Highest Density Difference Region Estimation with Application to Flow Cytometric Data

Tarn Duong[1], **Inge Koch**[2], and **M. P. Wand***[,3]

[1] Institut Pasteur, Groupe Imagerie et Modélisation; CNRS, URA 2582, F-75015 Paris, France
[2] School of Mathematics and Statistics, The University of New South Wales, Sydney 2052, Australia
[3] School of Mathematics and Applied Statistics, University of Wollongong, Wollongong 2522, Australia

Received 30 September 2008, revised 14 January 2009, accepted 2 February 2009

Motivated by the needs of scientists using flow cytometry, we study the problem of estimating the region where two multivariate samples differ in density. We call this problem *highest density difference region estimation* and recognise it as a two-sample analogue of highest density region or excess set estimation. Flow cytometry samples are typically in the order of 10 000 and 100 000 and with dimension ranging from about 3 to 20. The industry standard for the problem being studied is called *Frequency Difference Gating*, due to Roederer and Hardy (2001). After couching the problem in a formal statistical framework we devise an alternative estimator that draws upon recent statistical developments such as patient rule induction methods. Improved performance is illustrated in simulations. While motivated by flow cytometry, the methodology is suitable for general multivariate random samples where density difference regions are of interest.

Key words: Flow cytometry; Frequency difference gating; Highest density region; Multivariate density estimation; Patient rule induction method.

1 Introduction

Flow cytometry is a high-throughput technique by which multiple physical characteristics of single cells or other particles are simultaneously measured as they pass through a laser beam in a fluid stream (Shapiro, 2003). Its use in both basic and clinical research is experiencing rapid growth. A typical flow cytometry experiment produces several large multivariate samples; typically of dimension between 3 and 20. An example of flow cytometric data is shown in Fig. 1 (source: Roederer *et al.*, 2001b). Pairwise scatterplots for the CD3, CD8 and CD4 antigen levels of 1000 lymphocyte cells are shown for a human immunodeficiency negative (HIV−) patient and HIV positive (HIV+). A question of interest is which are the regions in three-dimensional space where the *difference of the densities* of the two underlying populations is high? Roederer and Hardy (2001) is at least one paper in the flow cytometry literature concerned with this question. In their introduction they list biomedical reasons for wanting to find regions of high differing density. Examples include "the analysis of phenotypic differences between subsets may elucidate differentiation pathways within a lineage" and "comparison of cells derived from different animals or people in order to identify cell populations that may correlate with clinically-relevant differences". In this article we critique the approach of Roederer and Hardy (2001), known as *Frequency Difference Gating* (FDG), and explore improvements based on recent developments in Statistics.

* Correspondence to: e-mail: mwand@uow.edu.au, Phone: +61242213556, Fax: +61242214845

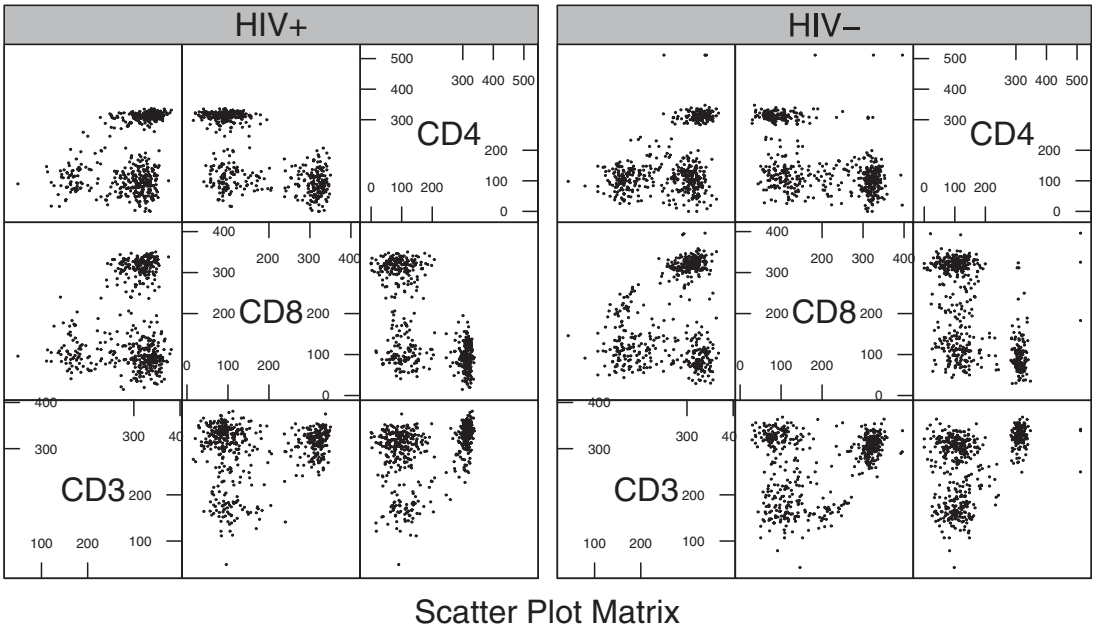

Figure 1 Scatter plot matrices for HIV+ (left) and HIV− (right) patients.

While driven by flow cytometric research, the resulting methodology is applicable to general settings involving large multivariate samples.

The above-mentioned question can be stated formally as estimation of the *highest density difference region* (HDDR) based on random samples from two multivariate distributions. This is made mathematically precise in Section 2. Pinpointing where two samples have differing density is an old problem in exploratory data analysis. An example is Tukey's hanging rootogram (*e.g.* Tukey, 1972) where histogram counts are replaced by their square-roots and compared graphically. Nevertheless, there is very little formal research on HDDR estimation.

The single sample analogue of HDDR estimation, highest density region estimation, has an established literature. Contributions include Hartigan (1987), Müller and Sawitzki (1991), Polonik (1995), Hyndman (1996), Tsybakov (1997), Baïllo, Cuesta-Albertos and Cuevas (2001), Cadre (2006) and Jang (2006). Alternative terminology includes estimation of the *density contours*, *density level sets* and *excess mass regions*. This literature is, however, mainly concerned with theoretical results. A number of practical issues in highest density region estimation, such as good data-driven rules for choosing smoothing parameters, are yet to be resolved.

The FDG method for HDDR estimation (Roederer and Hardy, 2001) involves the following two steps:

(i) Partition the space (*e.g.* $\mathbb{R}^3$) into box-shaped sub-regions using one of the samples (labelled the "control" sample by Roederer and Hardy, 2001).
(ii) Perform chi-squared tests on the resulting counts. Use the test statistics to estimate regions of differing density.

Roederer and Hardy (2001) achieve (i) *via* an algorithm called *Probability Binning* (Roederer *et al.*, 2001a, b) – a recursive binary partitioning strategy similar to that used by Classification and Regression Trees (CART) (Breiman *et al.*, 1984). However, CART is sometimes criticised for its "impatience" – committing to splits from the start of the procedure and fragmenting the data too

quickly. The Patient Rule Induction Method (PRIM) of Friedman and Fisher (1999) has been proposed as a remedy – being a more "patient" partitioning algorithm that allows boxes to be compressed and expanded as the procedure progresses. We modify PRIM for the HDDR estimation problem and show it to give improved performance over FDG on simulated data.

Baggerly (2001) critiqued the chi-squared methodology of Roederer and Hardy (2001). In particular, he showed some defects in their distribution theory and provided some remedies. An additional pitfall in flow cytometry applications is over-sensitivity of the chi-squared tests. This was pointed out by McLaren, Legler and Brittenham (1994) with the claim "With such large numbers of cells, the power of the Pearson χ^2-test may be so great that small deviations from a hypothesised model may be detected that are statistically but not practically significant". In the case of very large samples they propose the use of indifference regions in chi-squared testing and label the resultant procedure a *generalised chi-squared test*. The test statistics have non-central χ^2 distributions under the null hypothesis. Our procedure for HDDR estimation incorporates their advice.

Section 2 sets up the mathematical framework for HDDR estimation, while in Section 3 we review the FDG solution to the problem. Our PRIM-based approach to HDDR estimation is described in Section 4. In Section 5 we report on some simulations we carried out to compare PRIM-based HDDR estimation with FDG. We present a flow cytometry application in Section 6 and our conclusions in Section 7.

2 Highest Density Difference Region Estimation

We begin by formalising the problem addressed by the FDG algorithm of Roederer and Hardy (2001). Let f^+ and f^- be two density functions on $\mathbb{R}^d$ and, for some $0<\pi<1$, let

$$g \equiv \pi f^+ - (1-\pi) f^-$$

denote the *weighted density difference*. Hall and Wand (1988) worked with a general weighting coefficient π in the discrimination context. In the current context there is no compelling reason for the weights to differ; hence, from now on, we will work with $g \equiv (f^+ - f^-)/2$ and call it the *density difference*.

