

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

Peter Tiño, 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 essayed 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 endothelial 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 essayed 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. Tiño 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).

A two-dimensional representation of R, G and B intensities called *hexaMplot*, naturally suited for assessing the drug effect on essayed 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 essayed genes, i.e. ideally the gene representations in the *hexaMplot* should cluster around the slant axis. Deviations from 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].

A simple methodology to assess the overall therapeutic effect of the drug was proposed in [6]. The correlation coefficient of *hexaMplot* representations of the essayed 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 [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].

While such an approach represents a fruitful and 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 essayed 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.
- 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 essayed 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 endothelial cells in section III. 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

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

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

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

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:

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

The aim of the Hough Transform (HT) and its generalizations is to detect possible line¹ candidates along which some of the data points are aligned. The detection is performed in the parameter space (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 $\mathcal{D} = \{\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^N\}$, $\mathbf{x}^i \in \mathbb{R}^2$, $i = 1, 2, \dots, N$, we accumulate the evidence contributions in the Hough space \mathcal{H} as proposed in [13], [14], namely

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

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 $p(\alpha|\mathbf{x}^i)$, $i = 1, 2, \dots, N$, 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 $G = \{\tilde{\alpha}_1, \tilde{\alpha}_2, \dots, \tilde{\alpha}_M\}$ and turned the densities $p(\alpha|\mathbf{x})$ into probabilities $P(\tilde{\alpha}_j|\mathbf{x})$ over the G :

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

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:

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

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

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

¹extension to other parametrized objects such as circles is straightforward

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

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

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 and priors over the parameters

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 $X_2 = Y_2 - Y_3$. Even if we assume that the measurement errors of the three log intensities Y_1, Y_2 and Y_3 are independent, the hexaMplot coordinates X_1, X_2 will be negatively correlated. Assuming that the measurement noise of the log intensity Y_i is a zero mean Gaussian with variance $\sigma_i^2, i = 1, 2, 3$, (X_1, X_2) will be Gaussian distributed with covariance matrix

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

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

$$A = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}. \quad (10)$$

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

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

As explained in the previous section, for the hexaMplot we would naturally expect more observations in the (positive, negative) or (negative, positive) quadrants of the $(\log B/G, \log G/R)$ coordinate system than in the all positive or all negative quadrants. We express this insight through the piecewise linear prior $p(\alpha)$:

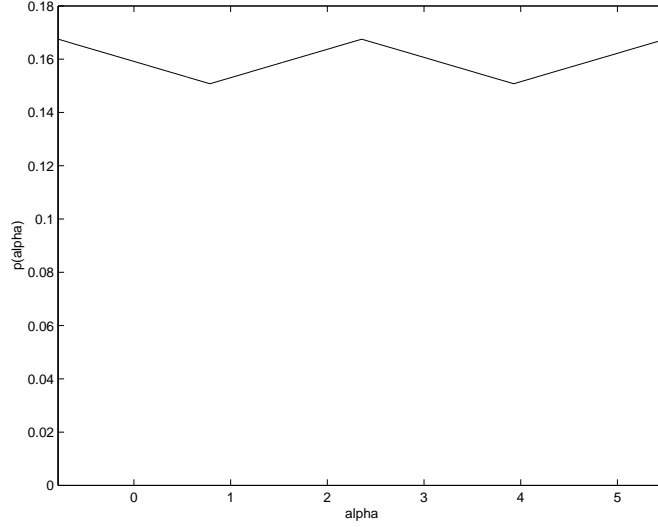


Fig. 1. Prior on angles α for the inhomogeneity parameter setting $C_\alpha = 0.9$.

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

where

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

and $C_\alpha \in [0, 1]$ is a constant determining the strength of inhomogeneity in the prior $p(\alpha)$. The prior is shown in figure 1 for a mild setting of the inhomogeneity parameter $C_\alpha = 0.9$ used in this study.

