

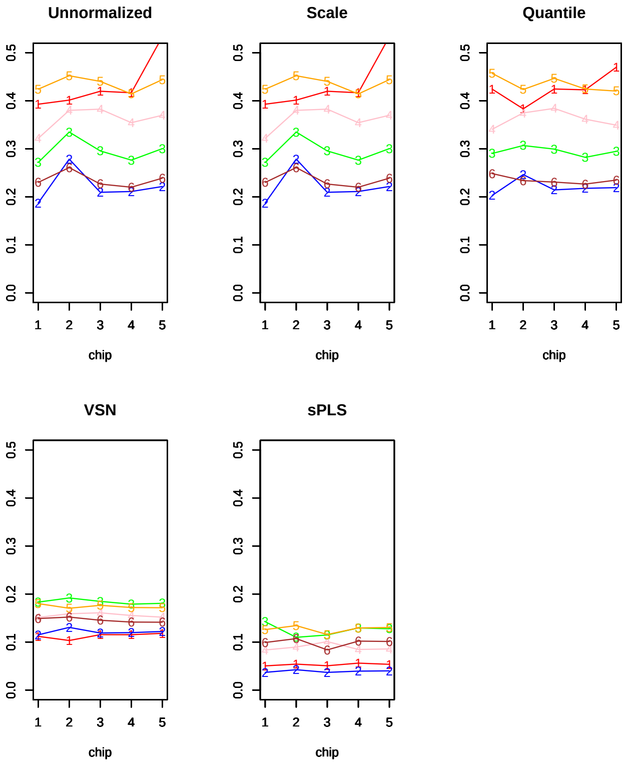

Figure 3: Coefficient of variation of $\log_2(pm)$ of six randomly selected genes at probe level.

5.5 to 9.5 in the last two plots). The *VSN* method has similar performance but does not reduce the range as much as the *sPLS* does. Other two methods do not reduce this variability. Thus, *sPLS* normalization helps to precisely summarize the gene expression indexes.

The purpose of the upstream pre-processing to the microarray data is to estimate precisely the gene expression index from the probe level intensities of a gene. If there were no systematic errors, not only should we expect small experimental variability of the measurements of probe level intensities for a gene in a single array, but also approximately the same variability and the same central tendency across replicated arrays. An efficient normalization method should be able to reduce the variability and equalize both the central tendency and the variability. In addition to the comparison among the intensity ranges in Figure 2, we investigate the boxplots of the probe level intensities and the plots of related standard deviations. We obtain the boxplots (not presented) of probe level intensities for four genes randomly selected from the liver dilution data. For each selected gene, the plot has five boxplots corresponding to five replicated

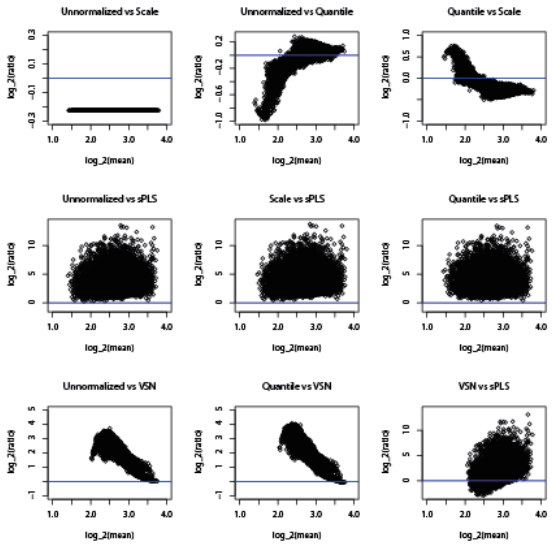

Figure 4: The plots of the log ratio of variances versus the log of the geometric average of means for array 94395hgu95a11.

arrays: 94394hgu95a11, 94395hgu95a11, 94396hgu95a11, 94397hgu95a11 and 94398hgu95a11, respectively. From the plots of original data, scale normalized, quantile normalized, *VSN* normalized and *sPLS* normalized data respectively, it is obvious that the original data show unequal medians and uniformly large variabilities for each gene across the replicated arrays, while the corresponding data after *sPLS* normalization indicates approximately the same medians and small variabilities. The reduced variabilities, by *sPLS* normalization, in each plot are also approximately the same across the replicated arrays. We also observe that the intensity values become larger after *VSN* normalization.

We also calculate and compare the standard deviations of the probe level intensities, before normalization, after scale normalization, quantile normalization, *VSN* normalization and *sPLS* normalization separately, for the randomly selected genes across replicated arrays. We obtain the plots (not presented) of standard deviations versus arrays and observe from the plots that the scale normalization does not reduce the magnitude of the standard deviation nor does it equalize the standard deviations across the arrays. The quantile normalization equalizes the standard deviations at certain level but does not reduce their values. Both *VSN* and *sPLS* normalization (*sPLS* method performs better) reduce and equalize these standard deviations across the arrays.

