Standardizing gene expression data from high-density oligonucleotide microarrays in a matched pair study

Hanga C. Galfalvy, Steven P. Ellis, J. John Mann, Etienne Sibille
New York State Psychiatric Institute and Columbia University
Department of Neuroscience, 1051 Riverside Dr, New York, NY 10032
(hanga@neuron.cpmc.columbia.edu)

Key Words: gene microarray, principal curve, variance function estimation.

1. Introduction

Quality of data is an important issue in analyzing data from gene microarrays. High-level data analysis is unlikely to yield results if the data hasn't been properly pre-processed. A variety of data normalization procedures have been developed for the high-density oligonucleotide microarrays, starting from probe level data. In this paper, we address a very specific problem: that of normalizing data from a matched pair study. When paired experimental and control samples are normalized to each other, we take advantage of the expected similarity of the majority of the gene expressions (so-called "housekeeping genes") in closely matched subjects, and also of the fact that paired samples have been analyzed under similar conditions.

One important difference between analyzing paired high-density microarrays and two-color cDNA microarrays is that the two samples are hybridized on two different arrays, and so may be affected differently by experimental error. Another is that the "reference" sample is replaced by a different control chip for each experimental chip, which makes the use of regression techniques for smoothing the scatterplot hard to defend, since the control sample itself is variable and is measured with error. The normalization technique has to treat intensity values from "experimental" and "control" samples in a symmetrical way.

We do not use the probe level data, but start from Affymetrix's Micro Array Suite 5.0 signal, and "straighten" pairwise scatterplots of gene intensities before using robust centering and scaling on the whole set of arrays. The problem of heterogeneous scatter at low vs. high intensity values is addressed by the use of variance function estimation.

This work was supported by PHS grants MH62185, MH48514, MH60995 and MH63559, and by the Theodore and Vada Stanley Foundation

2. A Matched Pair Study

Microarray technology has been extensively used for cancer research, where differences in gene expression levels are relatively big. With the advances in the array technology and the relevant data analysis methods it is possible to apply the technique to illnesses where the expected differences are smaller and subtler. One such area is psychiatric illnesses like major depression.

Major depression is a debilitating mood disorder that can lead to suicide, This study focuses on comparing gene expression in the brain samples of suicide victims who have been diagnosed with major depression to samples of depression-free psychiatric controls. Previous research in neuroscience indicates that a region of the cortex named Brodmann area 47 is connected to cognitive functions that change in patients with major depression. Samples from Brodmann area 47 for 9 suicide victims and matched controls were collected and hybridized onto Affymetrix Genechips. The samples were matched onage, sex, race and post-mortem interval as closely as possible. Samples in each pair were analyzed close in time and under similar conditions to minimize the effect of experimental error on the intra-pair variability. Gene intensities for 12,625 transcripts per array were computed using Affymetrix's MAS 5.0 system.

