PARETO DEPTH SAMPLING DISTRIBUTIONS FOR GENE RANKING

G. Fleury

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

A. Hero[†]*, S. Zareparsi*[*] *and A. Swaroop*[*]

University of Michigan
[†]Dept. of EECS and [*]Dept. of Ophthalmology and Visual Sciences
Ann Arbor MI, USA

ABSTRACT

In this paper we propose a method for gene ranking from microarray experiments using multiple discriminants. The novelty of our approach is that a gene's relative rank is determined according to the ordinal theory of multiple objective optimization. Furthermore, the distribution of each gene's rank, called Pareto depth, is determined by resampling over the microarray replicates. This distribution is called the Pareto depth sampling distribution (PDSD) and it is used to assess the stability of each ranking. Graphical representation of the PDSD as an image communicates information about the stability of each gene's rank. We illustrate on data from a mouse retina microarray experiment[0].

1. INTRODUCTION

Multicriteria gene filtering seeks to find genes whose expression profiles strike an optimal compromise between maximizing (or minimizing) several criteria. It is often easier for a molecular biologist to specify several criteria than a single criterion. For example the biologist might be interested in aging genes, which he might define as those genes having expression profiles that are increasing over time, have low curvature over time, and whose total increase from initial time to final time is large. Or one may have to deal with two biologists who each have different criteria for what features constitute an interesting aging gene.

In this paper we present a general method for rank ordering of genes based on a statistical version of the Pareto front partial order in multicriteria optimization. As a linear ordering of multiple criteria does not generally exist, an absolute ranking of the selected genes is generally impossible. However a partial ordering is often possible when formulated as a multicriterion optimization problem. This idea was used in our previous work [1, 2, 3] to obtain relative rankings of gene expression levels based on microarray experiments. We called our multiobjective approach to gene ranking *Pareto front analysis* (PFA). As pointed out in [3] the PFA approach is related to the notion of data depths and contours of depth in a multivariate sample [4]. In an analogous manner we will refer to the *Pareto depth* of a gene as the Pareto front on which the gene lies. Here we introduce the Pareto depth sampling distribution (PDSD) as a tool to both select high ranked genes and to visualize the stability of the gene rankings as an image. Gene microarray data from two experiments will be used to illustrate our analysis.

Mouse Retinal Aging Study: The experiment consists of hybridizing 24 retinal tissue samples taken from each of 24 age-sorted mice at 6 ages (time points) with 4 replicates per time point. These 6 time points consisted of 2 early development (Pn2, Pn10) and 4 late development (M2, M6, M16, M21) samples. RNA from each sample of retinal tissue was reverse transcribed to cDNA, amplified and hybridized to the 12,422 probes on one of 24 Affymetrix U74Av2 Mouse GeneChip microarrays. The data arrays from the GeneChips were processed by Affymetrix MAS5 software to yield log2 probe response data.

Human Retinal Aging Study: The experiment consists of hybridizing 16 retinal tissue samples taken from 8 young human donors and 8 old human donors. The ages of the young donors ranged from 16 to 21 years and the ages of the old donors ranged from 70 to 85 years old. The 16 tissue samples were hybridized to 16 Affymetrix U95A Human GeneChip microarrays each containing $N = 12,642$ probes.

2. GENE SCREENING AND RANKING

We assume that there are T populations (time samples) each consisting of M_t replicates, $t = 1, \ldots, T$. For each of the samples we assume an independent microarray hybridization experiment is performed yielding N gene probe responses extracted from the microarray. Define the measured response of the n-th probe on the m-th microarray acquired at time t

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

[0]This research was partially supported by National Institutes of Health grant NIH-EY11115 (including microarray supplements), Macula Vision Research Foundation, and Elmer and Sylvia Sramek Foundation.

