Searching for co-expressed genes in three-color cDNA microarray data using a probabilistic model based Hough Transform

Peter Tiňo, Hongya Zhao, Hong Yan

Abstract

The effects of a drug on the genomic scale can be assessed in a three-color cDNA microarray with the three color intensities represented through the so-called hexaMplot. In our recent study we have shown that the Hough Transform (HT) applied to the hexaMplot can be used to detect groups of co-expressed genes in the normal-disease-drug samples. However, the standard HT is not well suited for the purpose because: (1) the assayed genes need first to be hard-partitioned into equally and differentially expressed genes, with HT ignoring possible information in the former group; (2) the hexaMplot coordinates are negatively correlated and there is no direct way of expressing this in the standard HT and (3) it is not clear how to quantify the association of co-expressed genes with the line along which they cluster. We address these deficiencies by formulating a dedicated probabilistic model based HT. The approach is demonstrated by assessing effects of the drug Rg1 on homocysteine-treated human umbilical vein endothetial cells. Compared with our previous study we robustly detect stronger natural groupings of co-expressed genes. Moreover, the gene groups show coherent biological functions with high significance, as detected by the Gene Ontology analysis.

I. Introduction

MICROARRAY technology enables us to measure expression levels of thousands of genes simultaneously. The technology revolutionized research in systems biology, personalized treatment and drug development (e.g. [1], [2], [3]). Traditional dual-color cDNA microarrays employ two different fluorescence dyes corresponding to two samples (e.g. “normal” and “disease”). It has been recently demonstrated that it is possible to use a third dye associated with yet another sample hybridized to a single microarray [4], [5]. This opened up possibilities to assess effects of a drug in a three-color cDNA microarray essay hybridizing three samples: normal (dyed red), disease (dyed green) and drug-treated (dyed blue) [6]. After scanning and data processing, the intensity levels of the three dyes (R, G and B) are read out from every spot on the array. Each spot represents a gene from the pool of genes being assayed on the array and the intensities R, G and B reflect expression levels of the genes in the normal (healthy), disease and drug-treated samples, respectively.

P. Tiňo is with School of Computer Science, The University of Birmingham, Birmingham, B15 2TT, UK (e-mail: P.Tino@cs.bham.ac.uk).

H. Zhao and H. Yan are with Dept. Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong (e-mail: hyzhao,h.yan@cityu.edu.hk).

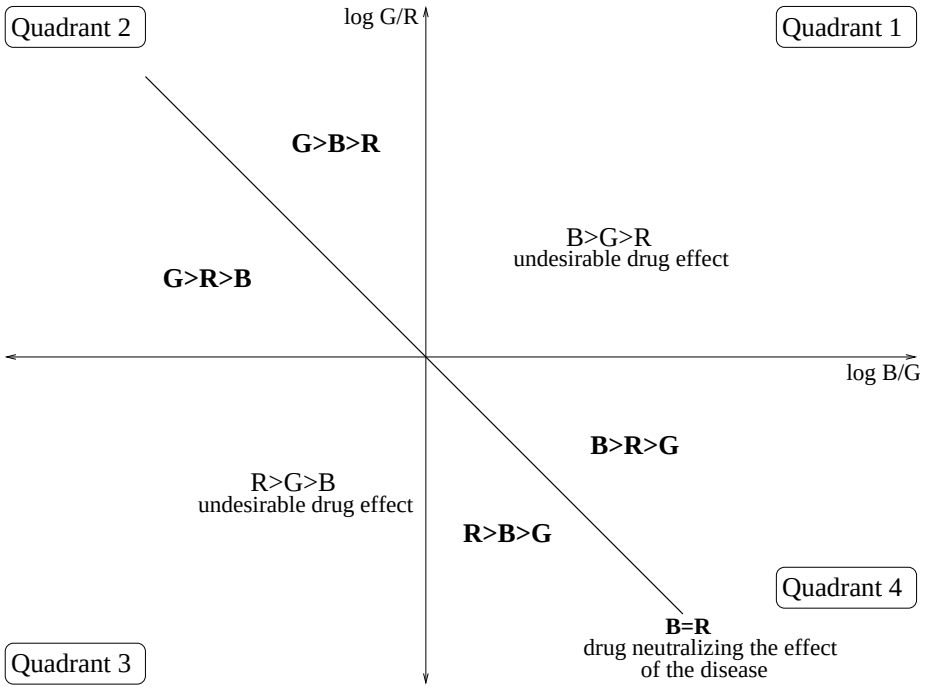

Fig. 1. Basic layout of hexaMplot. Points in quadrants 2 and 4 correspond to genes for which the effect of the disease has been largely neutralized by the drug. Gene representations in quadrants 1 and 3 point to undesirable drug effects. R, G and B represent gene expression levels of normal, diseased and drug-treated samples respectively.

A two-dimensional representation of R, G and B intensities called *hexaMplot*, naturally suited for assessing the drug effect on assayed genes was introduced in [6]. The hexaMplot coordinates represent the log ratios of intensity pairs: $x_1 = \log_2 B/G$ and $x_2 = \log_2 G/R$. Note that genes appearing in the upper and lower half-plane of the hexaMplot are up- and down-regulated, respectively, by the disease. Analogously, genes located in the left and right half-plane of the hexaMplot are up- and down-regulated, respectively, by the drug treatment, compared with the disease sample. Also note that along the slant axis $x_2 = -x_1$, we have $\log_2 B/R = 0$, meaning that the expression levels of genes in the normal and drug-treated samples are the same.

Naturally, one would like the drug to neutralize the effect of the disease on the assayed genes, i.e. ideally the gene representations in the hexaMplot should cluster around the slant axis. Deviations form the slant axis within the 4th and 2nd quadrants ($x_1 > 0, x_2 < 0$ and $x_1 < 0, x_2 > 0$, respectively) still represent drug effects in the right direction. However, genes falling into the 1st and 3rd quadrants of the hexaMplot ($x_1, x_2 > 0$ and $x_1, x_2 < 0$, respectively) show an undesirable effect of the drug, either further enhancing the up-regulation, or suppressing the down-regulation of the gene by the disease. Typically, most of the genes will not be effected by the disease or drug treatment [6] and their representations will cluster around the origin of the hexaMplot. There will be a sizable portion of genes in the 2nd and 4th quadrants of hexaMplot and a smaller portion of genes in the 1st and 3rd quadrants [6], [7]. The layout of hexaMplot is illustrated in figure 1. Points in quadrants 2 and 4, clustered around the slant axis, correspond to genes for which the effect of the disease has been largely neutralized by the drug. Gene representations in quadrants 1 and 3 point to situations of undesirable drug effect.

A simple methodology to assess the overall therapeutic effect of the drug was proposed in [6]. The correlation coefficient of hexaMplot representations of the assayed genes was calculated and assessed for statistical significance. A more involved analysis in [7] detects groups of genes with similar expression patterns relative to the disease and the drug proposed for its treatment. Each such group is aligned along a line ray starting in the hexaMplot origin. The direction of the ray signifies whether the drug has positive or negative effect on expression of the group of genes, while the angle measures the drug effect level [7]. The lines were detected through the Hough Transform (HT, see e.g. [8]) applied to differentially expressed genes. Among the detected lines, only the lines passing through the origin were considered . The biological function of the resulting gene groups was analyzed in the Gene Ontology (GO) framework[1] [9], [10], [11], [12]. GO provides a unique vocabulary across various genomic databases of diverse species. The driving organizational principle is to preserve essential functional features of genes shared among the organisms. In the GO analysis one assesses the significance of a group of genes by calculating the probability (p-value) that genes from the group will be associated with the GO category (node) by chance. If majority of genes in the group have the same biological function, such a probability will be very low [12].

The literature on techniques specifically designed for processing gene expression data from 3-color microarrays is very limited. Most studies related to 3-color microarrays deal with technical issues of array design/construction, or refer to our previous work on data analysis methods for assessing drug effects through 3-color microarrays [6], [7]. To our best knowledge, there are no techniques that could be considered direct alternatives to [6], [7] for gene expression analysis via 3-color microarrays in the context of assessing drug effects. However, while the approach of [7] represents a fruitful an interesting direction in mining gene-related effects of the drug under investigation, there are several problems associated with it:

1) The HT was applied to the differentially expressed genes only. Detection of the differentially expressed (and hence "interesting") genes was done through fitting a bi-variate Gaussian on hexaMplot representations of the whole gene sample and then applying a probability density threshold (critical value of the χ^2-distribution). The "hard" separation of assayed genes into equally vs. differentially expressed genes is not optimal, especially since there will typically be a high density of points (genes) around the separating confidence ellipse. The obtained results can be sensitive to the particular choice of the confidence value defining what is differentially expressed and what is not.
2) The HT implicitly imposes a noise model in the data space that does not fit the nature of hexaMplot representations well. First, the induced noise model depends on the line parametrization used, which is unsatisfactory. Second, the (x_1, x_2) hexaMplot representations are negatively correlated and there is no direct way of representing this fact in the standard HT.

[1]Of course, it is possible that genes not on the same line are strongly associated to a GO category due to some kind of nonlinear coherence in gene expression values. However, to our best knowledge, there is no such nonlinear model discussed in biology literature. The idea of a nonlinear model needs to be tested to exist consistently in many databases and interpreted biologically. The Hough Transform can been used to detect complicated curve shapes, so although our model is linear, the framework can be easily generalized to nonlinear ones if they are indeed biologically meaningful.

3) Determination of the quantization level in the Hough space should reflect the amount of "measurement" noise in the hexaMplot features. The quantization level determines the amount of smoothing in the Hough accumulator, which in turn has an effect on the number of distinct peaks (detected lines) in the Hough space. Also, given a detected line, there is no principled way of quantifying the strength of association of the points with that line.

In this paper, we address these shortcomings in the framework of a principled probabilistic model based formulation outlined in the next section. Briefly, all assayed genes are considered. The weaker and stronger contribution of equally and differentially expressed genes is obtained naturally in a "soft" manner from the probabilistic formulation of the model behind the hexaMplot. The model explicitly takes into account the size and the negatively correlated nature of the noise associated with hexaMplot gene representations. Both the strength of association of individual genes with a particular group (line ray in hexaMplot) and the support for the group by the selected genes can be quantified in a principled manner through posterior probabilities over the line angles, given the observations.

The paper has the following organization: After introducing our model based Hough Transform in section II, we apply the methodology to assess the effect of a drug Rg1 on homocysteine-treated human umbilical vein endothetial cells in section III. Our approach is compared with alternative clustering and correlation based approaches in section III-A. The main findings are summarized in section IV.