Let g^+ and g^- be the two non-negative functions defined by

$$g^+(\mathbf{x}) = \max(0, g(\mathbf{x})) \text{ and } g^-(\mathbf{x}) = -\min(0, g(\mathbf{x}))$$

so that $g = g^+ - g^-$. For $0 \le \tau \le 1$ the τ highest positive density difference region is

$$R_\tau^+ \equiv \{\mathbf{x} \in \mathbb{R}^d : g^+(\mathbf{x}) \ge g_\tau^+\} \text{ where } g_\tau^+ \text{ is the greatest number for which } \int_{R_\tau^+} g^+(\mathbf{x})\, d\mathbf{x} \ge 1-\tau.$$

The τ highest negative density difference region R_τ^- and corresponding threshold g_τ^- are defined analogously. The τ HDDR is then

$$R_\tau \equiv R_\tau^+ \cup R_\tau^-.$$

Figure 2 provides a graphical description of R_τ and its components. Note that this definition is analogous to the highest density region definition used by Hyndman (1996).

Now consider the problem of estimating R_τ from random samples

$$\mathbf{X}_1^+, \ldots, \mathbf{X}_{n^+}^+ \sim f^+ \text{ and } \mathbf{X}_1^-, \ldots, \mathbf{X}_{n^-}^- \sim f^-.$$

Let $\hat{R}_\tau^+$ and $\hat{R}_\tau^-$ respectively be estimators of R_τ^+ and R_τ^- so that $\hat{R}_\tau = \hat{R}_\tau^+ \cup \hat{R}_\tau^-$ is an estimator of R_τ. We quantify the error in $\hat{R}_\tau$ via

$$\text{err}(\hat{R}_\tau) = \mu(\hat{R}_\tau^+ \,\Delta R_\tau^+) + \mu(\hat{R}_\tau^- \,\Delta R_\tau^-) \tag{1}$$

for some measure μ on $\mathbb{R}^d$, where $A \Delta B$ denotes the symmetric difference between sets A and B. Note that defining $\text{err}(\hat{R}_\tau)$ to be simply $\mu(\hat{R}_\tau \Delta R_\tau)$ is problematic since, for example, $\hat{R}_\tau^+$ may intersect with

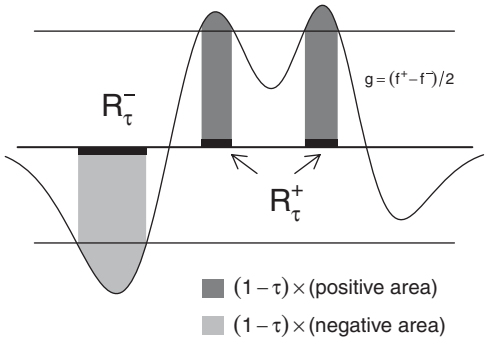

Figure 2 Graphical description of R_τ^+ and R_τ^- when g is defined on $\mathbb{R}$.

R_τ^-. There are several options for μ. In the single density context, Tsybakov (1997) considers the Lebesgue and Hausdorff measure. We propose to use the Lebesgue measure weighted by the average density

$$\mu(A) = \int_A \tfrac{1}{2}\{f^+(\mathbf{x}) + f^-(\mathbf{x})\}\,\mathrm{d}\mathbf{x}. \tag{2}$$

An obvious advantage of this measure is the ease with which it can be approximated by Monte Carlo:

$$\mu(A) \simeq N^{-1}\sum_{i=1}^{N} I(\mathbf{X}_i \in A), \tag{3}$$

where $\mathbf{X}_1,\ldots,\mathbf{X}_N$ is a random sample from the 50:50 mixture of f^+ and f^-. This feature is useful for our data-driven tuning parameter choice described in Appendix A.2, as well as for simulation studies for highest density difference estimation, as described in Section 5.

3 Review of Frequency Difference Gating

The two main steps of FDG (Roederer and Hardy, 2001) are

(i) Use an algorithm called *Probability Binning* to partition $\mathbb{R}^d$ into sub-regions and obtain counts for those sub-regions.
(ii) Use the chi-squared test statistics to estimate regions of high differing density.

Probability Binning treats one sample as the "control" and the other as the "test". Using the notation of Section 2 we will take the sample corresponding to f^+: $\mathbf{X}_1^+,\ldots,\mathbf{X}_{n^+}^+$ to be the control sample. The first stage of FDG involves dividing $\mathbb{R}^d$ into box-shaped regions so that the counts based on the $\mathbf{X}_i^+$'s are equal among the boxes. For example, if $n_+ = 100$ and there are 20 boxes then the boxes should be chosen such that each one contains 5 $\mathbf{X}_i^+$'s. The same boxes are used for the f^- sample and the counts for that sample obtained. For a K-box partition FDG leads to a $2 \times K$ contingency table. Figure 3 is a graphical description of FDG for simulated data on $\mathbb{R}$ where $n^+ = 100$ and $K = 10$. Note that the disparities within columns correspond to regions of high density difference.

For higher dimensional samples, the FDG boxes are generated *via* recursive binary partitioning. Starting with the smallest d-dimensional hyper-rectangle, or box, containing the control sample, the

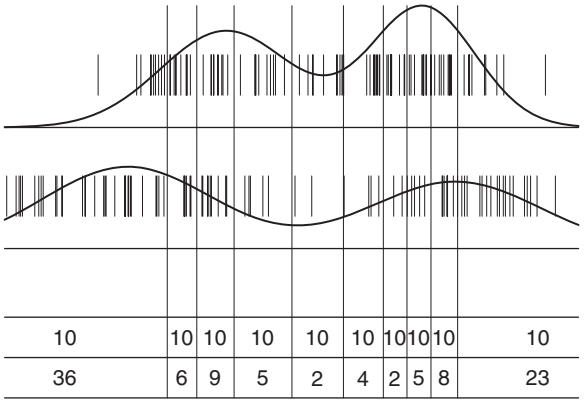

Figure 3 Graphical description of FDG for univariate data.

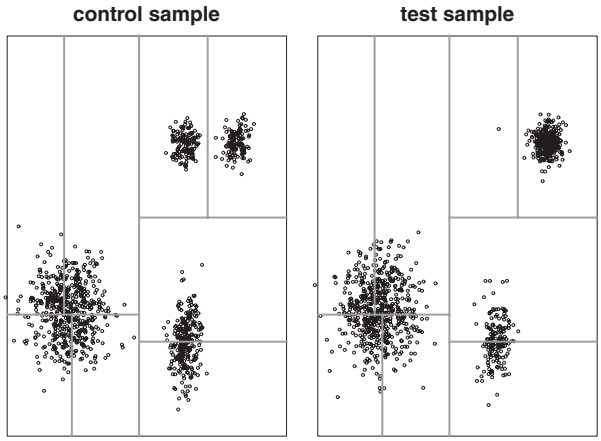

Figure 4 Illustration of FDG in $d = 2$ dimensions. In this case $L = 3$ levels of recursive binary partitioning have been performed to produce $2^3 = 8$ boxes.

procedure splits the original box along the $(d-1)$-dimensional hyperplane orthogonal to the dimension that has maximum marginal sample variance. The split point is chosen so that the two new boxes contain equal number of points. This binary splitting procedure is applied recursively for L levels to produce a partition of $r = 2^L$ boxes. Figure 4 illustrates $L = 3$ level FDG for some simulated data in $\mathbb{R}^2$.

The second phase of FDG involves forming the $2 \times r$ contingency table of counts based on the control and test samples induced by the partition. Chi-squared tests are then used to determine regions of significant density difference. Baggerly (2001) derived the theoretical distribution of the chi-squared statistic arising from Probability Binning and showed that the cut-offs used by Roederer and Hardy (2001) are overly conservative. He then described appropriate modifications to the chi-squared testing phase.

Although binary partitioning strategies have enjoyed enormous success in data mining contexts, especially in terms of interpretability, they have been criticised for being too "impatient"

(*e.g.* Hastie, Tibshirani and Friedman, 2001). Binary partitioning means that the data are fragmented quite quickly. Friedman and Fisher (1999) devised the PRIM to redress this problem. Section 4.1 explains how PRIM can be adapted for HDDR estimation.

Another problem with FDG is over-sensitivity of chi-squared tests for very large samples. This issue is discussed by Pederson and Johnson (1990) and McLaren *et al.* (1994) who recommended use of indifference regions. Details are given in Section 4.2.

4 A Patient Rule Induction Method-Based Algorithm for Highest Density Difference Region Estimation

We now describe our alternative to FDG for HDDR estimation. It calls upon PRIM to obtain subregions of the space where highest density differences seem plausible. Generalised chi-squared tests are then used to test for significant density difference. Before laying out the algorithm we briefly describe its main components.

4.1 Patient rule induction method

The PRIM was developed by Friedman and Fisher (1999) as a method for estimating maxima in multivariate regression functions based on noisy data. A concise description of PRIM is given in Section 9.3 of Hastie *et al.* (2001). The main input into PRIM is a set of regression-type data: $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$, where $\mathbf{X}_i \in \mathbb{R}^d$ and $Y_i \in \mathbb{R}$. The output is a set of boxes in $\mathbb{R}^d$ that estimate regions corresponding to maxima in $E(Y|\mathbf{X} = \mathbf{x})$.