0-7803-8388-5/04/$20.00 ©2004 IEEE

Consider the common problem of finding a set of genes whose mean expression levels are significantly different between a pair of time points ($T = 2$) [5]. The measured probe responses from such genes should exhibit small variability over population (intra-class dispersion) and high variability over time (inter-class dispersion). Two natural measures of intra-class dispersion ξ_1 and inter-class dispersion ξ_2, respectively, are the (scaled) absolute difference between sample means:

$$\xi_2(n) = \frac{1}{\sqrt{\frac{1}{M_1} + \frac{1}{M_2}}} \left| \bar{y}_{1.}(n) - \bar{y}_{2.}(n) \right|, \qquad (1)$$

where $\bar{y}_{t.}(n) = M_t^{-1} \sum_{m=1}^{M_t} y_{tm}(n)$, and the pooled sample standard deviation:

$$\xi_1(n) = \sqrt{\frac{(M_1 - 1)\sigma_1^2(n) + (M_2 - 1)\sigma_2^2(n)}{(M_1 - 1) + (M_2 - 1)}} \qquad (2)$$

where $\sigma_t^2(n) = (M_t - 1)^{-1} \sum_{m=1}^{M_t} \left(y_{tm}(n) - \bar{y}_{t.}(n) \right)^2$. The simple paired t-test can be used to separate the populations by thresholding the ratio $T_{\rm pt}(n) = \xi_2(n)/\xi_1(n)$ of the two dispersion measures and this could be used to rank the genes in decreasing order of $T_{\rm pt}$, or, equivalently, in increasing p-value.

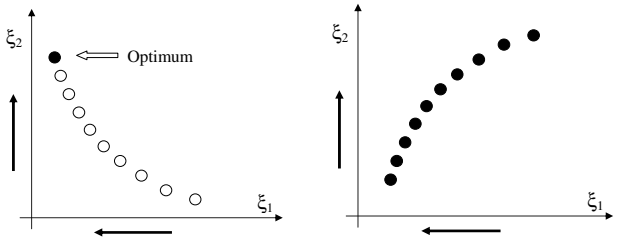

Fig. 1. *Left: a linear ordering exists and a single gene (optimum) dominates the others.Right: No non-trivial partial ordering exists.*

Multiple objective optimization captures the intrinsic compromises among possibly conflicting objectives. To illustrate, in the present context we consider the pair of criteria $\xi_2(n)$ (1) and $\xi_1(n)$ (2). A gene that maximizes ξ_2 and minimizes ξ_1 over all genes would be a very attractive gene indeed (Fig. 1.a). Unfortunately, such an extreme of optimality is seldom attained with multiple criteria. In rare cases there exists no non-trivial partial ordering and no sensible ranking is possible (see Fig. 1.b). However, in most cases, illustrated in Fig. 2.a, a partial ordering is possible. In the left panel of Fig. 2 gene A dominates gene C because both criteria are higher for A than for C. D, A and B are said to be non-dominated because improvement of one criterion in going from D to A to B corresponds to degradation of the other criterion. All the genes which are non-dominated constitute a curve which is called the Pareto front. A second Pareto front is obtained by stripping off points on the first front and computing the Pareto front of the remaining points (see Fig. 2). This process can be repeated to define a third front and so on. A gene that lies on the k-th Pareto front will be said to be at "Pareto depth" k.

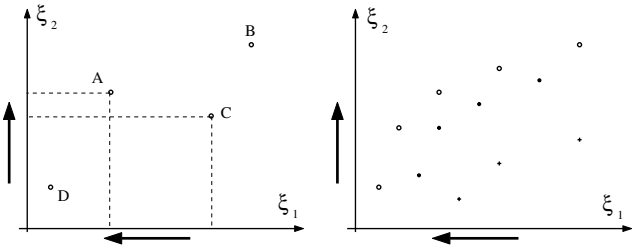

Fig. 2. *Left: A, B, D are non-dominated genes and form the Pareto front in the dual criteria plane where ξ_1 is to be minimized and ξ_2 is to be maximized. Right: successive Pareto fronts in dual criteria plane (o : first Pareto front, * : second Pareto front, + : third Pareto front).*

In practical cases there are multiple Pareto fronts each consisting of many genes. We illustrate in Figs. 3 and Fig. 4 where we show the scatterplots, called sample mean multicriteria scattergrams, of the empirical criteria $\{(\xi_1(n), \xi_2(n))\}_{n=1}^{N}$ defined in (1) and (2) for all gene probe responses extracted from microarrays in the mouse retina aging experiment and the human retina aging experiment, respectively.