II. Probabilistic model based Hough Transform

Consider a line ray in $\mathbb{R}^2$ (hexaMplot space) starting in the origin at an angle $\alpha \in [-\pi/4, 7\pi/4)$. We assume a bi-variate zero-mean Gaussian measurement noise with covariance matrix Σ_X. The density of possible measurements $\mathbf{x} = (x_1, x_2)^T \in \mathbb{R}^2$ corresponding to the point $(r\cos\alpha, r\sin\alpha)$ on the line is given by

$$p(\mathbf{x}|\alpha, r) = \frac{1}{2\pi|\Sigma_X|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}^T - (r\cos\alpha, r\sin\alpha))\ \Sigma_X^{-1}\ (\mathbf{x} - (r\cos\alpha, r\sin\alpha)^T)\right\}, \tag{1}$$

where $r > 0$ is the (Euclidean) distance of the point on the line from the origin.

One may have a prior knowledge about the parameter values $(\alpha, r) \in [-\pi/4, 7\pi/4) \times [0, \infty)$, summarized in the form of a prior distribution $p(\alpha, r)$. Given an observation $\mathbf{x}$, the induced uncertainty in the parameter space is given by the posterior

$$p(\alpha, r|\mathbf{x}) = \frac{p(\mathbf{x}|\alpha, r)\ p(\alpha, r)}{\int_{[\pi/4, 7\pi/4)\times[0,\infty)} p(\mathbf{x}|\alpha', r')\ p(\alpha', r')\ \mathrm{d}\alpha'\mathrm{d}r'}. \tag{2}$$

We are interested only in data points aligned (up to the measurement noise) along a common line passing through the origin. To obtain the amount of support for the angle parameter α given the observation $\mathbf{x}$, we integrate r from the posterior:

$$p(\alpha|\mathbf{x}) = \int_{[0,\infty)} p(\alpha, r|\mathbf{x})\ \mathrm{d}r. \tag{3}$$

The aim of the Hough Transform (HT) and its generalizations is to detect possible line[2] candidates along which some of the data points are aligned. The detection is performed in the parameter space

[2]Extension to other parametrized objects such as circles is straightforward.

(Hough space), where each observation induces a certain amount of mass on parameters compatible with the observation. For example, in the original HT one partitions the Hough space of line parameters (e.g. (bias, slope)) and increments each parameter pair by one if it turns out to be compatible with the observation. After running through all the observations, peaks in the Hough space indicate the lines with most support, e.g. lines along which many of the observations are aligned. In our case, the Hough space is the angle interval $\mathcal{H} = [-\pi/4, 7\pi/4]$. We do not discretize the Hough space; instead, each observation $\mathbf{x}$ induces a support kernel $p(\alpha|\mathbf{x})$ in $\mathcal{H}$. Given a set of observations $\mathcal{D} = \{\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^N\}$, $\mathbf{x}^i \in \mathbb{R}^2$, $i = 1, 2, ..., N$, we accumulate the evidence contributions in the Hough space $\mathcal{H}$ as proposed in [13], [14], namely

$$H(\alpha; \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} p(\alpha|\mathbf{x}^i). \tag{4}$$

Note that the probabilistic support kernels in [13], [14] are obtained in a different manner. Whereas in our model formulation we start with a generative model of the data (line rays starting in the origin and endowed with a measurement noise) and determine the (possibly non-Gaussian) support kernel for each observation $\mathbf{x}$ as the marginal posterior given $\mathbf{x}$, Ji and Haralick [13], [14] *impose* that the support kernel has a Gaussian form of a fixed shape that is determined from the image data to which the HT is applied.

Given that a line candidate with inclination angle α has been detected by inspecting the peaks of the Hough accumulator $H(\alpha; \mathcal{D})$, one can ask which points from $\mathcal{D}$ are strongly associated with it. This can be done by consulting the posteriors $p(\alpha|\mathbf{x}^i)$, $i = 1, 2, ..., N$, and selecting the points above some threshold value θ. To enhance the threshold interpretability, we discretized the angle space $\mathcal{H}$ into a regular grid $G = \{\tilde{\alpha}_1, \tilde{\alpha}_2, ..., \tilde{\alpha}_M\}$ and turned the densities $p(\alpha|\mathbf{x})$ into probabilities $P(\tilde{\alpha}_j|\mathbf{x})$ over the G:

$$P(\tilde{\alpha}_j|\mathbf{x}) = \frac{p(\tilde{\alpha}_j|\mathbf{x})}{\sum_{k=1}^{M} p(\tilde{\alpha}_k|\mathbf{x})}. \tag{5}$$

We then calculate the probability threshold $\theta \in (0, 1)$ as $\theta = \kappa/M$, $\kappa \in (0, M)$, meaning that only observations with posteriors at least κ times greater than the uninformative distribution $1/M$ will be considered. Given a probability threshold θ and a (discretized) angle $\tilde{\alpha}$, the set of selected points that support the line ray $\tilde{\alpha}$ reads:

$$S_\theta(\tilde{\alpha}) = \{\mathbf{x} \mid \mathbf{x} \in \mathcal{D},\ P(\tilde{\alpha}|\mathbf{x}) \geq \theta\}. \tag{6}$$

Once the set of selected points $S_\theta(\tilde{\alpha})$ for a particular line ray $\tilde{\alpha}$ is obtained, one can check how much the set as a whole supports that line ray by calculating the posterior

$$P(\tilde{\alpha}|S_\theta(\tilde{\alpha})) = \frac{p(S_\theta(\tilde{\alpha})|\tilde{\alpha})\ P(\tilde{\alpha})}{\sum_{\tilde{\alpha}' \in G} p(S_\theta(\tilde{\alpha})|\tilde{\alpha}')\ P(\tilde{\alpha}')}, \tag{7}$$

where $P(\tilde{\alpha}')$ is the prior distribution over the grid G and (assuming independence of observations)

$$\begin{aligned} p(S_\theta(\tilde{\alpha})|\tilde{\alpha}') &= \prod_{\mathbf{x} \in S_\theta(\tilde{\alpha})} p(\mathbf{x}|\tilde{\alpha}') \\ &= \prod_{\mathbf{x} \in S_\theta(\tilde{\alpha})} \int_0^\infty p(\mathbf{x}|r, \tilde{\alpha}')\ p(r|\tilde{\alpha}') \mathrm{d}r. \end{aligned} \tag{8}$$

Here, $p(\mathbf{x}|r, \tilde{\alpha}')$ is the noise model (1) and $p(r|\tilde{\alpha}')$ is the conditional prior on r.

A. Noise model

Our two-dimensional observations are hexaMplot representations of the 3-color intensities (R, G, B) measured in cDNA microarrays. It is usual to assume that the log intensities are normally distributed. Recall that the 2-dimensional hexaMplot representations read:

$$\mathbf{x} = (x_1, x_2)^T = \left(\log \frac{B}{G}, \log \frac{G}{R} \right)^T .$$

Now, consider three random variables (log intensities) Y_1, Y_2 and Y_3 representing $\log B$, $\log G$ and $\log R$, respectively. The hexaMplot representations (x_1, x_2) correspond to two random variables $X_1 = Y_1 - Y_2$ and $X_2 = Y_2 - Y_3$ coupled through Y_2. Even if we assume that the individual measurement errors of the three log intensities Y_1, Y_2 and Y_3 are independent, the implied noise in the hexaMplot coordinates X_1, X_2 will be negatively correlated. This simply follows from that fact that while Y_2 contributes negatively to X_1, its contribution to X_2 is positive. Assuming that the measurement noise of the log intensity Y_i is a zero mean Gaussian with variance σ_i^2, $i = 1, 2, 3$, (X_1, X_2) will be Gaussian distributed with covariance matrix

$$\Sigma_X = \begin{bmatrix} \sigma_1^2 + \sigma_2^2 & -\sigma_2^2 \\ -\sigma_2^2 & \sigma_2^2 + \sigma_3^2 \end{bmatrix}. \tag{9}$$

This can be seen by realizing that for an affine-transformed vector random variable $X = \mathbf{c} + AY$, where Y is a multivariate Gaussian distributed random variable with mean $\mathbf{m}$ and covariance Σ_Y, X will be Gaussian distributed with mean $\mathbf{c} + A\mathbf{m}$ and covariance $A\Sigma_Y A^T$. In our case

$$A = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}. \tag{10}$$

If (as done in this paper) we assume equal levels of measurement noise across the three colors, $\sigma^2 = \sigma_1^2 = \sigma_2^2 = \sigma_3^2$, we obtain

$$\Sigma_X = 2\sigma^2 \begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix}. \tag{11}$$

B. Priors over the parameters

As explained above, for the hexaMplot typically, most of the genes will not be effected by the disease or drug treatment making their representations cluster around the origin of the hexaMplot. There will be a sizable portion of genes in the 2nd and 4th quadrants of hexaMplot positively effected by the drug treatment. A smaller portion of genes negatively effected by the drug treatment will be represented in the 1st and 3rd quadrants. We express this insight through the piecewise linear prior $p(\alpha)$:

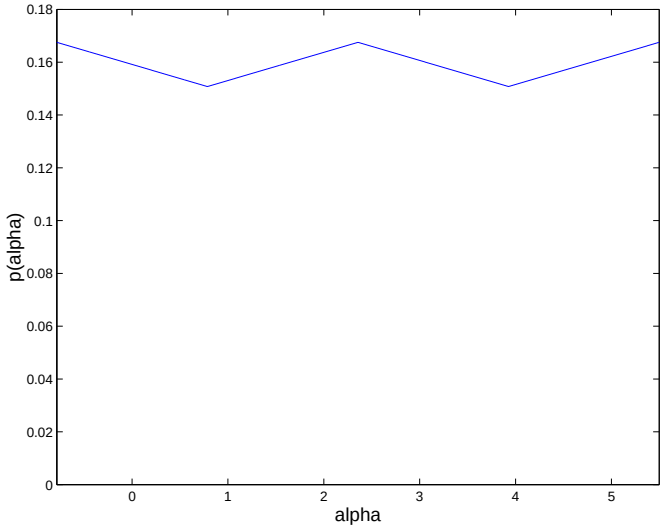

Fig. 2. Prior on angles α for the inhomogeneity parameter setting $C_\alpha = 0.9$.

$$p(\alpha) = \begin{cases} -\alpha \cdot \Delta_\alpha + \frac{1}{2\pi}, & \text{if } -\pi/4 \le \alpha < \pi/4 \\ \alpha \cdot \Delta_\alpha + \frac{3C_\alpha - 1}{2\pi(1+C_\alpha)}, & \text{if } \pi/4 \le \alpha < 3\pi/4 \\ -\alpha \cdot \Delta_\alpha + \frac{5-3C_\alpha}{2\pi(1+C_\alpha)}, & \text{if } 3\pi/4 \le \alpha < 5\pi/4 \\ \alpha \cdot \Delta_\alpha + \frac{7C_\alpha - 5}{2\pi(1+C_\alpha)}, & \text{if } 5\pi/4 \le \alpha < 7\pi/4 \end{cases} \tag{12}$$

where

$$\Delta_\alpha = \frac{2(1 - C_\alpha)}{\pi^2(1 + C_\alpha)} \tag{13}$$

and $C_\alpha \in [0, 1]$ is a constant determining the ratio between the minimum prior probability assigned to the angles $\alpha = \pi/4$ and $\alpha = 5\pi/4$ at the center of quadrants 1 and 3, respectively, and the maximum prior probability assigned to the angles $\alpha = 3\pi/4$ and $\alpha = 7\pi/4$ at the center of quadrants 2 and 4, respectively. As an example, the prior $p(\alpha)$ is shown in figure 2 for the inhomogeneity parameter C_α set to $C_\alpha = 0.9$. Note that when $C_\alpha = 1$, we obtain the maximum entropy uniform prior $p(\alpha) = 1/(2\pi)$ for all $\alpha \in [-\pi/4, 7\pi/4)$.

