Posterior Pareto Front Analysis for Gene Filtering

A. Hero[†], G. Fleury[◇]

[†]Depts. of EECS, BioMedical Eng., and Statistics, University of Michigan, Ann Arbor MI 49109, USA

[◇]Ecole Supérieure d'Electricité, Service des Mesures, 91192 Gif-sur-Yvette, France

Abstract

The massive scale and variability of microarray gene data creates new and challenging problems of signal extraction, gene clustering, and data mining, especially for temporal studies. Most data mining methods for finding interesting gene expression patterns are based on thresholding a single discriminant, e.g. a ratio of between-class to within-class variation or correlation to a template. We introduce a different approach for extracting information from gene microarrays which is based on a Bayesian formulation of multi-objective optimization which we call posterior Pareto front analysis. We will illustrate our methods by applying it to Fred Wright's GeneChip study.

I. Introduction

Microarray analysis of gene expression profiles offers one of the most promising avenues for exploring genetic factors underlying disease, regulatory pathways controlling cell function, organogenesis and development [16], [13], [15], [5]. The promise of microarrays is that the technology could allow researchers to accurately quantify expression in RNA levels of thousands of genes in a tissue sample, thereby providing valuable information about complex gene expression patterns. Recent advances in bioinformatics have brought us closer to realizing this promise. However, the massive scale and variability of microarray gene data creates new and challenging problems of clustering and data mining: the so-called *gene filtering problem*.

In [7], [8] we introduced a new approach to gene filtering, called Pareto gene filtering, which is based on multicriterion optimization and cross-validation. Pareto gene filtering allows the experimenter to isolate genes that achieve a good compromise between several competing gene-ranking criteria. Such genes lie on the so called *Pareto front* and are called non-dominated genes, see Sec. III for definitions. In this paper we present a Bayes posterior analysis approach to Pareto gene filtering which we call the method of *posterior Pareto fronts* (PPF). The main advantage of the PPF approach over the Pareto gene filtering approach is that it ranks each gene according to its posterior probability that it belongs to the Pareto front.

The outline of the paper is as follows. In Sec. II we briefly review and introduce our notation for microarray data and in III we recall elements of the Pareto gene filtering approach, in Sec. IV we introduce the general PPF gene filtering method and in Sec. V we consider specific contrast functions for PPF filtering. Finally in Sec. VII we apply PPF analysis to Fred Wright's Affymetrix mixing data set.

II. Gene Filtering in Microarrays

The ability to perform accurate genetic differentiation between two or more biological populations is a problem of great interest to geneticists and other researchers. For example, in a temporally sampled population of mice one is frequently interested in identifying genes that have interesting patterns of gene expression over time, called a gene expression profile. Gene microarrays, or chips, have revolutionized the field of experimental genetics by offering to the experimenter the ability to simultaneously measure thousands of gene sequences simultaneously. A gene chip consists of a large number N of known DNA probe sequences that are put in distinct locations, called wells, on a slide [11], [2], [6]. After hybridization of an unknown tissue sample to the gene chip, the abundance of each probe present in the sample can be estimated from the measured levels of hybridization (responses).

The study of differential gene expression between T populations requires hybridizing several (M) samples from each population to reduce response variability. Define the measured response at the n-th gene chip probe location for the m-th sample at time t

$$y_{tm}(n),\ n = 1, \ldots, N,\ m = 1, \ldots, M,\ t = 1, \ldots, T.$$

When several gene chip experiments are performed over time they can be combined in order to find genes with interesting expression profiles. This is a data mining problem for which many methods have been proposed including: multiple paired t-tests; linear discriminant analysis; self organizing (Kohonen) maps (SOM); principal components analysis (PCA); K-means clustering; hierarchical clustering (kdb trees, CART, gene shaving); and support vector machines (SVM) [10], [1], [3]. Validation methods have been widely used and include [19], [12]: significance analysis of microarrays (SAM); bootstrapping cluster analysis; and leave-one-out cross-validation. Most of these methods are based on filtering out profiles that maximize some criterion such as: the ratio of between-population-variation to within-population-variation; or the temporal correlation between a measured profile and a profile template. As contrasted to maximizing such *scalar* criteria, multi-objective gene filtering seeks to simultaneously maximize gene profiles [7]. This method is closely related to multi-objective optimization which has been used in for many applications [18], [20].

III. Multi-objective Gene Filtering

Multi-objective gene filtering can be motivated by the following simple example. Let there be $T = 2$ time points and define $\underline{\mu}(i) = [\mu_1(i), \mu_2(i)]^T$ the true unobserved expression levels of the i-th gene at each of these times. Let an experimenter have P gene selection criteria which, when applied to this gene response, gives the vector criterion:

$$\underline{\xi}(i) = [\xi_1(\underline{\mu}(i)), \ldots, \xi_P(\underline{\mu}(i))]^T.$$

Gene i is said to be better than gene j in the p-th criterion if $\xi_p(i) > \xi_p(j)$.

When it is desired to filter out strongly increasing gene profiles, one set of selection criteria might be $(P = 2)$:

$$\xi_1(\underline{\mu}) = \mu_2 - \mu_1, \xi_2(\underline{\mu}) = \mu_2 + \mu_1. \tag{1}$$

If $\mu_1$and μ_2 are positive valued and a proportional increase in the profile is more meaningful to the experimenter then she might prefer the criteria

$$\xi_1(\underline{\mu}) = \log \mu_2/\mu_1, \xi_2(\underline{\mu}) = \log \sqrt{\mu_2 \mu_1}. \tag{2}$$

If the measured profile of the i-th gene has vector mean $\underline{\mu} = \underline{\mu}(i)$ for which ξ_1 and ξ_2 are both large then this gene would be of interest to the experimenter. For filtering out such genes one might consider thresholding a compound scalar filtering criterion, e.g. the weighted arithmetic average of (1)

$$J_\alpha(\underline{\mu}) = \alpha(\mu_2 - \mu_1) + (1 - \alpha)(\mu_2 + \mu_1), \tag{3}$$

or of (2)

$$J_\alpha(\underline{\mu}) = \alpha \log(\mu_2/\mu_1) + (1 - \alpha) \log \sqrt{\mu_2 \mu_1}, \tag{4}$$

where $0 < \alpha < 1$. An obvious issue that arises in selecting such a scalar criteria is: what is the most suitable choice of the weight α? One way out of this dilemma is to filter out all genes which maximize J_α for some choice of α. It turns out that this set of genes are on the *first Pareto front* resulting from multiple-criterion optimization of the pair $[\xi_1(\underline{\mu}_i), \xi_2(\underline{\mu}_i)]^T$ [4].

Multi-criterion optimization captures the intrinsic compromises among possibly conflicting objectives. Consider Fig. 1 and suppose that ξ_1 and ξ_2 are to be maximized. It is obvious that genes A, B and C are "better" than genes D and E because both criteria are higher for the former than for the latter. Note that no gene among A, B and C dominates the other in both criteria ξ_1 and ξ_2. Multi-objective filtering uses this "non-dominated" property as a way to establish a preference relation among genes A, B, C, D and E. More formally, we say gene i is dominated if there exists some other gene $g \neq i$ such that for some $p = p_o$

$$\xi_p(i) < \xi_{p_o}(g) \text{ and } \xi_p(i) \leq \xi_p(g),\ p \neq p_o.$$

The set of non-dominated genes are defined as those genes that are not dominated. All the genes which are non-dominated constitute a curve which is called the Pareto front. A second Pareto front can obtained by stripping off points on the first front and computing the Pareto front of the remaining points - which for the example in Fig. 1 would be genes D and E.

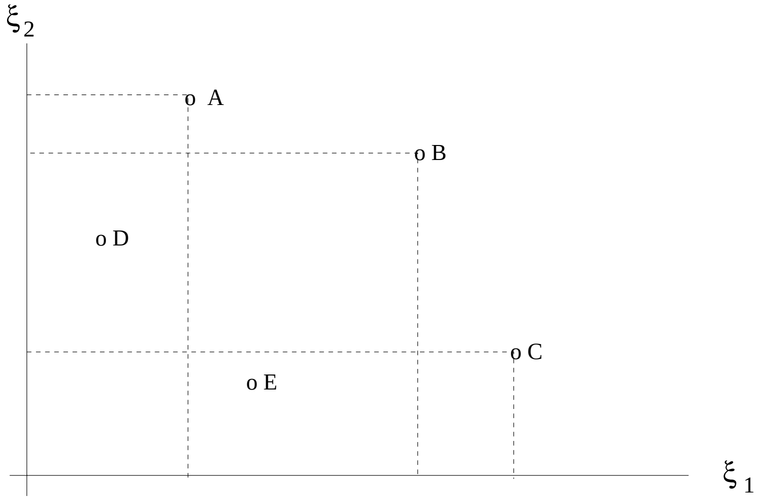

Fig. 1. *A, B, C are non-dominated genes relative to criteria* ξ_1 *and* ξ_2.

