University of Wollongong
Research Online

Centre for Statistical & Survey Methodology
Working Paper Series

Faculty of Engineering and Information Sciences

2008

Highest density difference region estimation with application to flow cytometric data

T. Duong
Institut Pasteur

I. Koch
University of New South Wales

M. P. Wand
University of Wollongong, mwand@uow.edu.au

Recommended Citation
Duong, T.; Koch, I.; and Wand, M. P., Highest density difference region estimation with application to flow cytometric data, Centre for Statistical and Survey Methodology, University of Wollongong, Working Paper 19-08, 2008, 19p.
http://ro.uow.edu.au/cssmwp/16

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au

Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

19-08

Highest density difference region estimation with application to flow cytometric data

Tarn Duong, Inge Koch and Matt P. Wand

Copyright © 2008 by the Centre for Statistical & Survey Methodology, UOW. Work in progress, no part of this paper may be reproduced without permission from the Centre.

Centre for Statistical & Survey Methodology, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

Highest density difference region estimation with application to flow cytometric data

TARN DUONG

Imaging and Modelling Group, Institut Pasteur, Paris, France

INGE KOCH

School of Mathematics and Statistics, The University of New South Wales, Sydney, Australia

AND M.P. WAND

School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, Australia

31st January, 2008

Abstract

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 & 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 (PRIM). 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.

Keywords: Flow Cytometry; Frequency Difference Gating; Highest density region; Multivariate density estimation; Patient Rule Induction Method; Probability Binning.

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 Figure 1 (source: Roederer, Moore, Treister, Hardy & Herzenberg, 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 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 & 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 of people in order to identify cell populations that may correlate with clinically-relevant differences'. In this article we critique the approach of Roederer & Hardy (2001), known as *Frequency Difference Gating*, and explore improvements based on recent developments in Statistics. While driven by flow cytometric research, the resulting methodology is applicable to general settings involving large multivariate samples.

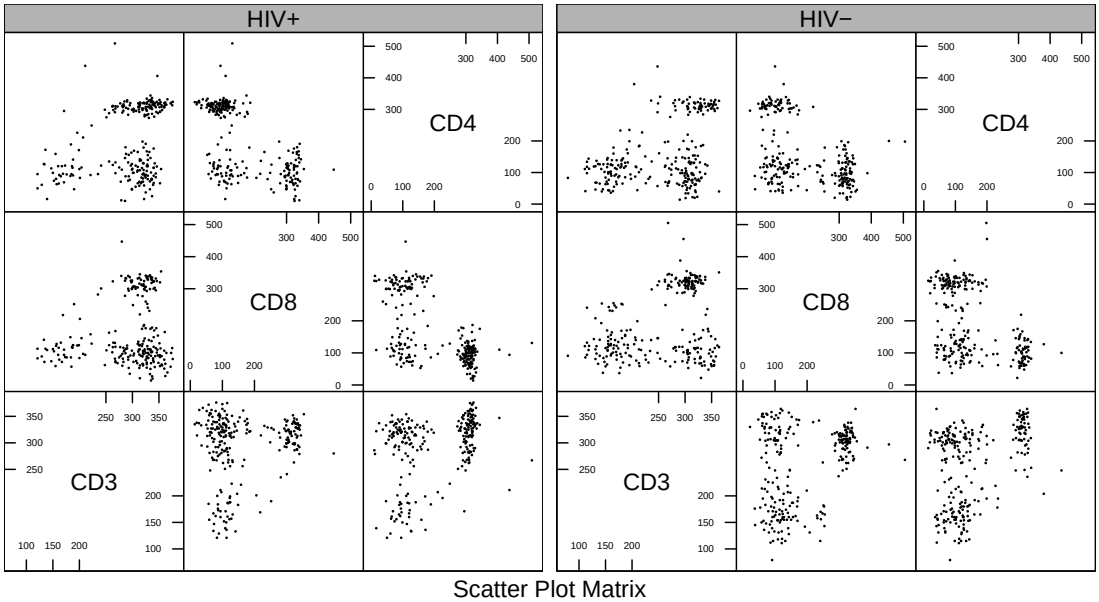

Figure 1: *Scatter plot matrices for HIV+ (left) and HIV– (right) patients.*

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, has an established literature. Contributions include Hartigan (1987), Müller & Sawitzki (1991), Polonik (1995), Tsybakov (1997), Baíllo, Cuesta-Albertos & 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 Frequency Difference Gating method for HDDR estimation (Roederer & Hardy, 2001) involves the following two steps:

(1) Partition the space (e.g. $\mathbb{R}^3$) into box-shaped sub-regions using one of the samples (labelled the 'control' sample by Roederer & Hardy, 2001).