It is also natural to expect that if measurements of many genes were performed, only a smaller portion of the data will show significant effects, i.e. there will be more points concentrated around the origin of the hexaMplot than further away from it, especially in quadrants 1 and 3 of the $(\log B/G, \log G/R)$ system. We express this by formulating a prior on $r \ge 0$ (conditional on the angle α) as a mixture of two truncated Gaussians:

$$p(r|\alpha) = (1 - \kappa(\alpha))\ p_1(r) + \kappa(\alpha)\ p_2(r), \tag{14}$$

where

$$p_i(r) = \frac{2}{\sqrt{2\pi}\omega_i} \exp\left\{-\frac{r^2}{2\omega_i^2}\right\}, \quad i = 1, 2$$

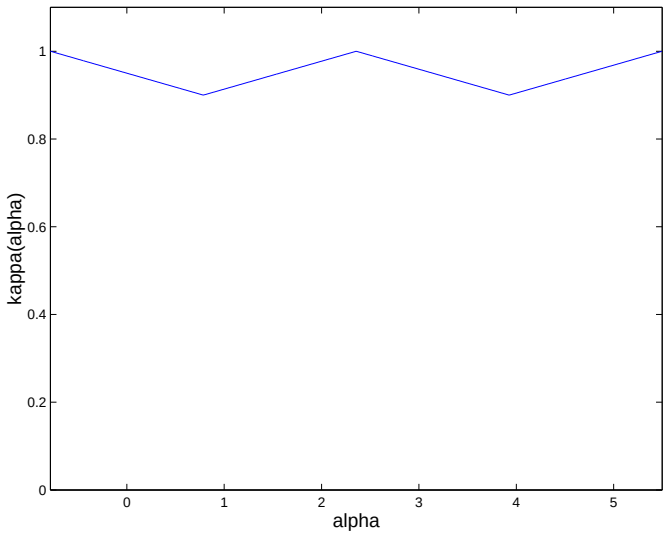

Fig. 3. Mixing coefficient $\kappa(\alpha)$ for $C_r = 0.9$.

with $0 < \omega_1 \leq \omega_2$, and

$$\kappa(\alpha) = \begin{cases} -\alpha \cdot \Delta_r + \frac{1+C_r}{2}, & \text{if } -\pi/4 \leq \alpha < \pi/4 \\ \alpha \cdot \Delta_r + \frac{3C_r-1}{2}, & \text{if } \pi/4 \leq \alpha < 3\pi/4 \\ -\alpha \cdot \Delta_r + \frac{5-3C_r}{2}, & \text{if } 3\pi/4 \leq \alpha < 5\pi/4 \\ \alpha \cdot \Delta_r + \frac{7C_r-5}{2}, & \text{if } 5\pi/4 \leq \alpha < 7\pi/4, \end{cases} \tag{15}$$

$$\Delta_r = \frac{2(1-C_r)}{\pi}. \tag{16}$$

The (truncated) Gaussians $p_i(r)$, $i = 1, 2$, operating on $r \geq 0$ reflect the assumption of greater concentration of gene representations around the origin of the hexaMplot than further away from it. Moreover, in quadrants 2 and 4 there may be greater variation of points than in quadrants 1 and 3. In the middle of quadrants 2 and 4, the mixing coefficients of the mixture prior $p(r|\alpha)$ (14) are equal to $1-\kappa(\alpha) = 0$ and $\kappa(\alpha) = 1$, making the prior $p(r|\alpha) = p_2(r)$. In the middle of quadrants 1 and 3, the mixing coefficients are determined by $\kappa(\alpha) = C_r \leq 1$ and the prior reads $p(r|\alpha) = (1 - C_r)\ p_1(r) + C_r\ p_2(r)$. Since $\omega_1 \leq \omega_2$, in quadrants 1 and 3 the prior assumes greater concentration around the origin than in quadrants 2 and 4. Note that when $C_r = 1$, we obtain a simple (truncated) Gaussian prior of standard deviation $\omega_r = \omega_2$ on r, independent of the angle α. We illustrate the angle-conditional mixing coefficient $\kappa(\alpha)$ for $C_r = 0.9$ in figure 3.

The joint prior $p(\alpha, r) = p(\alpha)\ p(r|\alpha)$ is illustrated in figure 4 for $C_\alpha = C_r = 0.9$ and $\omega_1 = 1.9$, $\omega_2 = 2$.

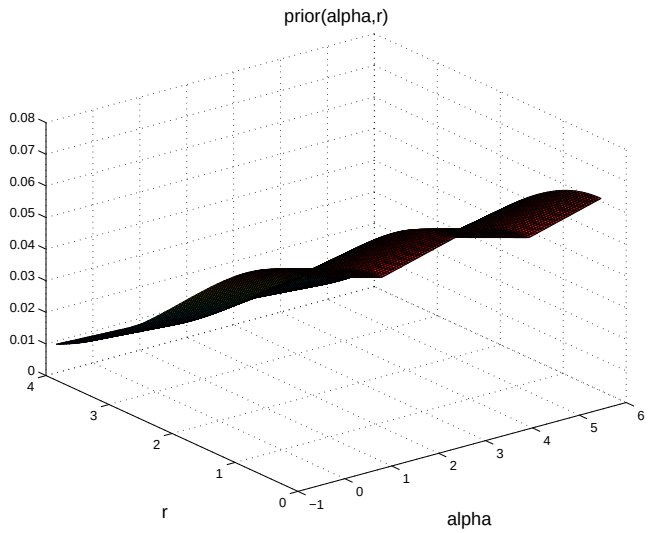

Fig. 4. Joint prior $p(\alpha, r) = p(\alpha)\ p(r|\alpha)$ for $C_\alpha = C_r = 0.9$ and $\omega_1 = 1.9$, $\omega_2 = 2$.

C. Interpretation of the model

The standard HT assumes that all points located within a "close" range of the hypothesis line contribute equally to support that line in the Hough space. The notion of closeness is determined implicitly by coarse-graining of the Hough space partition. It has been long recognized that such a *top hat* strategy to compute the contribution of each observation is inadequate since not all data points are equally reliable e.g. due to uncertainties induced by the image noise, edge orientation estimation etc. [15], [13], [14]. Several "soft" alternatives to the "hard" *top hat* kernels have been suggested, mainly in the explicit context of line detection in images, e.g. [16], [14], [17]. However, to our best knowledge, our approach is unique in that the support kernels in the Hough space are obtained from a principled generative model in the data space. A simple illustration of our approach is presented in figure 5(a). Three observations $\mathbf{x}^i$, $i = 1, 2, 3$, aligned along a line ray (bold solid line) with angle α are shown. The discs surrounding the observations signify the measurement noise. Of course, when it comes to estimating the angle α, the closer the observation is positioned towards the origin, the greater is the induced uncertainty about the actual angle in the Hough space $\mathcal{H}$. The uncertainty in the angle estimates associated with observations $\mathbf{x}^1$, $\mathbf{x}^2$ and $\mathbf{x}^3$ is illustrated by the pairs of dotted, dashed and solid lines, respectively. Consequently the support kernel $p(\alpha|\mathbf{x}^1)$ will be least informative (dotted bold line in figure 5(b)), while the support kernel $p(\alpha|\mathbf{x}^3)$ will be highly peaked (solid bold line in figure 5(b)). The standard HT and many of its modifications apply the same kernel on top of all estimates in the Hough space. Our model based formulation of the HT naturally treats the variable degrees of uncertainty in the support kernels.

As a specific example, consider three points $\mathbf{x}^1 = (0.1, -0.1)^T$, $\mathbf{x}^2 = (-1, 0)^T$ and $\mathbf{x}^3 = (-1.75, 1.75)^T$ lying on rays with angles $-\pi/4$, π and $3\pi/4$, respectively. The posteriors in the parameter space under parameter setting $C_\alpha = C_r = 0.9$, $\omega_1 = 1.9$, $\omega_2 = 2$ are shown in figure 6. The joint posteriors $p(\alpha, r|\mathbf{x}^i)$ for noise levels $\sigma = 0.2$ and $\sigma = 0.05$ are presented in figures 6(a) and (c), respectively. Note how the reduction in noise variance leads to more peaky (informative) posteriors in the (α, r)-space. Note also that