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 the (positive, negative) or (negative, positive) quadrants of the $(\log B/G, \log G/R)$ system. We express this by formulating a prior on $r \geq 0$ (conditional on the angle α) as a mixture of two truncated Gaussians:

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

where

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

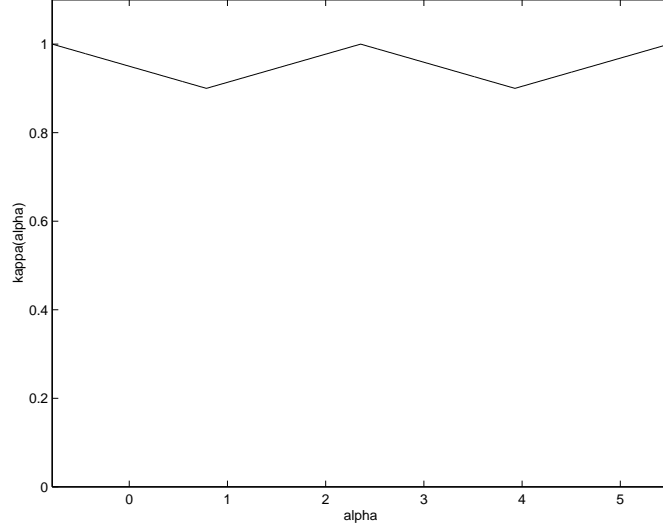


Fig. 2. Mixing coefficient $\kappa(\alpha)$ for $C_r = 0.9$.

with $0 < \sigma_2 < \sigma_1$, and

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

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

We illustrate the angle-conditional mixing coefficient $\kappa(\alpha)$ for $C_r = 0.9$ (used in this study) in figure 2.

The joint prior $p(\alpha, r) = p(\alpha) p(r|\alpha)$ is illustrated in figure 3 for $C_\alpha = C_r = 0.9$ and $\sigma_1 = 1.9$, $\sigma_2 = 2$.

B. Interpretation of the model

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.

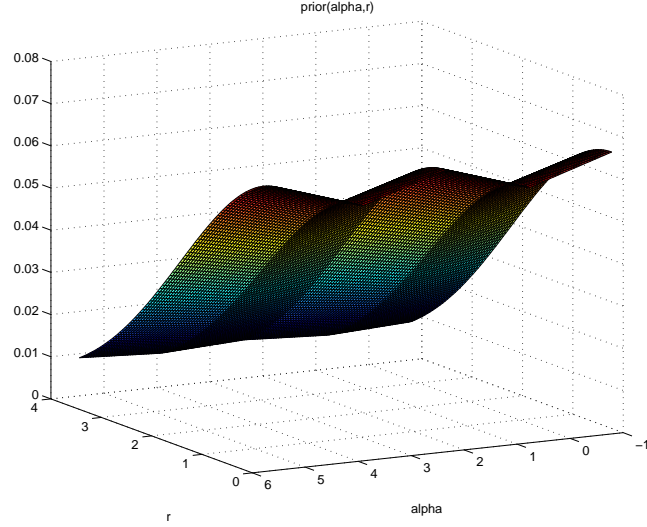


Fig. 3. Joint prior $p(\alpha, r) = p(\alpha) p(r|\alpha)$ for $C_\alpha = C_r = 0.9$ and $\sigma_1 = 1.9$, $\sigma_2 = 2$.

A simple illustration of our approach is presented in figure 4(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 4(b)), while the support kernel $p(\alpha|\mathbf{x}^3)$ will be highly peaked (solid bold line in figure 4(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 the parameter settings outlined above are shown in figure 5. The joint posteriors $p(\alpha, r|\mathbf{x}^i)$ for noise levels $\sigma = 0.2$ and $\sigma = 0.05$ are presented in figures 5(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 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 5(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 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