(2) Perform chi-squared tests on the resulting counts. Use the test statistics to estimate regions of differing density.

Roederer & Hardy (2001) achieve (1) via an algorithm called *Probability Binning* (Roederer, Moore, Treister, Hardy & Herzenberg; 2001a, 2001b) – a recursive binary partitioning strategy similar to that used by Classification and Regression Trees (CART) (Breiman, Friedman, Olshen & Stone, 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 & 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 Frequency Difference Gating on simulated data.

Baggerly (2001) critiqued the chi-squared methodology of Roederer & 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 & 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 highest density difference region estimation, while in Section 3 we review the Frequency Difference Gating 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 highest density difference region estimation with Frequency Difference Gating. 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 Frequency Difference Gating algorithm of Roederer & 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 so, 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

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

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

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

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

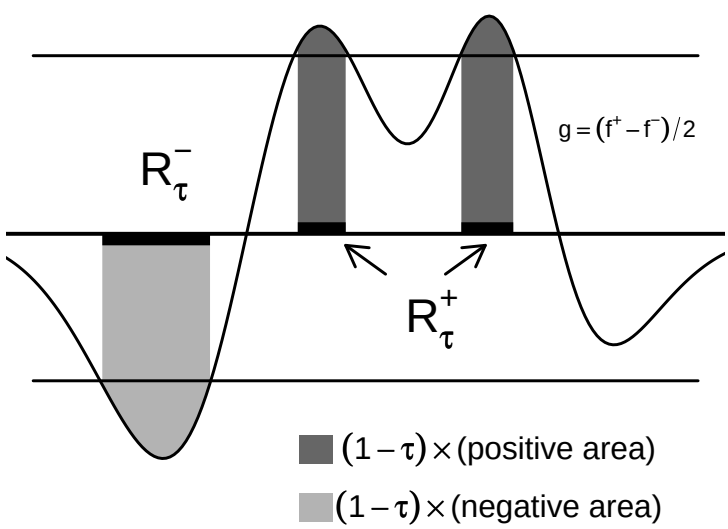

Figure 2: *Graphical description of R_τ^+ and R_τ^- when g is defined on $\mathbb{R}$.*

Now consider the problem of estimating R_τ from random samples

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

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

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

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

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

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

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

where $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ 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 Frequency Difference Gating (Roederer & Hardy, 2001) are

(1) Use an algorithm called *Probability Binning* to partition $\mathbb{R}^d$ into sub-regions and obtain counts for those sub-regions.

(2) 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^+: $\boldsymbol{X}_1^+, \ldots, \boldsymbol{X}_{n^+}^+$ to be the control sample. The first stage of Frequency Difference Gating involves dividing $\mathbb{R}^d$ into box-shaped regions so that the counts based on the $\boldsymbol{X}_i^+$'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 $\boldsymbol{X}_i^+$'s. The same boxes are used for the f^- sample and the counts for that sample obtained. For a K-box partition Frequency Difference Gating leads to a $2 \times K$ contingency table. Figure 3 is a graphical description of Frequency Difference Gating 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 Frequency Difference Gating boxes are generated via recursive binary partitioning. Starting with the smallest d-dimensional hyper-rectangle, or box, containing the control sample, the procedure splits the original box along the $(d-1)$-dimensional hyperplane orthogonal to the dimension which has maximum marginal sample variance. The split point is chosen so that the two new boxes contain equal numbers 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 Frequency Difference Gating for some simulated data in $\mathbb{R}^2$.