These methods are applicable when the criteria ξ_1 through ξ_P are observable. However, as these criteria depend on the true mean values $\underline{\mu}(i)$ of the i-th gene profile, the criteria are not observable. In [7], [8] we applied a non-parametric Pareto analysis for detecting interesting gene temporal profiles based on $\{y_{tm}(i)\}_{t,m,i}$, the measured abundances for each probe i, time point t and random sample m. First a set of T^M time trajectories were defined for each gene, corresponding to all possible time paths through the sets of M samples at each of T time points. For each trajectory we extracted the sign of the slope between each time point to capture instantaneous increase or decrease of each gene trajectory. The set of T^M sign profiles summarized the monotonic properties of a gene's temporal evolution pattern. For each gene several criteria were then computed including: the proportion of the T^M trajectories satisfying a specific evolution pattern, e.g. monotonicity of gene profile; the strength of the evolution pattern, e.g. the gene response difference between first and last time points; or the negative curvature of the profile. The Pareto fronts were cross-validated using simple leave-one-out resampling methods. The cross-validation was used for ranking the genes according to the number of resampling sets in which a specific gene appears on the first Pareto front.

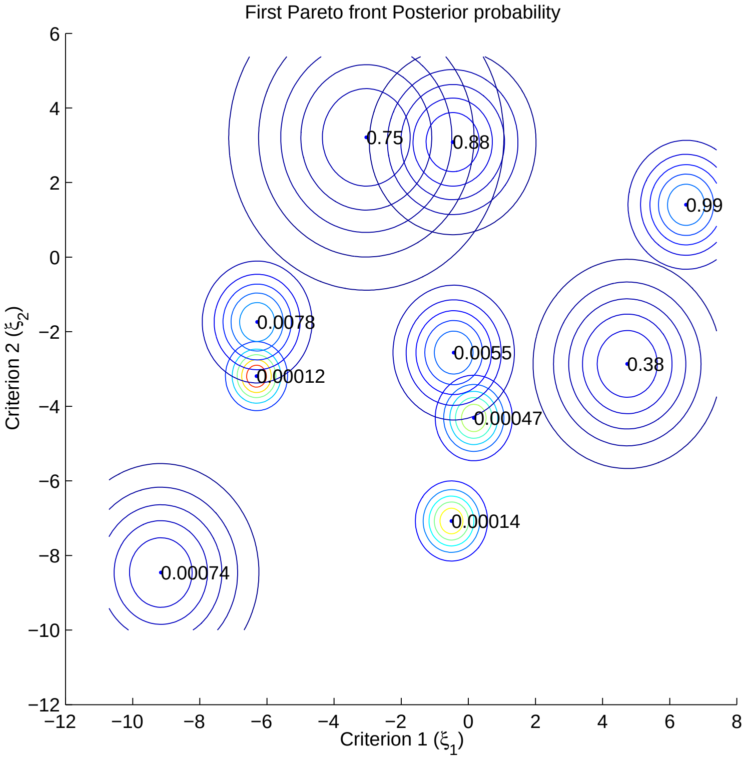

Fig. 2. *Graphical illustration of Posterior Pareto fronts analysis for a simulated population of genes having 3 time-point expression profiles with equal variances at each time.* ξ_1 *and* ξ_2 *correspond to slope and curvature of each profile. Spherical contours indicate inherent uncertainty in a given gene's placement on the* ξ_1, ξ_2 *plane (determined by pooled variance estimates of* ξ_1 *and* ξ_2 *for each gene). Number label to right of each gene location is posterior probability that given gene belongs to the first Pareto front.*

IV. Posterior Pareto Filtering

The posterior Pareto front analysis introduced here casts the ranking procedure of [7] in a Bayesian framework. Figures 2 and 3 illustrate the utility of our analysis.

The posterior probability $p(i|Y)$ that a particular gene i is on the first Pareto front is easily expressed using [17, Prop. 4.4]:

$$\begin{aligned}
& p(i|Y) && (5)\\
= \;& P(\xi_1(i) \geq \max_j \xi_1(j) \text{ or } \ldots \text{ or } \xi_P(i) \geq \max_j \xi_P(j)|Y) && (6)\\
= \;& \sum_{k=1}^{P} P(E_k(i)|Y) - \sum_{k_1<k_2} P(E_{k_1}(i), E_{k_2}(i)|Y) + \ldots && (7)\\
& +(-1)^{p+1} \sum_{k_1<\ldots<k_p} P(E_{k_1}(i), \ldots, E_{k_p}(i)|Y) && (8)\\
& +(-1)^{P+1} P(E_{k_1}(i), \ldots, E_{k_P}(i)|Y)
\end{aligned}$$

where the summation $\sum_{k_1<\ldots<k_p\leq p}$ is taken over the $\binom{P}{p}$ subsets of size p in $\{1, \ldots, P\}$, E_i denotes the event $\xi_1(i) \geq \max_j \xi_1(j)$ and $Y = \{y_{mt}(i)\}_{mti}$ is the entire observation extracted from the gene chip set.

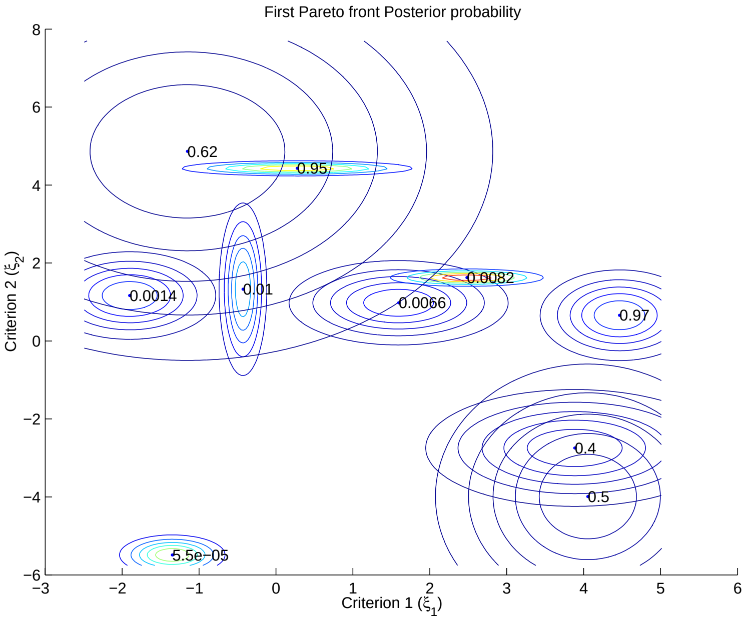

Fig. 3. *Same as in Fig. 2 except that the gene profile time samples have unequal variances.*

For $P = 2$ the expression (5) simplifies to:

$$\begin{aligned} & p(i|Y) \\ = \; & P(\xi_1(i) \geq \max_j \xi_1(j) \text{ or } \xi_2(i) \geq \max_j \xi_2(j)|Y) \\ = \; & P(\xi_1(i) \geq \max_j \xi_1(j)|Y) + P(\xi_2(i) \geq \max_j \xi_2(j)|Y) \\ & -P(\xi_1(i) \geq \max_j \xi_1(j), \xi_2(i) \geq \max_j \xi_2(j)|Y) \end{aligned}$$

In principle these probabilities can be computed when joint distributions of $\{\xi_1(i), \xi_2(j)\}_{i,j}$ are available.

In the special case that $\{\xi_p(i)\}_{ip}$ are conditionally independent given Y and that $\xi_p(i)$ has conditional (lebesgue) probability density function (p.d.f.) $f_{\xi_k(i)|Y}(u)$. Then in (5) $P(E_{k_1}(i), \ldots, E_{k_p}(i)|Y) = \prod_{j=1}^p P(E_{k_j}(i)|Y)$ and

$$P(E_k(i)|Y) = \int f_{\xi_k(i)|Y}(u) \prod_{j \neq k} F_{\xi_j(i)|Y}(u) du \tag{9}$$

where $F_{\xi_j(i)|Y}(u)$ is the conditional cumulative distribution function (c.d.f.) of $\xi_j(i)$.