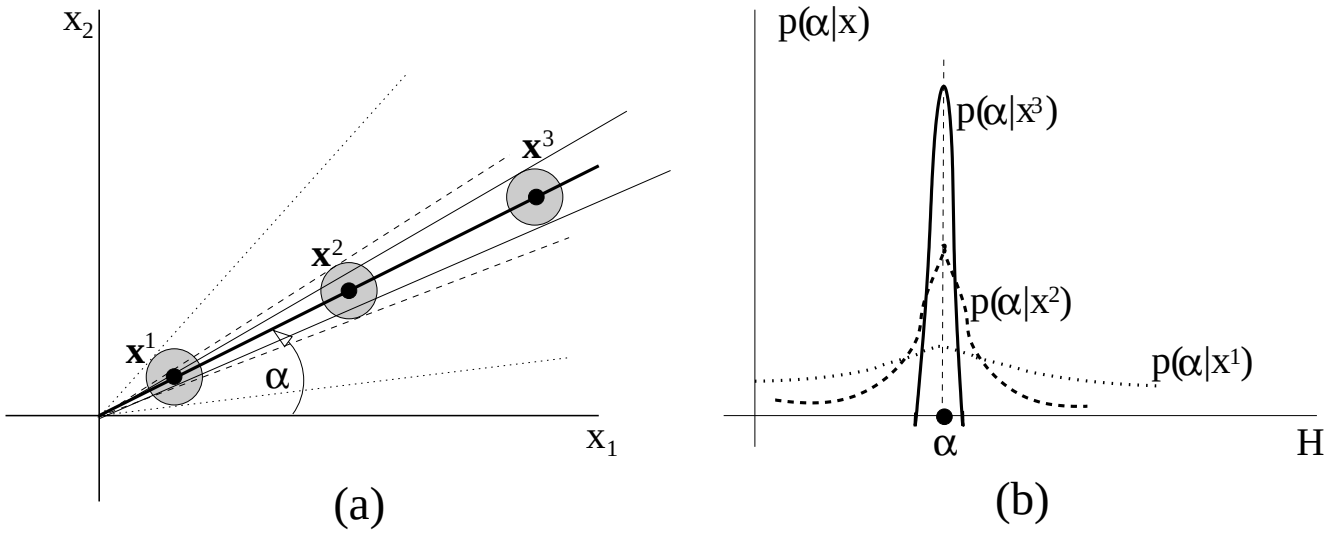

Fig. 5. Illustration of the model based approach to the HT used in this study. Three observations $\mathbf{x}^i$, $i = 1, 2, 3$, are aligned along a ray of line (bold solid line in (a)) with angle α. Discs surrounding the observations signify the measurement noise. The closer is the observation towards the origin, the greater is the induced uncertainty about the actual angle in the Hough space $\mathcal{H}$. The uncertainty in the angle estimates associated with observations $\mathbf{x}^1$, $\mathbf{x}^2$ and $\mathbf{x}^3$ is illustrated in (a) by the pairs of dotted, dashed and solid lines, respectively. The support kernel $p(\alpha|\mathbf{x}^1)$ of the first observation will be least informative (dotted bold line in (b)), while the support kernel $p(\alpha|\mathbf{x}^3)$ of the last observation will be most peaky (solid bold line in (b)).

the further away the observation is from the origin, the more informative the corresponding posterior over α becomes. The marginal contributions $p(\alpha|\mathbf{x}^i)$ to the Hough space "accumulator" are shown in figures 6(b) and (d) for $\sigma = 0.2$ and $\sigma = 0.05$, respectively.

One can interpret $H(\alpha; \mathcal{D})$ in (4) as a form of the Parzen window estimator of the density of the angle parameters in $\mathcal{H}$. Modes of $H(\alpha; \mathcal{D})$ detect angles with maximum support, given our model formulation. Of course, an alternative route may be to estimate the posterior distribution over $\mathcal{H}$ given the full data $\mathcal{D}$. However, this cannot be done under a single line model, as there will typically be more line candidates with substantial amount of points aligned along them. In that case we could have opted for a mixture model setting, with mixture components formulated as noisy line rays starting in the origin. Many points would still not be sufficiently explained by few focused line segments and a "garbage collector" mixture component would need to be invoked. This is reminiscent of a "mixture-like" approach recently suggested in [18], where elaborate sampling in the parameter and model spaces in a constrained setting is used in the model inference stage.

Whereas the standard HT and its variants were mostly designed for line detection in image processing, taking into account specific features such as intensity of gradient information, here we assume that the observed data represents a cloud of distinct points (not necessarily image related) and *some* of those points can be aligned along specific parametrized geometric shapes (e.g. ray of line starting in the origin). In addition, the number of such geometric objects with data aligned along them can be rather small with most data points "unexplained". In such cases, fitting of mixture formulations can be unstable and the HT strategy of searching for geometric objects by detecting peaks in the accumulator represents a more straightforward and robust approach. Unlike in the original HT, we formulate the model using a continuous

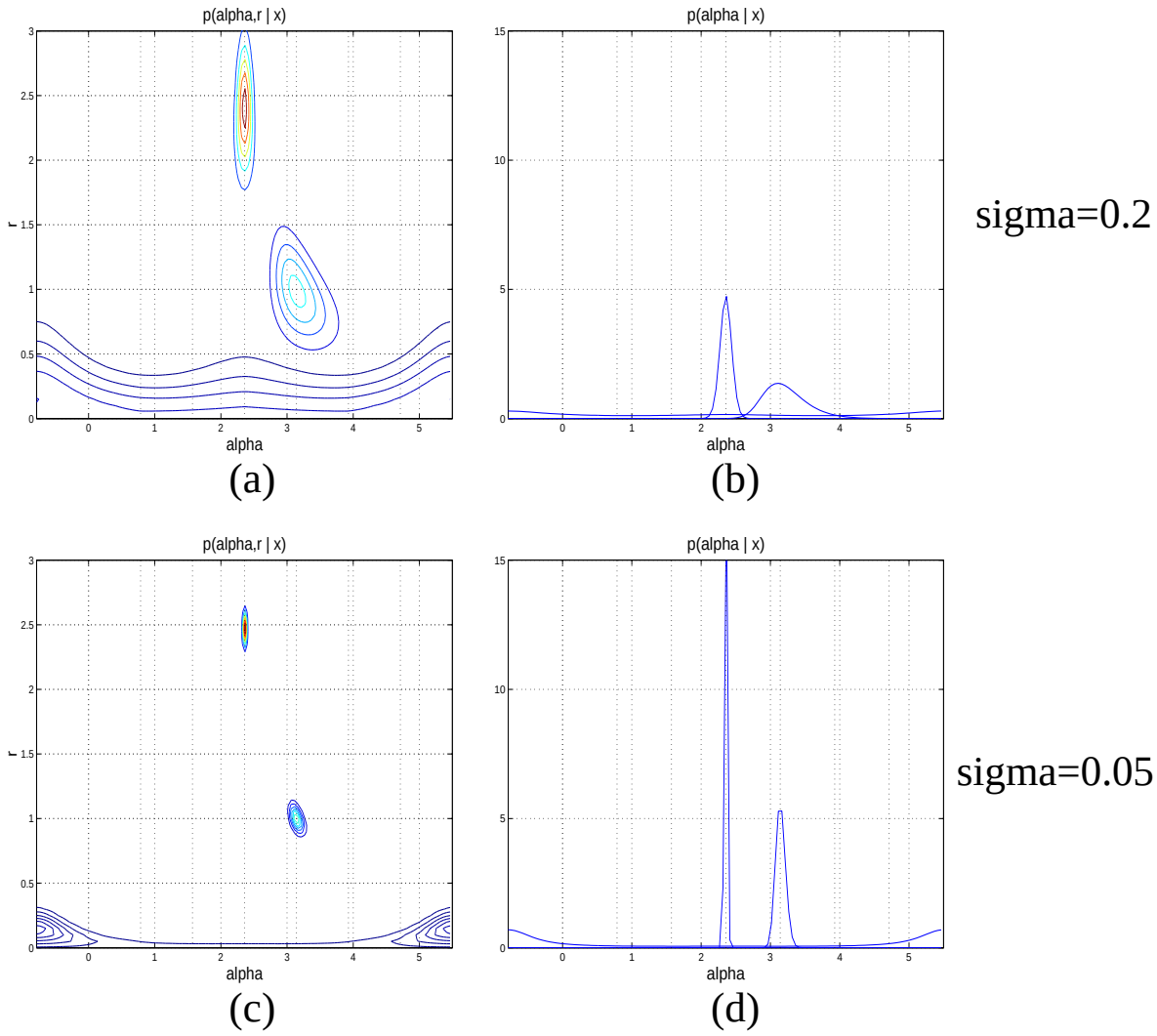

Fig. 6. Posteriors in the parameter space for observations $\mathbf{x}^1 = (0.1, -0.1)^T$, $\mathbf{x}^2 = (-1, 0)^T$ and $\mathbf{x}^3 = (-1.75, 1.75)^T$ lying on rays with angles $-\pi/4$, π and $3\pi/4$, respectively. (a) $p(\alpha, r|\mathbf{x}^i)$, $\sigma = 0.2$; (b) $p(\alpha|\mathbf{x}^i)$, $\sigma = 0.2$; (c) $p(\alpha, r|\mathbf{x}^i)$, $\sigma = 0.05$; (d) $p(\alpha|\mathbf{x}^i)$, $\sigma = 0.05$.

Hough space that can be discretized for practical purposes. In the HT, discretization of the Hough space is a essential and quantization interval length is crucial, as it implicitly determines robustness of HT to noise. In our model the quantization interval can be arbitrarily fine, computational cost permitting, without significantly effecting the model properties.

III. Assessing the effect of Rg1 on homocysteine-treated human umbilical vein endothetial cells

In this section we will apply the methodology developed above to the analysis of the drug Rg1 (dominant compound of the extract of ginsenosides in ginseng) on homocysteine-treated human umbilical vein endothetial cells (HUVEC). The data has been previously analyzed using the standard HT method applied to the hexaMplot in [7]. There are 1128 genes assayed in four microarrays obtained in four repeats under the same experimental conditions. The original microarray data was normalized using the nonlinear Loess

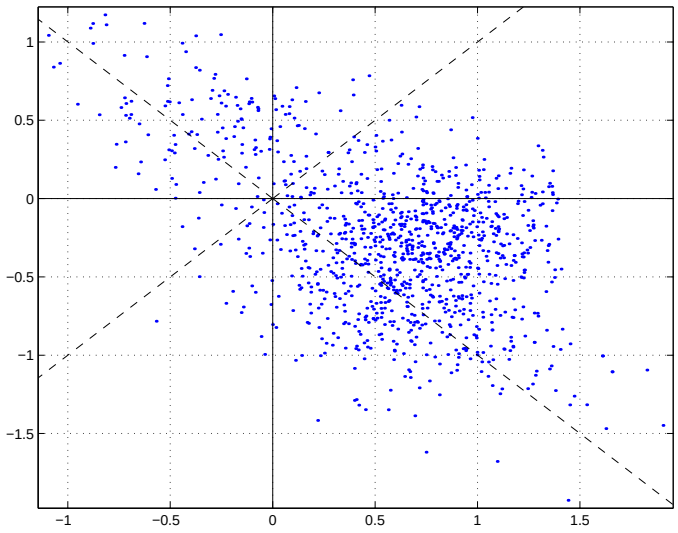

Fig. 7. Mean hexaMplot of the normalized data in the Rg1 drug experiment on homocysteine-treated human umbilical vein endothetial cells. The dashed lines show rays with angles $-\pi/4$, $\pi/4$, $3\pi/4$ and $5\pi/4$.

method[3] [20]. The mean hexaMplot of the normalized data (see section I) is shown in figure 7.