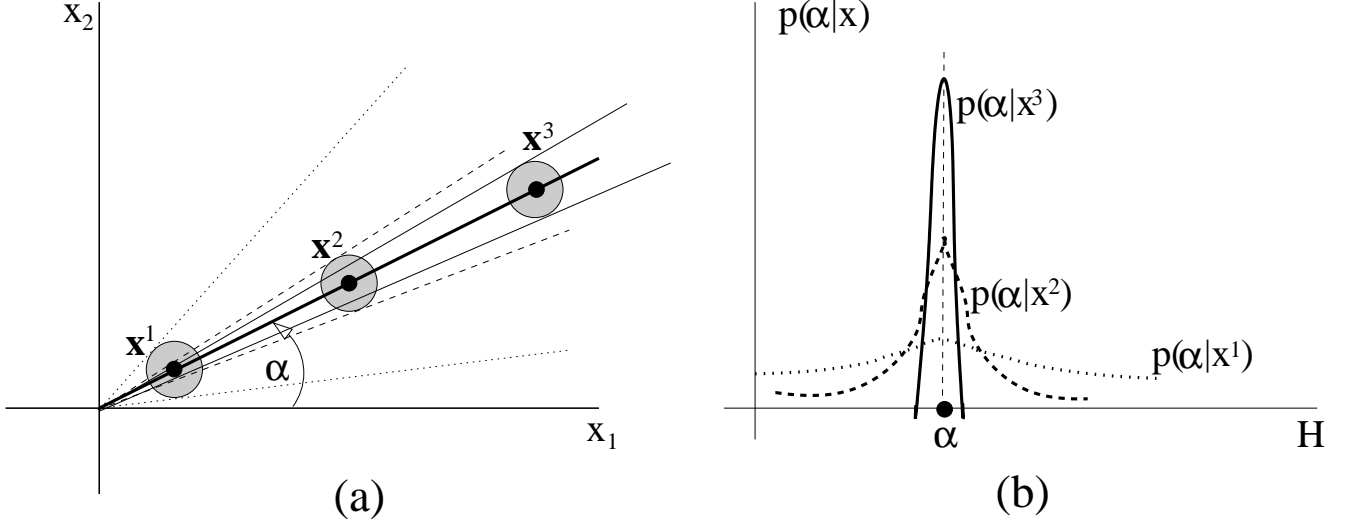


Fig. 4. Illustration of the model based approach to HT used in this study. Three observations \mathbf{x}^i , $i = 1, 2, 3$, are aligned along a ray of line (bold solid line in (a)) with angle α . Discs surrounding the observations signify the measurement noise. The closer is the observation towards the 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)).

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 edge 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 Hough Transform 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 Hough space that can be discretized for practical purposes. In HT, discretization of the Hough space is an 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 affecting the model properties.

III. ASSESSING THE EFFECT OF Rg1 ON HOMOCYSTEINE-TREATED HUMAN UMBILICAL VEIN ENDOTHELIAL 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

