Posterior Pareto Front Analysis for Gene Filtering

A. Hero[†], G. Fleury[◇]
[†]Depts. of EECS, BioMedical Eng., and Statistics, University of Michigan, Ann Arbor MI 49109, USA
[◇]Ecole Supérieure d'Electricité, Service des Mesures, 91192 Gif-sur-Yvette, France

Abstract—

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

I. Introduction

In [3], [4] we introduced a new approach to gene filtering, called Pareto gene filtering, which is based on multicriterion optimization and cross-validation. Pareto gene filtering allows the experimenter to isolate genes that achieve a good compromise between several competing gene-ranking criteria. Such genes lie on the so called *Pareto front* and are called non-dominated genes, see Sec. II for definitions. In this paper we present a Bayes posterior analysis approach to Pareto gene filtering which we call the method of *posterior Pareto fronts* (PPF). The main advantage of the PPF approach over the Pareto gene filtering approach is that it ranks each gene according to its posterior probability that it belongs to the Pareto front. We refer the reader to [7] for a more complete presentation of the work presented here.

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

II. Posterior Pareto Gene Filtering

A gene chip consists of a large number N of known DNA probe sequences that are put in distinct locations, called wells, on a slide [8], [1], [2]. After hybridization of an unknown tissue sample to the gene chip, the abundance of each probe present in the sample can be estimated from the measured levels of hybridization (responses). The study of differential gene expression between T populations requires hybridizing several (M) samples from each population to reduce response variability. Define the measured response at the n-th gene chip probe location for the m-th sample at time t

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

When several gene chip experiments are performed over time they can be combined in order to filter out those 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) [6]. As contrasted to maximizing such *scalar* criteria, multi-objective gene filtering seeks to simultaneously maximize gene profiles [3]. This method is closely related to multi-objective optimization which has been used for many applications [10], [11].

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 $\xi_p(i) > \xi_p(j)$. Multi-criterion optimization captures the intrinsic compromises among these possibly conflicting objectives. Consider Fig. 1 and suppose that ξ_1 and ξ_2 ($P = 2$) are to be maximized. It is obvious that genes A, B and C are "better" than genes D and E because both criteria are higher for the former than for the latter. Note that no gene among A, B and C dominates the other in both criteria ξ_1 and ξ_2. Multi-objective filtering uses this "non-dominated" property as a way to establish a preference relation among genes A, B, C, D and E. More formally, we say gene i is dominated if there exists some other gene $g \neq i$ such that for some $p = p_o$

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

All the genes which are non-dominated constitute a curve which is called the Pareto front. A second Pareto front can obtained by stripping off points on the first front and computing the Pareto front of the remaining points - which for the example in Fig. 1 would be genes D and E.

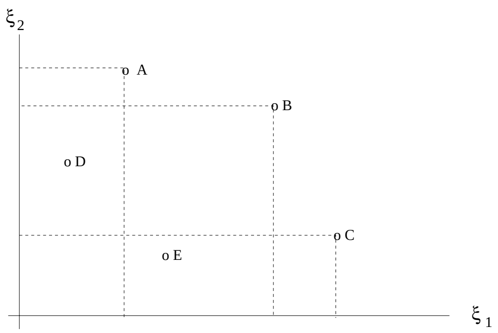

Fig. 1. *A, B, C are non-dominated genes relative to criteria ξ_1 and ξ_2.*

When the true means are unknown the criteria $\xi_p(i)$ can be estimated, e.g. by methods of moments, and the estimated Pareto front can be computed. To assign statistical confidence to these estimates cross-validation methods can be applied. The posterior Pareto front analysis introduced here casts the cross-validation procedure of [3] in a Bayesian framework. Although the theory can be developed for more general cases, here we assume an additive model for the (log) gene profile measurement:

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

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

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

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 $\sigma_t^2(i) = \sigma_\tau^2(i)$, $t, \tau = 1, \ldots, T$. For lack of space we only consider the latter here. Assume that: (i) $\{\mu_t(i)\}_{ti}$ 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. Then the joint posterior p.d.f. of $\underline{\mu}(i) = [\mu_1(i), \ldots, \mu_T(i)]^T$ takes the form of a multivariate Student-t density. We use a simple approximation to the associated c.d.f. via a multivariate L_∞ approximation to obtain

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

where $\hat{\sigma}^2(i) = T^{-1} M^{-1} \sum_t \sum_m (y_{tm}(i) - \hat{\mu}_t(i))^2$, $Y_i = \{y_{tm}(i)\}_{tm}$.