Recall that genes distributed along the same line ray starting in the origin show similar expression patterns and drug effects. Detection of obvious line rays in figure 7 would be problematic, if working directly in the hexaMplot coordinates. However, when concentrating on the accumulated support in the Hough space of angles from individual observations, the main tendencies can be picked up robustly.

We applied the mild inhomogeneity setting for prior[4] $p(\alpha, r)$ illustrated in figure 4, namely $C_\alpha = C_r = 0.9$ and $\omega_1 = 1.9$, $\omega_2 = 2$.

In figures 8(a),(b) and 9 we show the Hough accumulator $H(\alpha; \mathcal{D})$ (4) for three levels of noise standard deviation: $\sigma = 0.5$, $\sigma = 0.01$ and $\sigma = 0.05$, respectively. Vertical dashed lines indicate dominant peaks in the accumulator. As expected, lowering the noise variance results in more peaky support kernels $p(\alpha|\mathbf{x})$ and consequently in a less smooth accumulator. We ran an "annealing process" starting with a large noise variance $\sigma = 1$ and, as the variance decreases, we detect the emerging line ray (gene group) candidates. In figure 8(a) only two ray candidates can be detected, roughly corresponding to angles $\alpha = -\pi/4$ and $\alpha = 3\pi/4$. These line rays explain most of the data in the 4th and 2nd quadrants of the hexaMplot in figure 7. At $\sigma = 0.01$ (figure 8(b)) the Hough accumulator is too rugged with many ray candidates (gene groups) supported by few data points (genes). Figure 10 presents the six detected line rays (solid bold lines) for $\sigma = 0.05$ (figure 9), together with the selected points (stars) $\mathcal{S}_\theta(\tilde{\alpha})$ (6) supporting the lines (see

[3]The data was normalized by robust local regression in the MA plot [19] within arrays and then the scales were adjusted between the arrays as proposed in [20]. The within and between arrays normalization was performed using Limma package written in R language [21].

[4] One might wonder about the mismatch between our prior (centered at the origin of the hexaMplot) and the data distribution in figure 7. However, note that the role of prior distribution is to express our ideas about the distribution of items of interest (in our case - cylindric coordinates of the 2-dimensional gene hexaMplot representations) *prior to seeing the actual data measurements*. As explained in section II-B, one may naturally expect the gene representations to be centered around the origin (most of the genes will not be effected by the disease or drug treatment). There is no reason for the distribution of the actual measured data to follow every detail of the prior distribution. The data we work with simply reflect the gene selections made by the biologists when designing the experiment.

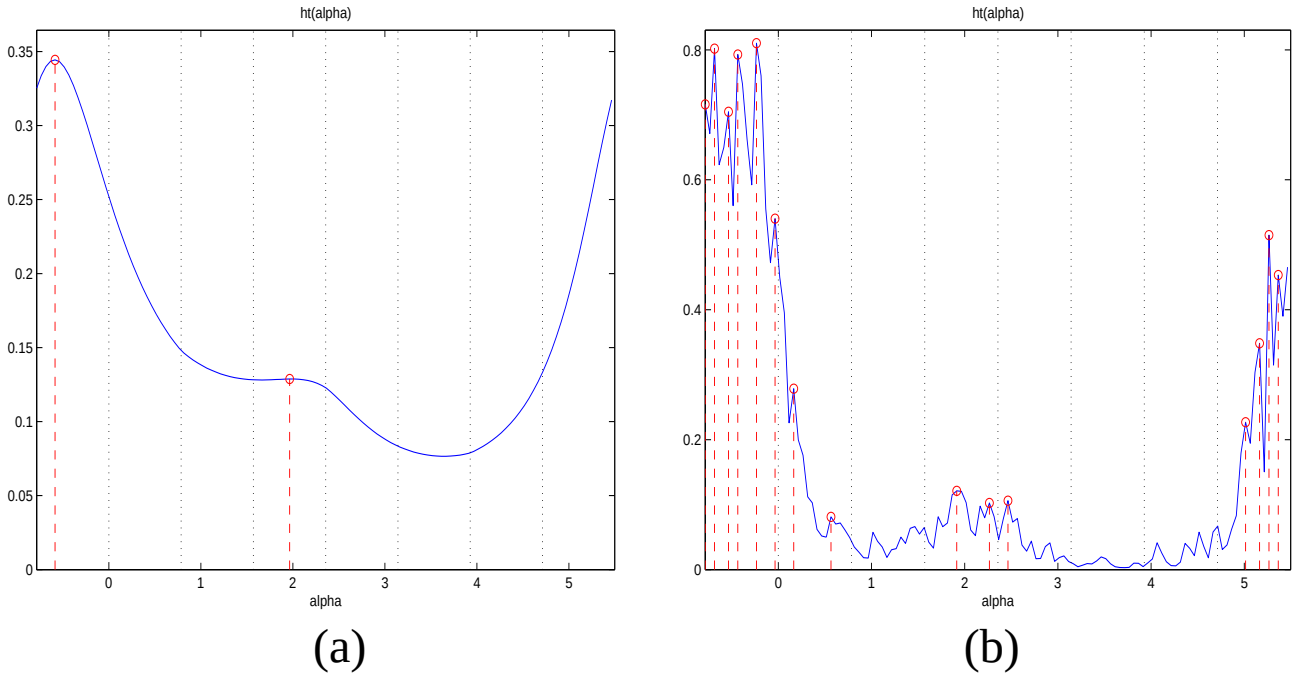

Fig. 8. Hough accumulator $H(\alpha; \mathcal{D})$ for 2 levels of noise standard deviation: $\sigma = 0.5$ (a) and $\sigma = 0.01$ (b).

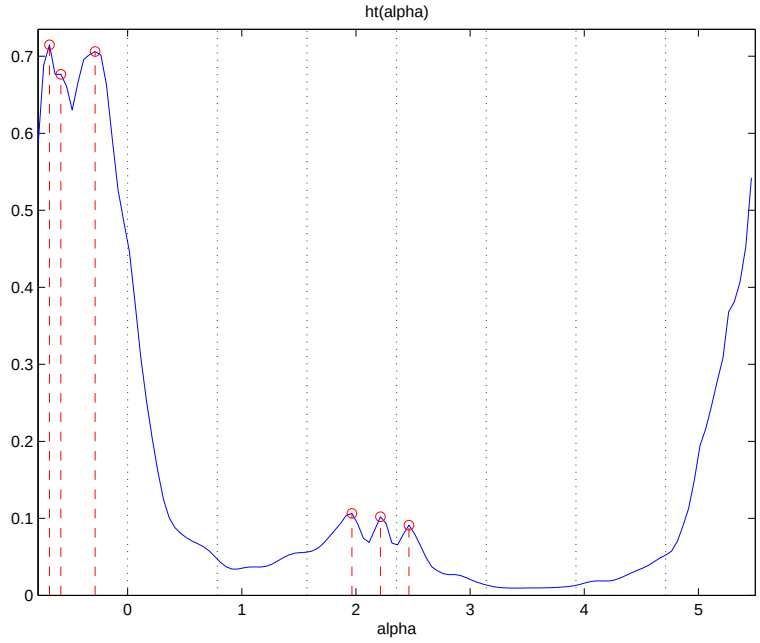

Fig. 9. Hough accumulator $H(\alpha; \mathcal{D})$ for standard deviation of the noise set to $\sigma = 0.05$.

section II). Here, $M = 126$ and $\kappa = 25$, meaning that the angle space $\mathcal{H} = [-\pi/4, 7\pi/4)$ was discretized into 126 values $\{\tilde{\alpha}_i\}$ and the selection threshold was $\theta = \kappa/M = 0.2$.

The posteriors $P(\tilde{\alpha}|S_\theta(\tilde{\alpha}))$ (7) of the detected lines are shown in figure 11(a). It is clear that the selected genes $S_\theta(\tilde{\alpha}))$ support the detected lines (gene groups) - especially those in the 4th quadrant of the hexaMplot - very strongly. We also plot the posteriors of those six detected gene groups for the case of a larger observational noise $\sigma = 0.3$ (figure 11(b)). Moreover, for three levels of observational

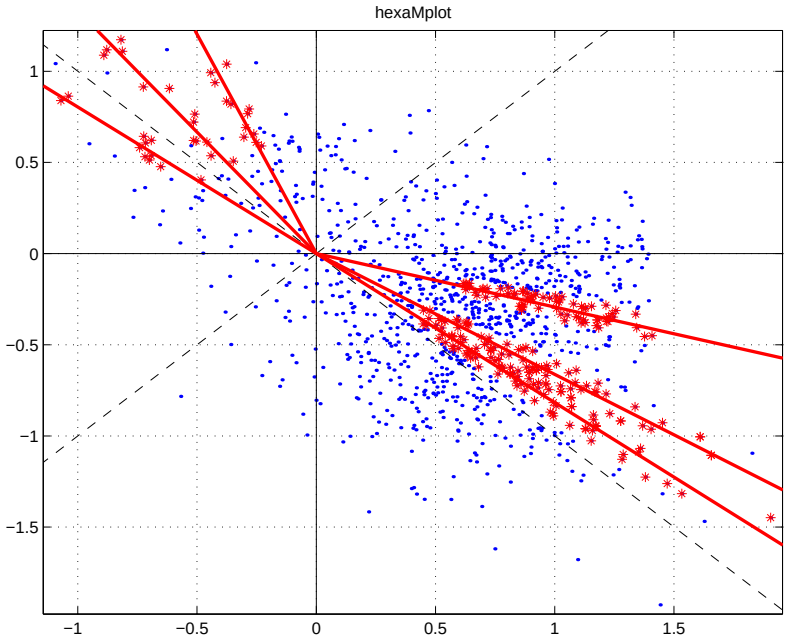

Fig. 10. Detected line rays (solid bold lines) for $\sigma = 0.05$ and selected points supporting those lines (stars). The selected points $\mathcal{S}_\theta(\alpha)$ were chosen using $M = 126$ and $\kappa = 25$.

line	$\tilde{\alpha}$	$(\tilde{\alpha}_-(0.05), \tilde{\alpha}_+(0.05))$	$(\tilde{\alpha}_-(0.3), \tilde{\alpha}_+(0.3))$	$(\tilde{\alpha}_-(1.0), \tilde{\alpha}_+(1.0))$
1	-0.685	(-0.735,-0.684)	(-0.736, -0.683)	(-0.885,-0.535)
2	-0.585	(-0.635,-0.584)	(-0.635, -0.535)	(-0.785,-0.435)
3	-0.285	(-0.335,-0.284)	(-0.385, -0.235)	(-0.685,-0.235)
4	1.965	(1.914,1.966)	(1.665, 2.265)	(-0.035, 3.015)
5	2.215	(2.165,2.216)	(2.015, 2.365)	(1.565, 2.765)
6	2.465	(2.415,2.466)	(2.265, 2.615)	(1.865, 3.015)

TABLE I

THE SHORTEST INTERVALS $(\tilde{\alpha}_-(\sigma), \tilde{\alpha}_+(\sigma))$ CONTAINING THE ESTIMATED LINE ANGLES AND 95% OF THE POSTERIOR MASS $P(\cdot|S_\theta(\tilde{\alpha}))$ AROUND THEM. WE SHOW THE INTERVALS FOR THREE LEVELS OF OBSERVATIONAL NOISE: $\sigma = 0.05$, $\sigma = 0.3$ AND $\sigma = 1.0$.

noise $\sigma = 0.05, 0.3, 1.0$, the shortest "quantile" intervals $(\tilde{\alpha}_-(\sigma), \tilde{\alpha}_+(\sigma))$ containing the estimated angles and 95% of the posterior mass $P(\cdot|S_\theta(\tilde{\alpha}))$ around them are reported in table I. Intervals $(\tilde{\alpha}_-(\sigma), \tilde{\alpha}_+(\sigma))$ represent the uncertainty one has in the point estimates of line angles, given the support of the selected points $S_\theta(\tilde{\alpha})$. Note that unlike in the traditional HT, this uncertainty measure follows from a principled model formulation that reflects our assumptions about the data generation process.