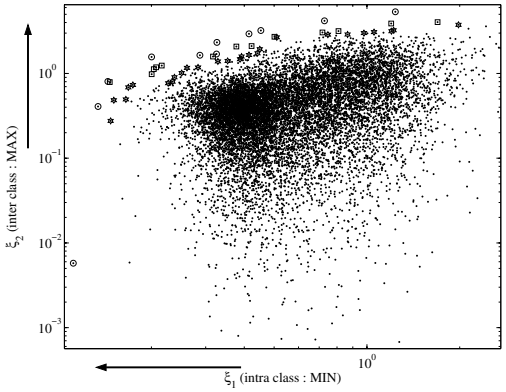

Fig. 3. *The sample mean* **multicriterion scattergram** *for the mouse retina aging experiment when comparing the populations at two time points M21 and M2. The first three Pareto fronts are indicated by circles, squares, and asterisks, respectively.*

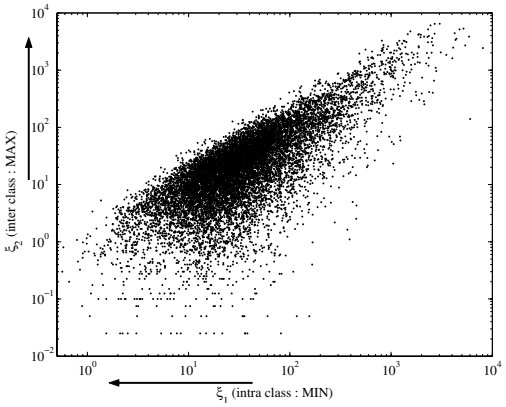

Fig. 4. *The sample mean multicriterion scattergram for the human retina aging experiment (analog to Fig. 3) when comparing young to old populations.*

3. PARETO DEPTH SAMPLING DISTRIBUTION

To account for sample variation we applied a simple leave-one-out cross-validation procedure to evaluate the sensitivity of the Pareto fronts to resampling the available samples. For each time point a sample is omitted leaving 2^M sets of $(M-1)^2$ pairs to be tested (here we set $M_t = M$, corresponding to the two data sets presented above). For each of these resampled set of genes the Pareto fronts are computed. The most resistant genes are those which remain on the top Pareto fronts throughout the resampling process. To quantify the movement of a given gene across the Pareto fronts as we resample, we introduce the Pareto depth sampling distribution (PDSD). For each gene this distribution corresponds to the empirical distribution of the Pareto front indexes visited during the resampling process:

$$\mathrm{Pdsd}_n(k) = M_{\mathrm{resamp}}^{-1} \sum_{j=1}^{M_{\mathrm{resamp}}} 1_n(j,k), k = 1, \ldots, N$$

where $M_{\mathrm{resamp}} = 2^M$ is the number of resampling trials, and $1_n(j,k)$ is an indicator function of the event: "j-th resampling of n-th gene is on k-th Pareto front." If K is the total number of Pareto fronts in the scattergram $(\xi_1(n), \xi_2(n)\}_{n=1}^N$ then, by convention, we define $\mathrm{Pdsd}_n(k) = 0$ for $k > K$. As the PDSD is a probability distribution $\mathrm{Pdsd}_n(k) \geq 0$ and $\sum_k \mathrm{Pdsd}_n(k) = 1$.