PRIM requires specification of (a) the maximum number of boxes and (b) the minimum box mean (MBM). The resulting boxes depend heavily on these tuning parameters. However, to date, little research has been done into data-driven values for their choice.

The `R` package `prim` (Duong, 2008) (`http://cran.r-project.org`) facilitates the practical use of PRIM.

4.2 Generalised chi-squared tests

Consider the classical $r \times c$ contingency table situation where O_{ij} is the observed count in cell (i,j). The usual chi-squared test on $(r-1)$ $(c-1)$ degrees of freedom applies to both the test for homogeneity of r multinomial distributions, each with c categories, and the test for independence of two factors with r and c levels (*e.g.* Rice, 1995). The former situation is relevant to our algorithm; hence, consider testing

$$H_0 : p_{1j} = \cdots = p_{rj},\ j = 1, \ldots, c \ \textit{versus}\ H_1 : \text{not}\, H_0 \tag{4}$$

where p_{ij} is the probability of the jth category of the ith multinomial distribution. The Pearson X^2 statistic is given by

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} (O_{ij} - E_{ij})^2 / E_{ij}. \tag{5}$$

Here $E_{ij} = O_{i\cdot} O_{\cdot j}/n$ is the expected count in cell (i,j) under H_0, where n is the total count and, for example, $O_{i\cdot} \equiv \sum_{j=1}^{c} O_{ij}$.

An alternative formulation for (4) is $H_0{:}\delta = 0$ *versus* $H_1{:}\delta > 0$ where

$$\delta \equiv \sum_{i=1}^{r} \sum_{j=1}^{c} (p_{ij} - \overline{p}_{\cdot j})^2 / \overline{p}_{\cdot j}$$

is the *discrepancy* between the p_{ij} and the $\overline{p}_{\cdot j} \equiv \sum_{i=1}^{r} p_{ij}/r$. For very large n Pederson and Johnson (1990) and McLaren *et al.* (1994) recommend working with the generalised hypothesis set-up

$$H_0 : \delta = \delta_0, \delta_0 > 0 \textit{ versus } H_1 : \delta > \delta_0$$

where the interval $(0,\delta_0)$ is an indifference region. According to Drost *et al.* (1989), the approximation

$$X^2 \overset{\text{approx.}}{\sim} \chi^2_{(r-1)(c-1)}(n\delta_0) \text{ under } H_0 \tag{6}$$

is good for $n \geq 100$, where $\chi^2_k(\nu)$ denotes the non-central chi-squared distribution with k degrees of freedom and non-centrality parameter ν. Patnaik (1949) is an early reference for derivation of (6). Generalised chi-squared tests thus involve setting the indifference parameter δ_0 and making inference on the basis of (6).

Now take $r = 2$ and re-label as follows: $p_j^+ \equiv p_{1j}$ and $p_j^- \equiv p_{2j}$. Also let $n_j^+ \equiv O_{1j}$, $n_j^- \equiv O_{2j}$, $n^+ \equiv O_{1\cdot}$ and $n^- \equiv O_{2\cdot}$. Then the Pearson X^2 statistic (5) can be expressed as

$$X^2 = \sum_{j=1}^{c} \frac{(\hat{p}_j^+ - \hat{p}_j^-)^2}{\hat{p}_j^+/n^- + \hat{p}_j^-/n^+} \tag{7}$$

where $\hat{p}_j^+ \equiv n_j^+/n^+$ and $\hat{p}_j^- \equiv n_j^-/n^-$. An approximate generalised chi-squared test for $H_0 : p_j^+ = p_j^-$, $j = 1, \ldots, c$ with indifference region $(0,\delta_0)$ involves rejection at level α if

$$X^2 > \chi^2_{c-1;1-\alpha}(n\delta_0) \tag{8}$$

where $\chi^2_{k;1-\alpha}(\nu)$ denotes the $1-\alpha$ quantile of the $\chi^2_k(\nu)$ distribution. If H_0 is rejected then the sub-hypotheses

$$H_{0j} : \frac{(\hat{p}_j^+ - \hat{p}_j^-)^2}{\hat{p}_j^+/n^- + \hat{p}_j^-/n^+} \leq \delta_0/c, \; j = 1, \ldots, c \tag{9}$$

can be tested using $\chi^2_1(n\delta_0/c)$ as an approximate null distribution.

Moore (1984) asserts that the chi-squared statistic computed with respect to partitions based on the data is asymptotically equivalent to that with fixed boundaries, provided the data-based partition boundaries converge in probability to the fixed boundaries. This is important for the PRIM-based algorithm, since it uses the data to perform partitioning.

4.3 Algorithm overview

Our new algorithm for HDDR estimation uses PRIM applied to the "regression" data sets $(\mathbf{X}_i, Y_i)$ and $(\mathbf{X}_i, -Y_i)$ where the $\mathbf{X}_i$'s are the $\mathbf{X}_i^+$ and $\mathbf{X}_i^-$ data pooled together and the $Y_i \in \{-1, 1\}$ are indicators of whether $\mathbf{X}_i$ is from the $-$ or $+$ sample. The first application of PRIM, using the $(\mathbf{X}_i, Y_i)$, produces a partition of $\mathbb{R}^d$, the elements of which are candidates for $\hat{R}_\tau^+$. Similarly, candidates for $\hat{R}_\tau^-$ are generated *via* application of PRIM to the $(\mathbf{X}_i, -Y_i)$. Figure 5 illustrates PRIM-based partitioning on some simulated bivariate data. The PRIM partitions lead to a $2 \times r$ contingency table of counts, where r is the partition size ($r = 7$ for the partition in Fig. 5). Generalised chi-squared tests applied to contingency table allow for inferentially sound fine-tuning. Full details of the algorithm are given in Appendix A.

A thorny problem concerns that of selecting PRIM's MBM parameter for estimation of a particular R_τ. Even for univariate highest density region estimation based on kernel density estimation there is scant literature on automatic bandwidth selection. In Appendix A.1 we describe a method for choosing the MBMs from the data.

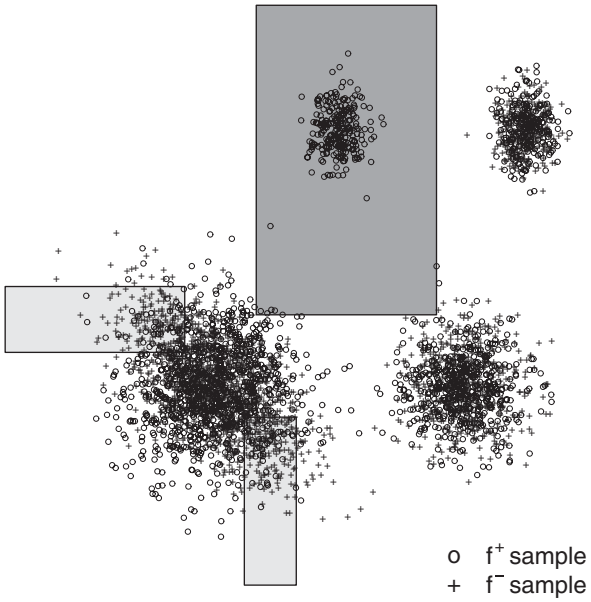

Figure 5 Illustration of PRIM-based partitioning. The shaded partition elements in the lower left region are candidates for estimation of R_τ^+; those in the upper middle region are candidates for estimation of R_τ^-. Grey-level shading is used to distinguish the partition elements.

4.4 Approximation of error in highest density difference region estimation

The algorithm summarised in the previous section requires determination of $R_\tau = R_\tau^+ \cup R_\tau^-$ when f^+ and f^- are multivariate normal mixture densities. The simulations of Section 5 also have this requirement. We now show how Monte Carlo methods can be used to approximate R_τ^+. An analogous approach applies to R_τ^-.

Let $\mathbf{Z}_1, \ldots, \mathbf{Z}_N$ be a random sample from $g^+ / \int_{\mathbb{R}^d} g^+$. Then, following the argument in Section 3.3 of Hyndman (1996), a consistent estimate of g_τ^+ is the τth sample quantile of $g^+(\mathbf{Z}_1), \ldots, g^+(\mathbf{Z}_N)$. Since

$$g^+ \leq |g| = |\tfrac{1}{2}f^+ - \tfrac{1}{2}f^-| \leq \tfrac{1}{2}f^+ + \tfrac{1}{2}f^-$$

we can use an accept–reject scheme (*e.g.* Robert and Casella, 1999) based on samples from the bounding density $\frac{1}{2}f^+ + \frac{1}{2}f^-$. This is a normal mixture density whenever f^+ and f^- are; hence, generation of the required samples is straightforward.

