Pareto-Optimal Methods for Gene Ranking

Alfred O. Hero[1] and Gilles Fleury[2]

[1]University of Michigan, Ann Arbor, MI and [2]Ecole Supérieure d'Electricité, Gif-sur-Yvette, France

June 8, 2004

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 gene profiles. Many 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 analysis (PFA). 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. Both a model-driven Bayesian Pareto method and a data-driven non-parametric Pareto method, based on rank-order statistics, are presented. The methods are illustrated for two gene microarray experiments.

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

Accepted for publication in *Journ. VLSI Sig. Proc. - Special Issue on Genomic Sig. Proc.*, June. 2003: Corresponding author:

Prof. Alfred O. Hero, III

Room 4229, Dept. of EECS University of Michigan

1301 Beal Avenue

Ann Arbor, MI 48109-2122

Tel. (734) 763-0564

Fax: (734) 763-8041

email: hero@eecs.umich.edu

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 [31, 25, 30] or [11] 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, called gene expression profiles. 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* which can be divided into two tasks: gene screening, aiming to specify a list genes with uncommon expression at some level of statistical significance, and gene ranking, aiming to rank order genes on this list. The most common approaches to gene screening are significance tests implemented by thresholding a set of standard test statistics, e.g. one sided tests of profile correlation to a template, 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 [40] or the Microarray Suite and Data Mining Tool (DMT) distributed by [2]. The present paper is concerned with applying multiple objective optimization to gene ranking. A companion paper [20] develops a similar approach for gene screening.

Significance tests can easily be extended to ranking the list of screened genes, e.g., in decreasing order of statistical significance according to observed p-values. For example, to rank gene profiles according to similarity to a given template one can rank genes in decreasing order of measured profile-to-template correlation coefficient. These types of ranking methods are based on optimizing single fitness criterion. A problem with this single-criterion ranking method is that it is often difficult for the molecular biologist or experimenter to articulate what attributes he is seeking in terms of a single quantitative criterion. It is usually easier for him to specify qualitative aspects of the profiles of interest to him, e.g., monotone increasing or decreasing in the beginning, bumpy in the middle, and flat at the end. In such a situation, it is more natural to try to capture each aspect separately with multiple criteria.

This paper proposes a systematic multiple criterion approach to gene ranking, denoted Pareto-optimal ranking, 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 [4, 5] and the Pareto website of the New School [44]). As discussed in Steuer [41] this principle has since been applied to many other fields. Since V. Pareto's name has

many other associations in probability and statistics, it is important to emphasize that the proposed method of Pareto-optimal gene ranking 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. Sobel [38] was the first to publish a statistical application of Pareto optimality. Here we propose an extension, using Bayesian and bootstrap formulations, with applications to gene ranking in microarray experiments.

To apply Pareto-optimal gene ranking the experimenter computes a number of fitness criteria for each gene, generating a point cloud of criterion vectors which we call the *multicriterion scattergram*. For example, to select the most monotonic profiles over time the fitness criteria might be chosen as the differences in gene expression level over successive time points. The objective of Pareto-optimal ranking is to isolate genes that achieve a compromise between maximizing (or minimizing) the competing gene-fitness 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, i.e., a partial ordering, of the highest scoring gene profiles. In two recent conference papers [14, 15] we applied Pareto-optimal ranking to discover young- and old- dominant mouse retina genes in Affymetrix GeneChip experiments and the discovered genes were validated using RT-PCR techniques. The purpose of the present paper is to present the general Pareto-optimal ranking methodology, introduce a Bayesian formulation of Pareto-optimal ranking, and to illustrate this approach on a widely available data set created expressly for testing algorithms for gene screening, classification, and quantification of differential expression [26].

As the microarrays are obtained from a random sample of the population there can exist substantial statistical sampling errors that complicate the Pareto-optimal analysis. These sampling errors can be handled by cross-validation producing what is called a *resistant Pareto front* (RPF) of genes, defined as those genes that land on the Pareto front with high relative frequency under re-sampling of the microarrays. The RPF method is completely data-driven and as such it does not rely on any distributional assumptions on the data. Thus it is very flexible, allowing treatment of arbitrary fitness criteria such as dependent and non-linear functions of the data. As an example we present a non-parametric RPF method which is computed on rank-order statistics of the probe responses of the microarrays. Of course when the data distribution can be characterized, even approximately, data-driven methods have 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 sizes. To address these drawbacks a Bayesian approach is presented for Pareto-optimal gene ranking: 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 asymptotic approximations to extreme-value distributions we obtain an expression for the posterior probability whose complexity increases in the number of fitness criteria and not in the number of samples. The Bayesian model that we use for the expression indices and their means and variances is similar to the conditionally Gaussian with conjugate prior model used recently by Lönnstedt and Speed in [32].

We apply our Bayesian PPF analysis to a set of fitness criteria defined as linear functions, a matrix of *profile contrasts*, of the prior mean expression levels of each gene profile. For illustration, PPF and RPF analyses are applied and compared on Fred Wright's data set, described in [26], for detection of the most aberrant genes violating linearity in the Affymetrix human fibroblast mixture experiment. The specification of the set of most aberrant non-linear genes could be useful for an experimenter who wants to choose a few egregious genes on which to perform an expensive followup study, e.g., RT-PCR analysis. Our results show concordance between the genes selected by RPF and PPF analysis which suggests that the PPF is insensitive to the fairly restrictive model assumptions made.

It is worthwhile mentioning that, despite some superficial similarities, the concept of Pareto fronts is fundamentally different from John Tukey's notion of data depths and contours of depth in a multivariate sample [45, 12]. Data depths are induced by a sequence of nested convex hulls which contain smaller and smaller proportions of the sample as the depth increases. Similarly to Pareto fronts, the contours of these successive convex hulls induce a (partial) ordering or fitness on points in the sample. However, the data depths and their contours differ from the Pareto fronts in several important respects. The Pareto front defines a partial ordering relative to the non-dominated points, as measured by user-specified fitness criteria, while the data depth defines a partial ordering relative to a single point at the center of the sample, the "multivariate median." For example, while the 1-st data depth defines the entire shape of the sample the 1-st Pareto front only describes the shape of a side of the sample, namely the side having points with higher fitness scores. Furthermore, Pareto fronts are not in general convex while data depth contours are always convex.

The outline of the paper is as follows. In Sec. 2 a brief review of microarray data analysis is presented and in Sec. 3 non-statistical Pareto-optimal gene ranking approach is introduced. In Sec. 4 the data-driven RPF analysis methodology is described. In Sec. 5 the general PPF gene ranking method is developed and in Sec. 6 different profile contrast functions are considered. Finally in Sec. 7 PPF analysis is applied to finding aberrant

genes in Fred Wright's human fibroblast mixing data.

2 Gene Analysis from Microarray Data

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 expression levels. 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 [23, 9, 6, 13]. 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. 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 [1] product line. An example of the later is the cDNA microarray protocol of the National Human Genome Research Institute (NHGRI) [35]. A suite of software tools are available from Affymetrix and elsewhere for extracting accurate estimates of abundance, called expression indices. Computation of these indices can range from simple unweighted sample averaging, as in the Affymetrix MAS4 software, to more sophisticated model-based analyses, such as the Li-Wong method [27, 28]. Many of the more sophisticated packages are available as freeware, e.g., see Strimmer's website [43] for links to relevant software written in the R software language.

The study of differential gene expression between T populations requires hybridizing several microarrays from each population to reduce response variability. Define the expression index extracted from the m-th microarray at time t and at the n-th gene chip probe location

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

When several microarray experiments are performed over time they can be combined in order to find genes with interesting expression profiles. This is a gene screening problem to 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 [19, 3] and [8]). Validation methods have been widely used and include: significance analysis of microarrays (SAM); bootstrapping cluster analysis; and leave-one-out cross-validation (See [46] and [24]). Many of these methods are based on optimizing some single fitness 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.

In a well designed gene microarray experiment gene screening methods, e.g., paired t-tests and Fisher test of variance, will generally result in a large list of genes and the biologist must next face the problem of selecting a few of the most "promising genes" for further investigation out of this list. Resolution of this problem is of great importance since validation of gene response requires more sensitive techniques, such as RT-PCR, which have lower throughput than microarrays and are thus much more time consuming and expensive [9, 6]. Some sort of rank ordering of the list of genes would help guide the biologist to a solution. Thus gene filtering almost always boils down to a two stage procedure: (1) gene screening to determine a statistically significant list of uncommon genes profiles; and (2) gene ranking to order this list in decreasing order of interest to the molecular biologist. The focus of this paper is (2). Multiple criterion approaches to (1) are the focus of other work [20].