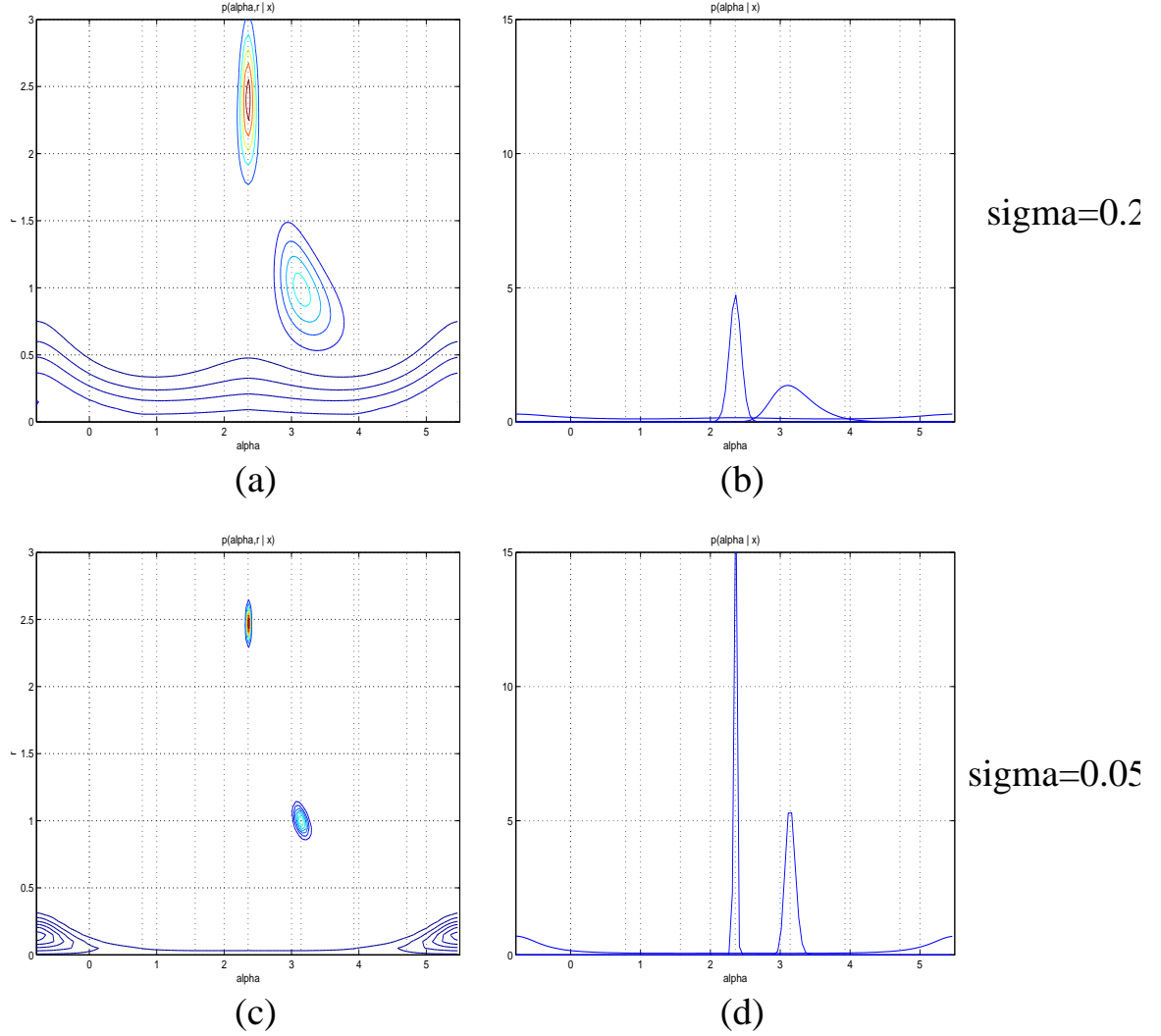


Fig. 5. Posteriors in the parameter space for observations $\mathbf{x}^1 = (0.1, -0.1)^T$, $\mathbf{x}^2 = (-1, 0)^T$ and $\mathbf{x}^3 = (-1.75, 1.75)^T$ lying on rays with angles $-\pi/4$, π and $3\pi/4$, respectively. (a) $p(\alpha, r | \mathbf{x}^i)$, $\sigma = 0.2$; (b) $p(\alpha | \mathbf{x}^i)$, $\sigma = 0.2$; (c) $p(\alpha, r | \mathbf{x}^i)$, $\sigma = 0.05$; (d) $p(\alpha | \mathbf{x}^i)$, $\sigma = 0.05$.

endothelial 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 method [19]. The mean hexaMplot of the normalized data is shown in figure 6.

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 6 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.

In figure 7 we show the Hough accumulator $H(\alpha; \mathcal{D})$ (4) for three levels of noise standard deviation: $\sigma = 0.5$, $\sigma = 0.05$ and $\sigma = 0.01$ (subplots (a), (b) and (c), 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

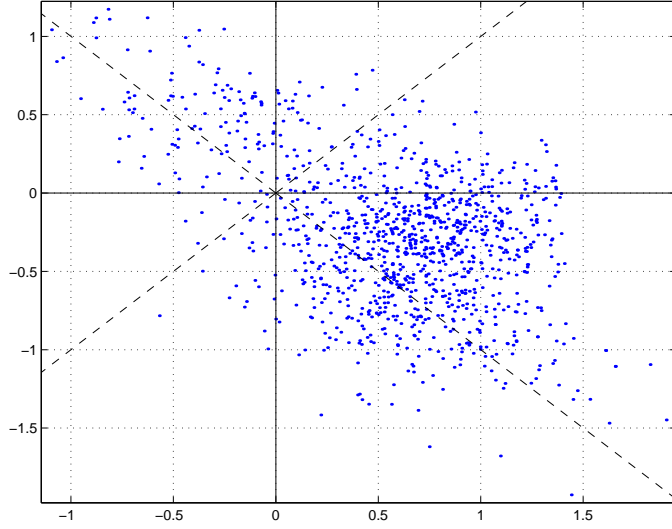


Fig. 6. Mean hexaMplot of the normalized data in the Rg1 drug experiment on homocysteine-treated human umbilical vein endothelial cells. The dashed lines show rays with angles $-\pi/4$, $\pi/4$, $3\pi/4$ and $5\pi/4$.