5 Simulations

We conducted a simulation study to assess the performance of the PRIM-based HDDR estimation procedure and compare it with FDG. Two different settings were considered, with dimension, sample sizes and τ values as follows:

$$\begin{array}{lllll} \text{Setting I}: & d = 2, & n = 1000, & \tau = 0.5 \\ \text{Setting II}: & d = 5, & n = 100\,000, & \tau = 0.9. \end{array}$$

Setting I permits some useful visual insights while Setting II aims to mimic a typical flow cytometry scenario. For each setting the true f^+ and f^- densities were taken to be normal mixtures. For Setting I we took f^+ to be

$$\tfrac{1}{2}N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}\frac{1}{8} & 0\\0 & \frac{1}{8}\end{bmatrix}\right)+\tfrac{1}{4}N\left(\begin{bmatrix}2\\0\end{bmatrix},\begin{bmatrix}\frac{1}{16} & 0\\0 & \frac{1}{16}\end{bmatrix}\right)+\tfrac{1}{8}N\left(\begin{bmatrix}1\\2\end{bmatrix},\begin{bmatrix}\frac{1}{64} & 0\\0 & \frac{1}{32}\end{bmatrix}\right)+\tfrac{1}{8}N\left(\begin{bmatrix}\frac{5}{2}\\2\end{bmatrix},\begin{bmatrix}\frac{1}{64} & 0\\0 & \frac{1}{32}\end{bmatrix}\right)$$

and f^- to be

$$\tfrac{5}{8}N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}\frac{1}{8} & -\frac{3}{40}\\-\frac{3}{40} & \frac{1}{8}\end{bmatrix}\right)+\tfrac{1}{4}N\left(\begin{bmatrix}2\\0\end{bmatrix},\begin{bmatrix}\frac{1}{16} & 0\\0 & \frac{1}{16}\end{bmatrix}\right)+\tfrac{1}{8}N\left(\begin{bmatrix}\frac{5}{2}\\2\end{bmatrix},\begin{bmatrix}\frac{1}{64} & 0\\0 & \frac{1}{32}\end{bmatrix}\right).$$

For Setting II, both densities are of the form

$$w_1N(\mu_1,\Sigma_1)+w_2N(\mu_2,\Sigma_2)+w_3N(\mu_3,\Sigma_3)+w_4N(\mu_4,\Sigma_4)$$

where each μ_k is a 5×1 vector and each Σ_k is a 5×5 covariance matrix ($1\le k\le 4$). We obtained the normal mixture parameters for f^+ and f^- by fitting normal mixtures to actual flow cytometry data; hence, for this application, Setting II has an element of realism with respect to the motivating application. Table 1 in Appendix B lists these parameters. Figure 6 provides visual displays of the Setting II data and the flow cytometry data on which it is based. Good agreement between the two is apparent.

One hundred replications were used in each setting. PRIM estimates of R_τ were obtained *via* the algorithm described in Appendix A. The FDG estimates required a choice of the level parameter L. This was achieved by the analogue of the normal mixture pilot approach to choosing MBMs as described in Appendices A.2 and A.3. In each case we approximated the error measure $\mathrm{err}(\hat{R}_\tau)$ as given by (1) and (2). For the Monte Carlo approximation (3) we used $N = 1\,000\,000$ throughout the study.

Figure 7 shows typical PRIM-based and FDG-based estimates of $R^+_{0.5}$ and $R^-_{0.5}$ for Setting I. Each estimate is that which results in the median $\mathrm{err}(\hat{R}_{0.5})$ value from the simulations. The PRIM-based estimator is reasonable, although not outstanding. For dimensions as low as 2 we would expect better performance from classical density estimation approaches. However, the superiority of PRIM compared with FDG is apparent. The former has more flexibility in placement of its boxes, while the latter is tied to those arising from the recursive binary splits.

Figure 8 summarises the results. In both settings the PRIM-based approach is the clear winner. Wilcoxon tests applied to the ratios are highly significant. The advantage of PRIM is much more pronounced in the higher-dimensional setting where it typically leads to a twofold to threefold reduction in the error measure. Both methods used the generalised chi-squared tests of Section 4.2; hence, the improvement is due to the PRIM partitioning. The Setting II results have much more relevance, since PRIM is designed for higher-dimensional data situations. For low d it is likely that more traditional density estimation approaches, such as kernel or k-nearest-neighbour estimators, will be better suited to the HDDR estimation problem.

6 Flow Cytometry Application

The PRIM-based algorithm was applied to flow cytometry data from a large-scale experiment involving patients that develop graft-*versus*-host disease (GvHD). The full data set involves blood samples of 31 patients and 10 anti-body cocktails, with flow cytometry measurements collected longitudinally for about 3 months, and is described and analysed by Brinkman *et al.* (2007). Brinkman *et al.* (2007) state that "It is likely that the outcome of GvHD could be improved if it were treated as early as possible" and "if the diagnosis of GvHD could be made more definitely, only those patients who absolutely required steroids … would be treated" and use longitudinal data analytic methods to identify biomarkers.

We conducted an investigation into the use of our HDDR estimation procedure for biomarker identification. One GvHD patient and one control patient were chosen among those that had flow cytometry analyses of their blood cells exactly 32 days after undergoing blood and marrow transplant. We focussed on the two six-dimensional samples corresponding to forward-scatter, side-scatter, and staining of the antibodies CD4-FITC, CD8β-PE, CD3-PerCP and CD8-APC for that particular day. To reduce skewness, the inverse sinh transformation was applied to all data before processing.

We applied the PRIM-based algorithm, with τ set at 0.5, to the two six-dimensional samples described in the previous paragraph. The positive and negative estimated HDDRs are shown in Figs. 9 and 10.

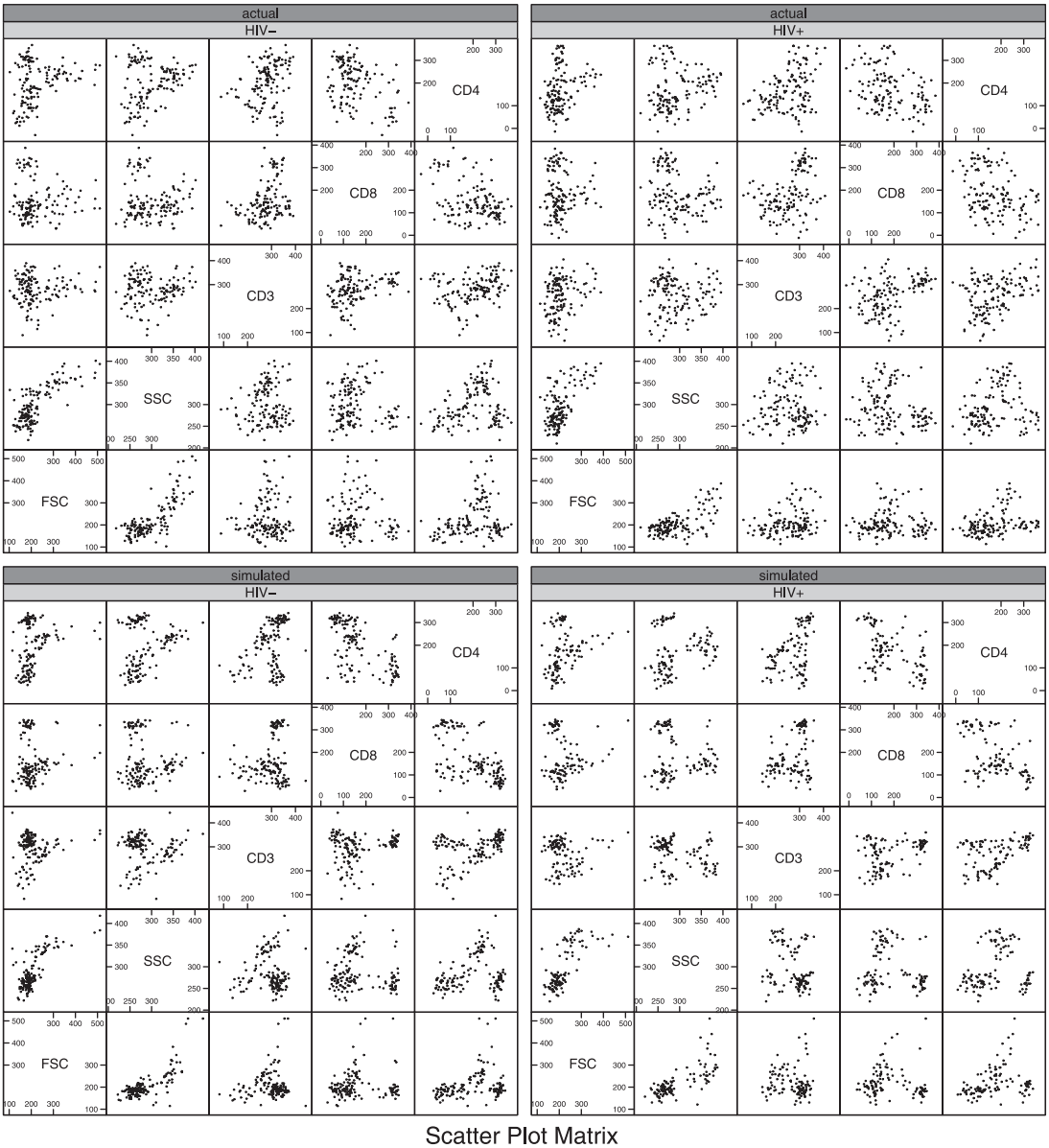

Figure 6 Pairwise scatterplots of a realisation of the Setting II normal mixture data. The actual flow cytometry data on which it is based are also shown.