Figure 5 corresponds to the (un-normalized) PDSDs over the first 40 Pareto depths for four different genes taken from the human data set under the dual criteria (ξ_1, ξ_2) of (1) and (2). The left and right panels of Fig. 6 show the PDSDs of the top 50 genes for the human retina data and mouse data, respectively. In each of these figures the top 50 genes

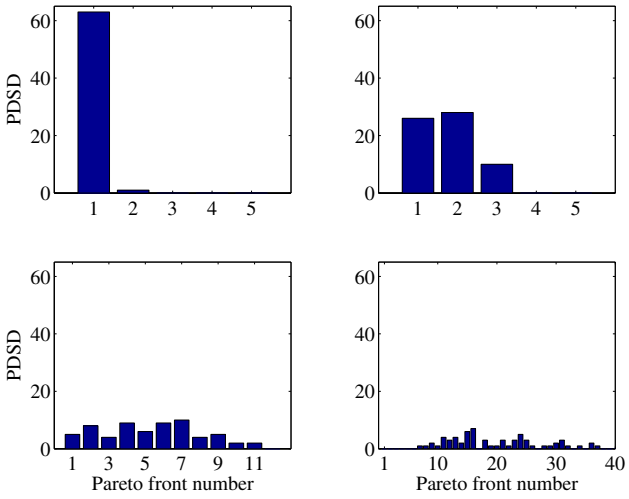

Fig. 5. *Unormalized PDSDs for four different genes taken from human retina experiment. These PDSDs are indexed by the Pareto depth, which is equivalent to Pareto front number.*

were ranked in order of increasing second PDSD moment $\sum_{k=1}^K k^2 \mathrm{Pdsd}_n(k)$. The PDSD images provide graphic indication of the Pareto variability of the human and mouse data sets. We note that even though the human data set has higher variance than the mouse data set, the top 50 human genes have lower Pareto variability since the human Pareto fronts are broader and contain more genes (compare Figs. 3 and 4).

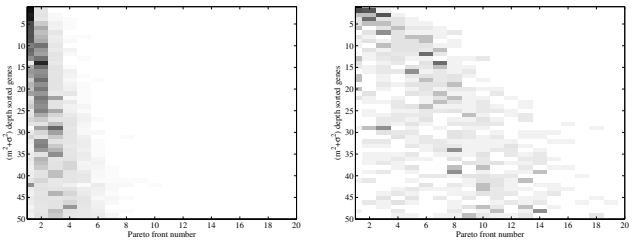

Fig. 6. *Left: An image of the PDSDs of the 50 top Pareto ranked human genes. Right: An image of the PDSDs of the 50 top Pareto ranked mouse genes. The magnitude of the PDSD is encoded in the false color range of black (PDSD=1) to white* $(PDSD = 0)$.

4. RANKING RATE COMPARISONS

We investigated the ranking performance of the second moment PDSD gene ranking procedure to the ranking performance of the paired t-test. Three hundred $(N = 300)$ different probe responses were simulated. Eight $(M = 8)$ replicates of the n-th gene probe response were generated

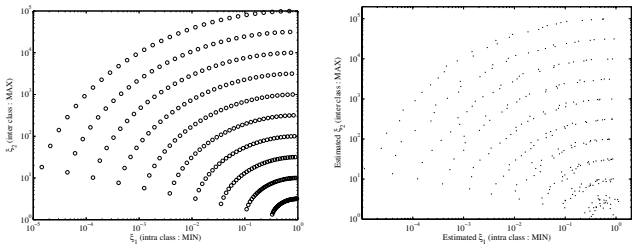

Fig. 7. *Left: ensemble mean scattergram (ground truth) for simulation study. Right: sample mean scattergram formed from a random realization.*

according to an i.i.d. Gaussian distribution with means and variances given by $(m_1(n), \sigma_1^2(n))$ and $(m_2(n), \sigma_2^2(n))$ for populations 1 and 2, respectively. The variances were made equal $\sigma_1^2(n) = \sigma_2^2(n) = \sigma^2(n)$ over both populations. The means and variances were set by the following formula:

$$\sigma(n) = \xi_2(n), \;\; m_1(n) = 0, \;\; m_2(n) = \xi_1(n)\xi_2(n)/2.$$

The values of $\xi_1(n), \xi_2(n)$ are illustrated in the ground truth scattergram in the left panel of Fig. 7. We designate the 90 genes on the first 3 fronts of this figure (depth increasing along -45^o diagonal) as *ground-truth-optimal* genes.