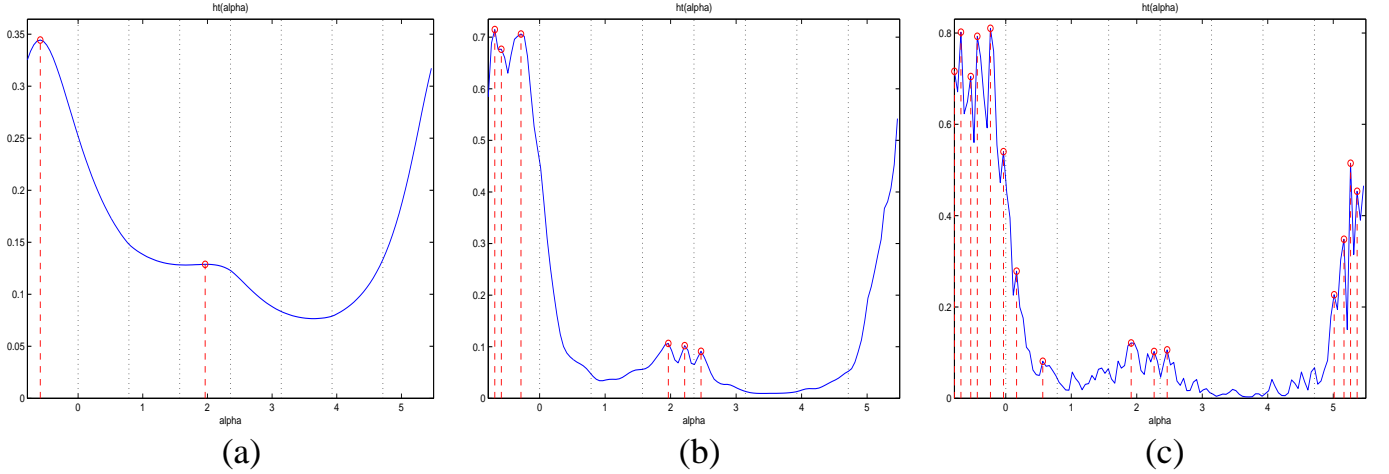


Fig. 7. Hough accumulator $H(\alpha; \mathcal{D})$ for 3 levels of noise standard deviation: $\sigma = 0.5$ (a), $\sigma = 0.05$ (b) and $\sigma = 0.01$ (c).

(gene group) candidates. In figure 7(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 6. Figure 8 presents the six detected line rays (solid bold lines) for $\sigma = 0.05$ (figure 7(b)), together with the selected points (stars) $S_\theta(\tilde{\alpha})$ (6) supporting the lines (see 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 $\theta = \kappa/M = 0.2$. We also show the joint posteriors $p(\alpha, r|\mathbf{x})$ for all the essayed gens (figure 9(a)), as well as their marginals $p(\alpha|\mathbf{x})$ - support kernels in the Hough space (figure 9(b)). At $\sigma = 0.01$ (figure 7(c)) the Hough accumulator is too rugged with many ray candidates (gene groups) supported by few data points (genes).