From the four available repeats of the microchips, the obtained rough estimates of standard deviation of the hexaMplot noise were below 0.3 for most genes. For such noise levels, the selected genes in the 4th quadrant of the hexaMplot still support the corresponding three gene groups quite strongly. In general, since the support kernels $p(\alpha|\mathbf{x})$ in the Hough accumulator $H(\alpha; \mathcal{D})$ (4) are posteriors over angles given a single observation, the resulting accumulator may be oversmooth. Hence, we detect the emerging dominant

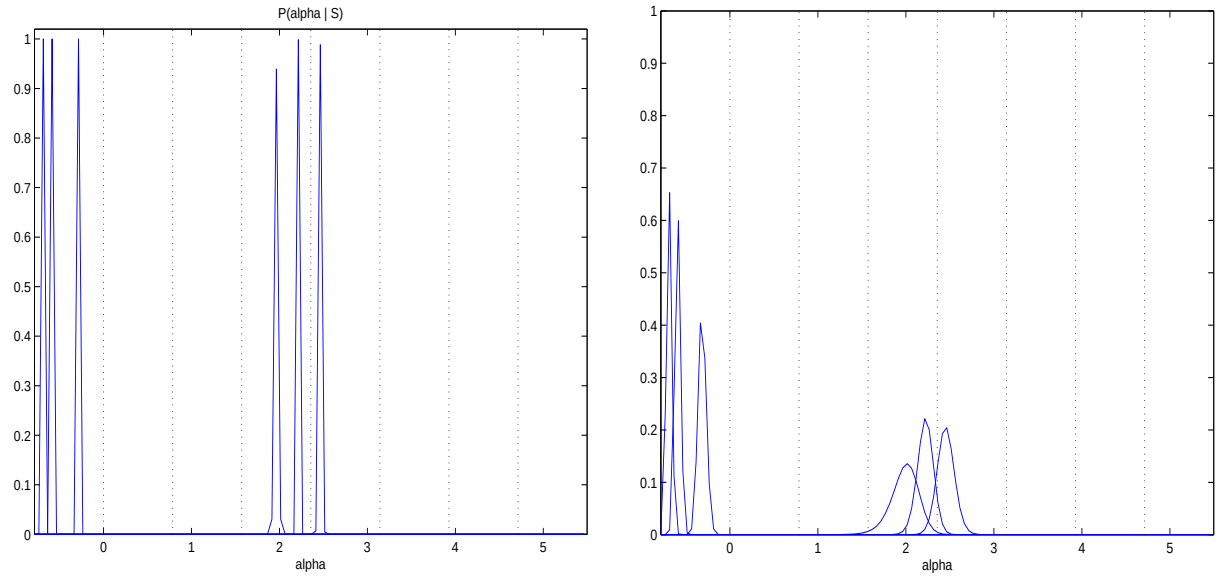

Fig. 11. Posteriors $P(\alpha|S_\theta(\alpha))$ of the six detected lines (gene groups) in figure 10 for two levels of observational noise: $\sigma = 0.05$ (a) and $\sigma = 0.3$ (b).

gene groups by annealing the HT through lowering the variance parameter σ^2 as described above, and for detected groups with solid gene support we calculate posteriors $P(\tilde{\alpha}|S_\theta(\tilde{\alpha}))$ given the full set $S_\theta(\tilde{\alpha}))$ of group members and a realistic estimate of the noise level.

We investigated the biological meaning of the six detected groups of genes (line rays in the hexaMplot) within the Gene Ontology (GO) framework [9], [10], [11], [12]. The results are summarized in table II. The table has the following organization: For each line we report the number $|\mathcal{S}_\theta(\tilde{\alpha})|$ of detected genes (second column). Representative GO terms for genes in $\mathcal{S}_\theta(\tilde{\alpha})$ are listed in the third column. For each GO term we report the number of genes annotated to that term in the GO (fourth column), as well as the number of genes from the gene group $\mathcal{S}_\theta(\tilde{\alpha})$ annotated to it (fifth column). In the sixth column we show the probability (p-value) that genes from $\mathcal{S}_\theta(\tilde{\alpha})$ would get annotated to the GO term by chance [12]. The p-values are calculated using hypergeometric distribution. If an annotation file contains n genes, a given GO term has m annotated genes, and a gene group contains q genes of interest, the probability of seeing k or more genes of interest annotated to that GO term is determined as[5] [12]:

$$\text{p-value} = \sum_{j=k}^{q} \frac{\binom{m}{j}\binom{n-m}{q-j}}{\binom{n}{q}}. \tag{17}$$

[5] Note that such calculations can lead to seemingly overly small p-values in cases of small number of genes from a gene group that are included in a GO term, e.g. line 4, GO: 0002439. Here $m = 86$ genes from the n related human genes are directly annotated to GO: 0002439 and $k = 1$ gene in line 4 cluster is annotated to that GO term. Due to the very large n and not very small m, the p value may be very small even if only one gene in our cluster is annotated to the GO term. This situation is common in GO annotation analysis of gene clusters. It is still an open problem how to select annotated GO terms of the genes of interest. Some researchers only consider GO terms with the smallest p-value. Others are not only interested in the GO term with the smallest p-value (e.g. < 0.01), but also require that the number of genes from the cluster annotated to the term be no less than a certain cutoff value (e.g. ≥ 5). In this study, the GO term with the smallest p-value is used without considering the number of genes annotated.

The first three lines with angles in $(-\pi/4, 0)$ represent genes with (R, G, B) intensities satisfying $G < R < B$. In other words, the disease decreases expression of a gene, compared with its normal expression level R, i.e. $G < R$. The drug eliminates this effect by overexpressing the genes, $B > R$. Genes in the group corresponding to the 1st line are related to acute inflammatory response (GO:0002675, GO:0002525) increasing for example the concentration of non-antibody proteins in the plasma (GO:0006953), or increasing the intra- or extra-cellular levels of prostaglandin (GO:0002539) and leukotriene (GO:0002540). Genes clustered along the 2nd line are related to cellular components and mechanisms effected by the disease. The 3rd line groups genes that are related to binding mechanisms (GO:0005488) and breakdown of neutral lipids (GO:0046461), membrane lipids (GO:0046466) and glycerolipids (GO:0046503). The disease also down-regulates genes related to pathways of the complement cascade which allow for the direct killing of microbes as well as regulation of other immune processes (GO:0001867, GO:0006957). The drug Rg1 corrects this situation by stimulating the pathways.

The 4th and 5th lines with angles in $(\pi/2, 3\pi/2)$ represent genes with (R, G, B) intensities satisfying $R < B < G$. Compared with its normal expression level, the expression of a gene is increased by the disease ($G > R$). The drug partially eliminates this effect by reducing the expression level to B, leaving B still above the normal expression R. Finally, the sixth line with $\alpha \in (3\pi/2, \pi)$ groups genes with (R, G, B) intensities satisfying $B < R < G$. The disease causes increased expression of a gene ($G > R$) and the drug compensates for this effect by driving the gene expression below the normal level ($B < R$). While genes grouped together by the 4th line are associated with immune and chronic inflammatory response, the genes corresponding to the 5th and 6th lines are again related to cellular components and mechanisms effected by the disease.

One can legitimately ask how stable are our results with respect to setting of the prior parameters[6]. We repeated the whole experiment under two less mild inhomogeneity settings $C_\alpha = C_r = 0.8$, $\omega_1 = 1.8$, $\omega_2 = 2$ and $C_\alpha = C_r = 0.7$, $\omega_1 = 1.7$, $\omega_2 = 2$. We also performed the experiments under uniform prior over the angles ($C_\alpha = 1$) and a simple Gaussian prior over r ($C_r = 1$) with standard deviation $\omega_r = \omega_2$ set to $\omega_r = 2, 3, 4, 10, 60$. Note that the setting $C_\alpha = C_r = 1, \omega_r = 60$ represents virtually uniform prior, as the radius of the data does not exceed 2.