Before commenting on these two figures it should be noted that they are subject to the limitations of visualising six-dimensional structure *via* two-dimensional projections. However, they do give some insight into the nature of the HDDR estimates. The PRIM-based estimate of $\hat{R}^{+}_{0.5}$ (Fig. 9) corresponds to a region in $\mathbb{R}^6$ where about one-third of the GvHD patient measurements are present but almost all the control patient measurements are absent. The dark points indicate inclusion in this region *via* the 15 (unique) bivariate views of the data.

The PRIM-based estimate of $\hat{R}^{-}_{0.5}$ is shown in Fig. 10. It contains about 20% of the control patient data, but none of the GvHD patient data. Overall, the PRIM-based estimates of $\hat{R}^{+}_{0.5}$ and $\hat{R}^{-}_{0.5}$ reveal some interesting sub-regions of $\mathbb{R}^6$ where the control and GvHD patients differ.

The actual HDDR estimates can be described mathematically *via* box constraints, even though we do not have a way of displaying them. We also obtained the PRIM-based HDDR estimates for other values of τ between 0.5 and 1. However, for these data the estimates exhibited little change.

Finally, we present the results of applying FDG to the same data (Figs. 11 and 12). The HDDR estimates based on FDG are considerably larger than those obtained from the PRIM-based algorithm. The positive FDG-estimated HDDR contains most of the data for the GvHD patient, while the negative FDG-estimated HDDR contains most of the data for the control patient. In short, the FDG-estimated HDDRs are not as "sharp" as those obtained from the PRIM-based algorithm. We conclude that PRIM and FDG can give quite different results for flow cytometry data. The small simulation study of Section 5 points to PRIM being the better option. Further work in this direction would be worthwhile.

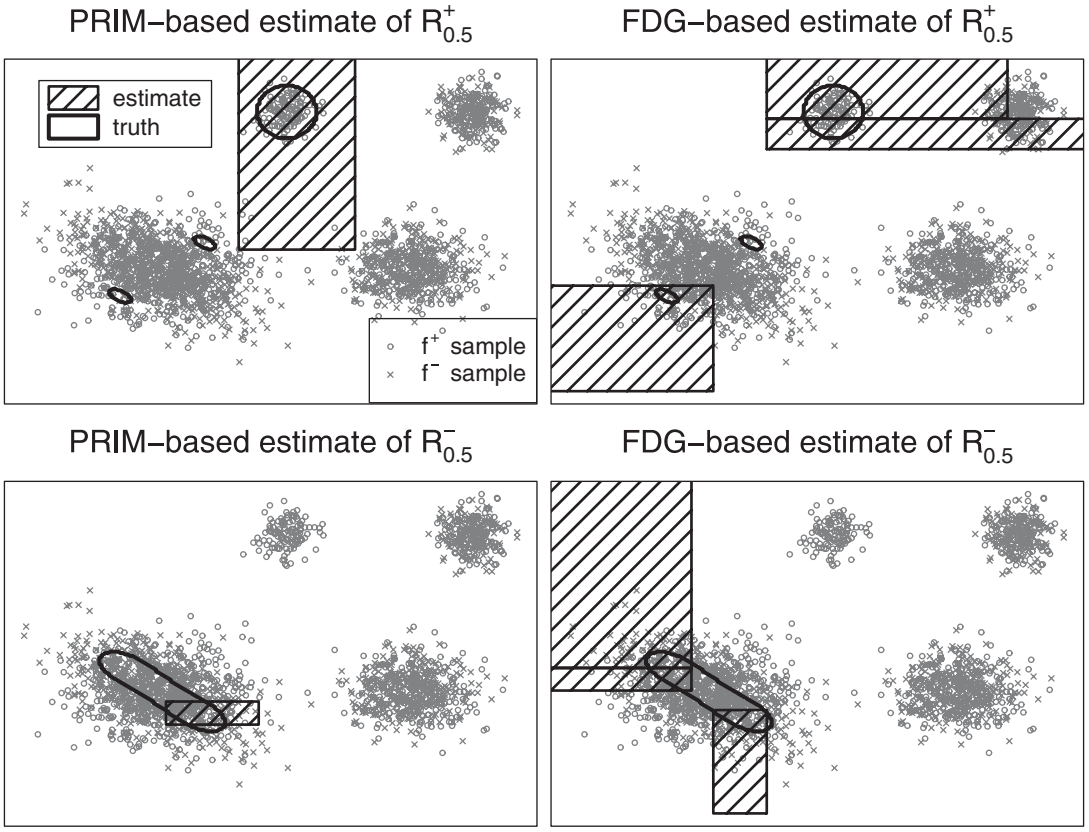

Figure 7 Median performance PRIM-based and FDG-based estimates of $R^{+}_{0.5}$ and $R^{-}_{0.5}$ for Setting I of the simulation study.

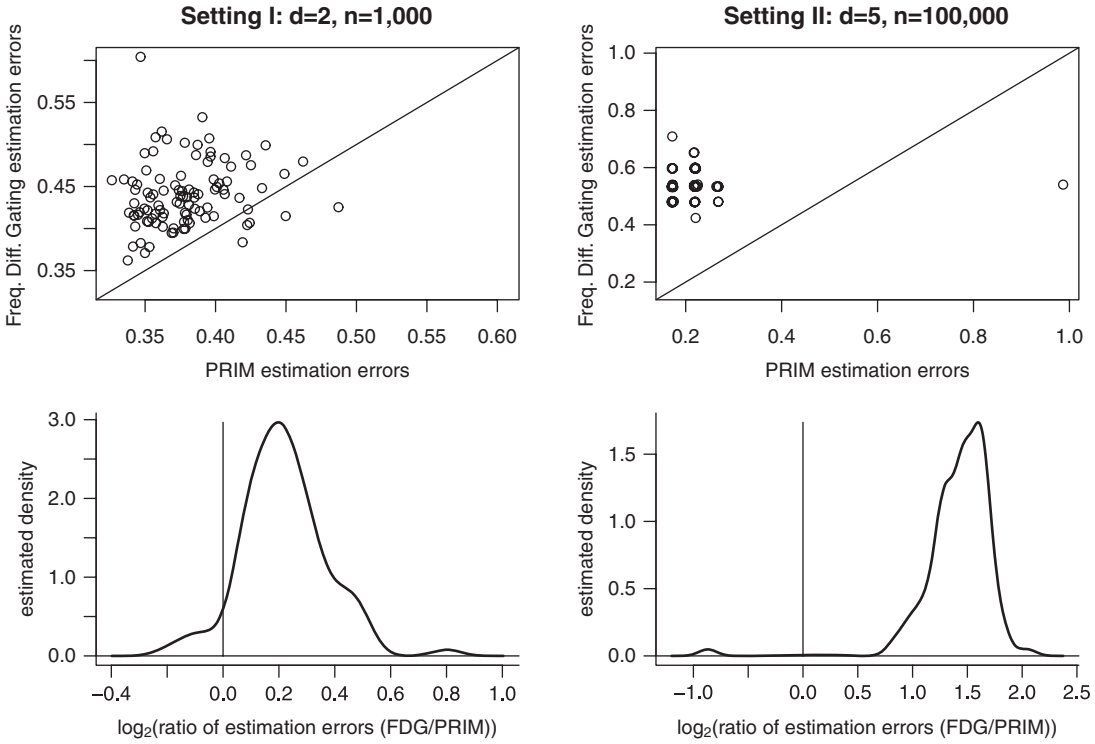

Figure 8 Summary of simulation results. The upper panels are scatterplots of the $\mathrm{err}(\hat{R}_{0.5})$ values for FDG *versus* PRIM-based estimation, together with the 1:1 line. The lower panels are kernel density estimates of the error ratios (with FDG error on the numerator).

7 Concluding Remarks

HDDR estimation is an area with enormous potential in many areas of application, including marketing, geology (*e.g.* Friedman and Fisher, 1999) and flow cytometric data analysis (*e.g.* Roederer and Hardy, 2001). However, there has been surprisingly little research on the topic. Even the single sample analogue, highest density region estimation, has a literature that is limited mainly to theoretical results of little practical benefit. It is our hope that this paper will be a catalyst for converting this field into a vibrant application-oriented area of research. We anticipate that the HDDR estimation structure laid out in Section 2 will serve as a foundation for ongoing research on this problem.