The posteriors $P(\tilde{\alpha}|S_\theta(\tilde{\alpha}))$ (7) of the detected lines are shown in figure 10(a). It is clear that the selected genes $S_\theta(\tilde{\alpha})$ support the detected lines (gene groups) - especially those in the 4th quadrant

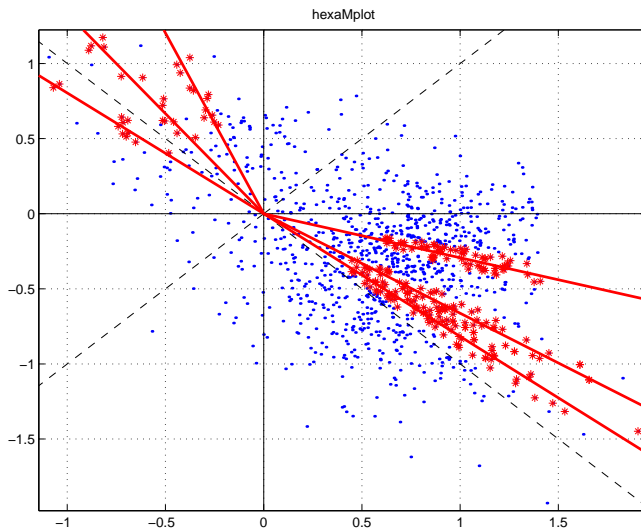


Fig. 8. Detected line rays (solid bold lines) for $\sigma = 0.05$ and selected points supporting those lines (stars). The selected points $S_\theta(\alpha)$ were chosen using $M = 126$ and $\kappa = 25$.

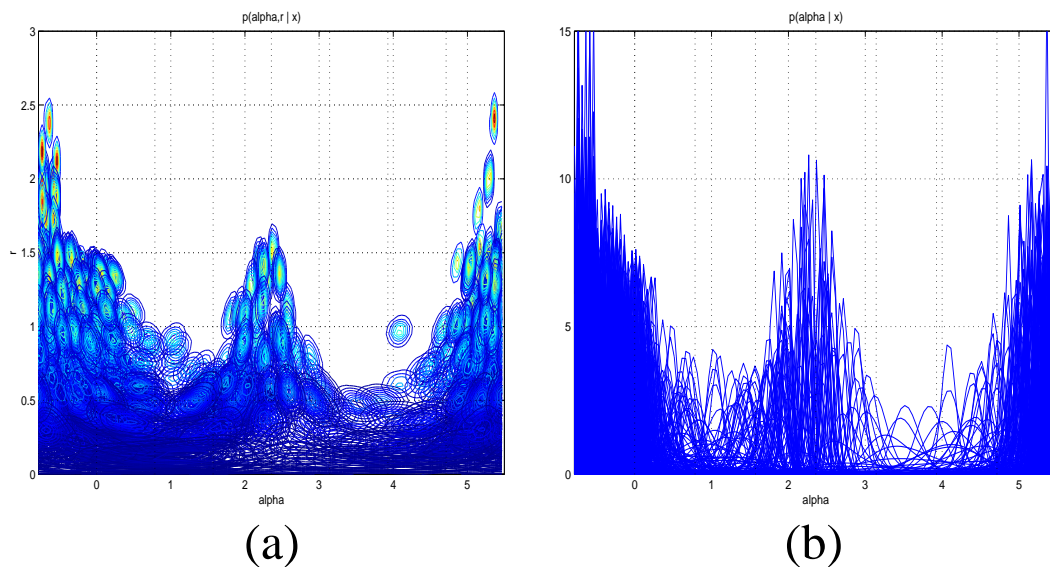


Fig. 9. Joint posteriors $p(\alpha, r | \mathbf{x})$ at $\sigma = 0.05$ for all the essayed gens (a) and their marginals $p(\alpha | \mathbf{x})$ - support kernels in the Hough space (b).

of the hexaMplot - very strongly. We also plot the posteriors of those six detected gene groups for two cases of larger observational noise: $\sigma = 0.3$ (figure 10(b)) and $\sigma = 1.0$ (figure 10(c)). For the three cases of observational noise level σ , 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 Hough Transform, 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

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 $\sigma = 1.0$.