A. Application to Filtering of Gene Expression Profiles

We start with an additive model for the (log) gene profile measurement:

$$y_{mt}(i) = \mu_t(i) + \epsilon_{mt}(i)$$

where $\epsilon_{mt}(i)$ are zero mean noise samples and $m = 1, \ldots, M$, $t = 1, \ldots, T$ and $i = 1, \ldots, N$. Given a prior $f(\mu_t(i), \sigma_t(i)^2)$ on the mean $\mu_t(i)$ and the variance $\sigma_t^2(i)$ of $y_{mt}(i)$ the posterior probabilities (5) can be computed. In the sequel we adopt the non-informative prior [9]

$$f_{\mu_t(i),\sigma_t^2(i)}(u,s) \;=\; \frac{c}{s^{a/2}},\;\; u\in\mathbb{R},\; s\in\mathbb{R}^+$$

where c is a positive normalizing constant and $a>0$.

Two special cases are of interest to us: (i) time varying variances $\{\sigma_t^2(i)\}_t$; and (ii) non-time varying variances $\sigma_t^2(i)=\sigma_\tau^2(i)$, $t,\tau=1,\ldots,T$. The former case is easier to treat than the latter.

A.1 Time varying variances

Consider the following model for $\mu_t(i)$ and $\epsilon_{mt}(i)$: (i) $\{\mu_t(i)\}_{ti}$ and $\{\sigma_t^2(i)\}_{ti}$ are independent sets of i.i.d. random variables; (ii) given these random variables $Y=\{y_{tm}(i)\}_{ti}$ are independent jointly Gaussian random variables with respective means $\{\mu_t(i)\}_{ti}$ and variances $\{\sigma_t^2(i)\}_{ti}$; (iii) $\{y_{tm}(i)\}_m$ are conditionally i.i.d.

It is easily shown that under the above assumptions the means $\{\mu_t(i)\}_{ti}$ are conditionally independent given Y with marginal posterior p.d.f. equal to the Student-t density

$$f_{\mu_t(i)|Y}(u) \;=\; k(Y_{ti})\left(1+\frac{1}{M+1}\frac{(u-\hat{\mu}_t(i))^2}{\hat{\sigma}_t^2(i)}\right)^{-(M-a+2)/2}, \tag{10}$$

where $\hat{\mu}_t(i) = M^{-1}\sum_m y_{tm}(i)$, $\hat{\sigma}_t^2(i) = M^{-1}\sum_m(y_{tm}(i)-\hat{\mu}_t(i))^2$, $Y_{ti}=\{y_{tm}(i)\}_m$, and $k(Y_{ti})$ is the measurement-dependent normalizing factor [9]:

$$k(Y_{ti}) = \frac{1}{\hat{\sigma}_t(i)\sqrt{\pi}}\frac{\Gamma(\frac{1}{2}(M-a+2))}{\Gamma(\frac{1}{2}(M-a+1))}. \tag{11}$$

The associated c.d.f. can be approximated using either the large M Gaussian approximation to the student-t or the L_∞ approximation $\left(\int_{-\infty}^{u} g^q(v)dv\right)^{1/q} \approx \sup_{v\le u} g(v)$, where $q>0$. The latter approximation improves as q gets large. The L_∞ approach has computational advantages as it yields a closed form expression - as contrasted with the Gaussian approximation that gives an expression involving integrals of the Gaussian density. Applying the L_∞ approximation to the integral of (10) yields

$$F_{\mu_t(i)|Y}(u) \;\approx\; \left(1+\frac{(\hat{\mu}_t(i)-u)_+^2}{\hat{\sigma}_t^2(i)}\right)^{-(M-a+2)/2}.$$

where $(x)_+$ equals x when $x>0$ and equals zero otherwise.

A.2 Non-time varying variances

Next consider the following model: (i) $\sigma_t^2(i)=\sigma^2(i)$; (ii) $\{\mu_t(i)\}_{ti}$ and $\{\sigma^2(i)\}_i$ are independent sets of i.i.d. random variables; (ii) given these random variables $Y=\{y_{tm}(i)\}_{ti}$ are independent jointly Gaussian random variables with respective means $\{\mu_t(i)\}_{ti}$ and variances $\{\sigma_t^2(i)\}_{ti}$; (iii) $\{y_{tm}(i)\}_m$ are conditionally i.i.d.

Due to (i) the mean profile $\{\mu_t(i)\}_t$ is no longer a conditionally independent sequence given Y. The joint posterior p.d.f. of $\underline{\mu}(i)=[\mu_1(i),\ldots,\mu_T(i)]^T$ takes the form of a multivariate Student-t

$$f_{\underline{\mu}(i)|Y}(u_1,\ldots,u_T) \;=\; k(Y_i)\left(1+\sum_{t=1}^{T}\frac{(u_t-\hat{\mu}_t(i))^2}{\hat{\sigma}^2(i)}\right)^{-(TM-a+2)/2},$$

where $\hat{\sigma}^2(i) = T^{-1}M^{-1}\sum_t \sum_m (y_{tm}(i) - \hat{\mu}_t(i))^2$, $Y_i = \{y_{tm}(i)\}_{tm}$, and $k(Y_i)$ is a similar scale factor to (11).

Analogously to the case of unequal variances, we can approximate the associated c.d.f. by a multivariate L_∞ approximation to (12):

$$F_{\underline{\mu}(i)|Y}(u_1, \ldots, u_T) \approx \left(1 + \sum_t \frac{(\hat{\mu}_t(i) - u_t)_+^2}{\hat{\sigma}^2(i)}\right)^{-(TM-a+2)/2}. \qquad (12)$$

V. Application to Profile Contrasts

A. Profile Amplitude Criterion

For ease of discussion we first adopt the time sampled means themselves $\xi_p(i) = \mu_p(i)$, $p = 1, \ldots T$, as the criteria of interest. We call this the profile amplitude criterion. We treat the case of time varying variances for concreteness. We generalize this to a set of contrast functions applied to the means in the next subsection. Using the expressions (10) and (12) in (9) gives an expression for $P(E_k(i)|Y)$ which only requires numerical evaluation of T one-dimensional integrals (as compared with T-dimensional integrals if we used the exact non-asymptotic c.d.f.).

By using the simple exponential lower bound $e^{-u^2} \leq 1/(1+u^2)$, it is possible to obtain a lower bound on $P(E_p(i)|Y)$, which, as $p(i|Y)$ is monotonic increasing in $P(E_p(i)|Y)$. $k = 1, \ldots, P$, yields lower bounds on the posterior Pareto front probabilities $\{p(i|Y)\}_i$. Specifically, for fixed i and t, let $\hat{\mu}_t(r_1) \leq \ldots \leq \hat{\mu}_t(r_{N-1})$ denote the rank ordered sample means from the set $\{\hat{\mu}_t(n)\}_{n \neq i}$ and define $\hat{\mu}_t(r_N) = \infty$. Then

$$\begin{aligned} P(E_t(i)|Y) \quad \geq \quad & \sum_{n=1}^{N-1} \frac{1}{\hat{\sigma}_t^2(i)\sqrt{\beta_{tn}}} \Big[\Phi([\hat{\mu}_t(r_{n+1}) - \gamma_{tn}]\sqrt{q\beta_{tn}}) \\ & -\Phi([\hat{\mu}_t(r_n) - \gamma_{tn}]\sqrt{q\beta_{tn}})\Big] \exp\left(-\frac{q}{2}(\alpha_{tn} - \gamma_{tn}^2/\beta_{tn})\right) \end{aligned}$$

where $q = m - a + 2$ and

$$\begin{aligned} \beta_{tn} &= \frac{1}{\hat{\sigma}_t^2(i)} + \sum_{k=1}^{n} \frac{1}{\hat{\sigma}_t^2(r_k)} \\ \gamma_{tn} &= \left(\frac{\hat{\mu}_t(i)}{\hat{\sigma}_t^2(i)} + \sum_{k=1}^{n} \frac{\hat{\mu}_t(r_k)}{\hat{\sigma}_t^2(r_k)}\right) / \beta_{tn} \\ \alpha_{tn} &= \frac{\hat{\mu}_t^2(i)}{\hat{\sigma}_t^2(i)} + \sum_{k=1}^{n} \frac{\hat{\mu}_t^2(r_k)}{\hat{\sigma}_t^2(r_k)}. \end{aligned}$$

In the above $\Phi(u) = (\sqrt{2\pi})^{-1}\int_{-\infty}^{u} e^{-u^2/2}du$.