3 Multiple Objective Gene Ranking

As contrasted to maximizing *scalar* criteria, multiple objective gene ranking 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 books by Stadler or [39] or Steuer [41]). Pareto optimality theory has been applied to a wide range of application areas including: economics, sociology, psychology, operations research, evolutionary computing and subset selection among multivariate populations (See above referenced books, and articles by Sobel [38], Zitler and Thiele [48] and Arrow and Hervé [5]).

Multi-objective gene ranking 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. When there is no risk of confusion we will use the simpler notation $\xi_p(i)$ for $\xi_p(\underline{\mu}(i))$. Let a group of experimenters agree on P gene selection criteria which, when applied to a given gene, 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 $\xi_p(\underline{\mu}(i)) > \xi_p(\underline{\mu}(j))$. When it is desired to filter out highly expressed and/or 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 the measured profile of the i-th gene has vector mean $\underline{\mu} = \underline{\mu}(i)$ for which ξ_1 and ξ_2 are the largest over all genes then this gene would be of obvious interest to the experimenter. However, there may be many other genes that could interest the experimenter, e.g. those where ξ_1 is large but ξ_2 is only moderate or vice-versa. Furthermore, if the criteria are in conflict then no single gene may simultaneously maximize ξ_1 and ξ_2. To capture a set of genes of interest, one might consider thresholding a compound scalar ranking criterion, e.g. the weighted arithmetic average of (1)

$$J_\alpha(\underline{\mu}) = \alpha(\mu_2 - \mu_1) + (1-\alpha)(\mu_2 + \mu_1). \tag{2}$$

Of course, if μ_1 and μ_2 are positive valued and a proportional increase in the profile is more meaningful to the experimenter then he might prefer the log criteria $\xi_1(\underline{\mu}) = \log \mu_2/\mu_1, \xi_2(\underline{\mu}) = \log\sqrt{\mu_2\mu_1}$, and $J_\alpha(\underline{\mu}) = \alpha \log(\mu_2/\mu_1) + (1-\alpha)\log\sqrt{\mu_2\mu_1}$. In either case, when $\alpha = 0$ or 1 maximizing this compound criterion would yield the two most fit genes under criteria ξ_1 or ξ_2.

An obvious issue that arises in selecting a scalar criterion J_α is: what is the most suitable choice of the weight α? Two experimenters, A and B, may not have selected the same weight factor α and therefore one of them would not necessarily be satisfied by the significance of the genes reported by the other. One way out of this dilemma is to find the entire set of genes which maximize J_α for some choice of α. This would give a set of genes that would be guaranteed to contain the favorite gene of all experimenters. It turns out that this set of genes are contained in a set called the *Pareto front* which results from multiple objective optimization of the pair $[\xi_1(i), \xi_2(i)]^T$ over i [10].

Multiple objective optimization captures the intrinsic compromises among possibly conflicting objectives in a natural way. Consider the multicriterion scattergram in Fig. 1 and suppose that fitness criteria ξ_1 and ξ_2 are to be maximized. Gene D is *dominated* by both gene A and gene B since gene D has lower fitness in both criteria ξ_1 and ξ_2. Likewise gene E is dominated by gene B and gene C. On the other hand genes A, B and C are not dominated by any other gene and are therefore preferable to genes D and E. Multi-objective ranking uses this non-dominated property as a way to establish a preference relation among genes given a set of criteria $\{\xi_q\}_q$. More formally, gene i is said to be dominated if there exists some other gene $g \neq i$ such that for at least one q

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

The set of non-dominated genes are defined as those genes that are not dominated. All the genes which are non-dominated constitute a set of points called the (first) Pareto front. A second Pareto front can be obtained by stripping off the points on the first front and computing the Pareto front on the remaining points. For the example in Fig. 1 the first Pareto front is $\{A, B, C\}$ and the second Pareto front is $\{D, E\}$.

The above multiple criterion ranking 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. Despite its obvious potential for improvement over single criteria optimization methods, to our knowledge Pareto front analysis has not been previously applied to gene ranking or to more general data mining problems. We speculate that this might be due to the unreliability of the non-statistical Pareto front technique when applied to noisy observations and to the lack of systematic methods for dealing with statistical uncertainty. We propose two methods for handling statistical uncertainty: cross-validation leading to resistant Pareto front (RPF) analysis, and Bayes smoothing, leading to posterior Pareto front (PPF) analysis.

4 Resistant Pareto Front Analysis

The idea behind resistant Pareto front (RPF) analysis is a simple case of leave-one-out cross validation but requires some notation to explain succinctly. Let $\hat{\underline{\xi}}_p^{(M_1,\cdots,M_T)}(n)$ denote an empirical estimator of the fitness criterion vector $\underline{\xi}(n)$ for the n-th gene using the entire sample population. Let $\hat{\underline{\xi}}^{(-m_1,\cdots,-m_T)}(n)$ denote the same empirical estimator computed on a reduced population obtained by omission of the m_t-th sample from each time point $t = 1, \ldots, T$, $m_t \in \{1, \ldots, M_t\}$. For a given $m_1, \ldots m_T$ we call this a leave-one-out estimator. When the sample population consists of independent sub-populations at different time points there will be a total of $\prod_{t=1}^{T} M_t$ different leave-one-out estimates of $\underline{\xi}(n)$. For each leave-one-out estimate $\hat{\underline{\xi}}^{(-m_1,\cdots,-m_T)}(n)$ find the Pareto front of genes. Define the indicator function $\Delta^{(-m_1,\ldots,-m_T)}(n)$ =1, if gene n is on the Pareto front and = 0, otherwise. Finally, compute the relative frequency scores

$$\mathrm{RF}(n) = \frac{\sum_{m_1=1}^{M_1} \cdots \sum_{m_T=1}^{M_T} \Delta^{(-m_1,\ldots,-m_T)}(n)}{\prod_{t=1}^{T} M_t}, \qquad n = 1, \ldots, N.$$

These relative frequency scores are then used to rank the genes in decreasing order of likelihood of belonging to the Pareto front. This procedure can be repeated for the second order and higher Pareto fronts to generate scores for the relative frequency that each gene lies on the first 2 or more Pareto fronts.

In [15] we applied the RPF procedure described above to filter a set of $N = 12,422$ genes obtained from an Affymetrix GeneChip study of retinal tissues of a population of 24 mice grouped into $T = 6$ time points (between postnatal 2 days (Pn2) through month 21 (M21)) each time point having data from $M_t = 4$ microarrays. A representative sample of the data for 4 different genes is shown in Fig. 2 which indicates a variety of gene expression profiles over time. The objective was to extract "aging genes," i.e. genes that demonstrated a marked and steady increase in expression level over time. First a set of $M^T = 4096$ time trajectories were defined for each gene, corresponding to all possible time paths through the sets of 4 samples at each of 6 time

points. For illustration three of these possible trajectories are shown for a specific gene in Fig. 3. 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 1296 sign profiles summarize the monotonic properties of a gene's temporal evolution pattern. For each gene three criteria were then computed including: 1) the proportion $\hat{\xi}_1$ of the 1296 trajectories that are monotonic; 2) the overall change $\hat{\xi}_2$ in expression level as measured by the difference between the first ($t = 1$) and last ($t = T$) time points; and 3) the negative curvature $\hat{\xi}_3$ of the profile computed as the average second order difference between all sets of three adjacent time points. The monotonicity criterion 1) is closely related to the well known Jonckheere-Terpstra (JT) test statistic [22, 21] for testing monotonic trends in multivariate samples. Like the JT test statistic, our monotonicity criterion is distribution free. However, our test is a more stringent test of monotonicity and does not suffer from the rank inversion property of the JT test [18].

The 3D multicriterion scattergram of the full-sample criterion vector $\hat{\underline{\xi}}^{(4,\ldots,4)}(n)$ is illustrated in Fig. 4 along with the (first) Pareto front consisting of over 100 genes. A more stringent gene ranking procedure is to intersect the Pareto fronts of all 3 possible 2D multicriterion scattergrams formed from pairs of fitness criteria, see Fig. 5 for illustration. When using all of the microarray data only one gene was found to lie on the intersection of these fronts. This Pareto-optimal gene trajectory is shown in Fig. 6. More genes were found by implementing the RPF cross-validation technique to determine the number of times each gene appears in one of the first ten intersecting Pareto fronts via the relative frequency scoring procedure described earlier. For more details see [15]. The result of this analysis yielded several strongly monotonic increasing genes which have been subsequently validated experimentally using RT-PCR analysis.

