Pareto-Optimal Methods for Gene Filtering

Alfred O. Hero III and Gilles Fleury[1]

June 10, 2002

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 single discriminants, e.g. the ratio of between-class to within-class variation or correlation to a template. Here a different approach is introduced for extracting information from gene microarrays. The approach is based on multiple objective optimization and we call it Pareto front (PF) analysis. This method establishes a ranking of genes according to estimated probabilities that each gene is Pareto-optimal, i.e., that it lies on the Pareto front of the multiple objective scattergram. For illustration the analysis is illustrated in the context of ranking the most aberrant non-linear genes in Fred Wright's GeneChip study.

Alfred Hero (`hero@eecs.umich.edu`) is Professor, with primary appointment in the Dept. of Electrical Engineering and Computer Science (EECS), and secondary appointments in the Dept. of BioMedical Eng. and the Dept. of Statistics, University of Michigan, Ann Arbor MI 48109-2122, USA. Gilles Fleury (`gfleury@freesurf.fr`) is Professor, Ecole Supérieure d'Electricité, Service des Mesures, 91192 Gif-sur-Yvette, France. This research was partially sponsored by a NATO grant to support Gilles Fleury during his sabatical at the University of Michigan over the summer of 2001.

Keywords: gene selection, gene screening, multicriterion scattergram, data mining, posterior Pareto fronts

1 Introduction

Microarray analysis of temporal gene expression profiles offers one of the most promising avenues for exploring genetic factors underlying disease, regulatory pathways controlling cell function, organogenesis and development; see Lockhart et al. (1996); Lee et al. (1999); Livesey et al. (2000) or DeRisi et al. (1997) for background. Gene microarrays can potentially identify RNA expression levels of thousands of genes in a time sequence of tissue samples, thereby providing valuable information about complex gene expression patterns over time. Recent advances in bioinformatics have brought us closer to realizing this potential. However, the massive scale and variability of microarray gene data creates new and challenging problems of clustering and data mining. One of these problems is the so-called *gene filtering problem*, also called gene screening and gene selection, which is to reliably extract genes exhibiting interesting expression profiles from the thousands of hybridization indices generated by the microarray. The most common approach to gene filtering are significance tests implemented by thresholding a set of test statistics, e.g. paired T-tests of mean differences, Fisher tests of variance, or Mann-Whitney rank tests. These can be found on most of the commercial and freeware packages used for statistical gene analysis such as the SAM MS Excel add-on distributed by Stanford University (2001) or the Microarray Suite and Data Mining Tool (DMT) distributed by Affymetrix (2002). Such approaches can yield a list of genes that are ranked in order of statistical significance according to observed p-values.

This paper describes a different approach to gene selection, denoted Pareto-optimal filtering, which is based on the ordinal theory of multiple objective optimization pioneered by the economist and sociologist Vilfreda Pareto (1848-1923). Pareto-optimality is a founding principle for social choice and decision-making in mathematical economics (See papers by Arrow et al. (2002); Arrow and Hervé (1986) and Pareto website of The New School (2001)). As discussed in Steuer (1986) this principle has since been applied to many other fields. Since V. Pareto's name has many other associations in probablity and statistics, it is important to emphasize that the proposed method of Pareto-optimal gene filtering is completely unrelated to Pareto analysis or Pareto graphs for statistical process control and quality assessment, to the Pareto principle of management science, or to the Pareto probability density, e.g., as in the Pareto model of income distribution.

To apply Pareto-optimal gene filtering the experimenter computes a number of ranking criteria for each gene, generating a point cloud of criterion vectors which is called the *multicriterion scattergram*. For example, to select the most monotonic profiles over time the ranking criteria might be chosen as the differences in gene expression level over successive time points. The objective of Pareto-optimal filtering is to isolate genes that achieve a compromise between maximizing (or minimizing) the competing gene-ranking criteria, i.e. to find the "winning" profiles. Such genes lie on the so-called *Pareto front* of the multicriterion scattergram and are the *non-dominated genes*, see Sec. 3 for definitions. Stripping off genes from successive Pareto fronts in the multicriterion scattergram yields a sequence of Pareto fronts at increasing depths in the data, called the first, second, third, ..., Pareto fronts, respectively. This sequence of fronts reveals a hierarchy of the highest scoring gene profiles. In two recent conference papers Fleury et al. (2002a,b) and a paper submitted to Nature-Genetics Yoshida et al. (2002), we applied Pareto-optimal filtering to discover young- and old- dominant mouse retina genes in GeneChip experiments and the discovered genes were validated using RT-PCR techniques. The purpose of the present paper is to present the general Pareto filtering methodology, introduce a Bayesian formulation of Pareto filtering, and to illustrate them on a widely available data set created expressly for testing gene filtering, classification, and differential expression estimation.

As the gene indices are randomly sampled from multiple subjects there can exist substantial statistical sampling errors that complicate the Pareto-optimal analysis. These sampling errors can be handled by cross-validation producing what can be called a *resistant Pareto front* (RPF) of genes, defined as those genes that land on the Pareto front with high relative frequency under resampling. As the RPF method does not rely on any distributional assumptions on the data it is very flexible, allowing treatment of arbitrary ranking criteria such as dependent and non-linear functions of the data. However, when the data distribution can be characterized, even approximately, RPF has obvious drawbacks. Principal among these drawbacks is the high computational load of cross-validation which can make RPF methods impractical to implement for large sample size. To address these drawbacks a Bayesian approach is presented for Pareto-optimal gene filtering: the *posterior Pareto front* (PPF) method.

As contrasted to the RPF method, the PPF method ranks each gene according to its posterior probability that it belongs to the Pareto front. This probability is computed using prior densities on various unknown parameters in the sampling error distribution. In particular, one can assume conditionally independent Gaussian gene indices and assign non-informative priors on the mean and variance for each time sampled gene. Using an asymptotic approximation to an extreme-value distribution yields an expression for the posterior probability whose complexity increases in the number of ranking criteria and not in the number of samples. The PPF analysis is applied to a set of ranking criteria defined as linear functions, called *profile contrasts*, of the prior mean expression levels of each gene profile. When the profile contrast matrix of coefficients of these contrasts is an orthogonal matrix and the mean expression levels are uncorrelated Gaussian the ranking criteria satisfy the statistical independence assumptions. For illustration, PPF and RPF analyses are applied and compared to Fred Wright's data set, described in Lemon et al. (2002), for detection of the most aberrant genes violating linearity in the Affymetrix human fibroblast mixture experiment.

The outline of the paper is as follows. In Sec. 2 a brief review of microarray data analysis is presented and in Sec. 3 the Pareto-optimal gene filtering approach is introduced. In Sec. 4 the general PPF gene filtering method is developed and in Sec. 5 linear contrast functions are considered. Finally in Sec. 6 PPF analysis is applied to finding aberrant genes in Fred Wright's human fibroblast mixing data.

2 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 have revolutionized the field of experimental genetics by offering to the experimenter the ability to simultaneously measure thousands of gene sequences simultaneously. A gene microarray consists of a large number N of known DNA probe sequences that are put in distinct locations on a slide. See one of the following references for more details: Kadota et al. (2001); Brown and Botstein (1999); Bassett et al. (1999); Fitch and Sokhansanj (2000). After hybridization of an unknown tissue sample to the gene microarrays, the abundance of each probe present in the sample can be estimated from the measured levels of hybridization (responses). Two main types of gene microarrays are in wide use: photo-lithographic gene chips and fluorescent spotted cDNA arrays. An example of the former is the Affymetrix (2000) product line. An example of the later is the protocol of the National Human Genome Research Institute (NHGRI) (2001).

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) (See Hastie et al. (2000); Allzadeh and etal (2000) and Brown et al. (2000)). Validation methods have been widely used and include: significance analysis of microarrays (SAM); bootstrapping cluster analysis; and leave-one-out cross-validation (See Tusher et al. (2001) and Kerr and Churchill (2000)). 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.

3 Multiple Objective Gene Filtering

As contrasted to maximizing *scalar* criteria, multiple objective gene filtering seeks gene profiles that strike an optimal compromise between maximizing several criteria. This is closely related to multiple objective optimization in which the concept of Pareto-optimal solutions play a crucial role. These solutions are almost never unique and are variously called the Pareto-optimal set, the Pareto front, the Pareto frontier, and the Edgeworth-Pareto front (See book by Stadler (1988) or Steuer (1986)). Pareto optimality theory has been applied to a wide range of application areas including: economics, sociology, psychology, operations research, and evolutionary computing (See above referenced books and articles by Zitzler and Thiele (1999) and Arrow and Hervé (1986) for examples).

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 investigate the entire set of genes which maximize

J_α for some choice of α. As shown by Das and Dennis (1997), it turns out that these genes are in a set called the *Pareto front* which results from multiple objective optimization of the pair $[\xi_1(\underline{\mu}_i), \xi_2(\underline{\mu}_i)]^T$ over i.

Multiple objective optimization captures the intrinsic compromises among conflicting objectives. Consider Fig. 1 and suppose that ranking criteria ξ_1 and ξ_2 are to be maximized. The collection of points in the figure are called the multicriterion scattergram. 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, gene i is said to be dominated if there exists some other gene $g \neq i$ such that for some $p = q$

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

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 (first) Pareto front. A second Pareto front can obtained by stripping off the points on the first front and computing the Pareto front on the remaining points - which for the example in Fig. 1 would be genes D and E.