In this paper, driven by the needs of flow cytometry research, we have concentrated on moderate-dimensional settings (roughly $5 \leq d \leq 15$). Our PRIM-based algorithm for HDDR estimation is seen to perform quite soundly and offer big improvements over the FDG approach of Roederer and Hardy (2001). It is expected that lower-dimensional settings will benefit from non-PRIM approaches such as kernel and kth-nearest-neighbour density estimation (*e.g.* Scott, 1992). Connections with recent machine learning research (*e.g.* Steinwart, Hush and Scovel, 2005) and general classification methodology (*e.g.* Hastie *et al.*, 2001) also await exploration. Finally, data-driven rules for the selection of smoothing and auxiliary parameters in the highest density and density difference region contexts remain virtually an open topic. Samworth and Wand (2008) deals with the univariate version of this problem.

Acknowledgements The authors would like to acknowledge the contributions of Robert Gentleman, Tony Rossini and Nolwenn Le Meur. This research is supported by an Australian Research Council Discovery

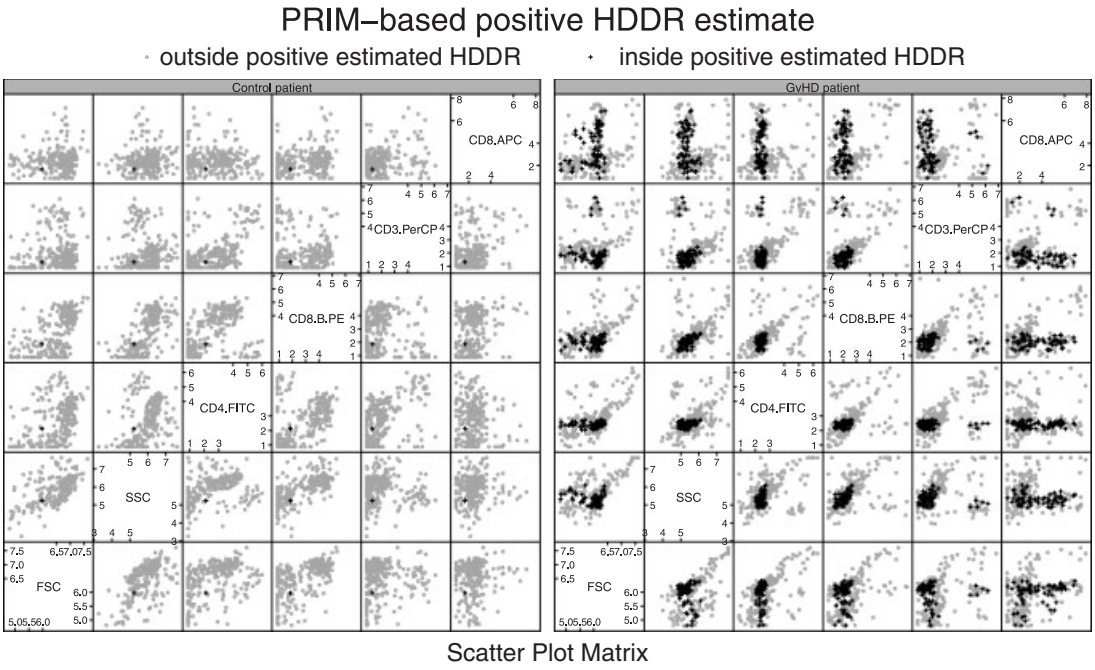

Figure 9 PRIM-based ($\tau = 0.5$) positive highest difference density region for the GvHD data. The dark points are those inside the region. A random sample of size 250 from each group has been taken to aid visual comparison. All data have undergone the inverse sinh transformation before plotting and processing.

Project grant (Project DP0556518) and Institut Pasteur through a Programme Transversal de Recherches grant (PTR 218).

Appendix A: Algorithmic Details

The inputs to the PRIM-based algorithm for HDDR estimation are:

- $\tau \in (0,1)$, corresponding to R_τ^+ and R_τ^-. This defines the target HDDR.
- $\mathbf{X}_1^+, \ldots, \mathbf{X}_{n_+}^+$ and $\mathbf{X}_1^-, \ldots, \mathbf{X}_{n^-}^-$, two multivariate random samples in $\mathbb{R}^d$, with underlying densities f^+ and f^-, respectively.
- α, the significance level in the generalised chi-squared tests. The default is 0.05.
- δ_0, the indifference region parameter for the generalised chi-squared tests. The default is $\delta_0 = 0.05^2$, which corresponds to differences of 5% in both absolute and relative terms.
- $K_{\max}$, the maximum number of boxes. The default is $K_{\max} = 20$.
- $\mathrm{MBM}^+(\tau)$, $\mathrm{MBM}^-(\tau)$, the MBMs for estimation of R_τ^+ and R_τ^-, respectively. We will suppose, for now, that they have been chosen from the data to be $\widehat{\mathrm{MBM}}^+(\tau)$ and $\widehat{\mathrm{MBM}}^-(\tau)$. Appendix A.1 discusses a data-driven rule for their choice.

Given $\widehat{\mathrm{MBM}}^+(\tau)$ and $\widehat{\mathrm{MBM}}^-(\tau)$, the full algorithm for estimation of R_τ is:

(i) Form the "regression" data set $(\mathbf{X}_i, Y_i)$, $1 \le i \le n$ $(n = n^+ + n^-)$, for input into PRIM. The $\mathbf{X}_i$'s are the $\mathbf{X}_i^+$ and $\mathbf{X}_i^-$ data pooled together. The $Y_i \in \{-1, 1\}$ are indicators of whether $\mathbf{X}_i$ is from the $-$ or $+$ sample.

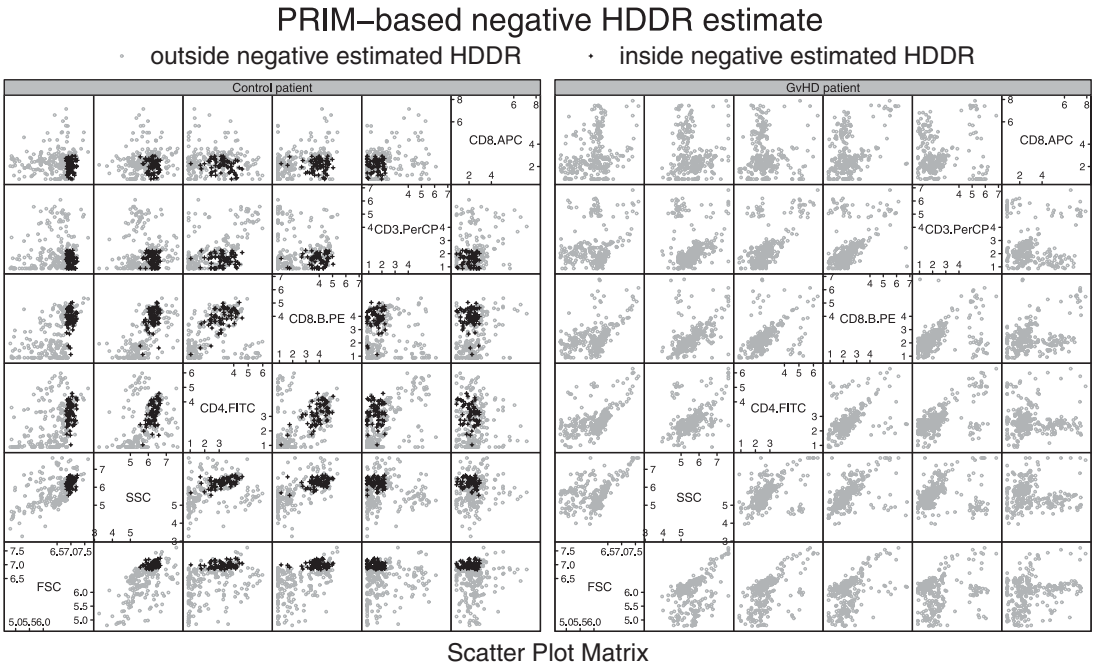

Figure 10 PRIM-based ($\tau = 0.5$) negative highest difference density region estimate for the GvHD data. The dark points are those inside the region. A random sample of size 250 from each group has been taken to aid visual comparison. All data have undergone the inverse sinh transformation before plotting and processing.

(ii) Feed the $(\mathbf{X}_i, Y_i)$ data into PRIM to obtain $K_{\max}$ (ordered) boxes $B_1, \ldots, B_{K_{\max}}$ in $\mathbb{R}^d$. Use these boxes to obtain a partition $\{P_1^+, \ldots, P_{K^+}^+\}$ of $\mathbb{R}^d$ as follows: $P_1^+ = B_1$, $P_k^+ = B_k - \cup_{j=1}^{k-1} B_j$, $2 \le k \le K^+$ where $K^+ = \max\{k : \text{ave}(Y; P_k^+) \ge \widehat{\text{MBM}}^+(\tau)\}$ and $\text{ave}(Y; P_k^+)$ is the average of the Y_i's in P_k^+.

(iii) Repeat Step ii, but with $(\mathbf{X}_i, -Y_i)$, fed into PRIM to obtain $P_1^-, \ldots, P_{K^-}^-$, with $K^- = \max\{k : \text{ave}(-Y; P_k^-) \le \widehat{\text{MBM}}^-(\tau)\}$.