5 Posterior Pareto Front Analysis

The posterior Pareto front (PPF) analysis introduced here is based on a Bayesian perspective and can offer a lower complexity alternative to the RPF procedure described in the previous section. 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:

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

$$= \int dP(\underline{\xi}(i)|Y) \prod_{j \neq i} P\left(\left\{\underline{\xi}(i) \leq \underline{\xi}(j)\right\}^c | Y, \underline{\xi}(i)\right)$$

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

$$p(i|Y) = \int d\underline{u} f_{\underline{\xi}(i)|Y}(\underline{u}) \prod_{j \neq i} \left[1 - P\left(\underline{u} \leq \underline{\xi}(j)|Y\right)\right]. \tag{3}$$

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

$$p(i|Y) = \int \int du_1 du_2 f_{\xi_1(i),\xi_2(i)|Y}(u_1, u_2) \prod_{j \neq i} \left[F_{\xi_1(j)|Y}(u_1) + F_{\xi_2(j)|Y}(u_2) - F_{\xi_1(j),\xi_2(j)|Y}(u_1, u_2)\right], \tag{4}$$

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

5.1 Application to Gene Ranking

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 (3) can be computed. This is a similar Bayesian setup as used in the empirical Bayes approach to microarray analysis of Lönnstedt and Speed [32]. However, as contrasted to the conjugate prior adopted in [32], here we will adopt the simpler non-informative prior as described in Geisser and Cornfield [17]:

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

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 case.

5.1.1 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

$$f_{\mu_t(i)|Y}(u) \quad = \quad k(Y_{ti}) \left(1 + \frac{(u - \hat{\mu}_t(i))^2}{\hat{\sigma}_t^2(i)}\right)^{-(M-a+2)/2}, \tag{5}$$

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 [17]:

$$k(Y_{ti}) = \frac{1}{\hat{\sigma}_t(i)\sqrt{\pi}} \frac{\Gamma(\frac{1}{2}(M-a+2))}{\Gamma(\frac{1}{2}(M-a+1))}. \tag{6}$$

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 (5) 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.

5.1.2 Constant 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

$$f_{\underline{\mu}(i)|Y}(u_1, \ldots, u_T) \quad = \quad k(Y_i) \left(1 + \sum_{t=1}^{T} \frac{(u_t - \hat{\mu}_t(i))^2}{\hat{\sigma}^2(i)}\right)^{-(TM-a+2)/2}, \tag{7}$$

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 (6).

Analogously to the case of unequal variances, the associated distribution function can be approximated by a multivariate L_∞ approximation to (7):

$$F_{\underline{\mu}(i)|Y}(u_1, \ldots, u_T) \approx \left(1 + \sum_t \frac{(\hat{\mu}_t(i) - u_t)_+^2}{\hat{\sigma}^2(i)}\right)^{-(TM-a+2)/2}. \tag{8}$$

6 Profile Contrasts

Linear contrasts have been advocated for many different problems of multivariate statistical inference and experimental design [34, 29]. Here we adopt linear contrasts as multiple criteria for Posterior Pareto ranking. The simplest contrasts are the time sampled means themselves $\xi_p(i) = \mu_p(i)$, $p = 1, \ldots, T$ which can be called the amplitude profile criterion. In the case of time varying variances using the expressions (5) and (7) in (4) 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).

A more flexible criterion are various contrasts between time means. In particular define the vector criterion $\underline{\xi}(i) = [\xi_1(i), \ldots, \xi_P(i)]^T$ as the linear function of the mean profile vector:

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

where $\mathbf{A} = ((a_{ij}))$ 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 $\mathbf{A}$ 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 ($\mathrm{var}(y_{tm}(i)) = \sigma^2(i)$) a weaker sufficient condition is that the rows of $\mathbf{A}$ be orthogonal since the joint density $f_{\underline{\mu}(i)|Y}(\underline{u})$ in (7) is invariant to orthogonal transformations of $\underline{u} - \underline{\hat{\mu}}(i)$.

As examples consider the following $P \times T$ contrast matrices

$$\mathbf{A}_3 = \begin{bmatrix} -1 & 0 & 1 \\ 1 & -2 & 1 \end{bmatrix}, \qquad \mathbf{A}_3^{'} = \begin{bmatrix} -1 & 1 & 0 \\ -1 & -1 & 2 \end{bmatrix}.$$

$$\mathbf{A}_4 = \begin{bmatrix} -1 & 1 & 0 & 0 \\ -1 & -1 & 2 & 0 \\ -1 & -1 & -1 & 3 \end{bmatrix}, \qquad \mathbf{A}_4^{'} = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & -1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}.$$

Applying posterior Pareto front analysis to $\underline{\xi}(i) = \mathbf{A}_3\underline{\mu}(i)$ will extract 3 time-point gene profiles which are end-to-end increasing (large ξ_1) and have large positive curvature (large ξ_2). If $\mathbf{A}_3$ is replaced with $\mathbf{A}_3^{'}$ then the analysis will find profiles which are monotonic increasing. For 4 time-points $\mathbf{A}_4$ will perform similar services as $\mathbf{A}_3$ while $\mathbf{A}_4^{'}$ will filter out "mexican hat" profiles. Note if the noises $\{\epsilon_{mt}(i)\}_{mt}$ are i.i.d. Gaussian then the linear contrasts are also independent and Gaussian as the above contrast matrices are orthogonal.

Of interest are general ways to construct meaningful contrast matrices $\mathbf{A}$ which are orthogonal, 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 $\mathbf{B}$ whose rows capture some set of desired linearly independent properties of the profile and then apply the PPF with the orthogonalized contrast matrix $\mathbf{A} = [\text{chol}(\mathbf{B}\mathbf{B}^T)]^{-1}\mathbf{B}$, where $\text{chol}(\mathbf{B}\mathbf{B}^T)$ is the Cholesky decomposition of $\mathbf{B}\mathbf{B}^T$. For example the following (non-orthogonal) matrix might be proposed as more natural for capturing strongly monotone increasing profiles

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

It turns out that the aforementioned Cholesky orthogonalization procedure yields the orthogonal matrix:

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

which is equal (up to a left multiplication by a positive diagonal matrix) to the contrast matrix $\mathbf{A}_3^{'}$.

7 Application to Dilution Experiment

To illustrate the application of Bayesian PPF analysis and data-driven RPF analysis we used these methods to find and rank the non-linear gene profiles in Fred Wright's dataset. This dataset is described in the paper by Lemon *etal* [26] and is available at the web address provided in the citation. Fred Wright's data set was obtained from a dilution experiment which the authors designed for empirically validating and comparing various differential gene expression methods of analysis. As explained in [26] 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. The probe responses (hybridization levels) on a total of 18 chips were processed. These 18 chips correspond to 6 replications within each of the 3 populations mentioned above. Each HuGeneFL chip contains the same 7129 gene probes. For each gene probe the sequence of probe responses from the "stimulated(t=1)," "50-50(t=2)," and "starved(t=3)," populations was defined, in that order, as a gene expression profile. For this type of dilution experiment 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 (most probably [47]) hybridization saturation. A typical set of expression indices is shown in Fig. 7.

We used the Li-Wong reduced expression indices derived in [26] for our analysis. Our objective is to determine the most peaked inverted V-shaped (concave) gene profiles in the dataset. The inverted V-shaped profiles are those genes whose expression increases over $t = 1$ to $t = 2$ followed by a decrease over $t = 2$ to $t = 3$. Genes whose profiles have the highest peak at $t = 2$ most severely violate the linearity assumptions among the concave

profiles. We applied a simple procedure to screen for non-linear gene profiles before performing the Pareto front analysis. Specifically, the probe responses in each gene expression profile were 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 assumed i.i.d. Gaussian additive noise with variance $\sigma^2(i)$ and a, b are undetermined 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 subspace $\{y \in \mathbb{R}^3 : y = a[1,2,3] + b[1,1,1]\}_{a,b \in \mathbb{R}}$, 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. The quantity $s(i)$ is the (pooled) sample variance estimate of σ^2. Under the linear profile hypothesis the statistic $F(i) = R(i)/s(i)$ is distributed as Fisher-F on 2 and $M - 3$ degrees of freedom [37]. Based on the observed p-values of the Fisher F statistics we determined that no genes fail the linearity test $F(i) = R(i)/s(i) \leq \gamma$ at any positive false discovery rate (FDR) [7] according to the FDR procedure of [36]. Nonetheless, as our purpose is to illustrate PFA, we first screened for genes that satisfied $F(i) > 5.5$. This threshold corresponds to a (single comparison) significance level of $p = 0.1$. This screening eliminated all but 98 profiles to which the PFA analysis was applied.