We further compare in Figure 3 the coefficients of variation (CV) of six randomly selected genes (corresponding to six curves in each plot). The five dots in each curve represent five arrays. We see once again that the *sPLS* normalization greatly reduces and equalizes the coefficients of variation across arrays. Note that the plot after scale normalization is the same as that before normalization – this is because of the fact that scaling does not change the value of the CV.

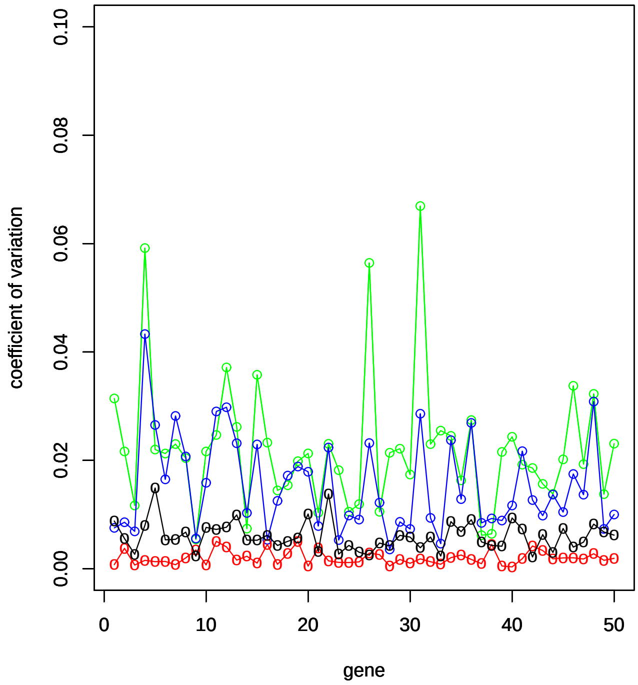

Figure 5: Coefficients of variation of the RMA gene expression indexes in five arrays from 94394hgu95a11 to 94398hgu95all. The green, blue, black and red curves are for scale normalized, quantile normalized, *VSN* and *sPLS* normalized data respectively.

At the end of this section, we present the observations on how the change of variation due to normalization depends on the gene intensity in the same array. We calculate the mean and the variance of the probe level intensities for every gene in one array. This is repeated for unnormalized data, data after scale normalization, data after quantile normalization, data after *VSN* normalization and data after *sPLS* normalization, respectively. The plot (not shown) of $\log_2(variance)$ versus $\log_2(mean)$ indicates that the probeset with larger mean intensity usually has larger variation (Huber *et al.* (2002) have the same observation). This will certainly lower the precision of the estimate of the large gene expression index. Thus we hope that the normalization step can reduce this variation. To see the effect of the normalization, in Figure 4, we plot the log with base 2 of the ratio of two variances versus the log with base 2 of the geometric average of two means from the array 94395hgu95a11 for the liver at dilution level 10 ug/(200 ul). It is clear that *sPLS* method significantly reduces the variances in both probesets with low intensities and those with high intensities. *VSN* method has similar, but less variance reduction. With small magnitude, the scale method increases all the variances while the quantile normalization reduces the variances in probesets with large intensities but increases the variances in probeset with low intensities. We should report that in the plots for the array 94394hgu95all (not shown), the performances of the *sPLS* method and the *VSN* are very similar to those in Figure 4. However, the scale method reduces all the variances while the quantile normalization reduces the variances in probesets with low intensities but increases the variances in probesets with high intensities.

In summary, we conclude that in terms of the central tendency and the variability of the probe level intensities, the *sPLS* normalization performs best among the four methods in that it equalizes the medians, reduces and equalizes standard deviations and coefficients of variation across the replicated arrays, while the scale normalization performs worst.

4.2 Gene expression assessment of normalization

In this section, we evaluate the performance of the *sPLS* normalization method at the probeset level. We obtain the gene expression indexes by the Robust Multichip Average (RMA) based upon a robust average of log background corrected PM intensities. One may refer to Irizarry *et al.* (2003) for detailed discussion on the RMA method. It should be remarked that RMA gene expression indexes in microarrays are obtained through the three sequential steps: RMA background correction, quantile normalization and Tukey's median polish. For the purpose of comparison, we also obtained the gene expression indexes by replacing the quantile normalization with the scale normalization, *VSN* normalization and *sPLS* normalization. Note that the background correction in RMA is applied on the data set in which every probeset contains 16 probe pairs and there is no background correction before *VSN* normalization.