For the case of equal variances, neither the joint p.d.f. (12) nor the joint c.d.f. (12) are separable functions and this complicates computation of $P(E_k(i)|Y)$ due to the need for T-dimensional integration. However, as above a multivariate Gaussian approximation can be applied to the c.d.f. and p.d.f. yielding a lower bound on $P(E_t(i)$ and hence on $p(i|Y)$.

B. Profile Constrast Criteria

Let the vector criterion $\underline{\xi}(i) = [\xi_1(i), \ldots, \xi_P(i)]^T$ be defined as a linear function of the mean profile vector:

$$\underline{\xi}(i) = A\underline{\mu}(i),$$

where $A = ((a_{ij}))$ is a $P \times T$ *contrast matrix.* We call $\underline{\xi}(i)$ a vector of *profile contrasts* for gene i. To retain the simplicity of our approximations to $p(i|Y)$, it is necessary that the component criteria in $\underline{\xi}(i)$ be statistically independent when conditioned on Y. At a minimum this requires $P \leq T$. Assume as above that the components of $\underline{\mu}$ are conditionally independent. A sufficient condition for independent ξ_p's is that non-zero elements of each of the rows of A do not overlap each other, i.e. $a_{ik}a_{jk} = 0$, for all $i \neq j$ and all k. When the variances are not time varying a weaker sufficient condition is that A be an orthogonal matrix, $AA^T = I$ since the joint p.d.f. $f_{\underline{\mu}(i)|Y}(\underline{u})$ in (12) is invariant to orthogonal transformations of $\underline{u} - \underline{\hat{\mu}}(i)$. Furthermore, as the Pareto fronts are invariant to monotonic increasing transformations of the ξ_p's, an even weaker sufficient condition is $AA^T = \text{diag}(a_{ii})$= a diagonal matrix. We illustrate this latter case below.

A Sampling of Profile Contrasts:

We specialize to the case of non-time-varying variances and $T = 2$, $T = 3$ and $T = 4$ for concreteness. Consider the corresponding candidate $T \times T$ contrast matrices

$$
\begin{aligned}
A_2 &= \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}, \\
A_2^{'} &= \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}, \\
A_3 &= \begin{bmatrix} -1 & 0 & 1 \\ 1 & -2 & 1 \\ 1 & 1 & 1 \end{bmatrix}, \\
A_3^{'} &= \begin{bmatrix} -1 & 1 & 0 \\ -1 & -1 & 2 \\ 1 & 1 & 1 \end{bmatrix}, \\
A_4 &= \begin{bmatrix} -1 & 1 & 0 & 0 \\ -1 & -1 & 2 & 0 \\ -1 & -1 & -1 & 3 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \\
A_4^{'} &= \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & -1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}.
\end{aligned}
$$

As all of these matrices satisfy AA^T = diagonal, we can apply the posterior Pareto analysis to any subset of ξ_p's in the vector $\underline{\xi} = A\underline{\mu}$ depending on the problem at hand. Applying the posterior Pareto front analysis to $\underline{\xi}(i) = A_2\underline{\mu}(i)$ will extract 2 time-point gene profiles which are monotonic increasing (large ξ_1) and/or have strong average expression levels (large ξ_2). When applied to $\underline{\xi}(i) = A_2^{'}\underline{\mu}(i)$ the analysis will extract strong monotonic decreasing genes from the 2 time-point profiles. Applying the posterior Pareto front analysis to $\underline{\xi}(i) = A_3\underline{\mu}(i)$ will extract strong 3 time-point gene profiles which are end-to-end increasing and have large positive curvature (large ξ_2). If A_3 is replaced with $A_3^{'}$ then the analysis will find strong profiles which are monotonic increasing. Using only the first two rows of $A_3^{'}$ will extract both string and weak monotonic increasing profiles. If the p.d.f. of $\xi_2(i)$ is truncated to zero over the range For 4 time-points A_4 will perform similar services as A_3 while $A_4^{'}$ will filter out "mexican hat" profiles.

Note that independence of these linear contrasts is preserved under non-linear transformations since the constrasts are conditionally Gaussian given $\underline{\mu}, \sigma^2$. The contrasts can also be constrained to satisfy positivity, lie in a interval, etc. Figures 4-7 illustrate the application of these contrasts to PPF extraction of monotonic increasing trajectories in 3 time-point and 4 time-point profiles.

Figure 4 shows a simulated 3 time-point data set with a pair of criteria corresponding to the first two elements of $A_3'\underline{\mu}$ - these are labeled "Contrast 1" and "Contrast 2," respectively. The PPF posterior probabilities are computed over the sector for which both criteria are strictly positive. The corresponding PPF ranked profiles are shown in the panel display in Fig. 5. One can extract the second Pareto front by rerunning the PPF analysis after removing the two genes having PPF probability greater than a threshold, e.g. 0.9. Figure 6 illustrates the PPF analysis for a simulated 4 time-point data set with three criteria selected as the first three elements of $A_4\underline{\mu}$. The PPF analysis is again performed on sectorized (positive) data and the ranked profiles are shown in Fig. 7.

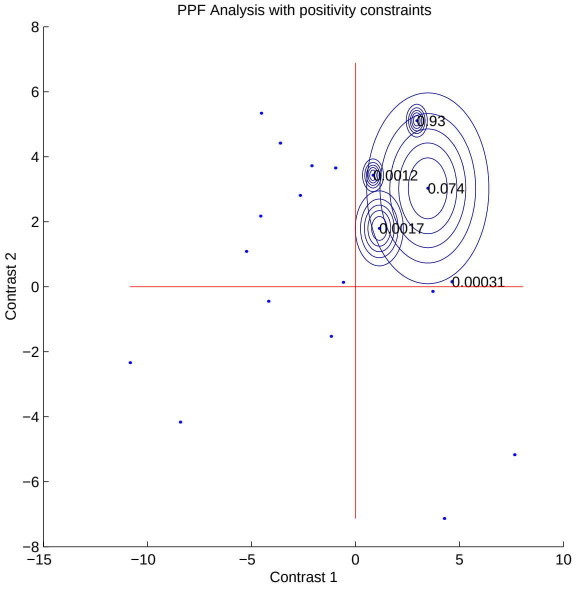

Fig. 4. *Sectorized PPF analysis with positivity constrained slope criteria which are given by the contrast functions extracted from the first two rows of A_3' (Contrast 1 and Contrast 2) applied to simulated 3 time-point gene profiles ($a = 2$). Constant contours around each point indicate standard errors.*

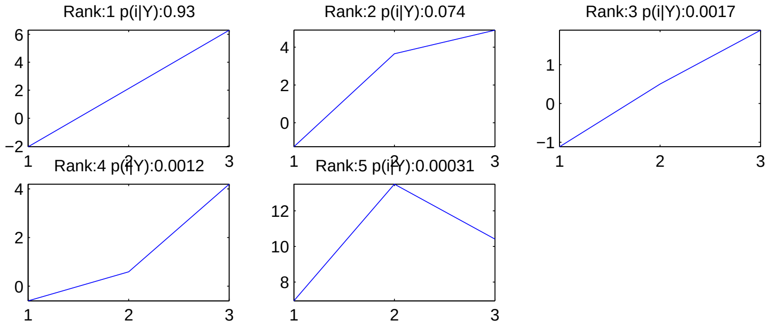

Fig. 5. *Ranked PPF profiles corresponding to the points in Fig. 4.*

VI. Extensions

There are several issues and extensions that should be explored. Some of these are:

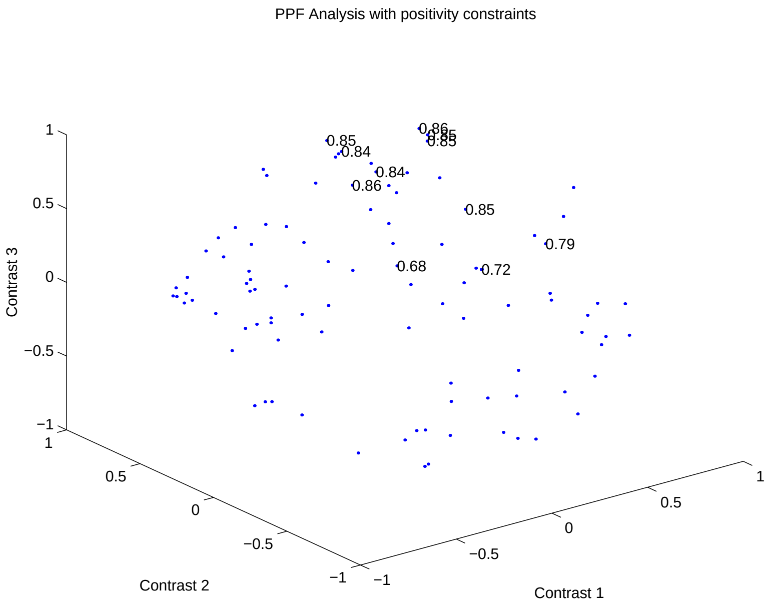

Fig. 6. *Sectorized PPF analysis with positivity constrained slope criteria which are given by the contrast functions extracted from the first three rows of* A_4 *applied simulated 4 time-point gene profiles.*

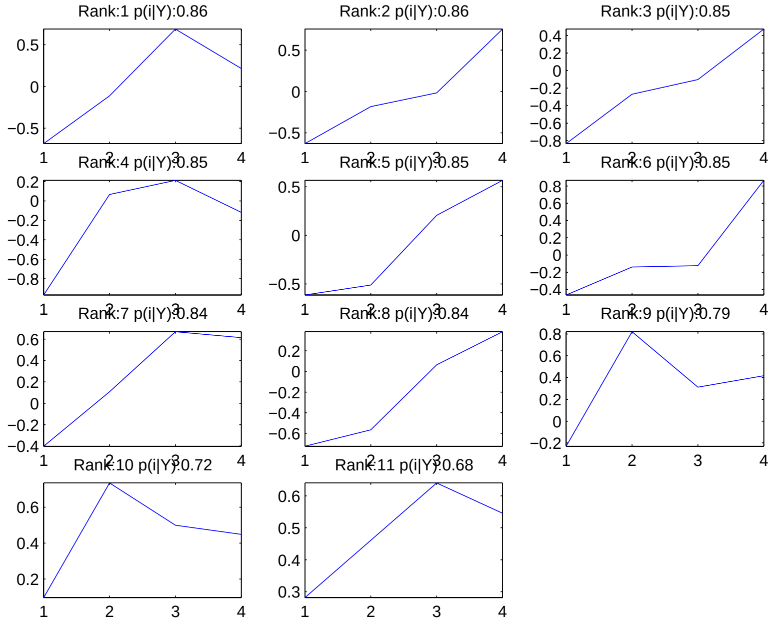

Fig. 7. *Ranked PPF profiles corresponding to points in Fig. 6.*

1. To come up with a magnitude independent sectorization, e.g. max and min curvature/end-end-slope. This would of course be data dependent and its effect on the statistical analysis needs to be considered.
2. Implement approximations to the posterior probabilities, e.g. upper and lower bounds, and compare to the exact expressions used here.
3. Implement the equal variance approximation using Gaussian or other approximations. The results here were obtained using a pooled variance estimate in the time varying variance pdf expression.
4. Compare this analysis to the non-parametric cross-validation Pareto method in [7].
5. Explore methodology, even approximate, for computing posterior probability of a gene being in the first K fronts.
6. Explore more general and systematic ways to come up with meaningful contrast matrices A which are unitary, so as to maintain independence, yet capture desired shape characteristics of teomporal exprssion profiles. A method, which we have not explored in depth, is to define a contrast matrix B whose rows capture some set of desired linearly independent properties of the profile and then apply the Pareto analysis with orthogonalized contrast matrix $A = [\text{chol}(BB^T)]^{-1}B$, where $\text{chol}(B * B^T)$ is the Cholesky decomposition of BB^T. For example the following matrix might be proposed as an alternative to $A_3^{'}$ in the previous section for capturing strong monotone increasing profiles given by

$$B = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

It turns out that the aforementioned orthogonalization procedure yields

$$A = \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} & 0 \\ -1/\sqrt{6} & -1/\sqrt{6} & 2/\sqrt{6} \\ 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \end{bmatrix},$$

which is equal (up to a left multiplication by a positive diagonal matrix) to the contrast matrix $A_3^{'}$.

VII. Experimental Results

We applied PPF analysis to Fred Wright's dataset described in the paper [14]. This data set is a mixing experiment which has been designed for empirically validating and comparing various differential gene expression methods of analysis. Three populations of genes were hybridized to Affymetrix HuGeneFL chips: starved human fibroblast cells; stimulated human fibroblast cells; and a 50-50 mixture of these cells. A total of 18 chips were processed corresponding to 6 replications within each of the three populations mentioned above. Each chip contains the same 7129 gene probes selected by Affymerix for the HuGeneFL chip. For each gene probe we arbitrarily defined the sequence of hybridization abundances from the "stimulated(t=1)," "50-50(t=2)," and "starved(t=3)," populations, in that order, as a gene expression profile. This provides a very nice test dataset for us since we know that the true profiles must be linearly increasing or decreasing over the three "time points." In Figs. 8 and 9 the 7129 mean contrasts are shown for the avgdiff and the Li-Wong reduced indices. These indices are extracted from the affymetrix .cel files and measure the differential expression levels between PM and MM oligonucleotides on the Gene Chip. See [14] for more details. Each point on this contrast plane is a vector containing the first two elements of vector $A_3^{'}\hat{\mu}(i)$ where $\hat{\mu}(i)$ is sample mean of over the six replicates in each group for a given gene. If the data were noiseless then all the contrast points would fall in the upper right and lower left sectors corresponding to monotonic increasing and monotonic decreasing gene expression profiles, respectively. One measure of the quality of the experiment is the proportion of genes falling outside of these two sectors, i.e. genes having non-monotonic profiles. The Li-Wong reduced indices are better in this quality measure.

Throughout this section we used the exponent $a = 2$ in the prior density input for the PPF analysis. We first applied the PPF analysis to the non-monotone convex cap profiles. For this we adopted the contrast

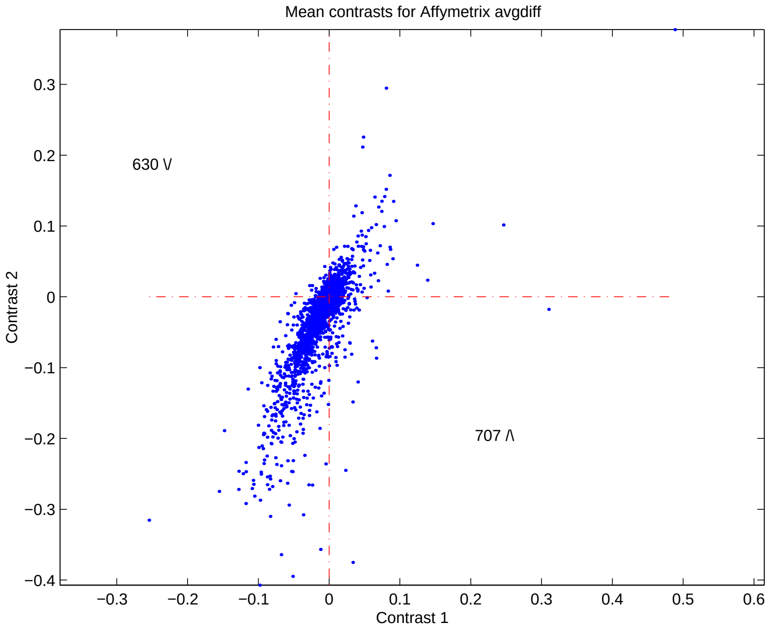

Fig. 8. *Scatterplot of slope contrasts (Sample mean contrasts defined from the first two rows of A_3') for avgdiff indices for Fred Wright's HuGeneFL mixture study. Annotations are the number of non-monotone genes with convex cup (upper left) and convex cap (lower right) profiles.*