7.1 Linear-Contrast Pareto Analysis

First we performed linear-contrast Pareto analysis using the following orthogonal contrast matrix

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

to generate two criteria with which to rank the inverted-V shaped profiles. Note that when applied to the i-th gene's sample-mean probe response vector $\underline{\hat{\mu}}(i) = [\hat{\mu}_1(i), \hat{\mu}_2(i), \hat{\mu}_3(i)]^T$, this yields the two statistically independent contrasts

$$\begin{aligned} \xi_1(i) &= \hat{\mu}_2(i) - \hat{\mu}_1(i) \\ \xi_2(i) &= -2(\hat{\mu}_3(i) - (\hat{\mu}_2(i) + \hat{\mu}_3(i))/2). \end{aligned}$$

As desired both contrasts are positive when the profile $\underline{\hat{\mu}}(i)$ is concave. Figure 8 displays the associated multi-criterion scattergram. The crosses in the figure indicate the 98 non-linear genes.

While we have investigated many different values for the prior PPF parameter a, we only present results for $a = 2$ here. We have observed that 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 sample mean contrasts of all microarray data. Figure 10 show the results of PPF 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 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. Figure 11 show the eight top scoring trajectories after PPF analysis. In each sub-panel the indicated piecewise linear line passes through the means of the 6 replicates of each of the 3 time samples.

A final remark concerns the relation between our contrast-based Pareto ranking approach and a simple template matching approach, which we call matched filtering (MF), to gene ranking. The MF approach ranks gene profiles according to their correlation to a template; a reasonable selection is the symmetric concave' profile $[-1, 2, -1]$. The MF approach is equivalent to using the single ranking criterion $\xi(i) = 2\hat{\mu}_2(i) - \hat{\mu}_1(i) - \hat{\mu}_3$. It is easily verified that $\xi = \alpha\xi_1(i) + \beta\xi_2(i)$ for positive constants $\alpha = 2.1213, \beta = 1.2247$, where ξ_1, ξ_2 are the contrasts that we used in the Pareto gene filter. Therefore, Pareto ranking is a generalization of matched filtering. Indeed, as explained above, Pareto ranking is a method that ranks profiles according to the whole family of templates described by the cone $\{\xi = \alpha\xi_1 + \beta\xi_2,\ \alpha, \beta > 0\}$.

7.2 Non-Parametric RPF analysis

For comparison we investigated a fully non-parametric data-driven Pareto analysis based on rank-order statistics. Rank order methods of microarray analysis are popular since they are distribution-free and avoid amplitude dependent biases and circumvent the need for microarray amplitude normalization. On the other hand such methods sometimes incur a loss in sensitivity for small sample sizes. The rank-order Pareto front procedure that we used is as follows. For each microarray we computed the rank-order of each gene according to its hybridization score, determined by the extracted Li-Wong indices as above, relative to all other genes on the microarray. Specifically, we used the Matlab command `[s,yr]=sort(y)` where `y` is the 7129×18 matrix whose columns are gene expression indices for each of the 18 microarrays (3 treatment groups of 6 samples each). The resulting 7129×18 matrix `yr` of integers from 1 to 7129 was then used to perform screening of non-linear genes, similarly to above, and subsequently to perform Pareto analysis under the following two criteria. The first criterion $\hat{\xi}_1(n)$ is the difference between the mid-point and the average of the two other points in the mean rank-order profile of gene n (Matlab command `[mean(yr(:,1:6)');mean(yr(:,7:12)');mean(yr(:,13:18)')]' * A'`). The second criterion is the number of possible rank-order profiles whose shapes match an inverted-V profile. Specifically, for each gene we generate all $6^3 = 216$ possible trajectories through the 3 sets of 6 replicated

measurements of hybridization levels. The proportion of these trajectories which have slope of positive sign followed by slope of negative sign is the second criterion $\hat{\xi}_2(n)$.

In Fig. 12 the multicriterion scattergram is displayed. Figure 13 shows the first five Pareto fronts computed on the full set of 3×6 non-linear gene samples indicated as crosses on Figure 12. Leave-one-out cross validation was performed to determine the resistant genes for which a high proportion of the 216 re-sampled 3×5 trajectories remained on the first Pareto front. Fig. 14 shows the top 8 resistant inverted V-shaped profiles ranked in terms of relative frequency of remaining on the first front.

The top ranked 25 gene profiles under each criterion are shown in Table 15 along with their probability scores. Also included for comparison to the PPF analysis is a linear-contrast RPF analysis. Only 17 genes obtained positive scores under the linear contrast RPF analysis (middle column of figure). The linear-contrast RPF analysis is a leave-one-out cross-validation procedure applied to the same linear contrasts (matrix $\mathbf{A}$) as adopted in PPF analysis. Observe that all of the 17 RPF genes appear in the first 25 of the PPF gene list: purely data-driven RPF (linear-contrast) analysis is concordant with the model-based Bayesian PPF analysis. This indicates that the performance of the PPF analysis is insensitive to the somewhat dubious assumptions (Gaussianity, independence, large M, and diffuse prior) under which the PPF posterior probabilities were derived. On the other hand, the non-parametric RPF analysis reveals 3 highly ranked genes (U23435-s-at, AFFX-PheX-M-at and AFFX-LysX-M-at) which are not in the list of top 25 PPF ranking genes.

8 Conclusion

This paper introduced a new method of gene ranking 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 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 a data-driven cross-validation method, called RPF, and a model-driven Bayesian posterior Pareto method, called PPF, were presented for gene ranking. In contrast to the cross validation method 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 cross-validation method requires fewer assumptions and may be more robust to dubious model assumptions.

As for possible future work, a full bootstrap implementation of the RPF method would undoubtedly make it more outlier resistant. However this would greatly increase computational complexity. Methods of multiple

comparisons [33], which have been previously applied to differential analysis of gene microarrays by Storey *etal* [42] and others, also appear applicable to multicriterion ranking and, in particular, to validating Pareto-optimal trajectories. Finally, the multiple objective optimization approach described in this paper may be applicable to the PIDEX method of Ge *etal* [16] for combining pairs of gene selection criteria.

Acknowledgments

The authors are grateful for illuminating discussions of this work with Anand Swaroop, Shigeo Yoshida and Debashis Ghosh at the University of Michigan. The authors also thank Terry Speed, UC Berkeley, and Fred Wright, UCLA, for their comments on this work. This research was partially supported by a NATO grant that funded Gilles Fleury's sabbatical at the University of Michigan during the summer of 2001.

References

[1] Affymetrix. *NetAffx User's Guide*, 2000. `www.netaffx.com/site/sitemap.jsp`.

[2] Affymetrix. *Genechip software*. Affymetrix, Inc, 2002. `www.affymetrix.com/products/software/index.affx`.

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

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

[5] K. J. Arrow and R. R. Hervé, *Social Choice and Multicriterion Decision Making*, MIT Press, Cambridge MA, 1986.

[6] 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.

[7] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. Royal Statistical Society*, vol. 57, pp. 289–300, 1995.

[8] 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.

[9] P. O. Brown and D. Botstein, "Exploring the new world of the genome with DNA microarrays," *Nature Genetics*, vol. 21, no. 1 Suppl, pp. 33–37, Jan 1999.

[10] 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.

[11] 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.

[12] D. Donoho and M. Gasko, "Breakdown properties of location estimates based on halfspace depth and projected outlyingness," *Annals of Statistics*, vol. 4, pp. 1803–1827, 1992.

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

[14] G. Fleury, A. O. Hero, S. Yosida, 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.*, volume IV, pp. 4024–4027, Orlando, FL, 2002.