The six peaks in the original Hough accumulator (figure 9) for $\sigma = 0.05$ dominated the Hough accumulators in all 7 prior settings described above. Being assured that the six lines in the hexaMplot are robustly recovered under a wide range of prior settings, we next investigated whether the leading GO terms in table II determined from gene groups $\mathcal{S}_\theta(\tilde{\alpha})$ would stay unchanged regardless of the prior settings. In table III we report for each investigated prior and each line $\tilde{\alpha}$ the number of genes N_δ by which the gene group $\mathcal{S}_\theta(\tilde{\alpha})$ differs from that of our original prior setting $C_\alpha = C_r = 0.9$, $\omega_1 = 1.9$, $\omega_2 = 2$. The gene groups corresponding to lines 1, 4, 5 and 6 are perfectly recovered for all tested prior settings. The gene groups for lines 2 and 3 differed by at most 2 genes, which is a negligible difference, given that the number of genes in original groups corresponding to lines 2 and 3 is 64 and 71, respectively. Indeed, the

[6]We are thankful to anonymous referees for suggesting to perform a systematic quantitative study of this issue.

line	$\lvert\mathcal{S}_\theta(\tilde{\alpha})\rvert$	GO term ID	# genes	# genes from $\mathcal{S}_\theta(\tilde{\alpha})$	p-value
1	80	GO:0002675	87	37	0.00173
		GO:0002525	85	36	0.00242
		GO:0006953	85	36	0.00242
		GO:0002527	86	36	0.00322
		GO:0002543	86	36	0.00322
		GO:0002539	60	27	0.00441
		GO:0002540	60	27	0.00441
2	64	GO:0044424	176	41	0.00000
		GO:0044444	175	41	0.00000
		GO:0044446	168	39	0.00294
		GO:0030117	165	38	0.00437
		GO:0045265	162	37	0.00536
3	71	GO:0005488	178	51	0.00000
		GO:0050794	163	55	0.00295
		GO:0046461	88	19	0.00381
		GO:0046466	88	19	0.00381
		GO:0046503	88	19	0.00381
		GO:0001867	104	24	0.00382
		GO:0006957	104	24	0.00382
4	12	GO:0005488	178	9	0.00000
		GO:0002439	86	1	0.00597
5	13	GO:0005488	178	9	0.00000
		GO:0044464	175	6	0.00000
		GO:0000502	164	5	0.00000
6	11	GO:0005488	178	6	0.00000
		GO:0043231	157	3	0.00340
		GO:0007242	140	2	0.00490

TABLE II

THE EFFECTS OF THE DRUG RG1 ON HCY-TREATED HUVE CELLS. FOR EACH GENE GROUP (LINE RAY IN FIGURE 10) WE SHOW REPRESENTATIVE GENE ONTOLOGY TERMS AND THE STRENGTH OF THEIR ASSOCIATION WITH THE GENE GROUP (P-VALUE).

Prior	line 1	line 2	line 3	line 4	line 5	line 6
$C_\alpha = C_r = 0.7$, $\omega_1 = 1.7$	0	2	2	0	0	0
$C_\alpha = C_r = 0.8$, $\omega_1 = 1.8$	0	1	1	0	0	0
$C_\alpha = C_r = 1$, $\omega_r = 2$	0	0	0	0	0	0
$C_\alpha = C_r = 1$, $\omega_r = 3$	0	0	0	0	0	0
$C_\alpha = C_r = 1$, $\omega_r = 4$	0	0	0	0	0	0
$C_\alpha = C_r = 1$, $\omega_r = 10$	0	0	1	0	0	0
$C_\alpha = C_r = 1$, $\omega_r = 60$	0	0	1	0	0	0

TABLE III

CARDINALITY OF THE SET DIFFERENCE BETWEEN THE GENE GROUPS $\mathcal{S}_\theta(\tilde{\alpha})$ DETECTED UNDER THE ORIGINAL PRIOR SETTING $C_\alpha = C_r = 0.9, \omega_1 = 1.9, \omega_2 = 2$ AND THE OTHER TESTED PRIOR SETTINGS. THE SIZES OF THE ORIGINAL SETS $\mathcal{S}_\theta(\tilde{\alpha})$ CORRESPONDING TO LINES 1, 2, 3, 4, 5 AND 6 ARE 80, 64, 71, 12, 13 AND 11.

GO analysis confirmed that all leading GO terms reported in table II are robustly recovered for all tested prior settings.

A. Comparison with alternative clustering and correlation based approaches

We conclude the experiments by comparing our approach with some other clustering and correlation based approaches to gene grouping based on gene expression profiles. There is, of course, a huge number of general approaches for gene grouping (see e.g. [22]). We used three representative approaches to highlight potential strengths/weaknesses of our approach for grouping genes in the context of assessing drug effects through 3-color cDNA arrays.

The genes were grouped using a probabilistic clustering method (Gaussian mixture model (GMM)), a non-probabilistic "hard" clustering method (K-means clustering) and a correlation based method (Average Correlation Clustering Algorithm (ACCA) [23]). The driving force behind gene grouping in clustering based approaches is a distance between expression profiles of individual genes (Euclidean distance in the case of K-means clustering, Mahanabolis distances in the case of GMM). On the other hand, in correlation based approaches it is the correlation ('angle') between expression profiles that determines the gene similarity. We considered two gene expression profiles for gene grouping:

- **3D** - the original three (log) intensities of gene expression under normal, disease and drug-treated conditions. Models fitted on such data will be referred to by GMM-3D, K-means-3D and ACCA-3D.
- **2D** - two-dimensional hexaMplot gene representations obtained from the original intensities as described in section I. Models built on hexaMplot gene representations will be denoted by GMM-2D, K-means-2D and ACCA-2D.

Gaussian mixture model is a generative probabilistic model of the data and so some sort of principled model selection can be applied to specify the number of components. We used cross-validated model likelihood [24] and the optimal number of 2- and 3-variate Gaussian components in GMM-2D and GMM-3D were 6 and 4, respectively. A 6-component GMM-2D fitted to the hexaMplot gene representations is

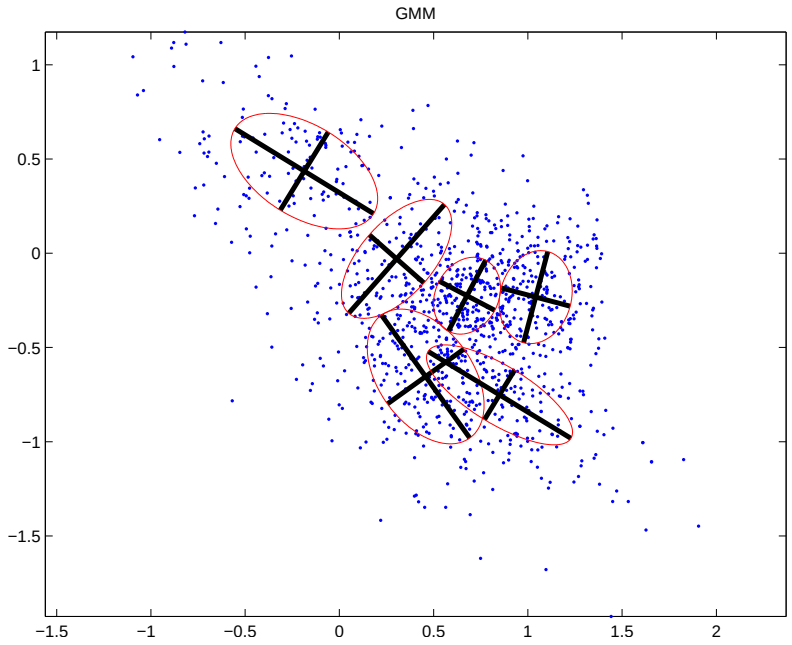

Fig. 12. Gaussian mixture model with 6 mixture components fitted on the hexaMplot gene representations (shown as dots). For each 2-variate Gaussian mixture component we show the axes of ellipses of constant density. The widths of the axes are given by one standard deviation.

illustrated in figure 12. The genes were formed around individual mixture components based on posterior probabilities of the components j, given a gene g, $P(j|g)$. If for a given gene g and a mixture component j, $P(j|g) > 3/4$, the gene g was assigned to group j. The other two methods, K-means clustering and ACCA, partition the data set in a "hard" manner. Model selection is less straightforward and for ease of comparison with our approach we constructed 6 gene groups in each case (2D and 3D).

As before, the detected gene groups were analyzed for biological meaning within the GO framework. The results are summarized in table IV. For each gene group we report the method under which it was constructed (first column), its size (third column) and the representative GO term[7] (fourth column) associated with the group most strongly (with the smallest p-value[8] (seventh column)). Also shown is the number of genes annotated to that GO term (fifth column) and the number of genes from the gene group annotated to the GO term (sixth column).

Comparing tables II and IV, it appears that the gene groupings by alternative approaches in general lead to larger, more general GO terms. Such GO terms are potentially less specific from the point of view of the underlying biological functionality. Interestingly, gene groups constructed by alternative techniques and the three main groups (1,2 and 3) discovered by our approach overlap with their representative GO terms in roughly similar numbers of genes. Hence, compared with alternative techniques, in our approach

[7]Note that groups 6 (GMM-2D) and 4 (GMM-3D) were not annotated with a GO term since their size was too small. This can happen in probabilistic clustering for 'internal' mixture components j surrounded by other mixture components. Even though many points g get explained by such a mixture component j in the sense that the posterior $P(j|g)$ maximizes $P(\cdot|g)$, $P(j|g)$ is not high enough to merit inclusion of g in the group j.

[8]Note from eq. (17) that the p-values can be very small due to very large number n of annotated human genes in GO and reasonably large GO term sizes m.

model	group	group size	GO term ID	# genes	# genes from group	p-value
ACCA-2D	1	18	GO:0006897	258	4	0.00450
	2	174	GO:0007169	278	19	0.00000
	3	275	GO:0043067	855	54	0.00000
	4	119	GO:0042981	847	27	0.00003
	5	430	GO:0008284	420	64	0.00000
	6	74	GO:0010033	730	16	0.00210
ACCA-2D	1	117	GO:0007169	278	13	0.00097
	2	45	GO:0048534	333	7	0.05800
	3	138	GO:0042981	847	37	0.00000
	4	158	GO:0048514	261	15	0.00009
	5	494	GO:0009611	584	69	0.00000
	6	138	GO:0042325	477	27	0.00000
GMM-2D	1	68	GO:0009611	584	14	0.00055
	2	134	GO:0042981	847	35	0.00000
	3	42	GO:0008406	109	6	0.00460
	4	85	GO:0009611	584	16	0.00053
	5	128	GO:0019220	497	22	0.00000
	6	3	–	–	–	–
GMM-3D	1	564	GO:0010033	730	97	0.00000
	2	131	GO:0007626	292	21	0.01100
	3	228	GO:0042981	847	62	0.00000
	4	1	–	–	–	–
K-means-2D	1	132	GO:0043085	535	24	0.00000
	2	227	GO:0010033	730	42	0.00000
	3	151	GO:0006468	767	28	0.00000
	4	112	GO:0043067	855	30	0.00000
	5	243	GO:0010941	858	47	0.00000
	6	225	GO:0016310	1102	44	0.00000
K-means-3D	1	103	GO:0006468	767	22	0.00006
	2	178	GO:0042981	847	52	0.00000
	3	46	GO:0045321	332	7	0.06500
	4	287	GO:0006955	962	48	0.00000
	5	83	GO:0007166	1976	23	0.05700
	6	393	GO:0006468	767	88	0.00000

TABLE IV

THE EFFECTS OF THE DRUG RG1 ON HCY-TREATED HUVE CELLS. FOR EACH GENE GROUP WE SHOW THE GROUP SIZE AND THE REPRESENTATIVE GENE ONTOLOGY (GO) TERM ASSOCIATED WITH THE GROUP WITH THE SMALLEST P-VALUE. WE ALSO REPORT THE NUMBER OF GENES IN THE GO TERM AND THE NUMBER OF GENES FROM THE GROUP ASSOCIATED WITH THE GO TERM.

smaller gene groups lead to more specific (smaller) representative GO terms (*# genes*), while the numbers of genes from the gene groups inside their representative GO terms are proportionally higher - even though we have smaller groups that yield smaller representative GO terms, the numbers of genes from groups in their representative GO terms (*# genes from* $\mathcal{S}_\theta(\tilde{\alpha})$ in table II) are comparable to their counterparts (*# genes from group* in table IV), that are however obtained from larger gene groups and GO terms. Hence, the gene groupings obtained by our approach appear to be more tightly linked to more specific biological functionalities, which is preferable from the point of view of systems biology.