The right panel of Fig. 7 shows a realization of the empirical scattergram obtained from sample mean and variance estimates derived from the replicates. Figure 8 shows the three first Pareto fronts and the boundaries of two acceptance regions for the paired t-test applied to the empirical scattergram of Fig. 7. The first three Pareto fronts do not capture all of the ground-truth-optimal genes but they have a very low (0%) false discovery rate (proportion of genes found which are not ground-truth-optimal). The solid line boundary of the paired t-test discovers the 90 genes with lowest p-value. Use of this acceptance region would result in discovery of more ground-truth-optimal genes than discovered by the first three Pareto fronts, but with a false discovery rate of approximately 15%. The dashed line boundary corresponds to a paired t-test threshold which would lead to discovery of all of the 90 ground-truth-optimal genes, however, the false discovery rate of this acceptance region is quite high ($> 40\%$).

In Fig.9 we plot the correct discovery rate and the false discovery rate, respectively, for the paired t-test ranking and the second moment PDSD ranking procedures. The latter Pareto depth test performed significantly better (higher correct discovery rate and lower false discovery rate) than the paired t-test for all M investigated.

5. CONCLUSION

This paper has presented a new method of Pareto analysis that can identify and rank genes that have both stable and

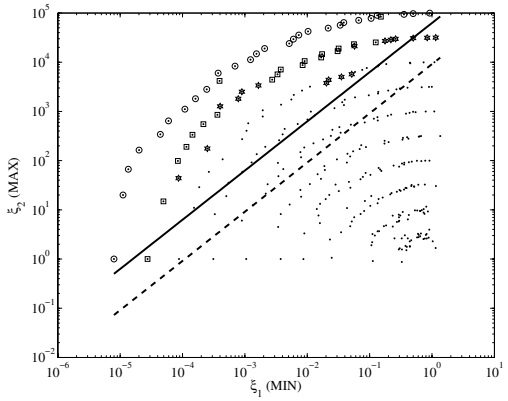

Fig. 8. *Three first Pareto fronts (circle, square and asterisk) and boundaries of paired t-test acceptance ragions.*

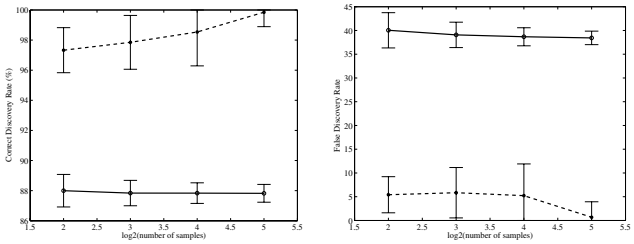

Fig. 9. *Correct discovery rate (left) and false discovery rate (right) as a function of the number of replicates for paired t-test (solid) versus Pareto depth test (dashed).*

low Pareto depths relative to the remaining genes. Additional genes discovered using this algorithm are now being validated by RT-PCR methods. The developed method has been implemented in Matlab and C and is sufficiently fast to be part of an interactive tool for gene screening, ranking, and clustering.

6. REFERENCES

[1] 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.*, Orlando, FL, 2002, vol. IV, pp. 4024–4027.

[2] 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.

[3] A. Hero and G. Fleury, "Pareto-optimal methods for gene analysis," *Journ. of VLSI Signal Processing*, vol. to appear, 2003, `www.eecs.umich.edu/~hero/bioinfo.html`.

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

[5] Terry P. Speed, *Statistical analysis of gene expression microarray data*, CRC Press, 2003.