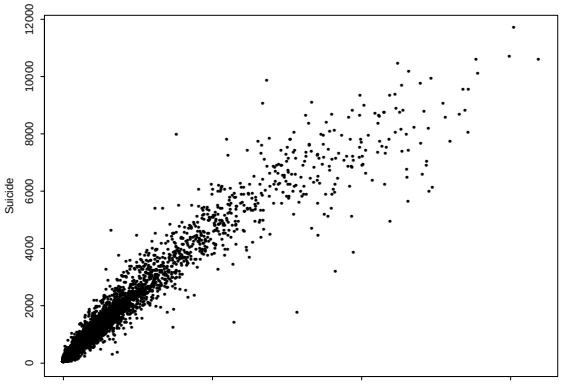

Figure 1. Gene intensities for a matched pair

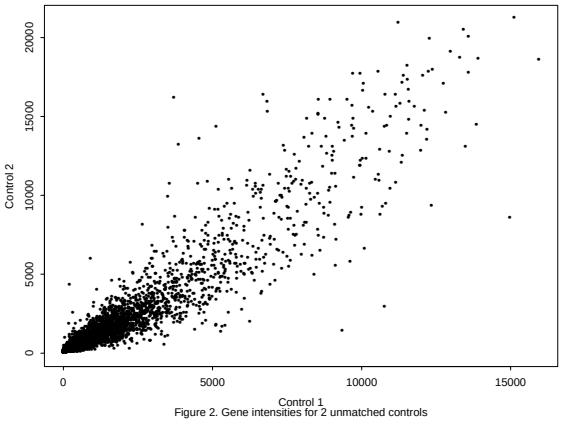

Figure 2. Gene intensities for 2 unmatched controls

Comparing the intensity scatterplots for matched pairs (on example is Figure 1) to those of unmatched controls (Figure 2) we see that there is less scatter in the matched pair plot. This may indicate that the matching protocol was successful and did indeed reduce the variation within the pair. However, the gene intensities from matched samples (or even unmatched ones, coming from humans) should be clustered around a diagonal line in the plot, because the majority of genes, called "housekeeping genes", should have similar expression levels in all human beings. In a matched pair, this similarity should be even stronger. The examples show clouds of points in a curve pattern rather than clustered in a line. This indicates that the arrays' intensity values are not properly "normalized". By centering and scaling all the arrays without taking into account the close relationship between matched pairs we would be loosing valuable information, the smart way is to exploit this relationship to "straighten" the scatterplots and scale afterward, or incorporate the multi-array scaling into the straightening protocol.

Another problem with the plots is the inhomogeneous scatter: lower intensity values seem to be less variable on an absolute scale that larger ones. Taking the log transform of the intensities is customary, but that does not solve the heterogeneous variance problem: on the log scale, the lower values will be more variable. Although this problem persists in all the pairs, dealing with it on pair-by-pair basis has technical advantages we will describe in the next sections.

3. Principal Curve Analysis

Existing software for microarray data analysis uses loess regression or smoothing splines to estimate the "central curve" in the data, than recomputes the intensity values so that they cluster around a diagonal line instead of the curve, preserving otherwise the original "pattern" of the data. This method is applied, for example, when normalizing the two-color cDNA microarrays intensities, where the "reference" sample intensities are plotted on the X axes, and the "experimental" sample intensities on the Y axis. When applying regression techniques to smooth the scatterplot, we automatically assume the intensities on the X axis fixed, i.e. that they are measured perfectly. In the case of our matched pair study, this is clearly an untenable assumption. Our data shows too much individual variation between controls to allow us to select a reference chip.

The alternative is to recognize that there is error in measurement for all the chips, and apply errors-in-variables regression methods instead of the least-squares based ones.

Principal curves are smooth, nonparametric curves that pass through the middle of a p-dimensional data set; they can be thought of as the generalization of the linear principal components. Here we will give only the definition of the principal curves, for details, see [2].

Let $\mathbf{X}$ be random vector in $\mathbf{R^P}$ with density h and finite second moments, $E(\mathbf{X}) = 0$. Let $\mathbf{f}$ be a smooth unit-speed curve in $\mathbf{R^P}$, parametrized by λ, that does not intersect itself and has finite length inside any finite ball.

Define the projection index as

$$\lambda_{\mathbf{f}}(\mathbf{x}) = \sup\{\lambda : ||\mathbf{x} - \mathbf{f}(\lambda)|| = \inf ||\mathbf{x} - \mathbf{f}(\mu)||\} \quad (1)$$

Then $\mathbf{f}$ is a principal curve of h if a.e. in λ

$$E(\mathbf{X}|\lambda_{\mathbf{f}}(\mathbf{X}) = \lambda) = \mathbf{f}(\lambda) \quad (2)$$

The definition essentially says that the average of all points that project onto a point on the curve is the point itself. The projection here is the usual projection in the p-dimensional Euclidean space, rather than the vertical projection from the LS regression.