III. Profile Contrasts

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

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

where $A = ((a_{ij}))$ is a $P \times T$ *contrast matrix* and $P \leq T$. Assume that the components of $\underline{\mu}$ are conditionally independent. As the Pareto fronts are invariant to monotonic increasing transformations of the ξ_p's, a sufficient condition for $\underline{\xi}(i)$ to have independent components is that $AA^T = \text{diag}(a_{ii})$= a diagonal matrix. Consider the corresponding candidate $T \times T$ contrast matrices

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

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

IV. Experimental Results

We applied PPF analysis to Fred Wright's dataset described in the paper [9]. This data set is a mixing experiment which has been designed for empirically validating and comparing various differential gene expression methods of analysis. Three populations of genes were hybridized to Affymetrix HuGeneFL chips: starved human fibroblast cells; stimulated human fibroblast cells; and a 50-50 mixture of these cells. A total of 18 chips were processed corresponding to 6 replications within each of the three populations mentioned above. Each chip contains the same 7129 gene probes selected by Affymerix for the HuGeneFL chip. For each gene probe we arbitrarily defined the sequence of hybridization abundances from the "stimulated(t=1)," "50-50(t=2)," and "starved(t=3)," populations, in that order, as a gene expression profile. Note that ideally the profiles are linearly increasing or linearly decreasing over these three "time points." We fixed the objective of finding the most aberrant non-linear profiles which display a peak at $t = 2$ (convex cap). As a preprocessing step a a standard Fisher test was applied to screen gene profiles having large residual linear regression errors inconsistent with a linearity hypothesis. Subsequently the posterior Pareto fronts of the most aberrant convex cap genes were computed using the contrast matrix:

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

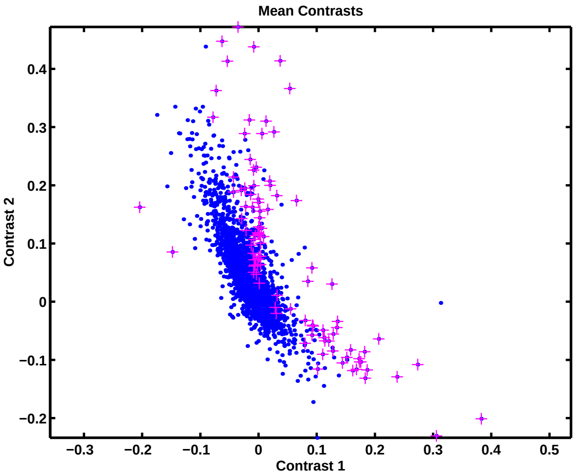

Fig. 2. *Multicriterion scattergram of sample mean contrasts $\{A\underline{\hat{\mu}}(i)\}_i$ with A given in (1) for Affymetrix Li-Wong reduced indices in Fred Wright's HuGeneFL mixture study). Crosses denote 98 genes that failed the Fisher linearity test at level $p = 0.1$.*

Throughout this section we used the exponent $a = 2$ in the prior density input for the PPF analysis. For this we adopted the contrast matrix

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

Figure 3 shows the results of PPF analysis. The contours around each point in the figure denotes the standard error (one standard deviation) circle and

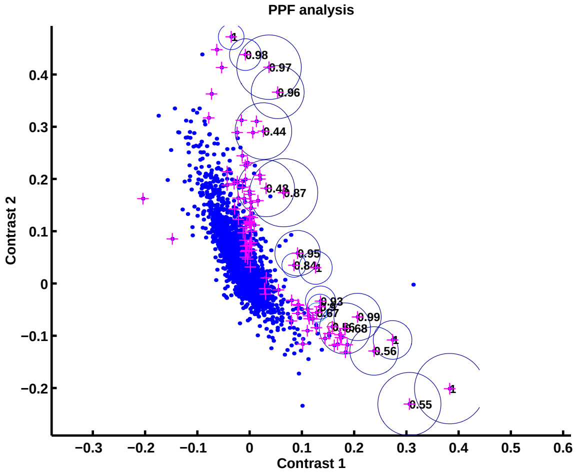

Fig. 3. *Non-linear genes with high posterior probability of belonging to the first Pareto front along with standard error constant contours and posterior probabilities. For clarity, only the first 20 top ranking genes are shown.*

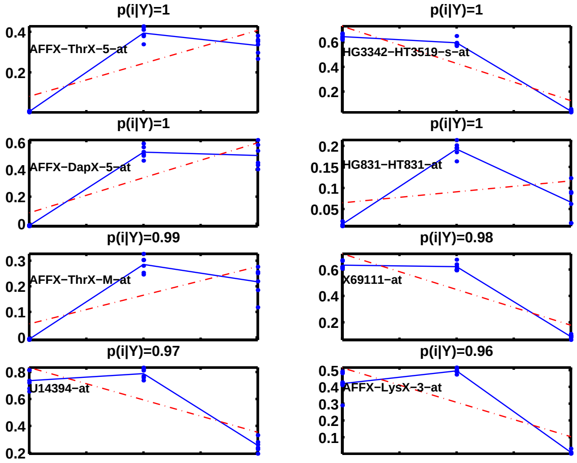

Fig. 4. *Eight top ranking genes according to the PPF analysis. $P(i|Y)$ denotes the Bayes posterior probability that each profile is Pareto-optimal according to the two linear contrast criteria.*

the annotation at the centers of the circles is the computed posterior probability (PPF) that the gene belongs to the first Pareto front. Figure 4 shows the eight top scoring trajectories under the PPF criterion. In each sub-panel the indicated piecewise linear line passes through the means of the 6 replicates for each of the 3 time samples.

V. Conclusion

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

References

[1] 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.

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

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

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

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

[6] 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.

[7] A. O. Hero and G. Fleury, “Pareto front analysis for gene filtering,” *J. Am. Statist. Assoc.*, p. submitted, 2002. `http://www.eecs.umich.edu/~hero/bioinfo.html`.

[8] 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.

[9] 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. `http://thinker.med.ohio-state.edu/projects/fbss/index.html`.

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

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