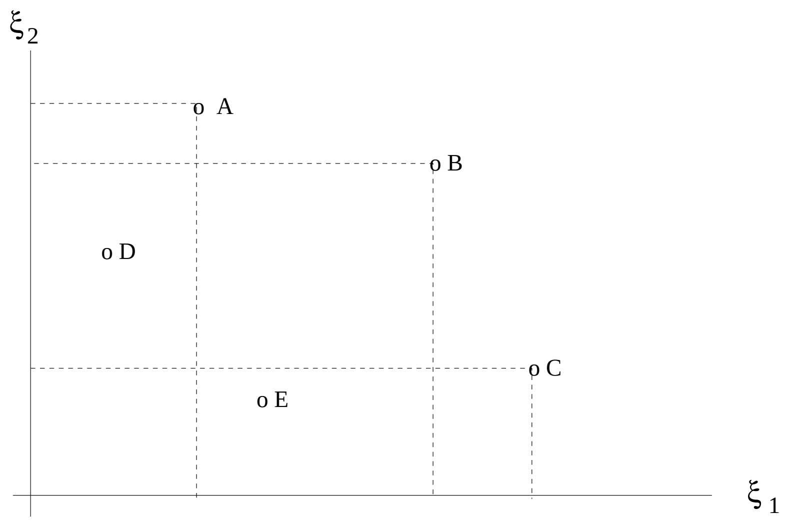

Figure 1: *A hypothetical multicriterion scattergram for genes A,B,C,D,E plotted as vectors in the plane described by a pair of ranking criteria ξ_1 and ξ_2. A, B, C are non-dominated genes and form the (first) Pareto front. A second Pareto front is formed by genes D,E.*

The above methods are applicable when the criteria ξ_1 through ξ_P are perfectly observable. However, as these criteria depend on the true mean values $\underline{\mu}(i)$ of the i-th gene profile, the criteria are only partially observed through a random sample from the underlying population. In Yoshida et al. (2002); Fleury et al. (2002a,b) we applied a non-parametric monotonicity criterion for detecting interesting monotonic 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 the sign of the slope between each time point was extracted 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. A cumulative cross-validation was also performed to determine the number of times a gene appears in one of the first ten Pareto fronts. The result of this analysis yielded a set which is called the *resistant Pareto fronts* (RPF).

4 Posterior Pareto Filtering

The posterior Pareto front analysis introduced here casts the cross-validation ranking procedure described above into a Bayesian framework. The posterior probability $p(i|Y)$ that a particular gene i is on the first Pareto front is easily expressed using the definition of non-dominance and the assumption that the criteria vectors $\{\underline{\xi}(j)\}_j$ are statistically independent given the chipset data Y. In the following expressions the notation $\underline{\xi}(i) \leq \underline{\xi}(j)$ means that $\xi_p(i) \leq \xi_p(j)$ for $p = 1, \ldots, P$, and E^c denotes the complement of event E:

$$\begin{aligned} & p(i|Y) \\ = \; & P\left(\cap_{j\neq i}\left\{\underline{\xi}(i) \leq \underline{\xi}(j)\right\}^c |Y\right) \\ = \; & \int dP(\underline{\xi}(i)|Y) \prod_{j\neq i} P\left(\left\{\underline{\xi}(i) \leq \underline{\xi}(j)\right\}^c |Y, \underline{\xi}(i)\right) \end{aligned}$$

or when the posterior density $f_{\underline{\xi}(i)|Y}(\underline{u})$ of $\underline{\xi}(i)$ is available

$$p(i|Y) \int d\underline{u} f_{\underline{\xi}(i)|Y}(\underline{u}) \prod_{j\neq i}\left[1 - P\left(\underline{\xi}(j) \geq \underline{u}|Y\right)\right]. \tag{5}$$

This expression requires evaluating a multidimensional integral over P-dimensions. For the case of two criteria ($P = 2$) the posterior probability reduces to:

$$p(i|Y) = \int\int du_1 du_2 f_{\xi_1(i),\xi_2(i)|Y}(u_1, u_2) \prod_{j\neq i}\left[F_{\xi_1(j)|Y}(u_1) + F_{\xi_2(j)|Y}(u_2) - F_{\xi_1(j),\xi_2(j)|Y}(u_1, u_2)\right], \tag{6}$$

where $F_{\xi_1(i),\xi_2(i)|Y}(u_1, u_2)$ is the bivariate conditional distribution function of $\xi_1(i), \xi_2(i)$: $F_{\xi_1(i),\xi_2(i)|Y}(u_1, u_2) = \int_{-\infty}^{u_1} dv_1 \int_{-\infty}^{v_2} du_2 F_{\xi_1(i),\xi_2(i)|Y}(v_1, v_2)$.

4.1 Pareto Filtering of Gene Expression Profiles

Start with the 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, the non-informative prior described in Geisser and Cornfield (1963) is adopted

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

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.

4.1.1 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 density equal to the student-t density

$$f_{\mu_t(i)|Y}(u) \;=\; k(Y_{ti}) \left(1 + \frac{(u - \hat{\mu}_t(i))^2}{\hat{\sigma}_t^2(i)}\right)^{-(M-a+2)/2}, \tag{7}$$

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 given in Geisser and Cornfield (1963):

$$k(Y_{ti}) = \frac{1}{\hat{\sigma}_t(i)\sqrt{\pi}} \frac{\Gamma(\frac{1}{2}(M - a + 2))}{\Gamma(\frac{1}{2}(M - a + 1))}. \tag{8}$$

The associated distribution function 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 (7) yields

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

where $(x)_+$ is the function equal to x when $x > 0$ and equal to zero otherwise.

4.1.2 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 density of $\underline{\mu}(i) = [\mu_1(i), \ldots, \mu_T(i)]^T$ takes the form of a multivariate student-t

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

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 scale factor similar to (8).

Analogously to the case of unequal variances, the associated distribution function can be approximated by a multivariate L_∞ approximation to (9):

$$F_{\underline{\mu}(i)|Y}(u_1, \ldots, u_T) \approx \left(1 + \sum_t \frac{(\hat{\mu}_t(i) - u_t)_+^2}{\hat{\sigma}^2(i)}\right)^{-(TM-a+2)/2}. \qquad (9)$$

5 Profile Contrasts

5.1 Profile Amplitude Criterion

To simplify the presentation the time sampled means $\xi_p(i) = \mu_p(i)$, $p = 1, \ldots T$ are initially adopted as the criteria of interest. This will be called the profile amplitude criterion and we focus on the case of time varying variances for concreteness. This will be generalized to a set of contrast functions applied to the means in the next subsection. Using the expressions (7) and (9) in (6) gives an expression for $p(i|Y)$ which only requires numerical evaluation of one-dimensional integrals (as compared with T-dimensional integrals if the exact non-asymptotic distribution function was used).

5.2 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.* The vector $\underline{\xi}(i)$ will be called the ***profile contrasts*** for gene i. To retain the simplicity of the 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 density $f_{\underline{\mu}(i)|Y}(\underline{u})$ in (9) is invariant to orthogonal transformations of $\underline{u} - \hat{\underline{\mu}}(i)$. 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. This latter case is illustrated below.

5.3 Examples of Profile Contrasts

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

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

$$A_4 = \begin{bmatrix} -1 & 1 & 0 & 0 \\ -1 & -1 & 2 & 0 \\ -1 & -1 & -1 & 3 \\ 1 & 1 & 1 & 1 \end{bmatrix},$$

$$A_4^{'} = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & -1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}.$$

As all of these matrices satisfy AA^T = diagonal, the posterior Pareto analysis can be applied 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 strong and weak monotonic increasing profiles. If the density 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 sector, etc.

Of interest are general ways to construct meaningful contrast matrices A which are unitary, so as to maintain multiple criteria independence for computational simplicity, yet to capture desired shape characteristics of temporal expression profiles. One possible method is to define a contrast matrix B whose rows capture some set of desired linearly independent properties of the profile and then apply the PPF with the 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 Cholesky 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 $A_3^{'}$.

6 Pareto Filtering Application

The PPF and RPF analysis methods were applied to Fred Wright's dataset described in the paper by Lemon et al. (2002) and available at the web address provided in the citation. The analysis software was written and implemented in Matlab. Fred Wright's data set was obtained from a mixing experiment which the authors designed for empirically validating and comparing various differential gene expression methods of analysis. As explained in Lemon et al. (2002) three populations of genes were hybridized to Affymetrix HuGeneFL chips: serum starved human fibroblast cells; serum 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 HuGeneFL chip contains the same 7129 gene probes. For each gene probe the sequence of hybridization levels from the "stimulated(t=1)," "50-50(t=2)," and "starved(t=3)," populations was defined, in that order, as a gene expression profile. This provides a suitable test dataset for testing PPF and RPF since the true profiles should be linearly increasing or decreasing over the three "time points." Any extracted non-monotone gene profiles must either be due to statistical estimation errors, uncontrolled fluctuations in sample concentrations during hybridization, or other experimental errors.

In Figs. 3 and 4 the multicriterion scattergram of 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 paper by Lemon et al. (2002) 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. the aberrant genes having non-monotonic profiles. As expected the Li-Wong reduced indices are better in this quality measure.

6.1 Sign-based Pareto analysis

The objective is to determine the most aberrant inverted V-shaped gene profiles. These are genes whose means lie within the lower right sector of the multicriterion scattergram in Fig. 4. As a preprocessing step a standard non-linear profile filter was applied using a Fisher test to screen gene profiles having large residual linear regression errors inconsistent with a linearity hypothesis. Specifically, each gene profile was regressed onto the linear model

$$y_{tm}(i) = a(i)t + b(i) + \epsilon_{tm}(i), \quad t = 1, 2, 3,$$

where $\{\epsilon_{tm}(i)\}_{tm}$ is i.i.d. Gaussian additive noise with variance σ^2 and a, b are undetermined linear-model coefficients. The regression gives an error residual for the i-th gene

$$R(i) = [\underline{\overline{y}}_{**}(i)]^T [I - \Pi][\underline{\overline{y}}_{**}(i)],$$

where Π is the 3×3 matrix which orthogonally projects $\mathbb{R}^3$ onto the affine subpace $\{y \in \mathbb{R}^3 : y = a[1,2,3] + b[1,1,1]\}_{a,b \in \mathbb{R}}$, and $[\underline{\overline{y}}_{**}(i)]^T = \frac{1}{M}\sum_{m=1}^{M}[y_{1m}, y_{2m}, y_{3m}]^T$ is the mean vector for the i-th gene profile. With $s(i)$ the (pooled) sample variance estimate of σ^2 the statistic $F(i) = R(i)/s(i)$ is distributed as Fisher-F on 2 and $M - 3$ degrees of freedom. The $1 - p$ quantile of this distribution gives a threshold on $F(i)$ above which a gene is classified as having a non-linear profile at significance level p. This preprocessing eliminated all but 98 genes from the 7129 total number of genes studied. In the sequel these will be called the "non-linear" gene profiles.