Then, based upon the RMA expression indexes, the coefficient of variation for each gene in all replicated microarrays is calculated for the scale normalized data, the quantile normalized data, *VSN* normalized data and the *sPLS* normalized data. Figure 5 presents the coefficients of variation of 50 randomly selected genes. We have the same observations from this figure as those for the probe level intensities – the data of gene expression indexes using *sPLS* normalization have uniformly and significantly smallest coefficients of variation, and the data using scale normalization have the largest ones.

4.3 Microarrays other than HG-U95

The *sPLS* normalization method can also apply to the microarrays with different number of probe pairs from 16, but needs to modify the cross-validation procedure in the method. For example, one HG-U133 plus 2.0 chip has 11 probe pairs in each probeset. To apply cross-validation for finding the number, k, of factors in *PLS* regression, we split the data into three groups, two of which form $(4 \times p)$ matrixs and the other forms a $(3 \times p)$ matrix (p is the number of genes). The other steps remain the same as those for HG-U95. We apply *sPLS* normalization to the HG-U133 plus 2.0 data produced by arrays u1332plus_ivt_breast_A, u1332plus_ivt_breast_B and u1332plus_ivt_breast_C. The MA plots (not presented) evidently show the success of the *sPLS* normalization to the HG-U133 plus 2.0 data set.

Acknowledgement

We thank the Editor, Professor Min-Te Chao, and an anonymous referee for their helpful comments.

References

Abdi, H. (2003). Partial least squares (PLS) regression. In *Encyclopedia of social sciences research methods* (Edited by Lewis-Beck, M., Bryman, A., Futing, T.). Sage.

Alm, A. *et al.* (2009). Partial least squares analysis in electrical brain activity. *Journal of Data Science*, **7**, 99-110.

Bolstad, B. M. *et al.* (2003). A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics* **19**, 185-193.

Chen, Z. *et al.* (2006). A distribution free summarization method for Affymetrix GeneChip® arrays. *Bioinformatics* **23**, 321-327.

Delaguardian, M. *et al.* (1996). Simultaneous kinetic spectrophotometric determination of 5 phenolic-compounds by reaction with p-aminophenol, using partial least-squares data treatment. *Analyst* **121**, 1321-1326.

Eriksson, L. *et al.* (1995). Multivariate-analysis of Aquatic toxicity data with PLS. *Aquatic Sciences* **57**, 217-241.

Garthwaite, P. H. (1994). An interpretation of partial least squares. *Journal of the American Statistical Association* **89**, 122-127.

Faller, D. *et al.* (2003). Normalization of DNA-microarray data by non-linear correction maximization. *Journal of Computational Biology* **10(5)**, 751-762.

Hartemink, A. *et al.* (2001). Maximum likelihood estimation of optimal scaling factors for expression array normalization. *Proceedings of SPIE* **4266**, 132-140.

Hoskuldsson, A. (1988). PLS regression methods. *Journal of Chemometrics* **2**, 211-228.

Huber W. *et al.* (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18**, Suppl. 1, S96-S104.

Irizarry, R. A. (2003). Exploration, normalization and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249-264.

Kays, S. E., Windham, W. R. and Barton, F. E. (1996). Prediction of total dietary fiber in cereal products using near-infrared reflectance spectroscopy. *Journal of Agricultural and food chemistry* **44**, 2266-2271.

Li, C. and Wong, W. H. (2001a). Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error applications. *Genome Biology* **2**, 1-11.

Li, C. and Wong, W. H. (2001b). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci USA 2001* **98**, 31-36.

Rosipal, R. and Krämer, N. (2006). Overview and recent advances in partial least squares. In *Subspace, Latent Structure and Feature Selection Techniques* (Edited by Saunders C., Grobelnik, M., Gunn, S., and Shawe-Taylor J.), 34-51. Springer.

Schadt, E. E. (2001). Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry Supplement* **37**, 120-125.

Wold, S. (1975). Soft modelling by latent variables; the nonlinear iterative partial least squares approach. In *Perspectives in Probability and Statistics: Papers in Honour of M. S. Bartlett* (Edited by J. Gani). Academic Press.

Workman, C. *et al.* (2002). A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biology* **3**: research0048.1-0048.16.

Received January 19, 2009; accepted May 6, 2009.

Zhide Fang
Biostatistics Program
School of Public Health
Louisiana State University Health Sciences Center – New Orleans
1615 Poydras Street, New Orleans, LA, 70115, USA
zfang@lsuhsc.edu

Xiaohu Li
Department of Mathematics
University of New Orleans
New Orleans, LA, 70148 USA
mathxhli@hotmail.com

Lizhe Xu
APHIS USDA PIADC
P. O. Box 848
Greenport, NY 11944-0848 USA
lizhe.xu@aphis.usda.gov