The Splus function `principal.curve`, written by Trevor Hastie and downloadable from the Statlib library, computes principal curves with two options for "smoothing": loess or smoothing spline. Initial values for $\mathbf{f}_i$ are the values from the first principal component line. The projection index λ_i is the arc length of polygon from $\mathbf{f}_1$ to $\mathbf{f}_i$. The essence of the algorithm is to alternate projection and expectation steps until relative change in distance is below threshold:

$$|D^2(h, \mathbf{f}^{(j-1)}) - D^2(h, \mathbf{f}^{(j)})|/D^2(h, \mathbf{f}^{(j-1)}) < \epsilon \quad (3)$$

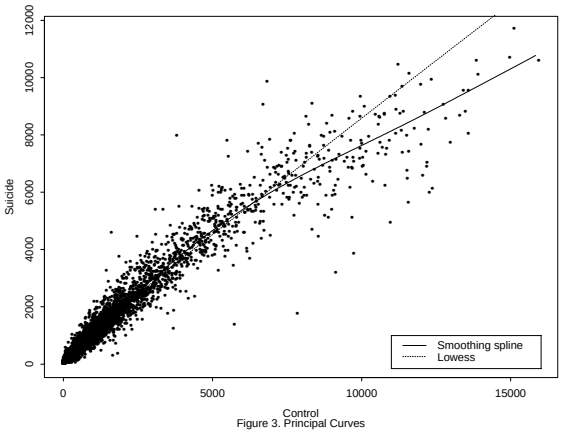

Figure 3. Principal Curves

For each point the algorithm returns the projection index and the distance from the principal curve. Then the "straightened" intensity values can be computed using the formula:

$$(X^{st}, Y^{st}) = (\lambda_i^* \mp d_i)/\sqrt{2} \quad (4)$$

The star over the projection index indicates the possibility of using a scaling step; since the length of the straightened curve will be longer than that of the diagonal line across our original data rectangle, the range for the straightened data will be larger, unless we shrink the λ_i by a scaling factor. This scaling factor can be computed to shrink the data back to its original range, but a more ambitious plan is to use it to scale all chips to the same range. Note that this is not equivalent to scaling all chips to the same variance.

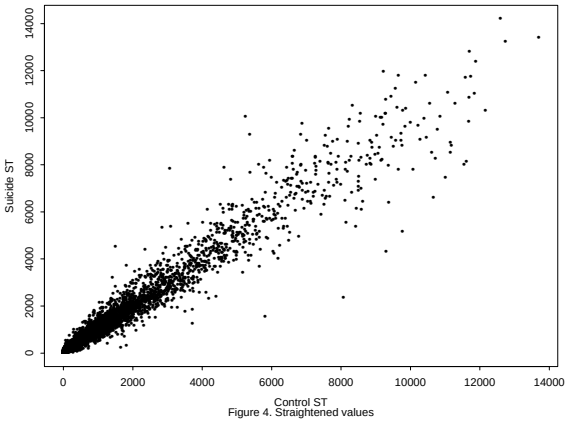

Figure 4. Straightened values

This normalization method can readily be extended from the matched pair case to straightening a whole set of p microarrays, by computing the first principal curve in p dimensions and straightening the p sets of intensity values accordingly. However, our estimate of the principal curve will depend on how similar the samples are to begin with.

4. Variance function estimation

When working with the MAS 5.0 signal (intensity) values, regardless of whether they are plotted on the original or the log scale, the variability or the amount of scatter will depend on the intensity. This poses problems when the goal of the analysis is to look for outliers: the definition of what constitutes an outlier will depend on the intensity value. In order to compute intensity-dependent confidence bounds for the data, the form of the dependence needs to be estimated. We will use the residuals from the straightening step to compute an initial estimate of the variance function, and will iterate the principal curve analysis with the variance function estimation until the resulting residuals will no longer depend on the intensity value. Typically, a few iterations will be sufficient.