A simple modification of the sign-based Pareto analysis method we adopted in Yoshida et al. (2002); Fleury et al. (2002b,a) can be applied to finding the most aberrant non-linear profiles. In Fig. 5 the multicriterion scattergram is displayed. The non-linear genes are displayed with crosses. The first criterion in the figure is the contrast defined by $A = [-1, 2, -1]$, which measures twice the difference between the middle point and the average of the two other points in each profile. The second criterion is the number of "virtual" profiles whose shapes match a convex cap profile. Specifically, for each gene generate all $6^3 = 216$ possible trajectories through the 3 sets of 6 replicated measurements of hybridization levels. The ranking is defined from the proportions of these trajectories which have slope of positive sign followed by slope of negative sign. This ranking criterion will be called "non-parametric" since the sign-based shape criterion does not depend on the sharpness or assymetry of the inverted V profile shape. Figure 6 shows the first five Pareto fronts computed on the full set of 3×6 non-linear gene samples indicated as crosses on Figure 5. These fronts were computed by successively stripping off genes found to lie on the previous Pareto fronts and rerunning Pareto analysis on the remaining points. Finally leave-one-out cross validation was performed to determine the resistant genes that for which a high proportion of the 216 resampled 3×5 trajectories remained on the first Pareto front. Fig. 7 shows the top 8 resistant profiles ranked in terms of relative frequency of remaining on the first front.

6.2 Contrast-based Pareto Analysis

Linear contrast criteria on the non-linear gene profiles were implemented to determine a ranking of the most aberrant inverted V-shaped profiles. For this the following contrast matrix is adopted

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

which takes large values for the inverted-V shaped profiles. Figure 8 displays the multicriterion scattergram which is simply a 90^o rotation of that in Fig. 4. The crosses in the figure indicate the 98 non-linear genes.

Both Bayesian (PPF) and cross-validation leave-one-out (RPF) methods for the contrast functions were investigated. While many different values for the prior parameter a have have been investigated for Bayesian PPF analysis, we only present results for $a = 2$ here. Increasing a makes the computed posterior probabilities more conservative (smaller) as the tails of the posterior densities become heavier. Figure 9 shows the first five Pareto fronts computed on the full data set without any cross-validation. Figures 10 and 12 show the results of PPF and RPF analysis in the multiple criteria plane. The contours around each point denotes the standard error (one standard deviation) circle and the annotation at the centers of the circles is the computed posterior probability (PPF), or relative frequency (RPF), that the gene belongs to the first Pareto front. These plots illustrate how statistical uncertainty in the multiple criteria plane (standard error contours) translates to probability that a gene lies on the first Pareto front.

Figures 11 and 13 show the eight top scoring trajectories under PPF and cross-validated RPF analysis, respectively. In each sub-panel the indicated piecewise linear line passes through the means of the 6 replicates for each of the 3 time samples. The top ranked 25 gene profiles under each criterion are shown in Table 2 along with their probability scores. Note that in the RPF linear contrast analysis all 17 of the positively ranked genes appear in the first 25 top PPF-ranked gene list. Note also that in the RPF non-parametric analysis there is a highly ranked gene (D63880-at) which is low ranked (not among first 25) by linear contrast PPF and not at all ranked by RPF. This discrepancy can be explained by the large variance of this gene profile at the midpoint - see top right panel in Fig. 7. This variance affects both of the linear contrast criteria but has less effect on the sign criterion used by the non-parametric RPF method.

7 Conclusion

This paper introduced a new method of gene filtering based on analysis of the Pareto fronts of a specified multiple criterion objective function applied to each gene. These techniques also have applicability to general

PPF linear contrast	P(i|Y)	RPF linear contrast	P(i|Y)	RPF non-parametric	P(i|Y)
AFFX-ThrX-5-at	0.999	AFFX-DapX-5-at	1	AFFX-LysX-3-at	1
HG3342-HT3519-s-at	0.998	AFFX-ThrX-5-at	1	D63880-at	1
AFFX-DapX-5-at	0.998	AFFX-ThrX-M-at	1	HG831-HT831-at	1
HG831-HT831-at	0.996	HG3342-HT3519-s-at	1	U73379-at	1
AFFX-ThrX-M-at	0.986	HG831-HT831-at	1	V00594-at	1
X69111-at	0.984	U14394-at	1	U14394-at	0.847
U14394-at	0.974	V00594-at	1	AFFX-ThrX-5-at	0.431
AFFX-LysX-3-at	0.962	X69111-at	1	AFFX-DapX-5-at	0.245
V00594-at	0.955	U45285-at	0.944	AFFX-PheX-3-at	0.222
U45285-at	0.932	AFFX-LysX-3-at	0.917	AFFX-HSAC07/X00351-5-at	0.208
AB000115-at	0.899	AFFX-HSAC07/X00351-5-at	0.806	AB000115-at	0.167
AFFX-HSAC07/X00351-5-at	0.866	AB000115-at	0.417	U00954-at	0.167
U73379-at	0.837	U73379-at	0.13	U45285-at	0.167
AFFX-DapX-M-at	0.678	V00594-s-at	0.074	U75362-at	0.167
Y09912-rna1-at	0.67	U75362-at	0.037	AFFX-ThrX-M-at	0.157
U75362-at	0.56	AFFX-PheX-5-at	0.028	HG1980-HT2023-at	0.032
AFFX-DapX-3-at	0.555	U03399-at	0.009	AFFX-PheX-M-at	0.028
V00594-s-at	0.554			U30998-at	0.028
HG1980-HT2023-at	0.483			Y09912-rna1-at	0.028
HG3044-HT3742-s-at	0.441				
D43636-at	0.389				
L27624-s-at	0.387				
U03399-at	0.378				
S69370-s-at	0.321				
AFFX-PheX-5-at	0.315				

Figure 2: *The top scoring genes (Affymetrix nomenclature) resulting from PPF and RPF analysis of the most non-monotone convex cap profiles for Fred Wright's data (Li-Wong reduced indices).* $P(i|Y)$ *denotes estimated probability that given gene belongs to first Pareto front obtained from Bayesian analysis (PPF) or leave-one-out cross-validation (RPF).*

data mining problems involving shape analysis and general selection criteria. The method is very flexible and involves choosing a set of appropriate profile contrasts which display desired characteristics of the expression profiles. Both cross-validation and Bayesian posterior Pareto methods were presented for ranking genes in order of the probability that the gene profile is Pareto optimal. In contrast to the cross validation methods the Bayesian method assigns positive probability to all genes and has lower complexity than the non-parametric cross-validation method for large sample size. On the other hand the non-parametric cross-validation method may be more robust to outliers which might have more influence on the sample mean profile shape used by the contrast methods.

As for possible future work, a full bootstrap implementation of the contrast based RPF method would undoubtedly make it more outlier resistant. However this would greatly increase computational complexity. Methods of multiple comparisons (Miller (1981)), which have been previously applied to differential analysis of gene microarrays by Storey and Tibshirani (2001) and others, also appear applicable to multicriterion filtering and, in particular, to validating Pareto-optimal trajectories. Finally, the multi-criteria methods described in this paper may be applicable to the PIDEX method of Ge et al. (2001) who propose different ways of combining pairs of gene selection criteria.

Acknowledgements

The authors are grateful for illuminating discussions of this work with Dr. Anand Swaroop and Dr. Shigeo Yoshida, of the Depts. of Ophthalmology and Human Genetics, with Prof. Debashis Ghosh of the Dept. of Biostatistics at UM Ann Arbor. The authors also thank Prof. Terry Speed, of UC Berkeley, for making us aware of Frank Wright's data. This research was partially supported by a NATO grant that funded Gilles Fleury's sabatical at the University of Michigan during the summer of 2001.

References

Affymetrix (2000), "NetAffx User's Guide," `http://www.netaffx.com/site/sitemap.jsp`.

— (2002), "Genechip software," Affymetrix, Inc, `http://www.affymetrix.com/products/software/index.affx`.

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

Arrow, K., Sen, A., and Suzumura, K. (2002), *Handbook of social choice and welfare*, Elsevier/North Holland, Amsterdam.

Arrow, K. J. and Hervé, R. R. (1986), *Social Choice and Multicriterion Decision Making*, MIT Press, Cambridge MA.

Bassett, D., Eisen, M., and Boguski, M. (1999), "Gene expression informatics–it's all in your mine," *Nature Genetics*, 21, 51–55.

Brown, M., Grundy, W. N., Lin, D., Cristianini, N., Sugent, C., Furey, T., Ares, M., and Haussler, D. (2000), "Knowledge-based analysis of microarray gene expression data by using support vector machines," *Proc. of Nat. Academy of Sci. (PNAS)*, 97, 262–267.

Brown, P. O. and Botstein, D. (1999), "Exploring the new world of the genome with DNA microarrays," *Nature Genetics*, 21, 33–37.

Das, I. and Dennis, J. (1997), "A Closer Look at Drawbacks of Minimizing Weighted Sums of Objectives for Pareto Set Generation in Multicriteria Optimization Problems," *Structural optimization*, 14.