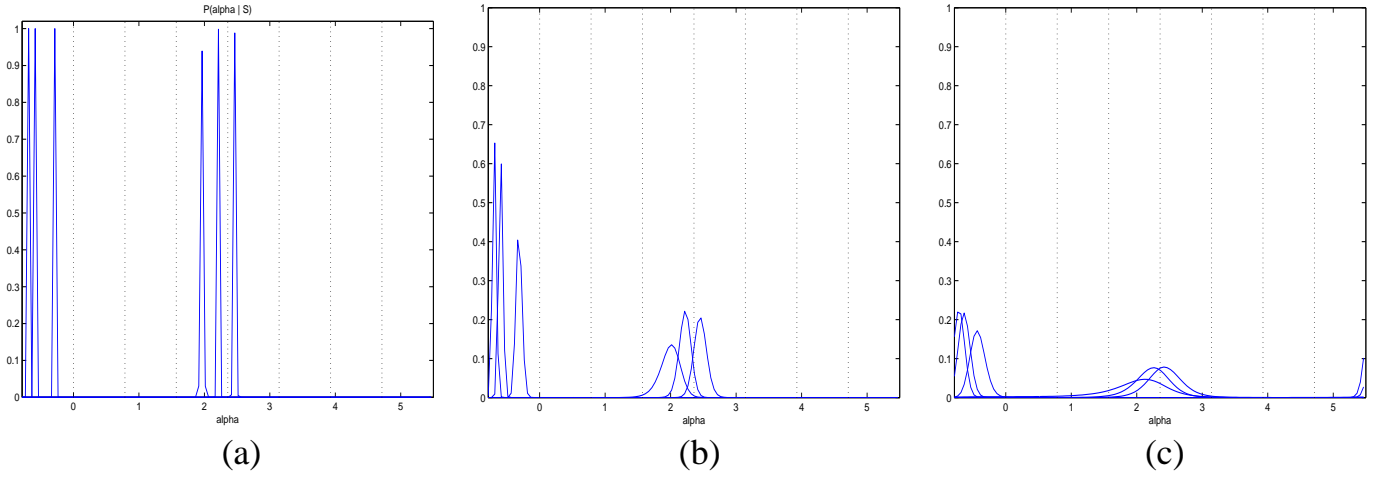


Fig. 10. Posteriors $P(\alpha | S_\theta(\alpha))$ of the six detected lines (gene groups) in figure 8 for three levels of observational noise: $\sigma = 0.05$ (a), $\sigma = 0.3$ (b) and $\sigma = 1.0$ (c).

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 gene groups by annealing the Hough transform 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 $|S_\theta(\tilde{\alpha})|$ of detected genes (second column). Representative GO terms for genes in $S_\theta(\tilde{\alpha})$ are listed in the third column. For each GO term we report the number of genes annotated to that term in the Gene Ontology (fourth column), as well as the number of genes from the gene group $S_\theta(\tilde{\alpha})$ annotated to it (fifth column). In the sixth column we show the probability (p-value) that genes from $S_\theta(\tilde{\alpha})$ would get annotated to the GO term by

chance [12].

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.

We conclude by mentioning that we tried slight variations of parameters of the priors $p(\alpha)$ and $p(\alpha|r)$, but the core six gene groups (line rays) analyzed above were always detected robustly.

IV. CONCLUSION

It has been shown that Hough Transform (HT) applied to hexaMplot gene representations can be used to detect groups of co-expressed gene in the normal-disease-drug samples [7]. However, the standard HT is not well suited for the purpose because: **(1)** the essayed genes need first to be hard-partitioned into equally and differentially expressed genes, with 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

line	$ \mathcal{S}_\theta(\tilde{\alpha}) $	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 8) WE SHOW REPRESENTATIVE GENE ONTOLOGY TERMS AND THE STRENGTH OF THEIR ASSOCIATION WITH THE GENE GROUP (P-VALUE).

into account, but the contribution of genes less differentially expressed² 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. [20]). 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 Hough Transform 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 Gene Ontology analysis. 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.

Acknowledgments

Peter Tiño 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 CITYU122506).

REFERENCES

- [1] D. Amaratunga and J. Cabrera, *Exploration and Analysis of DNA Microarray and Protein Array Data*. Hoboken, NJ: Wiley-IEEE, 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. and F. 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::TermFinder open 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.

²closer to the origin of the hexaMplot

- [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. Illingworth, 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] 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, 1994.
- [20] J. Buhmann, "Learning and data clustering," in *Handbook of Brain Theory and Neural Networks*, M. Arbib, Ed. Bradford Books, MIT Press, 1995.