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

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

[17] S. Geisser and J. Cornfield, "Posterior distributions for multivariate normal parameters," *J. Royal Statistical Society, Ser. B*, vol. 25, pp. 368–376, 1963.

[18] W. Hager, "On testing a priori hypotheses about quantitative and qualitative trends," *Methods of psychological research*, vol. 1, no. 4, pp. 1–23, 1996.

[19] 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.

[20] A. Hero, G. Fleury, A. Mears, and A. Swaroop, "Multicriteria gene screening for analysis of differential expression with DNA microarrays," *EURASIP Journ. of Applied Signal Processing*, vol. 2004, no. 1, pp. 43–52, 2004. `www.eecs.umich.edu/~hero/bioinfo.html`.

[21] M. Hollander and D. A. Wolfe, *Nonparametric statistical methods (2nd Edition)*, Wiley, New York, 1991.

[22] A. R. Jonckheere, "A distribution free k-sample test against ordered alternatives," *Biometrika*, vol. 41, pp. 133–145, 1954.

[23] 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.

[24] 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, 2000. `citeseer.nj.nec.com/414709.html`.

[25] 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.

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

[27] C. Li and W. Wong, "Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection," *Proc. of Nat. Academy of Sci. (PNAS)*, vol. 98, pp. 31–36, 2001.

[28] C. Li and W. Wong, "Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application," *Genome Biology*, vol. 2, pp. 1–11, 2001.

[29] H. R. Lindeman, *Analysis of variance in experimental design*, Springer, New York, 1992.

[30] 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.

[31] 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.

[32] I. Lönnstedt and T. Speed, "Replicated microarray data," *Statistica Sinica*, vol. 12, pp. 31–46, 2002.

[33] R. G. Miller, *Simultaneous Statistical Inference*, Springer-Verlag, NY, 1981.

[34] D. F. Morrison, *Multivariate statistical methods*, McGraw Hill, New York, 1967.

[35] National Human Genome Research Institute (NHGRI). *cDNA Microarrays*, 2001. www.nhgri.nih.gov/DIR/Microarray.

[36] T. Nichols. *Software: FDR inference in Matlab*, 2002. www.sph.umich.edu/~nichols/FDR/.

[37] C. R. Rao, *Linear Statistical Inference and Its Applications*, Wiley, New York, 1973.

[38] M. Sobel, "On selecting the Pareto-optimal subset of a class of populations," *Commun. Statist. - Theory Meth.*, vol. 21, no. 4, pp. 1085–1102, 1992.

[39] W. Stadler, *Multicriteria optimization in engineering and the sciences*, chapter Fundamentals of multicriteria optimization, Plenum, New York, 1988.

[40] Stanford University. *SAM: Significance analysis of microarrays*. Stanford Office of Technology and Licencing, 2001. www-stat.stanford.edu/~tibs/SAM/.

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

[42] J. D. Storey and R. Tibshirani, "Estimating false discovery rates under dependence, with applications to dna microarrays," Technical Report 2001-28, Department of Statistics, Stanford University, 2001.

[43] K. Strimmer. *R Packages for Gene Expression Analysis*. www.stat.uni-muenchen.de/~strimmer/rexpress.html.

[44] The New School. *Vilfred Pareto, 1848-1923*. History of Economic Thought, New School, 2001. cepa.newschool.edu/het/profiles/pareto.htm.

[45] J. Tukey, *Exploratory Data Analysis*, Wiley, NY NY, 1977.

[46] 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.

[47] F. Wright. personal communication, 2002.

[48] E. Zitzler and L. Thiele, "Multiobjective evolutionary algorithms: A comparative case study and the strength pareto approach," *IEEE Transactions on Evolutionary Computation*, vol. 3, no. 4, pp. 257–271, Nov. 1999.

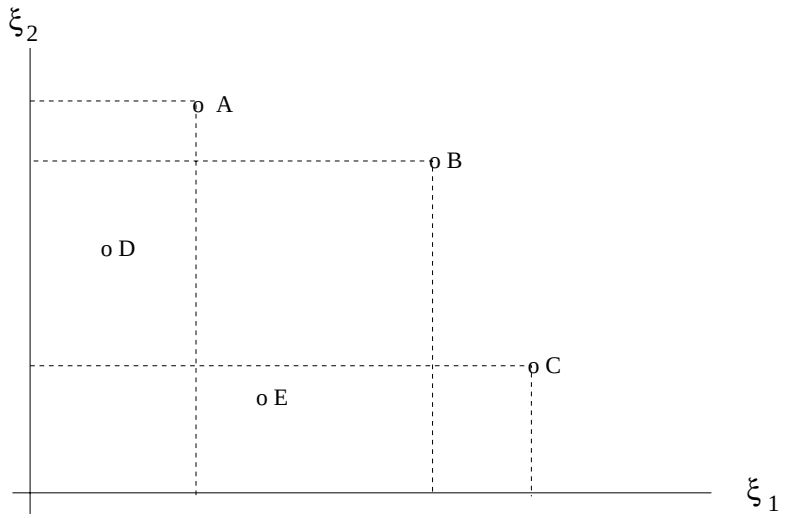

Figure 1: *A hypothetical multicriterion scattergram for genes A,B,C,D,E plotted as vectors in the plane described by a pair of fitness criteria ξ_1 and ξ_2. A, B, C are non-dominated genes and form the (first) Pareto front. A second Pareto front is formed by genes D,E.*

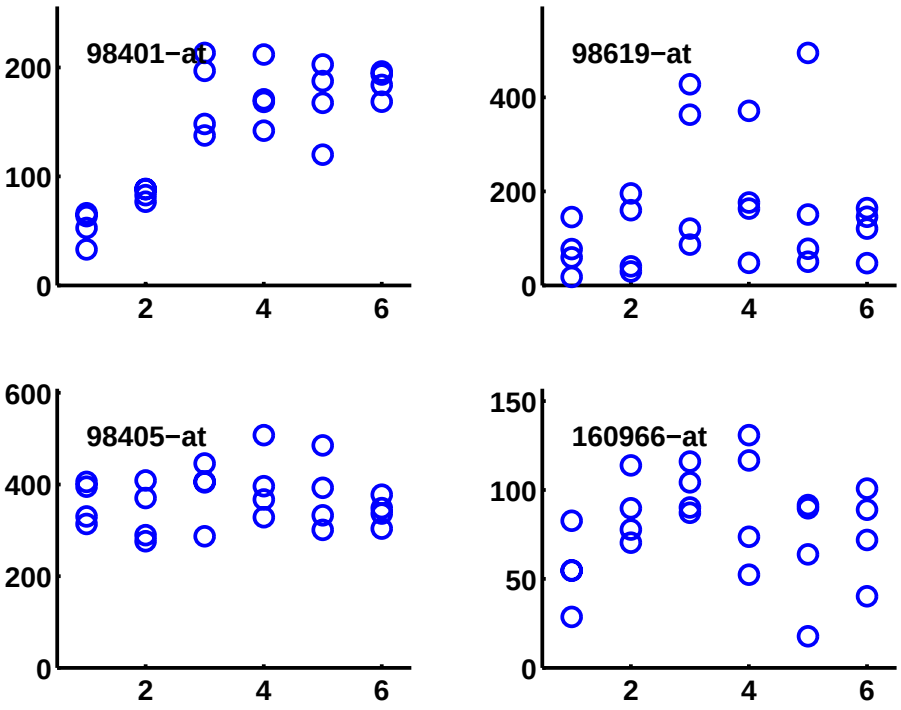

Figure 2: *Microarray data (MAS4) for 4 randomly selected gene hybridization profiles among the 14,222 genes encoded on the 24 microarrays in the 6 time-point mouse retinal aging study.*