DeRisi, J., Iyer, V., and Brown, P. (1997), "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science*, 278, 680–686.

Fitch, P. and Sokhansanj, B. (2000), "Genomic engineering: moving beyond DNA sequence to function," *IEEE Proceedings*, 88, 1949–1971.

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

— (2002b), "Pareto analysis for gene filtering in microarray experiments," in *European Sig. Proc. Conf. (EUSIPCO)*, Toulouse, FRANCE.

Ge, N., Huang, F., Shaw, P., and Wu, C. (2001), "PIDEX: a statistical approach for screening differentially expressed genes using microarray analysis," *Preprint*.

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

Hastie, T., Tibshirani, R., Eisen, M., Brown, P., Ross, D., Scherf, U., Weinstein, J., Alizadeh, A., Staudt, L., and Botstein, D. (2000), "Gene Shaving: a new class of clustering methods for expression arrays," Tech. rep., Stanford University.

Kadota, K., Miki, R., Bono, H., Shimizu, K., Okazaki, Y., and Hayashizaki, Y. (2001), “Preprocessing implementation for microarray (PRIM): an efficient method for processing cDNA microarray data,” *Physiol Genomics*, 4, 183–188.

Kerr, K. and Churchill, G. (2000), “Bootstrapping Cluster Analysis: Assessing the Reliability of Conclusions from Microarray Experiments,” *Proc. of Nat. Academy of Sci. (PNAS)*, 98, 8961–8965, `citeseer.nj.nec.com/414709.html`.

Lee, C., Klopp, R., Weindruch, R., and Prolla, T. (1999), “Gene expression profile of aging and its retardation by caloric restriction,” *Science*, 285, 1390–1393.

Lemon, W. J., Palatini, J. T., Krahe, R., and Wright, F. A. (2002), “Theoretical and experimental comparison of gene expression estimators for oligonucleotide arrays,” *Bioinformatics*, `http://thinker.med.ohio-state.edu/projects/fbss/index.html`.

Livesey, F., Furukawa, T., Steffen, M., Church, G., and Cepko, C. (2000), “Microarray analysis of the transcriptional network controlled by the photoreceptor homeobox gene,” *Crx. Curr Biol*, 6, 301–10.

Lockhart, D., Dong, H., Byrne, M., Follettie, M., Gallo, M., Chee, M., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. (1996), “Expression monitoring by hybridization to high-density oligonucleotide arrays,” *Nat. Biotechnol.*, 14, 1675–80.

Miller, R. G. (1981), *Simultaneous Statistical Inference*, Springer-Verlag, NY.

National Human Genome Research Institute (NHGRI) (2001), “cDNA Microarray Protocols,” `http://www.nhgri.nih.gov/DIR/Microarray/protocols.html`.

Stadler, W. (1988), *Multicriteria optimization in engineering and the sciences*, New York: Plenum, chap. Fundamentals of multicriteria optimization.

Stanford University (2001), “SAM: Significance analysis of microarrays,” Stanford Office of Technology and Licencing, `http://www-stat.stanford.edu/~tibs/SAM/`.

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

Storey, J. D. and Tibshirani, R. (2001), “Estimating false discovery rates under dependence, with applications to DNA microarrays,” Tech. Rep. 2001-28, Department of Statistics, Stanford University.

The New School (2001), “Vilfred Pareto, 1848-1923,” History of Economic Thought, New School, `http://cepa.newschool.edu/het/profiles/pareto.htm`.

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

Yoshida, S., Carter, T., Fleury, G., Hiriyanna, S., Othman, M., Lockhart, D., Hero, A., Barlow, C., and Swaroop, A. (2002), "Altered expression of immune- and stress- response genes in aging retina; implications for aging associated retinopathies," *Nature Genetics*, submitted.

Zitzler, E. and Thiele, L. (1999), "Multiobjective Evolutionary Algorithms: A Comparative Case Study and the Strength Pareto Approach," *IEEE Transactions on Evolutionary Computation*, 3, 257–271.

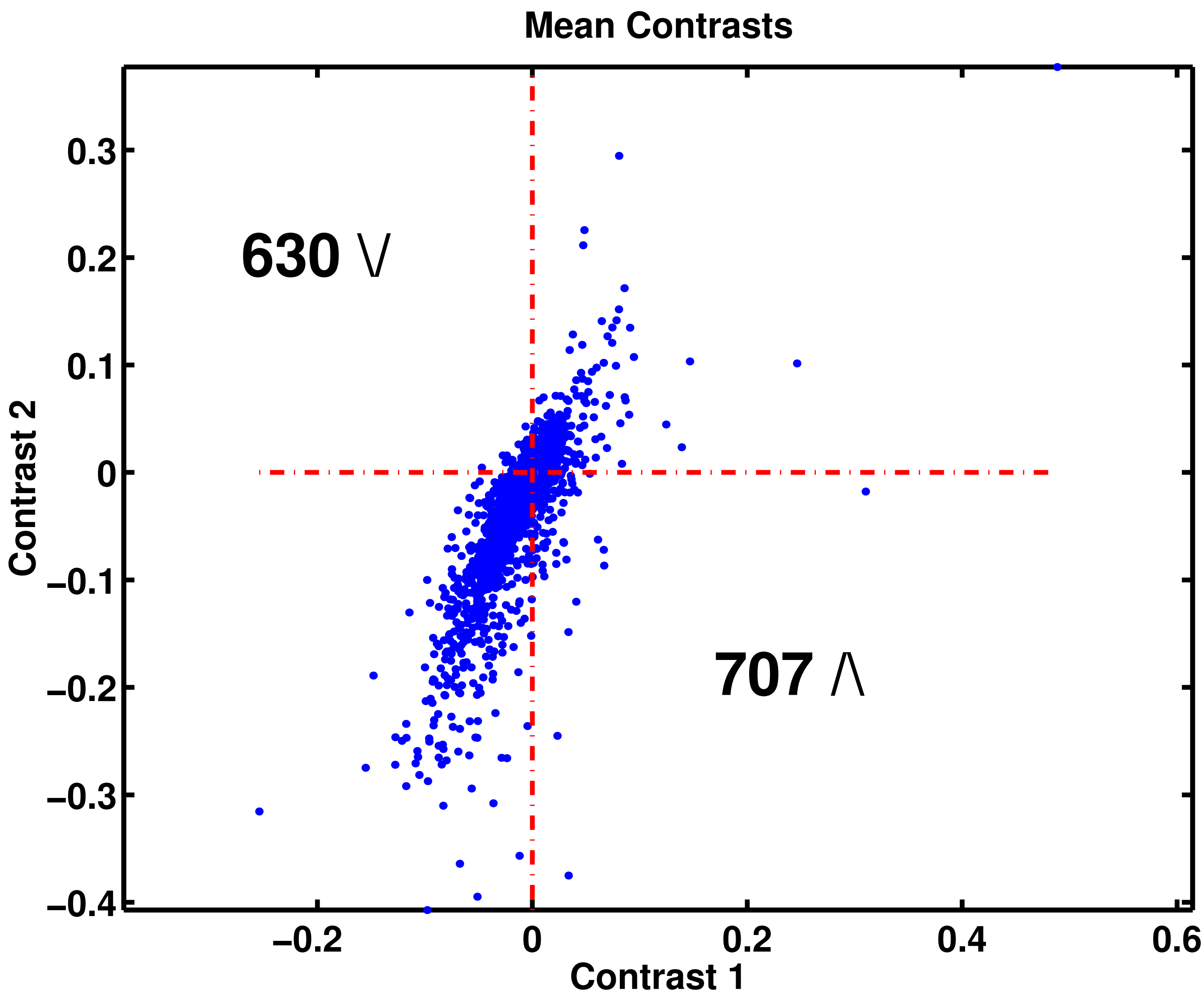

Figure 3: *Multicriterion scattergram of linear contrasts (sample mean contrasts defined from the first two rows of* A_3'*) for Affymetrix avgdiff indices for Fred Wright's HuGeneFL mixture study). Annotations are the number of non-monotone V-shaped profiles (convex cup profiles in upper left) and inverted V-shaped profiles (convex cap profiles in lower right).*