The second phase of Frequency Difference Gating involves forming the $2 \times r$ contingency table of counts based on the control and test samples induced by the partition. Chi-squared tests

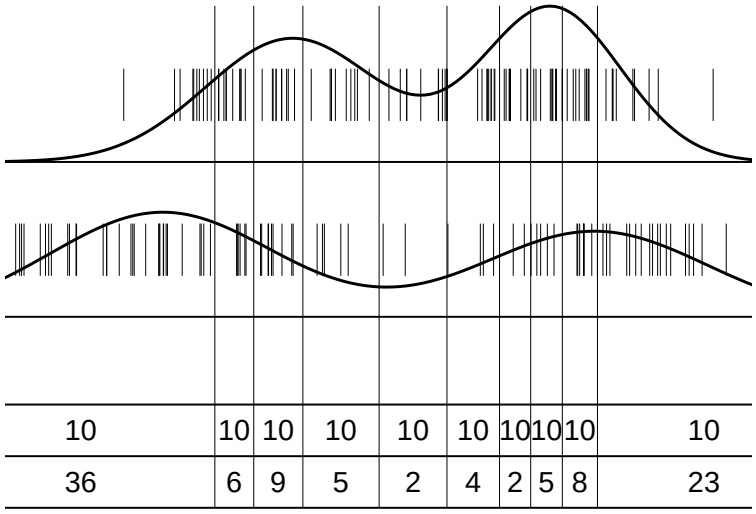

Figure 3: *Graphical description of Frequency Difference Gating for univariate data.*

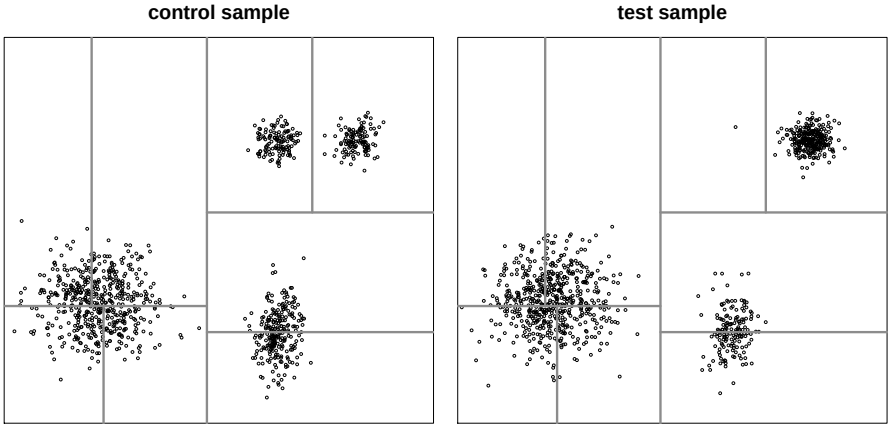

Figure 4: *Illustration of Frequency Difference Gating in $d = 2$ dimensions. In this case $L = 3$ levels of recursive binary partitioning have been performed to produce $2^3 = 8$ boxes.*

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 & Hardy (2001) are overly conservative. He then described appropriate modifications to the chi-squared testing phase.

While 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 & Friedman, 2001). Binary partitioning means that the data are fragmented quite quickly. Friedman & Fisher (1999) devised the Patient Rule Induction Method (PRIM) to redress this problem. Section 4.1 explains how PRIM can be adapted for highest density difference region estimation.

Another problem with Frequency Difference Gating is over-sensitivity of chi-squared tests for very large samples. This issue is discussed by Pederson and Johnson (1990) and McLaren, Legler

& Brittenham (1994) who recommended use of indifference regions. Details are given in Section 4.2.

4 A PRIM-based Algorithm for HDDR Estimation

We now describe our alternative to Frequency Difference Gating for HDDR estimation. It calls upon PRIM to obtain sub-regions 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 PRIM

The Patient Rule Induction Method (PRIM) was developed by Friedman & 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, Tibshirani & Friedman (2001). The main input into PRIM is a set of regression-type data: $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$, where $\boldsymbol{X}_i \in \mathbb{R}^d$ and $Y_i \in \mathbb{R}$. The output is a set of boxes in $\mathbb{R}^d$ that estimate regions corresponding to maxima in $E(Y|\boldsymbol{X} = \mathbf{x})$.

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, 2007) 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, so consider testing

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

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

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

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 & Johnson (1990) and McLaren *et al.* (1994) recommend working with the generalised hypothesis set-up

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

where the interval $(0, \delta_0)$ is an indifference region. According to Drost, Kallenberg, Moore & Oosterhoff (1989), the approximation

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

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 written 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 $H_0 : p_j^+ = p_j^-, \quad j = 1, \cdots, c$ 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

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

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 that 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 $(\boldsymbol{X}_i, Y_i)$ and $(\boldsymbol{X}_i, -Y_i)$ where the $\boldsymbol{X}_i$'s are the $\boldsymbol{X}_i^+$ and $\boldsymbol{X}_i^-$ data pooled together and the $Y_i \in \{-1, 1\}$ are indicators of whether $\boldsymbol{X}_i$ is from the $-$ or $+$ sample. The first application of PRIM, using the $(\boldsymbol{X}_i, Y_i)$, produces a partition of $\mathbb{R}^d$, the elements of which are candidates for $\widehat{R}_\tau^+$. Similarly, candidates for $\widehat{R}_\tau^-$ are generated via application of PRIM to the $(\boldsymbol{X}_i, -Y_i)$. 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 Figure 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 minimum box mean 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 rudimentary method for choosing the minimum box means from the data. There is clearly room for further research on this problem.