In terms of the the distances d_i from the principal curve analysis, the heterogeneous scatter means that their variance depends on the position of the projection $lambda_i$. To scale them to the same variance, estimate a function σ that smoothes the scatterplot $(\lambda_i, |d_i|)$. To facilitate the automatic analysis of a large number of chips, we will not try to find a parametric relationship, instead, we'll use lowess regression to compute the nonparametric variance function estimate $\hat{\sigma}$ from the model

$$d_i = \sigma(\lambda_i) + e_i \text{ where } e_i \text{ are iid normal errors} \quad (5)$$

Alternatively, L1 regression or smoothing splines could be used to estimate the variance.

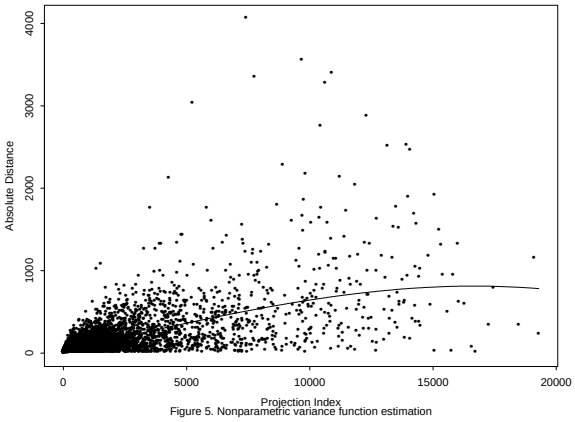

Figure 5. Nonparametric variance function estimation

Then the scaled distances will be computed as:

$$d_i^* = d_i/\hat{\sigma}(\lambda_i) \quad (6)$$

The scaled distances themselves can then be used as a diagnostic measure for identifying outliers. However, caution must be exercised, as the distribution of the d_i^* is usually not normal. In our study we

found the distribution to be well approximated by a t distribution with 4 degrees of freedom for a majority of the chips. Pointwise confidence limits can also be constructed using the quantiles of the t distribution. We emphasize, however, the exploratory nature of this analysis, since it does not take into account the correlation between genes.

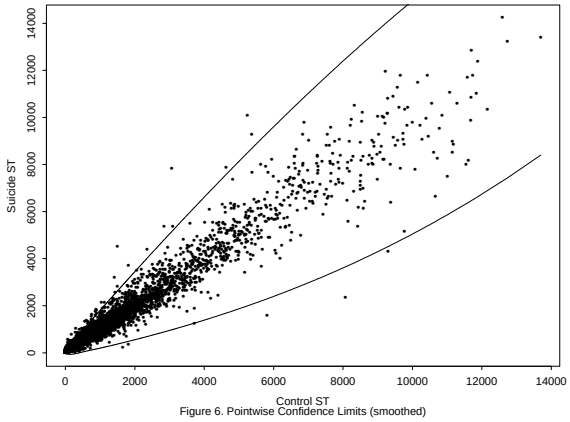

Figure 6. Pointwise Confidence Limits (smoothed)

5. Conclusions

When analyzing gene microarray data from matched pair studies, it makes statistical sense to take advantage of the similarity between the paired samples even in the normalization protocol. We used principal curve analysis to straighten the pairwise gene intensity scatterplots, and adapted variance function estimation techniques from the least squares case to address the problem of heterogeneous variance between genes.

It is important to note that probe-level analysis techniques that use probe data from a set of arrays to estimate the intensity values, like the dChip software described in [3], will return intensity values where the similarity between matched pairs is greatly reduced. It is not clear at this point how much of the difference is due to eliminated experimental error that was common to the two paired arrays. Also, the curved pattern of the pairwise scatterplots all but disappears in that case, making the use of the principal curve analysis redundant.

6. References

1. Davidian, M., and Carroll, R. J. (1987), *Variance function estimation*, Journal of the American Statistical Association, 82, pp. 1079-1091

2. Hastie, Trevor and Stuetzle, Werner (1989): *Principal Curves*. JASA, Vol 84, No. 406, pp 502-516.

3. Li, Cheng and Wong, Wing Hung (2001):*Model-based analysis of oligonucleotide microarrays: Expression index computation and outlier detection*. Proc. Natl. Acad. Sci. USA 98, no. 1, pp. 31-36.