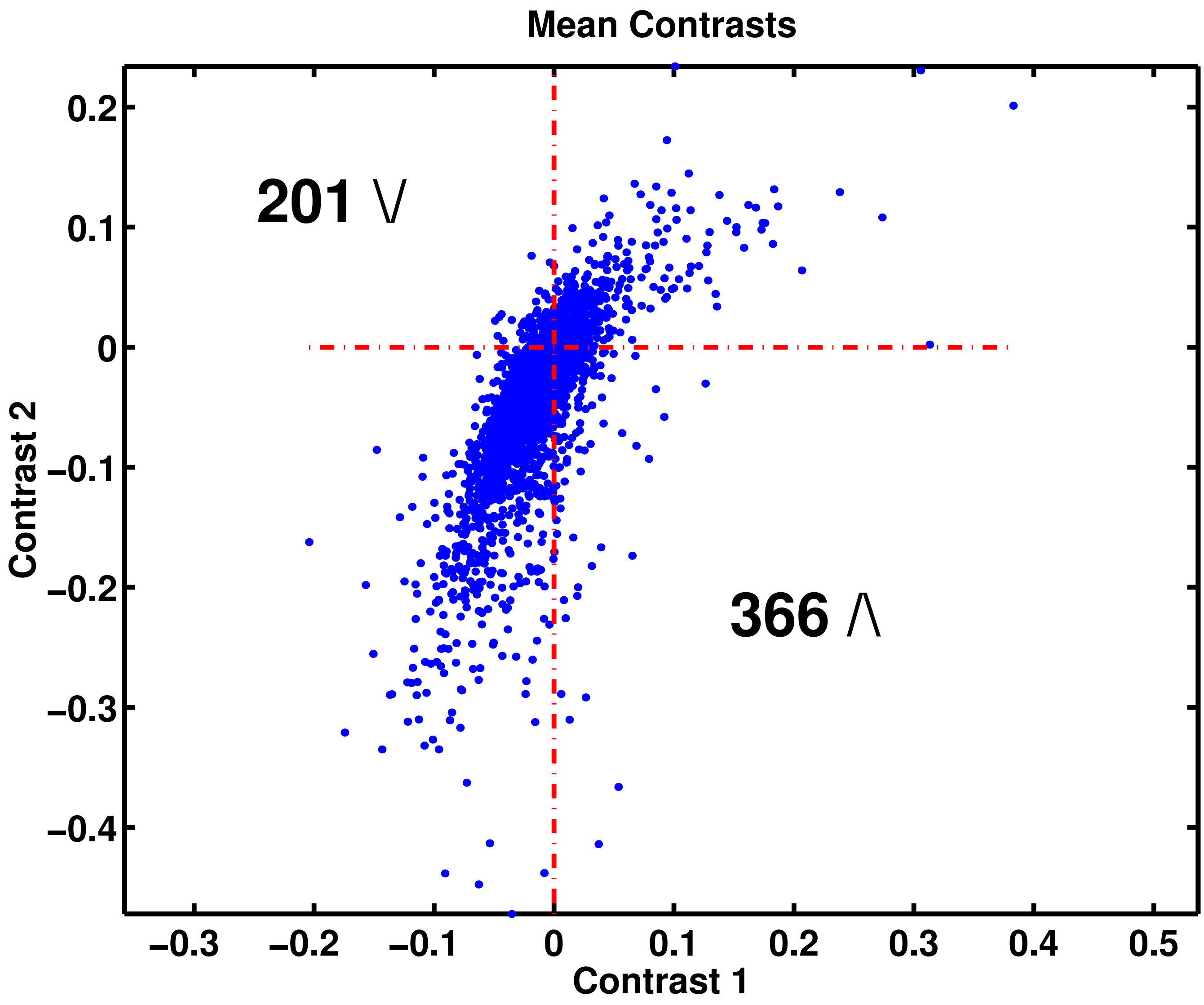

Figure 4: *Multicriterion mean scattergram of linear contrasts (sample mean contrasts defined from the first two rows of* A_3'*) for Li-Wong reduced indices in Fred Wright's HuGeneFL mixture study).*

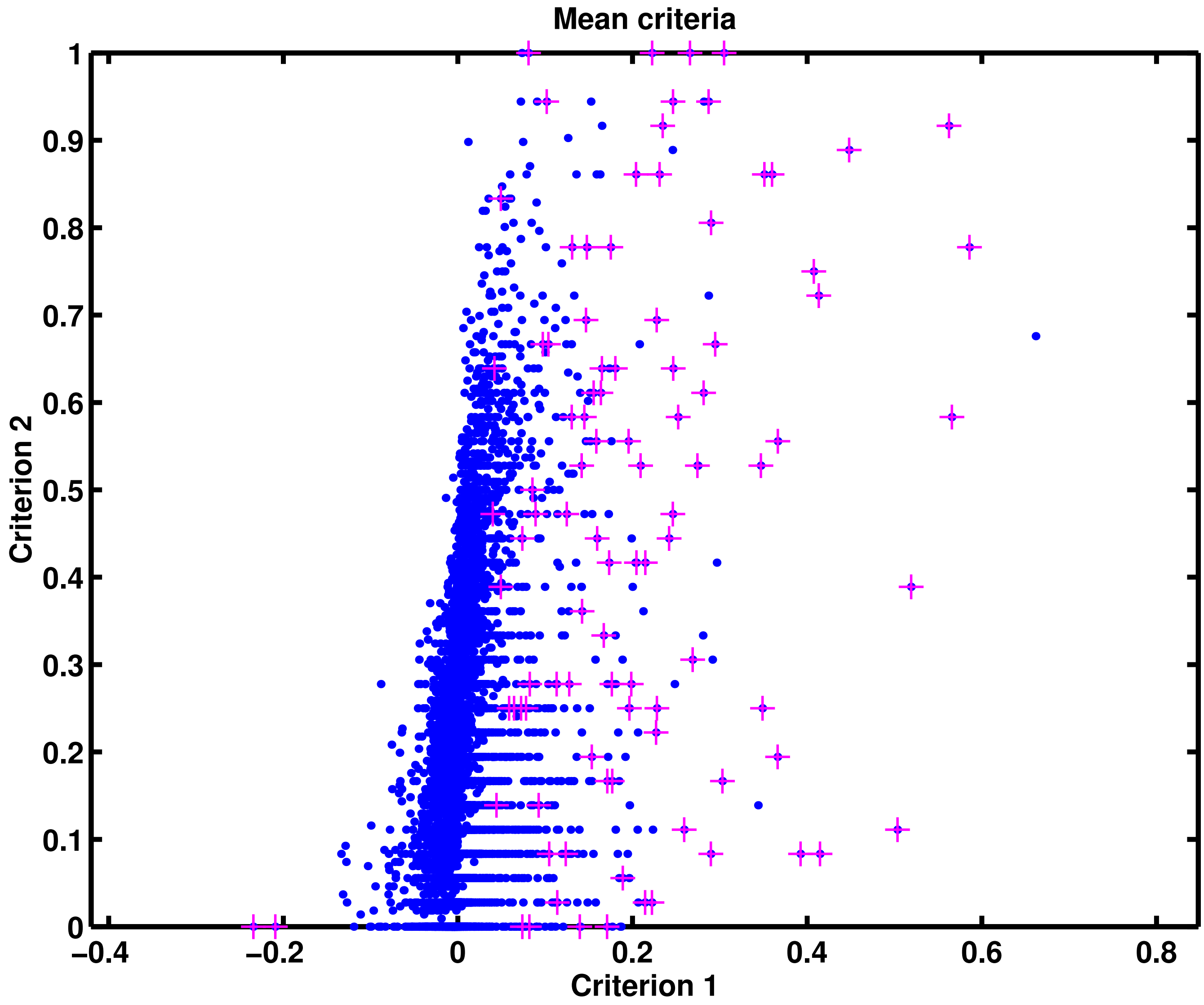

Figure 5: *Multicriterion mean scattergram of the non-parametric slope-sign ranking criterion for filtering the most aberrant inverted V-shaped gene profiles for Li-Wong reduced indices in Fred Wright's HuGeneFL mixture study. Crosses denote the 98 non-linear genes failing the Fisher linear profile test at a p-value of 0.1.*

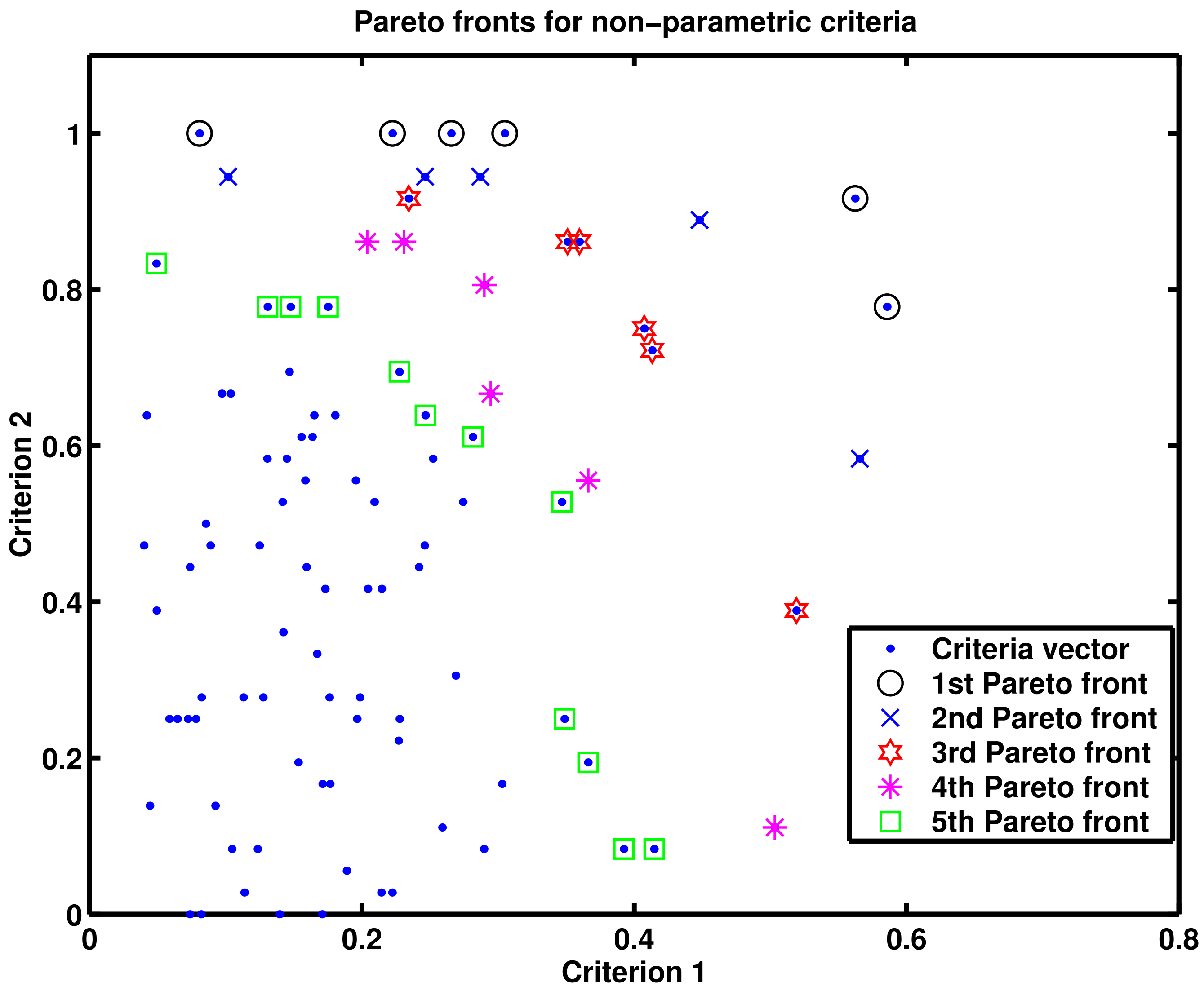

Figure 6: *The first five Pareto fronts (no cross-validation) of the non-parametric inverted V-shape criteria for the non-linear genes indicated by crosses in Fig. 5.*