(iv) Check that there is no overlap between the P_k^+ and the P_k^- (for differing densities this is unlikely for reasonable choices of $\widehat{\text{MBM}}^+(\tau)$ and $\widehat{\text{MBM}}^-(\tau)$. If there is some overlap then increase $\widehat{\text{MBM}}^+(\tau)$ and decrease $\widehat{\text{MBM}}^-(\tau)$ until the P_k^+ and the P_k^- are disjoint. If the boxes become null then the estimate of R_τ is null.

(v) Form the following partition of $\mathbb{R}^d$: $P_k = P_k^+$, $1 \le k \le K^+$; $P_{K^+ + k} = P_k^-$, $1 \le k \le K^-$, $P_{K+1} = \mathbb{R}^d - \cup_{j=1}^{K} P_k$ where $K = K^+ + K^-$.

(vi) Obtain the counts among the $\mathbf{X}_i^+$ and $\mathbf{X}_i^-$ samples, respectively, over the partition $P_1, \ldots, P_{K+1}$. Combine the counts into a $2 \times (K+1)$ contingency table.

(vii) Test for an overall difference between f^+ and f^- *via* a level α generalised chi-squared test on the $2 \times (K+1)$ contingency table with indifference region $(0, \delta_0)$. Details are provided by (7) and (8) with c set to $K+1$.

(viii) If the hypothesis of no overall difference is rejected then test which columns in the contingency table contribute significantly to the difference between the two samples. Details are provided by (9) with c set to $K+1$.

Take $\hat{R}_\tau^+$ to be the union of all P_k for which significance is achieved and $\hat{p}_k^+ > \hat{p}_k^-$.
Take $\hat{R}_\tau^-$ to be the union of all P_k for which significance is achieved and $\hat{p}_k^+ > \hat{p}_k^-$.

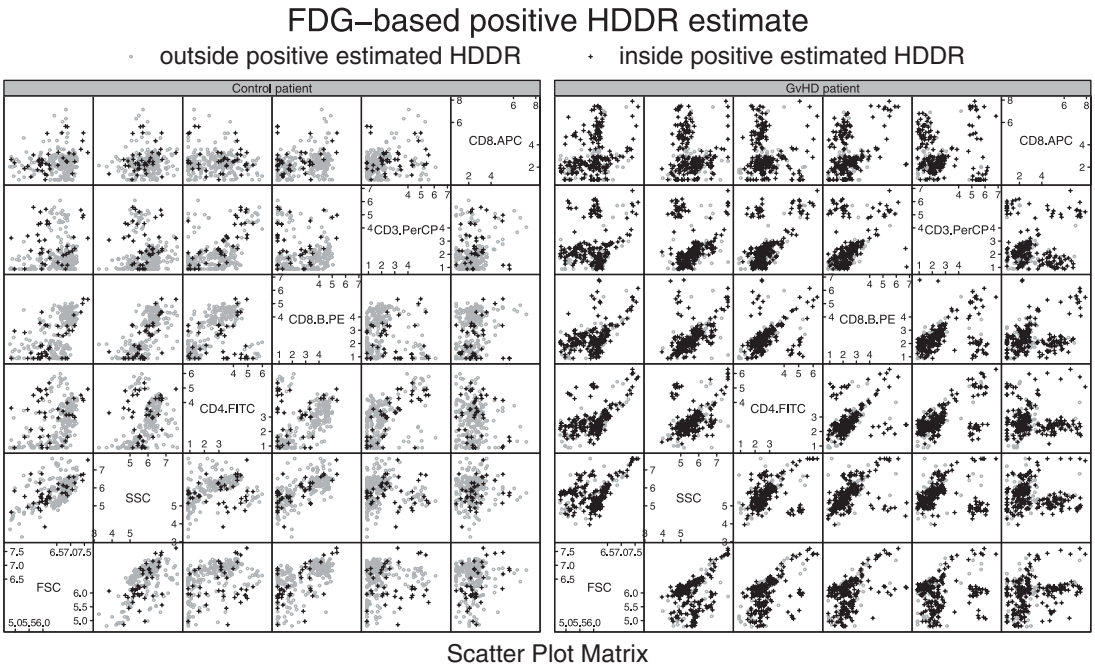

Figure 11 FDG -based positive highest difference density region estimate for the GvHD data. The dark points are those inside the region. A random sample of size 250 from each group has been taken to aid visual comparison. All data have undergone the inverse sinh transformation before plotting and processing.

A.1 Data-driven choice of MBM

(i) Obtain pilot estimates of R_τ^+ and R_τ^- based on fitting normal mixture densities to the $\mathbf{X}_i^+$ and $\mathbf{X}_i^-$. Let these pilot estimates be denoted by $R_{\tau,\text{pilot}}^+$ and $R_{\tau,\text{pilot}}^-$. Details on the normal mixture fitting are deferred to Appendix A.2.

(ii) Apply PRIM to the $(\mathbf{X}_i, Y_i)$ with MBM $= \overline{Y}$ to give a sequence $P_1^+, \ldots, P_{K_0^+}^+$, with MBMs $\text{ave}(Y; P_g^+), g = 1, \ldots, K_0^+$.

(iii) Apply PRIM to the $(\mathbf{X}_i, Y_i)$ K_0^+ times each with MBM equal to $\text{ave}(Y; P_g^+)$ to obtain estimates $\hat{R}_{\tau,g}^+$ for $1 \le g \le K_0^+$.

(iv) Take $\widehat{\text{MBM}}^+(\tau)$ to be $\arg\min_{1 \le g \le K_0^+} \text{err}(\hat{R}_{\tau,g}^+, \widehat{R}_{\tau,\text{pilot}}^+)$.

(v) Repeat Steps ii–iv; but using the $(\mathbf{X}_i, -Y_i)$ data in PRIM to obtain $P_1^-, \ldots, P_{K_0^-}^-$, with MBMs $\text{ave}(-Y; P_g^-), g = 1, \ldots, K_0^-$ and take $\widehat{\text{MBM}}^-(\tau)\, \arg\min_{1 \le g \le K_0^+} \text{err}(\hat{R}_{\tau,g}^+, \widehat{R}_{\tau,\text{pilot}}^+)$.

A.2 Details of Normal Mixture Fitting

The inputs to the normal mixture fitting procedure used in A.1 are

A continuous sample in $\mathbb{R}^d$.

$k_{\max}$: The maximum number of components in the normal mixture. The default is 10.

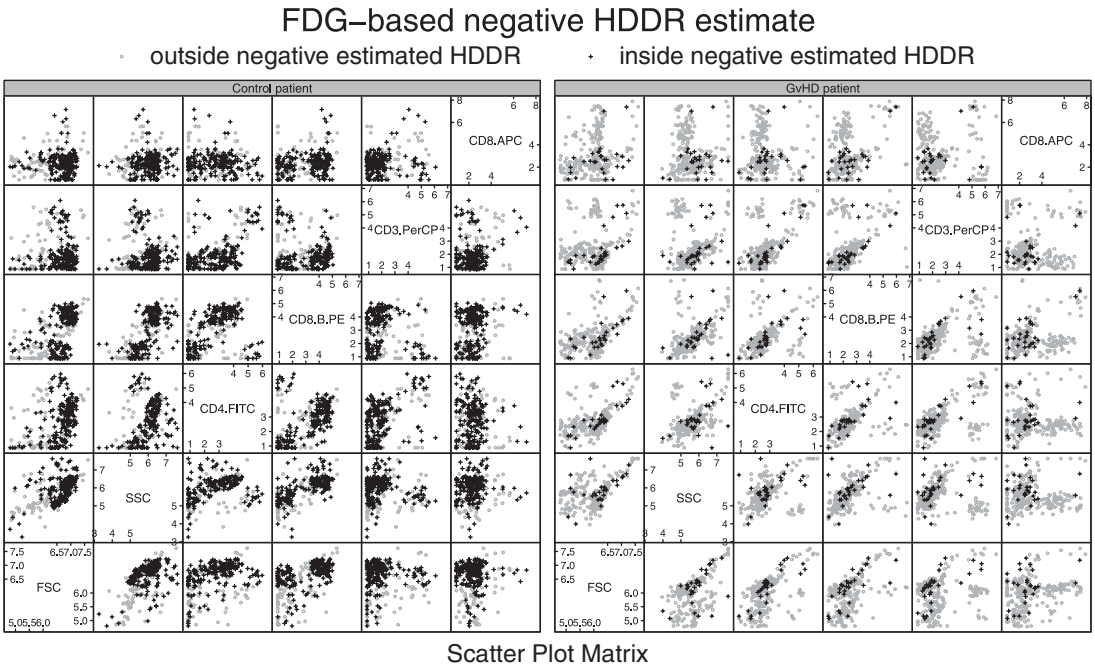

Figure 12 FDG-based negative highest difference density region estimate for the GvHD data. The dark points are those inside the region. A random sample of size 250 from each group has been taken to aid visual comparison. All data have undergone the inverse sinh transformation before plotting and processing.