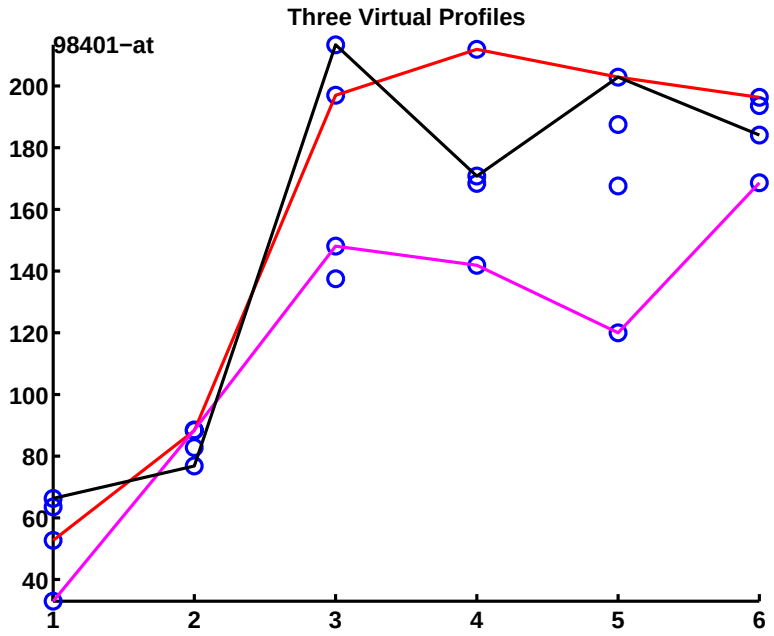

Figure 3: *Three of the 4096 possible virtual trajectories passing through the 6 time points of the upper left profile in Fig. 2.*

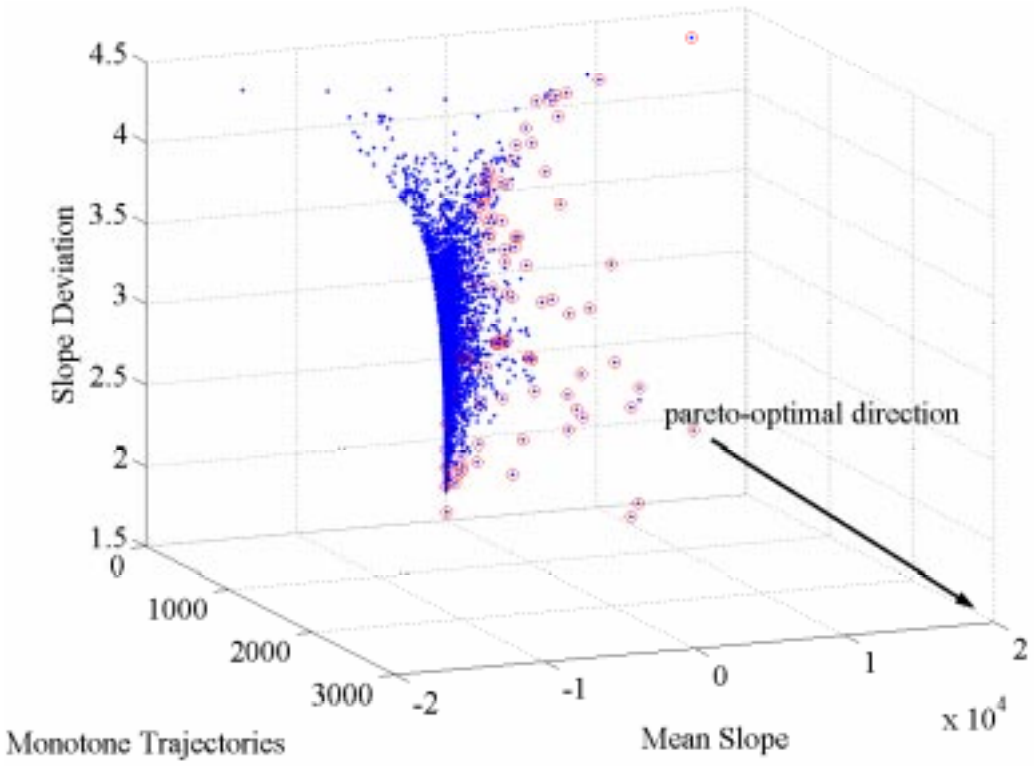

Figure 4: *The multicriterion scattergram (population averaged hybridization levels) and the Pareto front for the 24 mouse retinal aging study.*

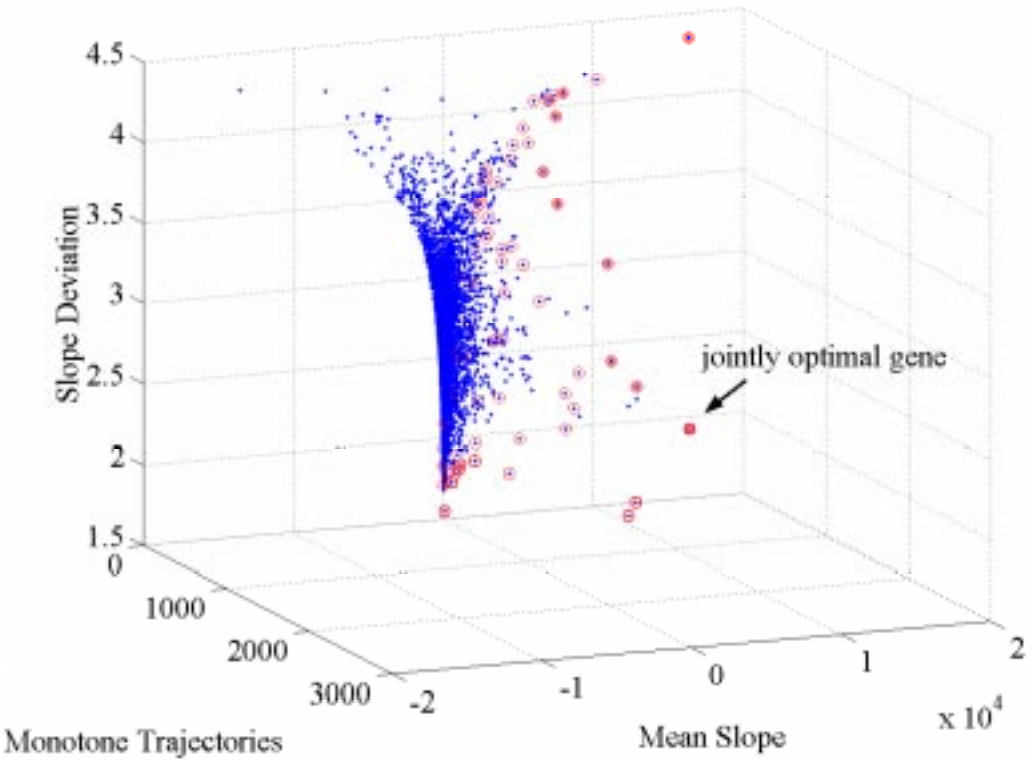

Figure 5: *The multicriterion scattergram (population averaged hybridization levels) and intersection of the 3 possible pairwise Pareto fronts (respectively denoted by box, circle, and asterisk) for the 24 mouse retinal aging study. Only one gene lies on intersection.*

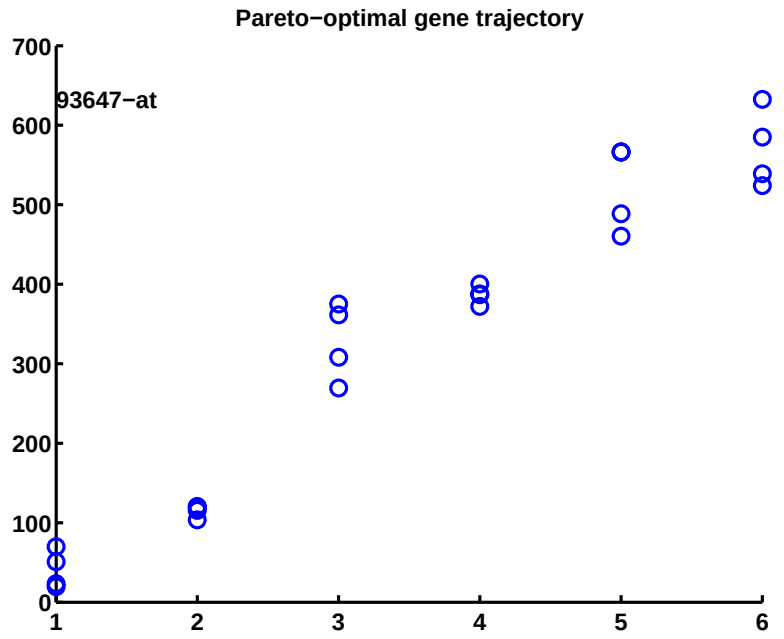

Figure 6: *The trajectory of the Pareto-optimal gene lying on the intersection of the three fronts in Fig. 5 in the 24 mouse retinal aging study.*