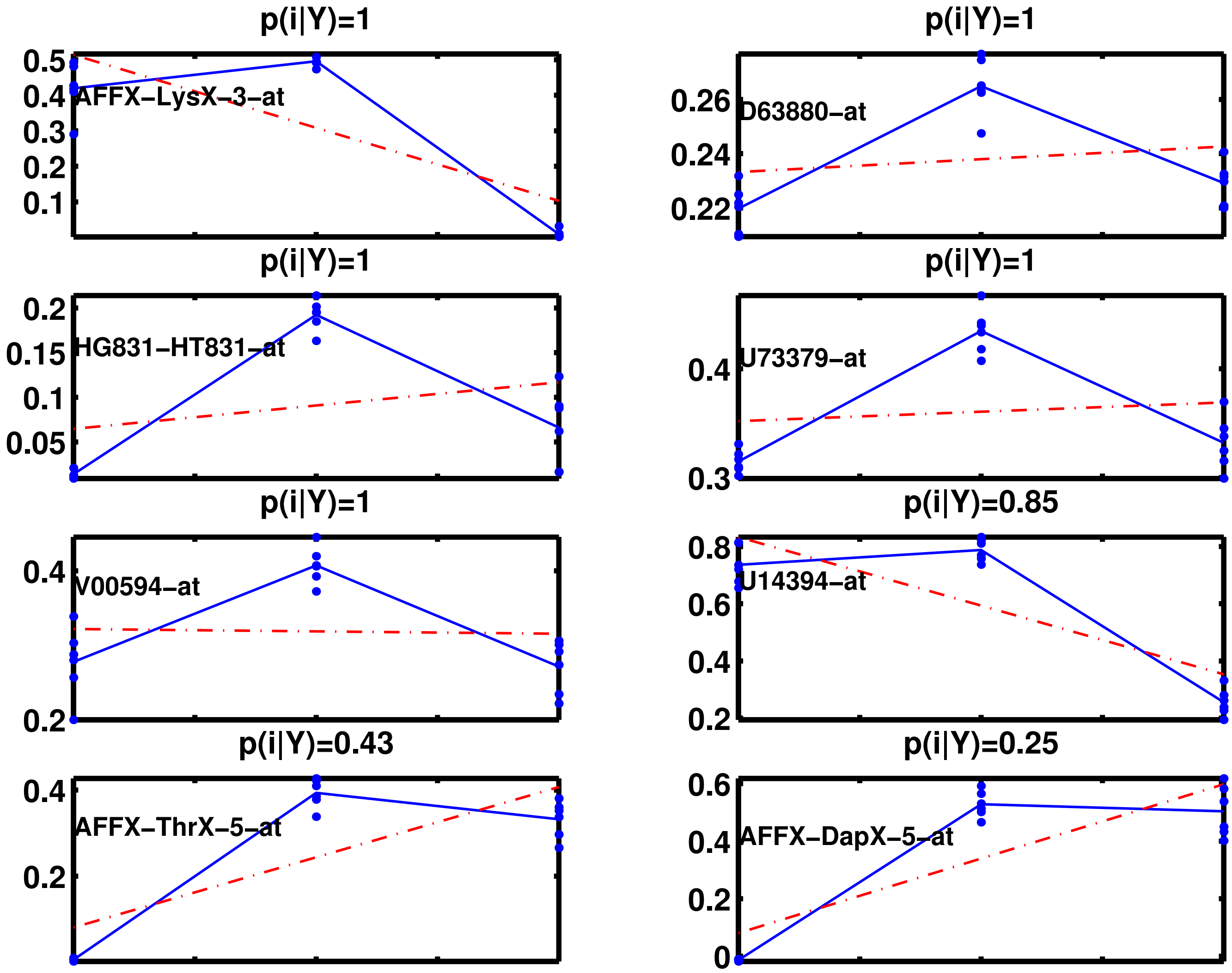

Figure 7: *The 8 top ranked cross-validated gene profiles remaining on the first Pareto front among the non-linear genes in Fig. 6. $P(i|Y)$ denotes the relative frequency that each resampled (leave-one-out resamping) profile is Pareto-optimal according to the non-parametric slope-sign criteria. Dashed line is the linear regression line.*

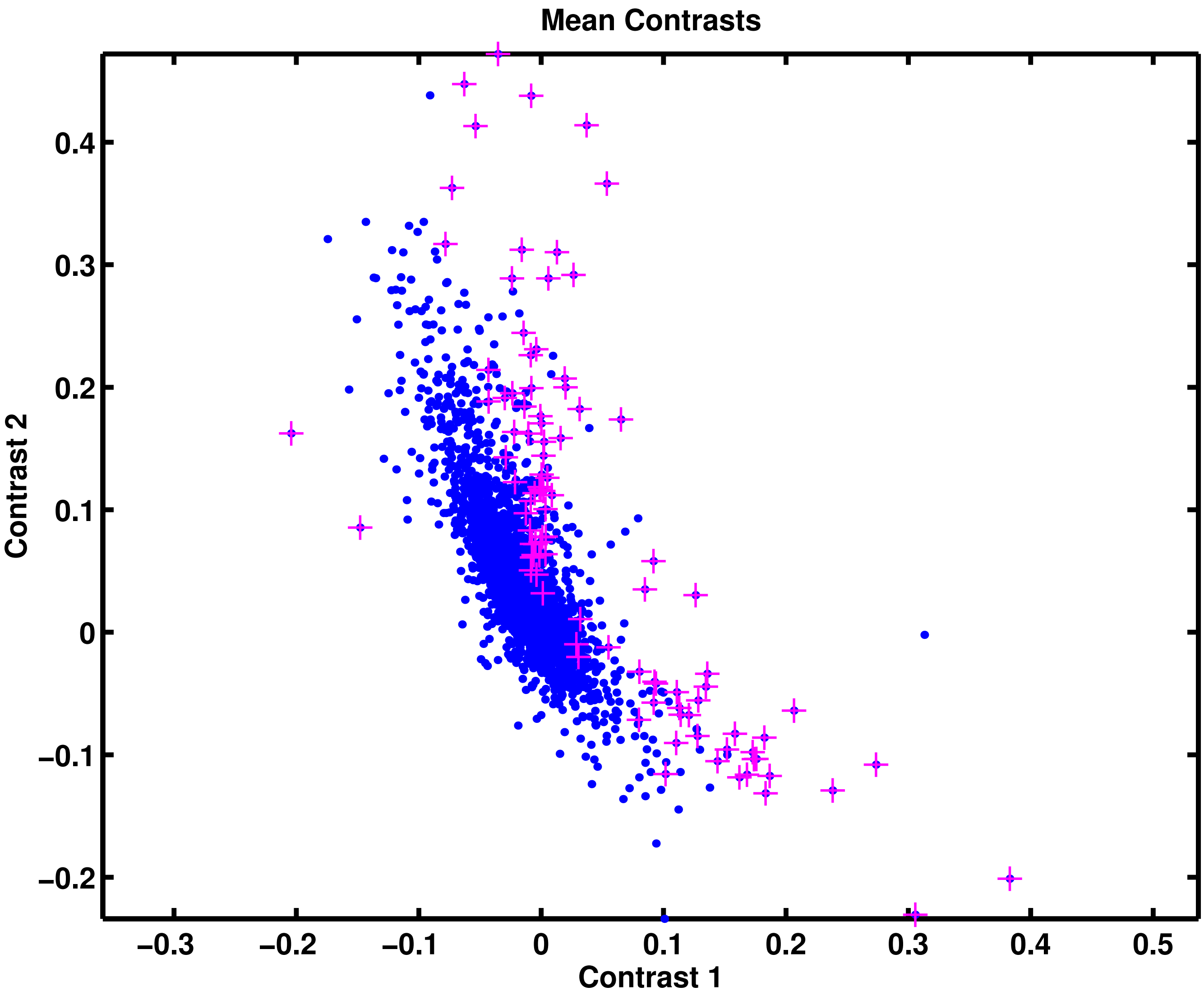

Figure 8: *Multicriterion scattergram corresponding to 4 with contrast matrix* $A = [-1, 1, 0; -1, -1, 2]$. *Crosses again indicate the 98 genes having non-linear profiles at a p-value of 0.1. The contrast A is designed to filter genes with significant inverted-V shaped profiles and the scattergram simply corresponds to rotating Fig. 4 by* 90^o.