(i) For each $k = 1, \ldots, k_{\max}$ and $\ell = 1, \ldots, 20$ random starts use k-means clustering to fit k clusters to the data. Let $n_i = n_i(k, \ell)$ be the size of the ith cluster ($1 \le i \le k$) and $\mathbf{c}_{ij} = \mathbf{c}_{ij}(k, \ell)$ be the jth point in the ith cluster ($1 \le j \le n_i$; $1 \le i \le k$). Compute the within-cluster variabilities

$$W_\ell(k) \equiv \sum_{i=1}^{k} \sum_{j=1}^{n_i} ||\mathbf{c}_{ij} - \overline{\mathbf{c}}_{i\cdot}||^2$$

where $\overline{\mathbf{c}}_{i\cdot} = \sum_{j=1}^{n_i} \mathbf{c}_{ij}/n_i$ and $||\cdot||$ is the Euclidean norm. Obtain the sequence $W(1), \ldots, W(k_{\max})$ where each $W(k)$ is chosen to be smallest among the $W_\ell(k)$ subject to the restriction that the sequence is monotonically decreasing. Also set $B(k) \equiv \sum_{i=1}^{k} ||\mathbf{c}_{i\cdot} - \overline{\mathbf{c}}_{\cdot\cdot}||^2$ with $\overline{\mathbf{c}}_{\cdot\cdot} = \sum_{i=1}^{k} \overline{\mathbf{c}}_{i\cdot}/k$.

(ii) Choose the number of clusters $k^* = \min\{k_{\mathrm{CH}}, k_{\mathrm{KL}}, k_{\mathrm{WV}}\}$ where

$$k_{\mathrm{CH}} = \underset{1 \le k \le k_{\max}}{\arg\max}\ \mathrm{CH}(k),\ k_{\mathrm{KL}} = \underset{1 \le k \le k_{\max}}{\arg\max}\ \mathrm{KL}(k) \text{ and } k_{\mathrm{WV}} = \max\{k : \mathrm{WV}(k) > 1.2\},$$

$$\mathrm{CH}(k) = \frac{B(k)/(k-1)}{W(k)/(n-k)},\ \mathrm{KL}(k) = \frac{D(k)}{D(k+1)},\ \mathrm{WV}(k) = \frac{W(k)}{W(k+1)}$$

and $D(k) = (k-1)^{2/d} W(k-1) - k^{2/d} W(k)$. $\mathrm{CH}(k)$ is due to Calinski and Harabasz (1974), $KL(k)$ to Krzanowski and Lai (1985) and $\mathrm{WV}(k)$ to Bumgarner (2007). This results in a partition of the data of size k^*.

(iii) Fit a multivariate normal distribution to each of the k^* partition sub-sets *via* maximum likelihood. Form a normal mixture density with weights corresponding to the relative cluster sizes.

Appendix B: Normal Mixture Parameters for Simulation Setting II

Table 1 lists the normal mixture parameters for Setting II of the simulation study described in Section 5. For a symmetric matrix Σ, vech(Σ) is the vector of entries of Σ on and below the diagonal, stacked in order from left to right.

Table 1 Normal mixture parameters for simulation Setting II.

w_1	μ_1	vech(Σ_1)	w_2	μ_2	vech(Σ_2)	w_3	μ_3	vech(Σ_3)	w_4	μ_4	vech(Σ_4)
Normal mixture parameters for f^+											
0.08	420	6609	0.31	194	888	0.36	223	2244	0.25	203	2003
	367	1312		271	312		303	1222		292	1258
	327	334		312	113		278	−625		179	626
	255	−808		314	85		127	652		124	447
	337	433		122	248		301	−223		155	340
		889			390			1670			1840
		421			93			−722			551
		−247			126			901			428
		243			293			−349			564
		2624			716			2911			2655
		1584			142			−592			97
		800			129			377			147
		6789			917			2196			1822
		−189			171			−55			−204
		4133			3384			197			3271
Normal mixture parameters for f^-											
0.17	193	671	0.33	189	605	0.27	183	1001	0.23	300	6658
	270	226		267	179		273	277		343	1241
	319	80		327	96		163	429		256	2216
	316	36		94	29		116	273		144	1264
	116	228		314	113		158	118		300	1701
		321			604			676			623
		72			27			371			696
		15			114			91			330
		268			145			112			431
		621			1445			2669			2477
		52			−80			62			1054
		123			−53			82			1354
		1340			1565			2167			2731
		279			186			−440			495
		3503			1218			4131			2460

References

Baïllo, A., Cuesta-Albertos, J. and Cuevas, A. (2001). Convergence rates in nonparametric estimation of level sets. *Statistics and Probability Letters* **53**, 27–35.

Baggerly, K. A. (2001). Probability binning and testing agreement between multivariate immunofluorescence histograms: extending the chi-squared test. *Cytometry* **45**, 141–150.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*, Wadsworth Publishing, Belmont, CA.

Brinkman, R. R., Gasparetto, M., Lee, S.-J. J., Ribickas, A. J., Perkins, J., Janssen, W., Smiley, R. and Smith, C. (2007). High-content flow cytometry and temporal data analysis for defining a cellular signature of graft-versus-host disease. *Biology of Blood and Marrow Transplantation* **13**, 691–700.

Bumgarner, S. (2007). *Use of Clustering in HIV Detection*. Master of Statistics Thesis, The University of New South Wales, Sydney, Australia.

Cadre, B. (2006). Kernel estimation of density level sets. *Journal of Multivariate Analysis* **97**, 999–1023.

Calinski, R. B. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics – Theory and Methods*, **3**, 1–27.

Drost, F. C., Kallenberg, W. C. M., Moore, D. S. and Oosterhoff, J. (1989). Power approximations to multinomial tests of fit. *Journal of the American Statistical Association* **84**, 130–141.

Duong, T. (2008). prim 1.0.6. R package.

Friedman, J. H. and Fisher, N. I. (1999). Bump-hunting for high dimensional data. *Statistics and Computing* **9**, 123–143.

Hall, P. and Wand, M. P. (1988). On nonparametric discrimination using density differences. *Biometrika* **75**, 541–547.

Hartigan, J. A. (1987). Estimation of a convex density contour in two dimensions. *Journal of the American Statistical Association* **82**, 267–270.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer-Verlag, New York.

Hyndman, R. J. (1996). Computing and graphing highest density regions. *The American Statistician* **50**, 120–126.

Jang, W. (2006). Nonparametric density estimation and clustering in astronomical sky surveys. *Computational Statistics and Data Analysis* **50**, 760–774.

Krzanowski, W. J. and Lai, Y. T. (1985). A criterion for determining the number of groups in a data set using sum of squares clustering. *Biometrics* **44**, 23–34.

McLaren, C. E., Legler, J. M. and Brittenham, G. M. (1994). The generalised χ^2 goodness-of-fit test. *The Statistician* **43**, 247–258.

Moore, D. S. (1984). Measures of lack of fit from test of chi-squared type. *Journal of Statistical Planning and Inference* **10**, 151–166.

Müller, D. W. and Sawitzki, G. (1991). Excess mass estimates and tests for multimodality. *Journal of the American Statistical Association* **86**, 738–746.

Patnaik, P. B. (1949). The non-central χ^2- and F-distributions and their applications. *Biometrika* **36**, 202–232.

Pederson, S. P. and Johnson, M. E. (1990). Estimating model discrepancy. *Technometrics* **32**, 305–314.

Polonik, W. (1995). Measuring mass concentrations and estimating density contour clusters–an excess mass approach. *The Annals of Statistics* **23**, 855–881.

Rice, J. A. (1995). *Mathematical Statistics and Data Analysis*, 2nd Edn.. Duxbury Press, Belmont, CA.

Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer-Verlag, New York.

Roederer, M. and Hardy, R. R. (2001). Frequency difference gating: a multivariate method for identifying subsets that differ between samples. *Cytometry* **45**, 56–64.

Roederer, M., Moore, W., Treister, A., Hardy, R. R. and Herzenberg, L. A. (2001a). Probability binning comparison: a metric for quantitating univariate distribution differences. *Cytometry* **45**, 37–46.

Roederer, M., Moore, W., Treister, A., Hardy, R. R. and Herzenberg, L. A. (2001b). Probability binning comparison: a metric for quantitating multivariate distribution differences. *Cytometry* **45**, 47–55.

Samworth, R. J. and Wand, M. P. (2008). Asymptotics and optimal bandwidth selection for highest density region estimation. Unpublished manuscript.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, New York.

Shapiro, H. M. (2003). *Practical Flow Cytometry*, 4th Edn. Wiley, New York.

Steinwart, I., Hush, D. and Scovel, C. (2005). Density level detection is classification. *In Neural Information Processing Systems* **17**, 1337–1344.

Tsybakov, A. B. (1997). On nonparametric estimation of density level sets. *The Annals of Statistics* **25**, 948–969.

Tukey, J. W. (1972). Some graphic and semigraphic displays. In: T. A. Bancroft (Ed.), *Statistical Papers in Honor of George W. Snedecor*, Iowa State University Press, Ames, Iowa, pp. 293–316.