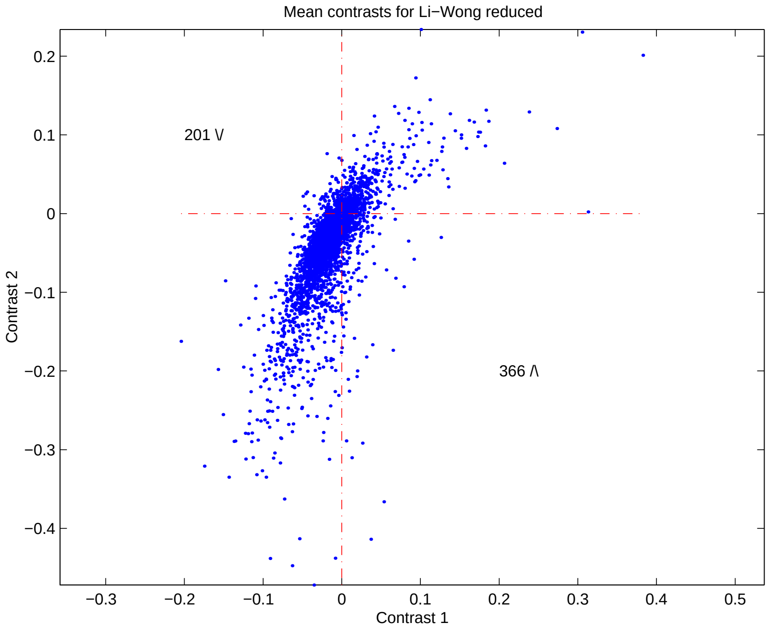

Fig. 9. *Scatterplot of slope contrasts (Sample mean contrasts defined from the first two rows of A_3') for Li-Wong reduced indices for Fred Wright's HuGeneFL mixture study. Annotations are the number of non-monotone genes with convex cup (upper left) and convex cap (lower right) profiles.*

matrix

$$A = \begin{bmatrix} -1 & 1 & 0 \\ 1 & 1 & -2 \end{bmatrix}.$$

The results are shown in Figs. 11-14. In the Figs. 11 and 12 are the PPF planes over a small sector (indicated) containing only six gene profiles along with their posterior probabilities for the Affymetrix avgdiff index and the Li-Wong index, respectively. The contour around each point denotes the standard error (one standard deviation) circle and the annotation at the center is the posterior probablity $P(i|Y)$. While the posterior scores for the six genes are different in each figure the genes are the same - avgdiff and Li-Wong contain identical genes in this sector of the contrast plane. Figs. 13 and 14 show eight top scoring trajectories among the top fifty trajectories ranked by the contrast matrix A and and PPF analysis. In each subpanel the piecewise linear line passes through the means of the 6 replicates for each of the 3 groups. Note that that the ranking of genes in avgdiff and Li-Wong indices is different but many commonalities exist. Many, but not all, genes are ranked strongly non-monotonic in both avgdiff and Li-Wong indices. The complete list is given in Fig. 10.

AFFX−BioDn−3−at	AFFX−BioDn−3−at
AFFX−CreX−3−at	AFFX−CreX−3−at
AFFX−HSAC07/X00351−5−st	AFFX−HSAC07/X00351−5−at
AFFX−HUMGAPDH/M33197−3−at	AFFX−HUMRGE/M10098−3−at
AFFX−HUMRGE/M10098−3−at	AFFX−HUMRGE/M10098−5−at
AFFX−HUMRGE/M10098−5−at	AFFX−LysX−3−at
AFFX−LysX−3−at	AFFX−LysX−M−at
AFFX−LysX−5−at	AFFX−PheX−3−at
AFFX−M27830−5−at	AFFX−PheX−5−at
AFFX−PheX−5−at	D49728−at
AFFX−ThrX−M−at	D49824−s−at
D49824−s−at	D76435−at
D83174−s−at	D86976−at
HG1980−HT2023−at	HG1800−HT1823−at
HG3044−HT3742−s−at	HG1980−HT2023−at
J00073−at	HG3044−HT3742−s−at
J04823−rna1−at	HG831−HT831−at
L06505−at	J00073−at
L21954−at	J03756−at
L24559−at	J04823−rna1−at
L37368−at	L06505−at
L77701−at	M14328−s−at
M14328−s−at	M19267−s−at
M21142−cds2−s−at	M19311−s−at
M24485−s−at	M26708−s−at
M35878−at	M35878−at
M55998−s−at	M55998−s−at
M60752−at	M60752−at
M80563−at	M88461−s−at
M81181−s−at	M95712−at
U03057−at	S54005−s−at
U12404−at	U03057−at
U12465−at	U12404−at
U14394−at	U12465−at
U27325−s−at	U14394−at
U45285−at	U25034−s−at
U51004−at	U27325−s−at
U52101−at	U52101−at
U58516−at	U70063−at
U90915−at	U73379−at
X03689−s−at	U78027−rna3−at
X13973−at	V00594−at
X16064−at	X02152−at
X67247−rna1−at	X16064−at
X86809−at	X67247−rna1−at
X95404−at	X90780−rna1−at
Y09912−rna1−at	X95404−at
Z21507−at	Z23090−at
Z24727−at	Z50022−at
Z69043−s−at	Z69043−s−at

Fig. 10. *The 50 top scoring genes (Affymetrix nomenclature) resulting from PPF analysis of the most non-monotone convex cap profiles for Fred Wright's data using Affymatrix avgdiff (left) and Li-Wong reduced (right) indices.*

We also applied PPF analysis to the lower left (monotone decreasing profiles) sector of the contrast plane in Fig. 8. The six top ranked gene profiles for each of avgdiff and Li-Wong reduced indices are shown in

Figs. 15 and 16.

VIII. Conclusion

This paper introduced a new method of Pareto gene filtering based on posterior analysis of the Pareto fronts of the multi-objective vector. This offers an alternative to non-parametric cross-validation approaches to Pareto filtering introduced by us in earlier work. The method is very flexible and involves choosing a set of appropriate profile contrasts which display desired characteristics of the expression profiles. These techniques also have applicability to general data mining problems. An issue that must be addressed is reduction in computational complexity which will be necessary for these, and other, validation techniques to be peformed in "real time."

References

[1] A. A. Allzadeh and etal, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, pp. 503–511, 2000.

[2] D. Bassett, M. Eisen, and M. Boguski, "Gene expression informatics–it's all in your mine," *Nature Genetics*, vol. 21, no. 1 Suppl, pp. 51–55, Jan 1999.

[3] M. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. Sugent, T. Furey, M. Ares, and D. Haussler, "Knowledge-based analysis of microarray gene expression data by using support vector machines," *Proc. of Nat. Academy of Sci. (PNAS)*, vol. 97, no. 1, pp. 262–267, 2000.

[4] I. Das and J. Dennis, "A closer look at drawbacks of minimizing weighted sums of objectives for pareto set generation in multicriteria optimization problems," *Structural optimization*, vol. 14, no. 1, , 1997.

[5] J. DeRisi, V. Iyer, and P. Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science*, vol. 278, no. 5338, pp. 680–686, Oct 24 1997.

[6] P. Fitch and B. Sokhansanj, "Genomic engineering: moving beyond DNA sequence to function," *IEEE Proceedings*, vol. 88, no. 12, pp. 1949–1971, Dec 2000.

[7] G. Fleury, A. O. Hero, S. Yoshida, T. Carter, C. Barlow, and A. Swaroop, "Clustering gene expression signals from retinal microarray data," in *Proc. IEEE Int. Conf. Acoust., Speech, and Sig. Proc.*, Orlando, FL, 2002.

[8] G. Fleury, A. O. Hero, S. Yoshida, T. Carter, C. Barlow, and A. Swaroop, "Pareto analysis for gene filtering in microarray experiments," in *European Sig. Proc. Conf. (EUSIPCO)*, Toulouse, FRANCE, 2002.

[9] S. Geisser and J. Cornfield, "Posterior distributions for mutlivariate normal parameters," *J. Royal Statistical Society, Ser. B*, pp. 368–376, 1963.

[10] T. Hastie, R. Tibshirani, M. Eisen, P. Brown, D. Ross, U. Scherf, J. Weinstein, A. Alizadeh, L. Staudt, and D. Botstein, "Gene shaving: a new class of clustering methods for expression arrays," Technical report, Stanford University, 2000.