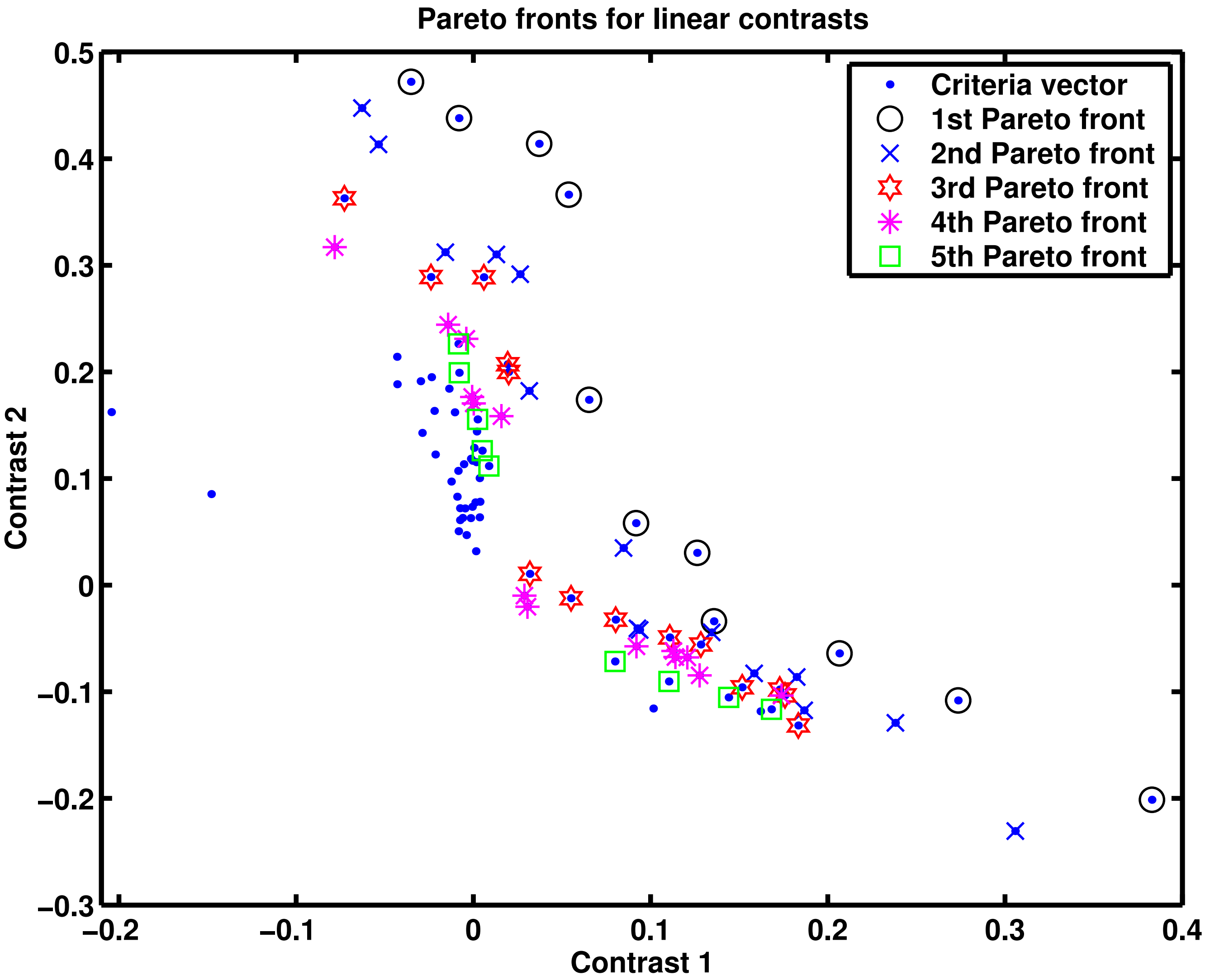

Figure 9: *The first five Pareto fronts for the genes with non-linear profiles shown in Fig. 8.*

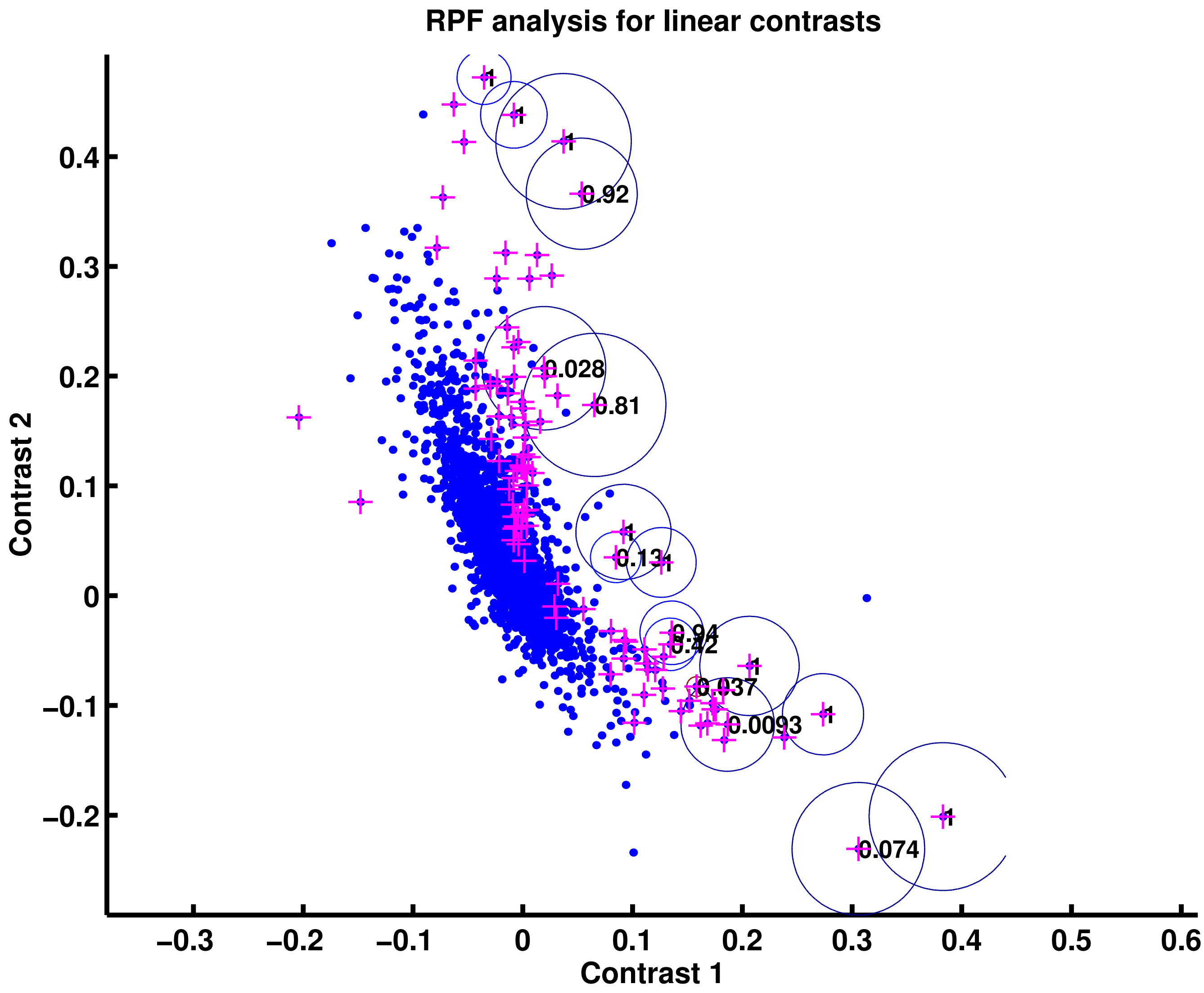

Figure 10: *17 genes which belong to the first Pareto front with non-zero probability, computed by cross-validation analysis applied to Fig. 9. Constant contours around each point indicate standard errors under equal variance hypothesis and the relative frequencies of lying on the first Pareto front are indicated at the center of relevant contours.*