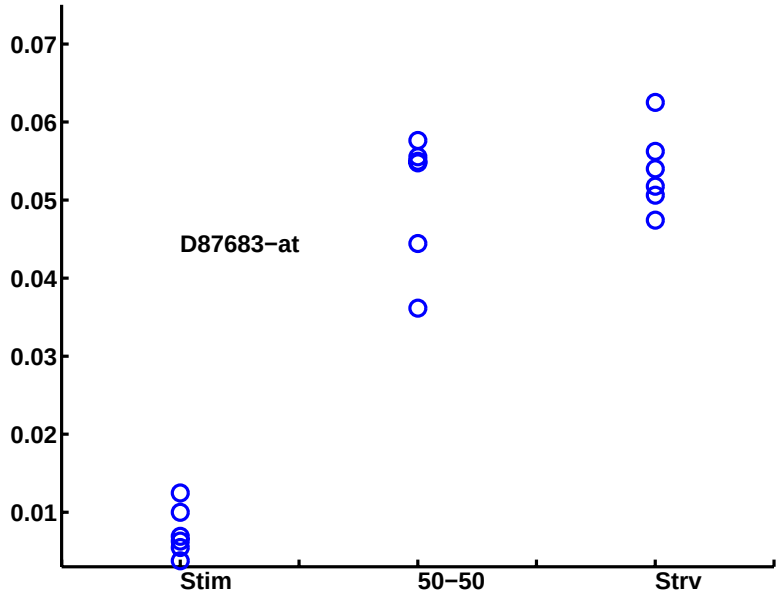

Figure 7: *Microarray data for a gene in human fibroblast mixture study.*

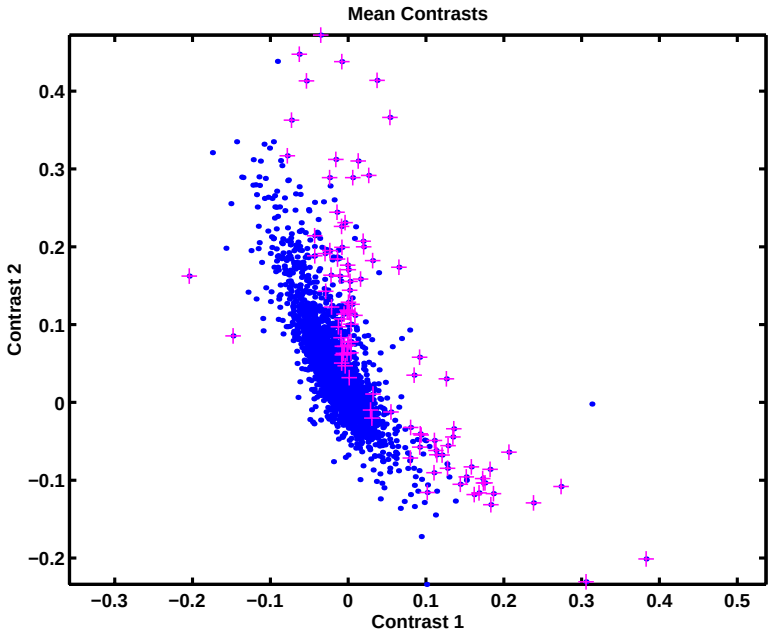

Figure 8: *Multicriterion scattergram corresponding to contrast matrix* $A = [-1, 1, 0; 1, 1, -2]$ *applied to the mean expression levels over 18 microarrays in human fibroblast study. Crosses indicate the 98 genes selected for analysis. The contrast matrix* A *is designed to find genes with inverted-*V *shaped profiles.*

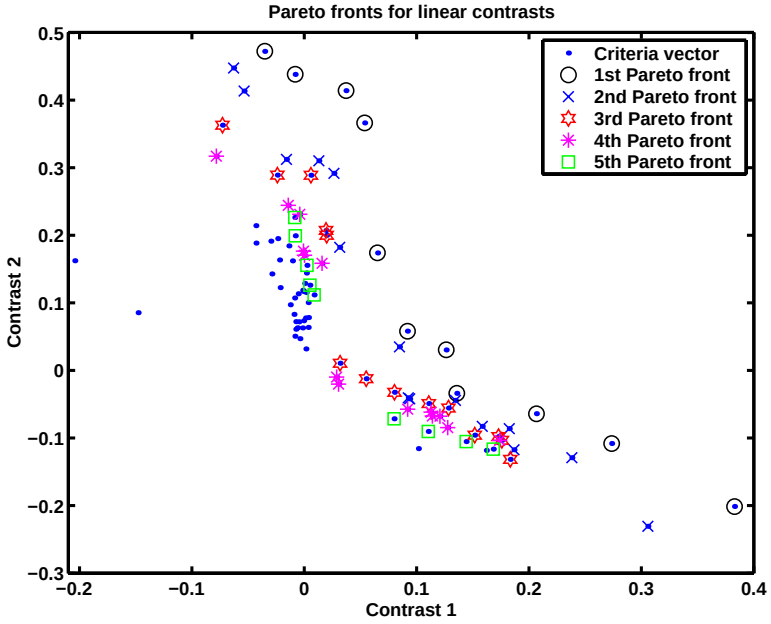

Figure 9: *The first five Pareto fronts (no cross-validation) for the genes with non-linear profiles shown in Fig. 8.*

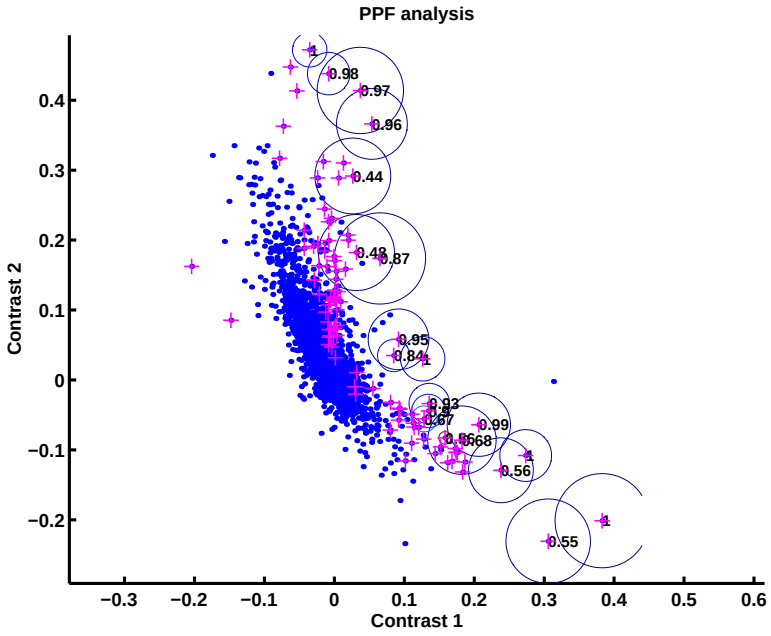

Figure 10: *Results of applying PPF analysis of human fibroblast study along with standard error constant contours and posterior probabilities of a given gene belonging to the first Pareto front. For clarity, only the first 20 top ranked genes are shown.*

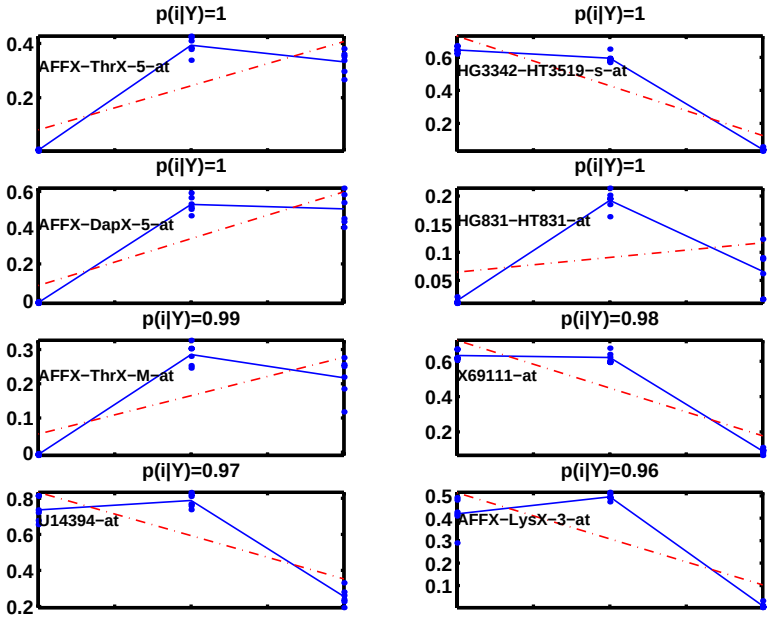

Figure 11: *Some top ranked gene profiles in human fibroblast study according to computed PPF posterior probabilities shown on Fig. 10. $P(i|Y)$ denotes the Bayes posterior probability that each profile is on the Pareto-front.*