[11] K. Kadota, R. Miki, H. Bono, K. Shimizu, Y. Okazaki, and Y. Hayashizaki, "Preprocessing implementation for microarray (prim): an efficient method for processing cdna microarray data," *Physiol Genomics*, vol. 4, no. 3, pp. 183–188, Jan 19 2001.

[12] K. Kerr and G. Churchill, "Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments," *Proc. of Nat. Academy of Sci. (PNAS)*, vol. 98, pp. 8961–8965. `citeseer.nj.nec.com/414709.html`.

[13] C. Lee, R. Klopp, R. Weindruch, and T. Prolla, "Gene expression profile of aging and its retardation by caloric restriction," *Science*, vol. 285, no. 5432, pp. 1390–1393, Aug 27 1999.

[14] W. Lemon, J. J. Palatini, R. Krahe, and F. Wright, "Theoretical and experimental comparison of gene expression estimators for oligonucleotide arrays," *Bioinformatics*, To appear 2002. `http://thinker.med.ohio-state.edu/projects/fbss/index.html`.

[15] F. Livesey, T. Furukawa, M. Steffen, G. Church, and C. Cepko, "Microarray analysis of the transcriptional network controlled by the photoreceptor homeobox gene," *Crx. Curr Biol*, vol. 6, no. 10, pp. 301–10, Mar 23 2000.

[16] D. Lockhart, H. Dong, M. Byrne, M. Follettie, M. Gallo, M. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. Brown, "Expression monitoring by hybridization to high-density oligonucleotide arrays," *Nat. Biotechnol.*, vol. 14, no. 13, pp. 1675–80, 1996.

[17] S. Ross, *A first course on probability*, Macmillan, 1988.

[18] R. E. Steuer, *Multi criteria optimization: theory, computation, and application*, Wiley, New York N.Y., 1986.

[19] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proc. of Nat. Academy of Sci. (PNAS)*, vol. 98, pp. 5116–5121, 2001.

[20] E. Zitler and L. Thiele, "An evolutionary algorithm for multiobjective optimization: the strength Pareto approach," Technical report, Swiss Federal Institute of Technology (ETH), May 1998.

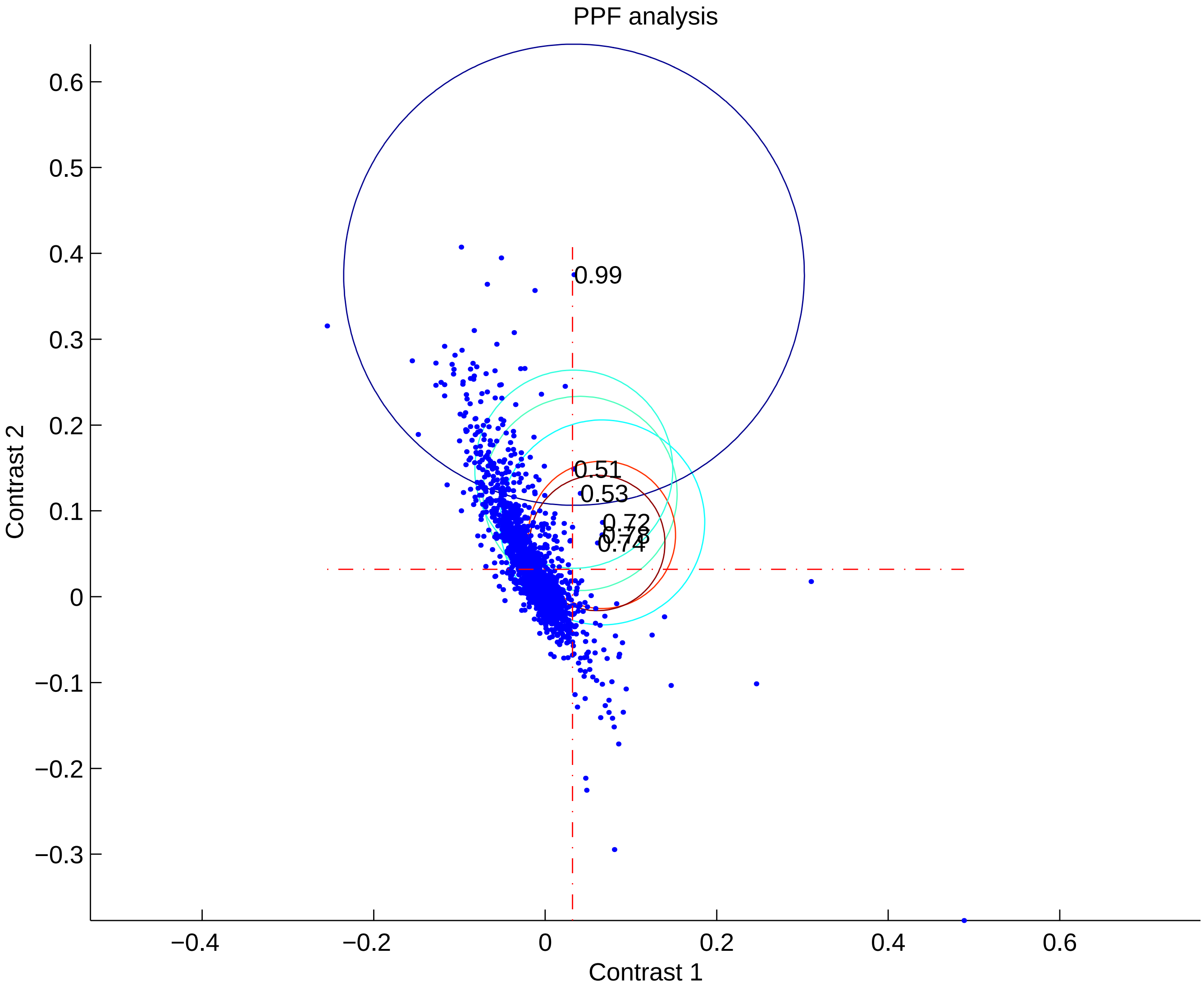

Fig. 11. *The PPF scores of 6 genes in the indicated restricted sector computed for Fred Wright's data using Affymatrix avgdiff indices. Constant contours around each point indicate standard errors. The contrast function A is as given in the text and corresponds to rotating the scatter plot in Fig. 8 by 90^o.*