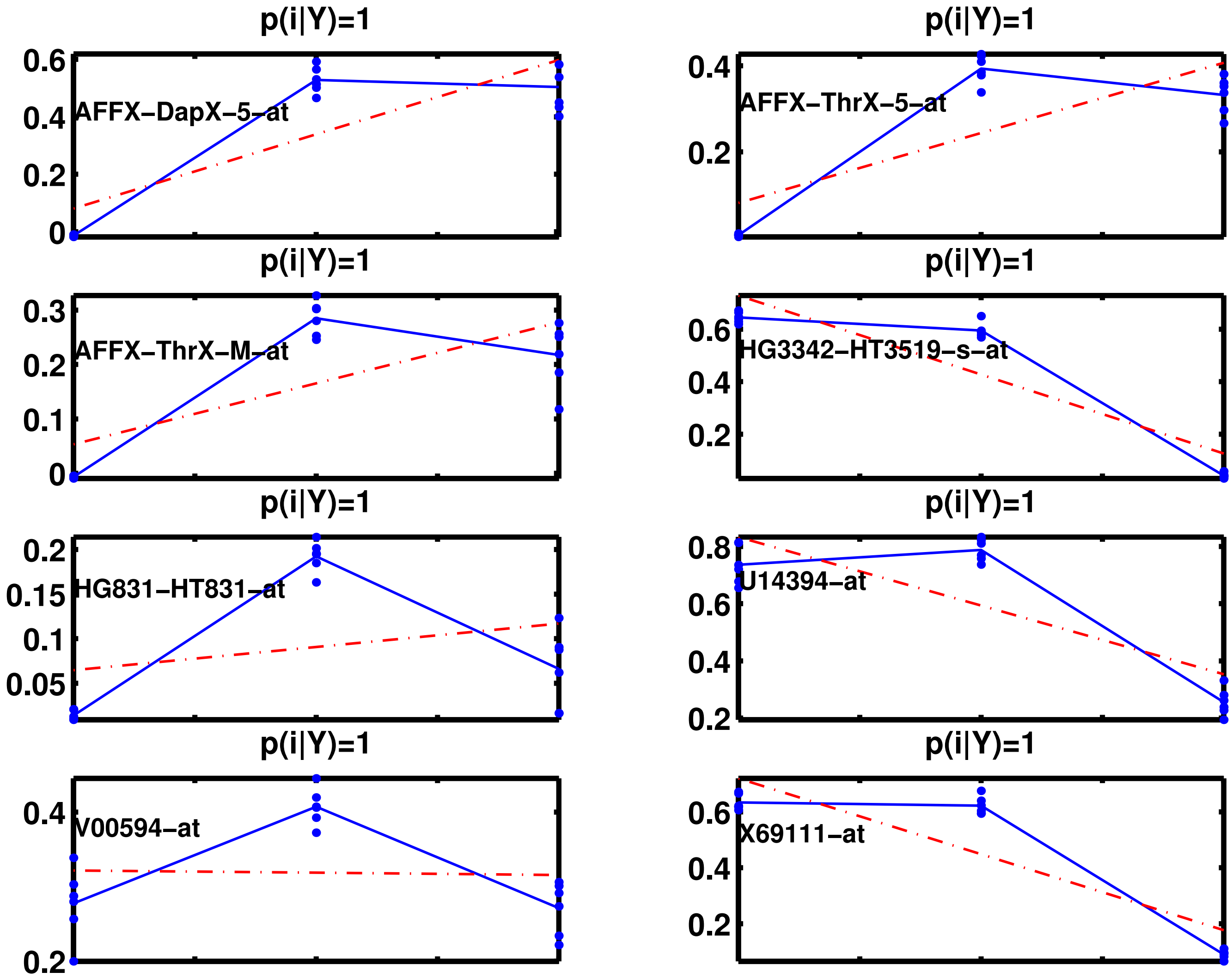

Figure 11: *The 8 top ranked cross-validated gene profiles remaining on the first Pareto front among the non-linear genes in Fig. 8. $P(i|Y)$ denotes the relative frequency that each resampled (leave-one-out resamping) profile is Pareto-optimal according to the two linear contrast criteria. Dashed line is the linear regression line.*

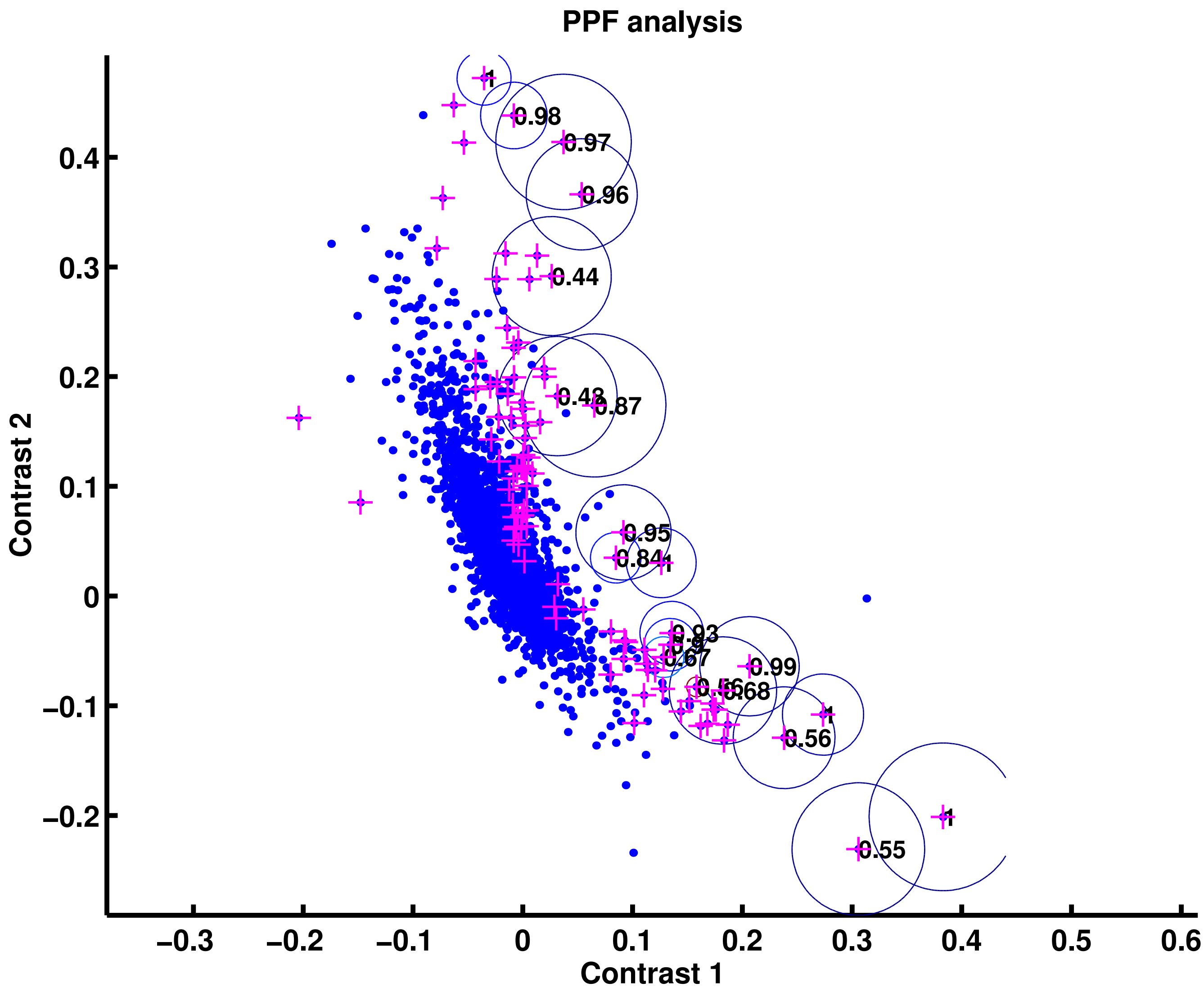

Figure 12: *Same as in Fig. 10 for the linear contrast PPF method along with standard error constant contours and posterior probabilities of belonging to the first Pareto front (prior parameter $a = 2$ used to compute posteriors). For clarity, only the first 20 top ranking genes are shown.*

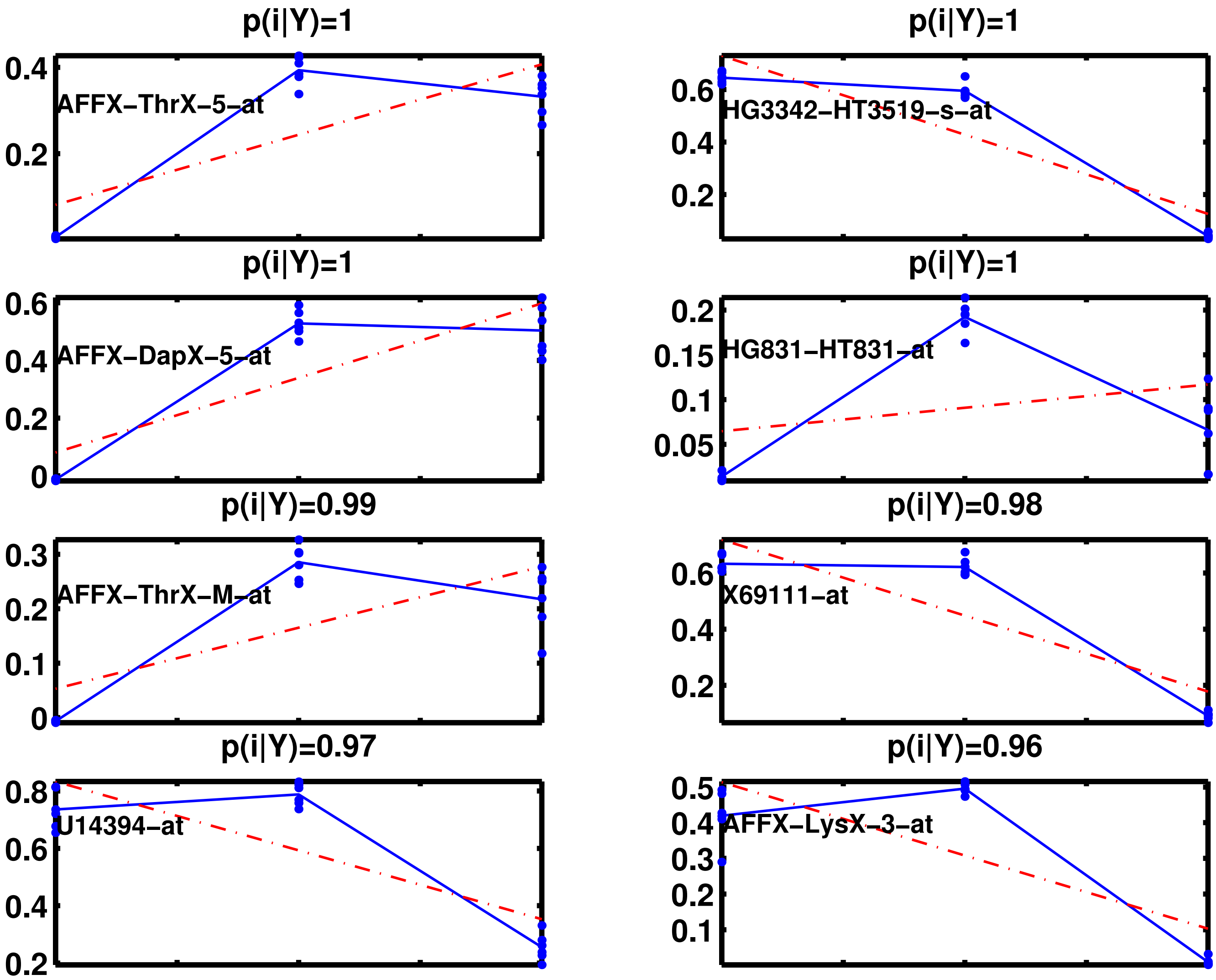

Figure 13: *Same as Fig. 11 except that gene profile ranking is according to computed PPF posterior probabilities shown on Fig. 12. $P(i|Y)$ denotes the Bayes posterior probability that each profile is Pareto-optimal according to the two linear contrast criteria under the non-informative prior ($a = 2$).*