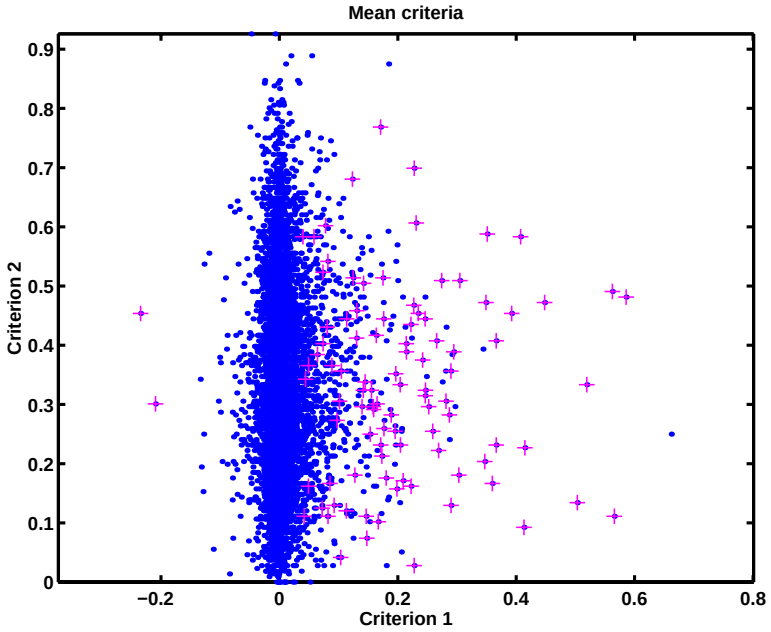

Figure 12: *Multicriterion mean scattergram for the non-parametric rank-order criteria for human fibroblast study.*

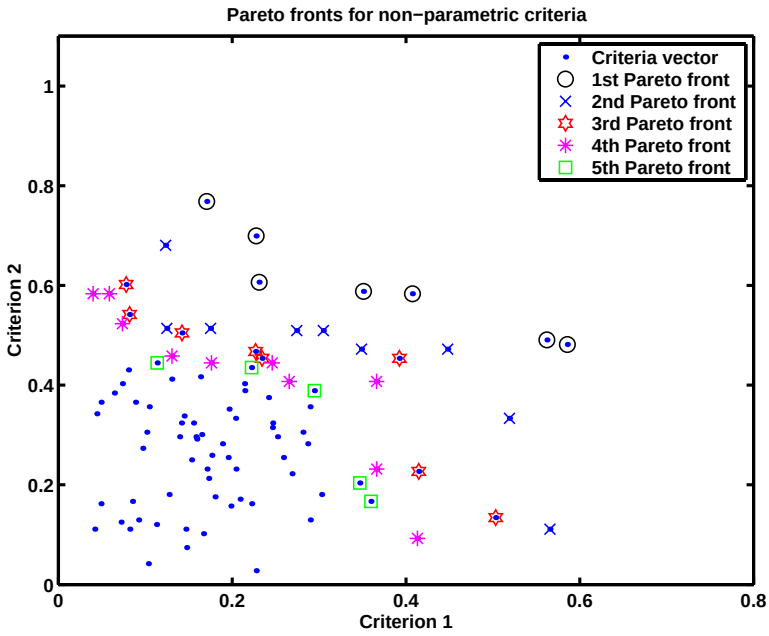

Figure 13: *The first five Pareto fronts (no cross-validation) of the non-parametric criteria for the non-linear genes indicated by crosses in Fig. 12.*

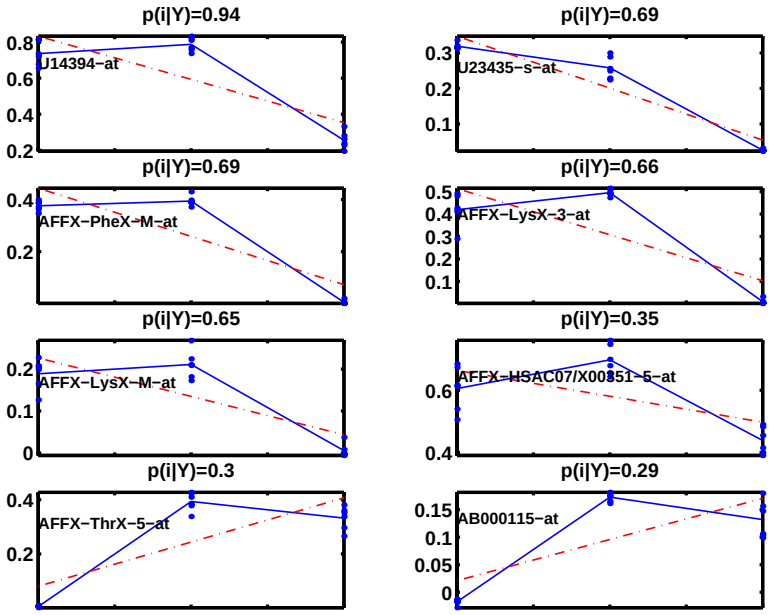

Figure 14: *The 8 top ranked cross-validated gene profiles remaining on the first Pareto front among the non-linear genes in Fig. 13. $P(i|Y)$ denotes the relative frequency that each re-sampled (leave-one-out cross-validation) profile is Pareto-optimal according to the non-parametric slope-sign criteria. Dashed line is the linear regression on t.*

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	U14394-at	0.944
HG3342-HT3519-s-at	0.998	AFFX-ThrX-5-at	1	U23435-s-at	0.694
AFFX-DapX-5-at	0.998	AFFX-ThrX-M-at	1	AFFX-PheX-M-at	0.685
HG831-HT831-at	0.996	HG3342-HT3519-s-at	1	AFFX-LysX-3-at	0.662
AFFX-ThrX-M-at	0.986	HG831-HT831-at	1	AFFX-LysX-M-at	0.648
X69111-at	0.984	U14394-at	1	AFFX-HSAC07/X00351-5-at	0.352
U14394-at	0.974	V00594-at	1	AFFX-ThrX-5-at	0.301
AFFX-LysX-3-at	0.962	X69111-at	1	AB000115-at	0.287
V00594-at	0.955	U45285-at	0.944	AFFX-DapX-5-at	0.245
U45285-at	0.932	AFFX-LysX-3-at	0.917	U53003-at	0.176
AB000115-at	0.899	AFFX-HSAC07/X00351-5-at	0.806	M92934-at	0.111
AFFX-HSAC07/X00351-5-at	0.866	AB000115-at	0.417	D29992-at	0.083
U73379-at	0.837	U73379-at	0.13	HG831-HT831-at	0.069
AFFX-DapX-M-at	0.678	V00594-s-at	0.074	S79522-at	0.042
Y09912-rna1-at	0.67	U75362-at	0.037	V00594-s-at	0.042
U75362-at	0.56	AFFX-PheX-5-at	0.028	D43636-at	0.032
AFFX-DapX-3-at	0.555	U03399-at	0.009	U22377-at	0.032
V00594-s-at	0.554			U75362-at	0.028
HG1980-HT2023-at	0.483			S70585-rna1-at	0.014
HG3044-HT3742-s-at	0.441			L02320-at	0.009
D43636-at	0.389			L05515-at	0.009
L27624-s-at	0.387			V00594-at	0.009
U03399-at	0.378			X69111-at	0.009
S69370-s-at	0.321			AFFX-PheX-5-at	0.005
AFFX-PheX-5-at	0.315			HG174-HT174-at	0.005

Figure 15: *The top scoring genes (Affymetrix nomenclature) resulting from PPF and RPF analysis of the most non-monotone concave profiles for Fred Wright's data (Li-Wong reduced indices). In the case of PPF,* $P(i|Y)$ *denotes the posterior probability that given gene belongs to first Pareto front with respect to the non-informative prior. In the case of RPF* $P(i|Y)$ *denotes the relative frequency that the gene belongs to the Pareto front with respect to re-sampling.*