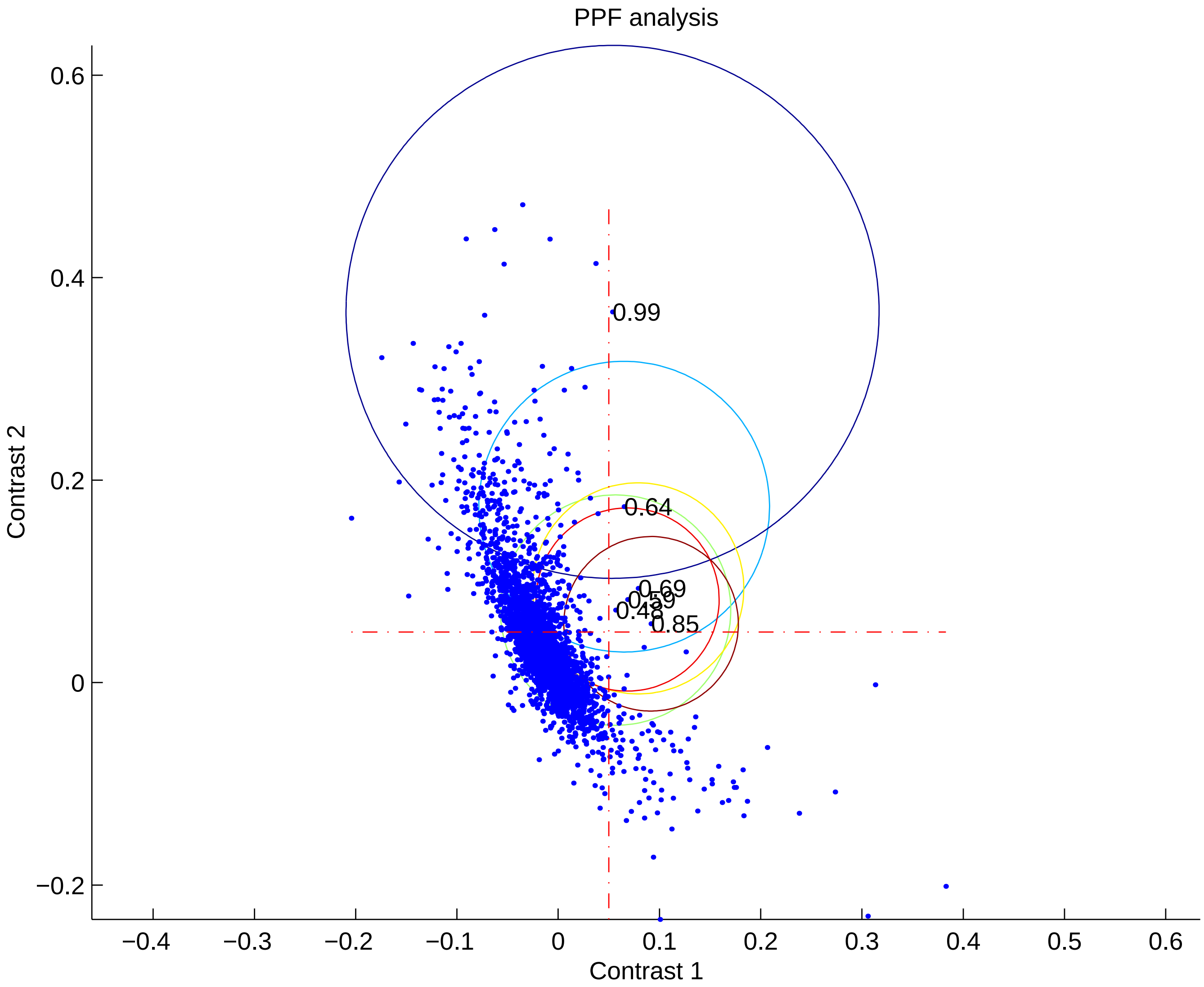

Fig. 12. *Same as Fig. 11 except that the PPF analysis is applied to Fred Wright's computed Li-Wong reduced indices.*

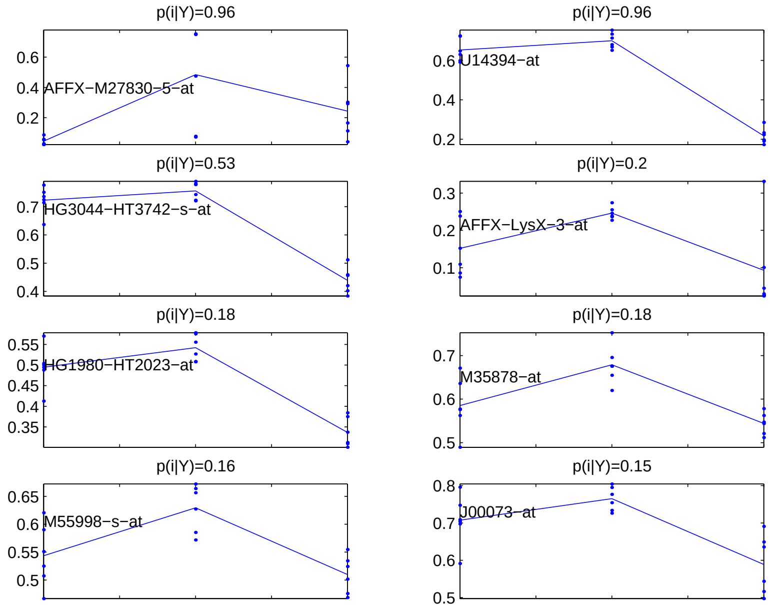

Fig. 13. *First 8 rank ordered convex cap gene profiles in an enlarged sector containing the one illustrated in Fig. 11.*

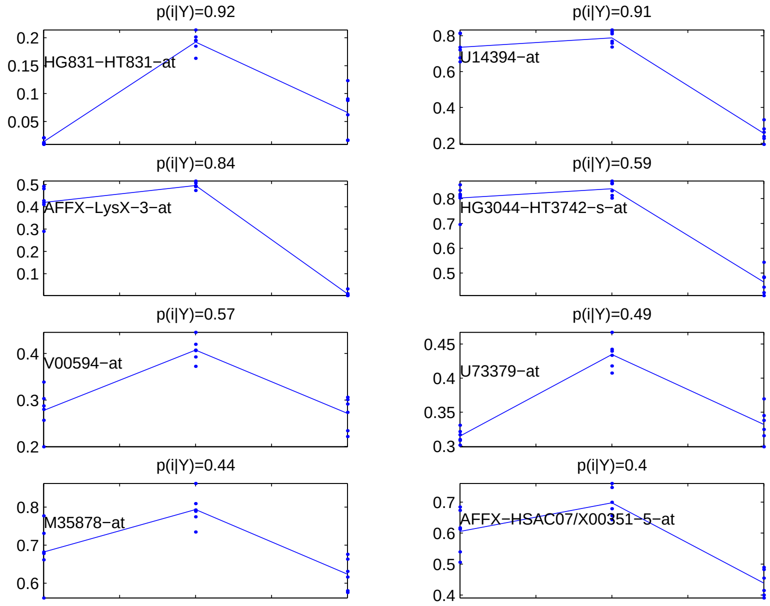

Fig. 14. *Same as in Fig. 13 except with Li-Wong reduced indices.*

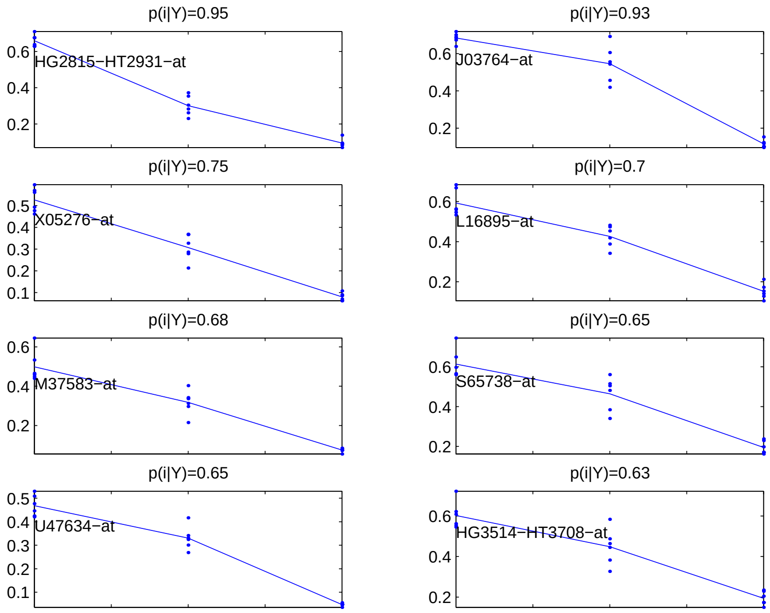

Fig. 15. *First 8 ranked PPF monotone decreasing profiles for avgdiff indices.*

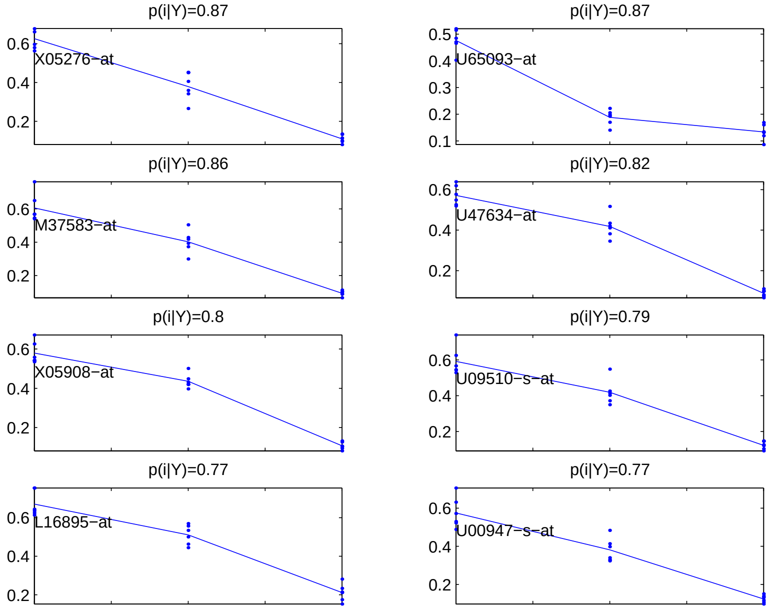

Fig. 16. *First 8 ranked PPF monotone decreasing profiles for Li-Wong reduced indices.*