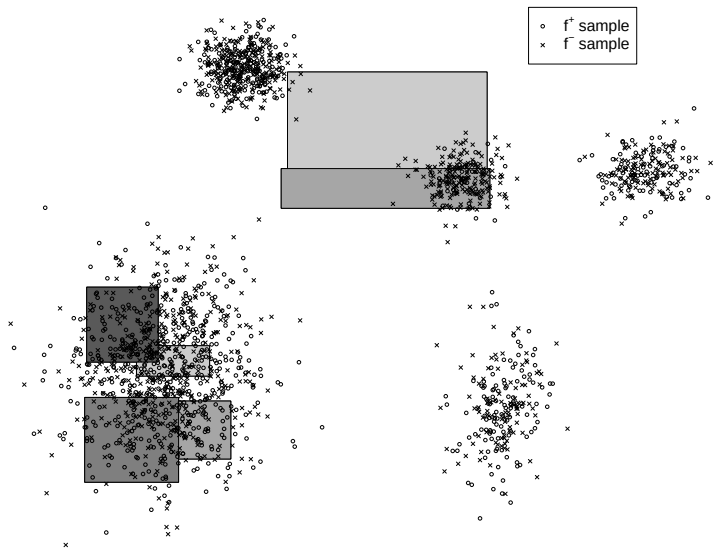

Figure 5: *Illustration of PRIM-based partitioning. The shaded partition elements in the lower left region are candidates for estimation of R_τ^+; those in the upper middle region are candidates for estimation of R_τ^-. Grey-level shading is used to distinguish the partition elements.*

4.4 Approximation of Error in HDDR 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 this can 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 $g^+ \Big/ \int_{\mathbb{R}^d} g^+$. Then, following the argument in Section 3.3 of Hyndman (1996), a consistent estimate of g_τ^+ is the τth sample quantile of $g^+(\mathbf{Z}_1)$,...,$g^+(\mathbf{Z}_N)$. 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 & 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, so 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 Frequency Difference Gating. Throughout the study we fixed $\tau = 0.5$. Two different settings were considered; with dimension and sample sizes as follows:

$$\begin{array}{ll} \text{Setting I}: & d = 2, \quad n = 1,000 \\ \text{Setting II}: & d = 5, \quad n = 100,000. \end{array}$$

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

$$w_1N(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)+w_2N(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)+w_3N(\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3)+w_4N(\boldsymbol{\mu}_4,\boldsymbol{\Sigma}_4)$$

where each $\boldsymbol{\mu}_k$ is a 5×1 vector and each $\boldsymbol{\Sigma}_k$ 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 so, for this application, Setting II has an element of realism with respect to the motivating application. Table 1 in Appendix B lists these parameters.

One hundred replications were used in each setting. PRIM estimates of $R_{0.5}$ were obtained via the algorithm described in Appendix A. The Frequency Difference Gating estimates required a choice of the level parameter L. This was achieved by the analogue of the normal mixture pilot approach to choosing minimum box means as described in Appendices A.2 and A.3. In each case we approximated the error measure $\mathrm{err}(\widehat{R}_{0.5})$ as given by (1) and (2). For the Monte Carlo approximation (3) we used $N=1,000,000$ throughout the study.

Figure 6 shows typical PRIM-based and Frequency Difference Gating-based estimates of $R^+_{0.5}$ and $R^-_{0.5}$ for Setting I. Each estimate is that which results in the median $\mathrm{err}(\widehat{R}_{0.5})$ 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 Frequency Difference Gating 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 7 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 two-fold to three-fold reduction in the error measure. 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 in Brinkman *et al.*(2007). Brinkman *et al.*(2007) state: "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.