Of course, one can always modify the basic clustering and correlation based approaches so that smaller and smaller groupings are obtained. One can then hope that even in smaller gene groups there will still be sufficiently many genes related to some important biological function (as detected by the Gene Onthology). Besides great computational expense of such an approach, of more concern is the question of the origin of such groupings. Where do they come from? In what sense do they appear naturally from the data? In our approach, line rays in the hexaMplot group coherently expressed genes as clusters, where the coherence is modeled as linear relations, whose biological interpretations have been studied extensively in biclustering. Indeed, GO appears to provide some evidence for such coherence. In view of drug evaluation, the desirable genes should accumulate along the slant axis of the hexaMplot. Considering the different effect of a drug, the lines passing through the origin of the hexaMplot are used to measure the drug effect levels. The direction of a line can be used to determine whether the drug has positive or negative effect on a group of genes. The probabilistic Hough Transform introduced in this study provides a natural mechanism for finding groups of related genes buried inside a background cloud of other gene representations. Other clustering/correlation based approaches are usually forced to *partition* (in a hard or soft manner) *the whole set of genes*, without being able to focus on 'interesting' groupings, while not paying attention to the rest (see figure 12). On the contrary, we pose a drug-effect related linear relation between gene representations in a group and detect significant groups modulo measurement noise, ignoring the the background cloud of points not showing significant linear groupings (see figure 10).

As mentioned above, it is difficult to apply a principled model selection in the case of K-means clustering, but a principled model selection could be applied to Gaussian mixture model (a "soft" version of K-means): it returned 6 and 4 components on the 2-D and 3-D gene representations, respectively. These gene groupings appear inferior to the ones discovered by our methodology. Of course it would be possible to modify e.g K-means to return clusters of the size of groupings detected by our approach. But the question is, why that particular setting of K-means and not another (smalled/larger clusters)? In our method, the group sizes emerge naturally from the data, they are not dictated by another unrelated method or imposed prior to model building.

We finish by stressing that these results, however encouraging, come from one specific study and the true value of the presented approach can only be determined by its use by the wider community of users and practitioners.

IV. Conclusion

It has been shown that the HT applied to hexaMplot gene representations can be used to detect groups of co-expressed genes in the normal-disease-drug samples [7]. However, the standard HT is not well suited for the purpose because: **(1)** the assayed genes need first to be hard-partitioned into equally and differentially expressed genes, with the HT applied only to the latter ones, ignoring possible information in the former group; **(2)** the hexaMplot coordinates are naturally negatively correlated and there is no direct way of expressing this in the standard HT and **(3)** it is not clear how to calculate in a principled and consistent manner the strength of association of a group of co-expressed genes with the line along which the genes cluster.

In this study we have addressed these deficiencies by formulating a dedicated probabilistic model based HT for detecting gene groups aligned along line rays starting in the origin in the hexaMplot. The nature of noise in the hexaMplot representations is specifically accounted for. All genes are taken into account, but the contribution of genes less differentially expressed[9] to detection of co-expressed gene groups is naturally suppressed. When finding the co-expressed gene groups we apply the annealing process driven by the decreasing noise variance parameter. As the noise variance decreases, more and more gene groups emerge analogously to the emergence of increasing number of increasingly detailed clusters in deterministic annealing (e.g. [25]). When a gene group with a solid gene support emerges, we quantify the confidence of detecting the group by calculating the full posterior of the corresponding line ray, given the gene group, for realistic noise estimates. To represent the uncertainty about point estimates of line angles, given the support of the selected points, we also calculate the shortest intervals containing the estimated line angles and 95% of the posterior mass around them. No such “confidence/reliability” quantitative measures follow naturally from the standard HT formulation.

Inclusion of all assayed genes in our analysis enabled us to robustly detect stronger natural groupings of co-expressed genes than those found in the previous study [7]. Whereas [7] reported 15 gene groups of size 4–12, we have found a smaller number of naturally emerging groups (6), three of which have significantly stronger gene support (64–80 genes in a group). The posteriors of these three groups under realistic noise estimates show solid support for their detection and, perhaps more importantly, the three gene groups show coherent biological functions with high significance, as detected by the GO analysis. Moreover, when compared with some general clustering and correlation based gene grouping techniques (Gaussian mixture modeling, K-means clustering, Average Correlation Clustering Algorithm), gene groups obtained by our approach appear to be more tightly linked to more specific biological functionalities. Robust detection of larger gene groups with coherent biological function is potentially of great importance for robust analysis of drug effects via 3-color cDNA normal-disease-drug sample microarrays. However, our encouraging results come from one specific study. To assess the true value of our approach it will need to be applied in a range of contexts by the wider community of users and practitioners.

Acknowledgments

[9] Closer to the origin of the hexaMplot.

Peter Tiňo was supported by the DfES UK/Hong Kong Fellowship for Excellence. This work was also supported by a grant from the Hong Kong Research Grant Council (Project CITYU123809).

REFERENCES

[1] D. Amaratunga and J. Cabrera, *Exploration and Analysis of DNA Microarray and Protein Array Dat*. New Jersey: Wiley-Interscience, 2004.

[2] C. Debouck and P. Goodfellow, "DNA microarrays in drug discovery and development," *Nature Genetics*, vol. 21, pp. 48–50, 1999.

[3] D. Gresham, M. Dunham, and D. Botstein, "Comparing whole genomes using DNA microarrays," *Nature Reviews Genetics*, vol. 9, pp. 291–302, 2008.

[4] Y. Cho, J. Meade, J. Walden, X. Chen, Z. Guo, and P. Liang, "Multicolor fluorescent differential display," *Biotechniques*, vol. 30, pp. 562–572, 2001.

[5] G. Tsangaris, A. Botsonis, and I. P. andF. Tzortzatou-Stathopoulou, "Evaluation of cadmium-induced transcriptome alterations by three color cDNA labeling microarray analysis on a T-cell line," *Toxicology*, vol. 178, no. 2, pp. 135–160, 2002.

[6] H. Zhao, N. Wong, K.-T. Fang, and Y. Yue, "Use of three-color cDNA microarray experiments to assess the therapeutic and side effect of drugs," *Chemometrics and Intelligent Laboratory Systems*, vol. 82, no. 1-2, pp. 31–36, 2006.

[7] H. Zhao and H. Yan, "HoughFeature, a novel method for assessing drug effects in three-color cDNA microarray experiments," *BMC Bioinformatics*, vol. 8, pp. 256–266, 2007.

[8] J. Illingworth and J. Kittler, "A survey of the Hough transform," *Computer Vision, Graphics, and Image Processing*, vol. 44, pp. 87–116, 1988.

[9] E. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J. Cherry, and G.Sherlock, "GO::TermFinderopen source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes," *Bioinformatics*, vol. 20, no. 18, pp. 3710–3715, 2004.

[10] S. Draghici, P. Khatri, R. Martins, G. Ostermeier, and S. Krawetz, "Global functional profiling of gene expression," *Genomics*, vol. 81, pp. 98–104, 2003.

[11] M. A. et al., "Gene Ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, pp. 25–29, 2000.

[12] R. Sealfon, M. Hibbs, C. Huttenhower, C. Myers, and O. Troyanskaya, "GOLEM: an interactive graph-based gene-ontology navigation and analysis tool," *BMC Bioinformatics*, vol. 7, p. 443, 2006.

[13] Q. Ji and R. Haralick, "An optimal bayesian hough transform for line detection," in *Proceedings of the International Conference on Image Processing*, 1999, pp. 691–695.

[14] ——, "Error propagation for the Hough transform," *Pattern Recognition Letters*, vol. 22, no. 6-7, pp. 813–823, 2001.

[15] A. Bonci, T. Leo, and S. Longhi, "A Bayesian approach to the Hough transform for line detection," *IEEE Transactions on Systems, Man and Cybernetics, Part A*, vol. 35, no. 6, pp. 945–955, 2005.

[16] J. Illingworth, G. Jones, J. Kittler, M. Petrou, and J. Princen, "Robust statistical methods of 2D and 3D image description," *Annals of Mathematics and Artificial Intelligence*, vol. 10, no. 1-2, pp. 125–148, 1994.

[17] J. Princen, J. Illingowrth, and J. Kittler, "Hypothesis testing: A framework for analyzing and optimizing Hough transform performance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 4, pp. 329–341, 1994.

[18] N. Toronto, B. Morse, D. Ventura, and K. Seppi, "The Hough transform's implicit Bayesian foundation," in *Proceedings of the International Conference on Image Processing*, 2007, pp. 377–380.

[19] S. Dudoit, Y. Yang, M. Callow, and T. Speed, "Statistical methods for identifying genes with differential expression in replicated cDNA microarray experiments," *Statistica Sinica*, vol. 12, no. 1, pp. 111–139, 2002.

[20] Y. Yang, S. Dudoit, P. Luu, D. Lin, V. Peng, J. Ngai, and T. Speed, "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation," *Nucleic Acids Research*, vol. 30, no. 4, p. e15, 2002.

[21] G. Smyth, "Limma: linear models for microarray data," in *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber, Eds. New York: Springer, 2005, pp. 397–420.

[22] D. Jiang, C. Tang, and A. Zhang, "Cluster analysis for gene expression data: a survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 11, pp. 1370–1386, 2004.

[23] A. Bhattacharya and R. K. De, "Average correlation clustering algorithm (acca) for grouping of co-regulated genes with similar pattern of variation in their expression values," *Journal of Biomedical Informatics*, online PubMed preprint, PMID: 20144735, 2010.

[24] P. Smyth, "Model selection for probabilistic clustering using cross-validatedlikelihood," *Statistics and Computing*, vol. 10, no. 1, pp. 63–72, 2000.

[25] J. Buhmann, “Learning and data clustering,” in *Handbook of Brain Theory and Neural Networks*, M. Arbib, Ed. Bradfort Books, MIT Press, 1995.