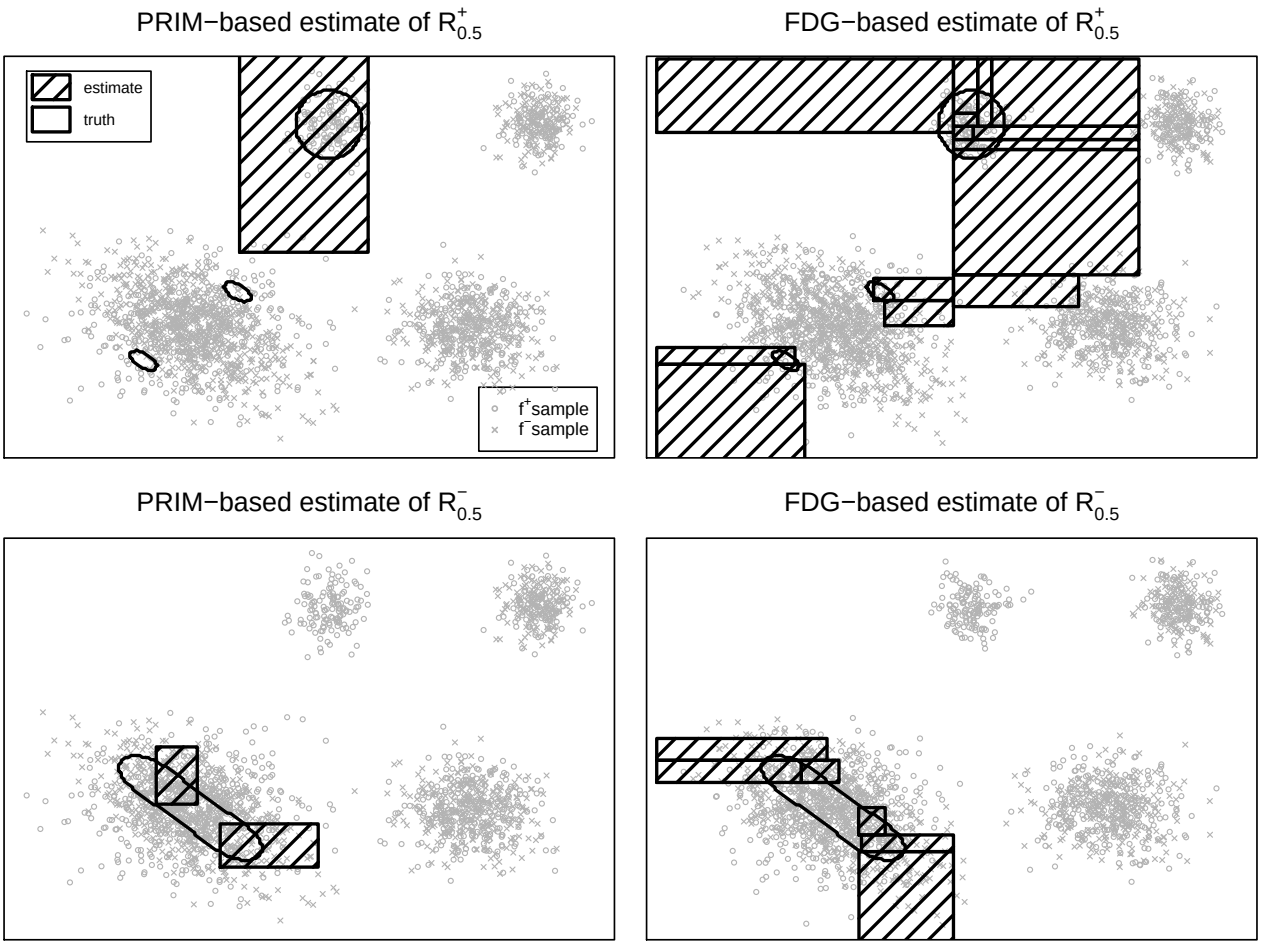

Figure 6: *Median performance PRIM-based and Frequency Difference Gating (FDG)-based estimates of $R^+_{0.5}$ and $R^-_{0.5}$ for Setting I of the simulation study.*

We conducted a cursory 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 6-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 6-dimensional samples described in the previous paragraph. We then gated on lymphocytes using forward-scatter and side-scatter. The positive and negative estimated HDDRs in the ensuing antibodies space are shown in Figures 8 and 9. It is seen that there is quite a striking difference in the two samples with these HDDR gates. These type of gates have the potential to aid the discovery of new biomarkers for GvHD.

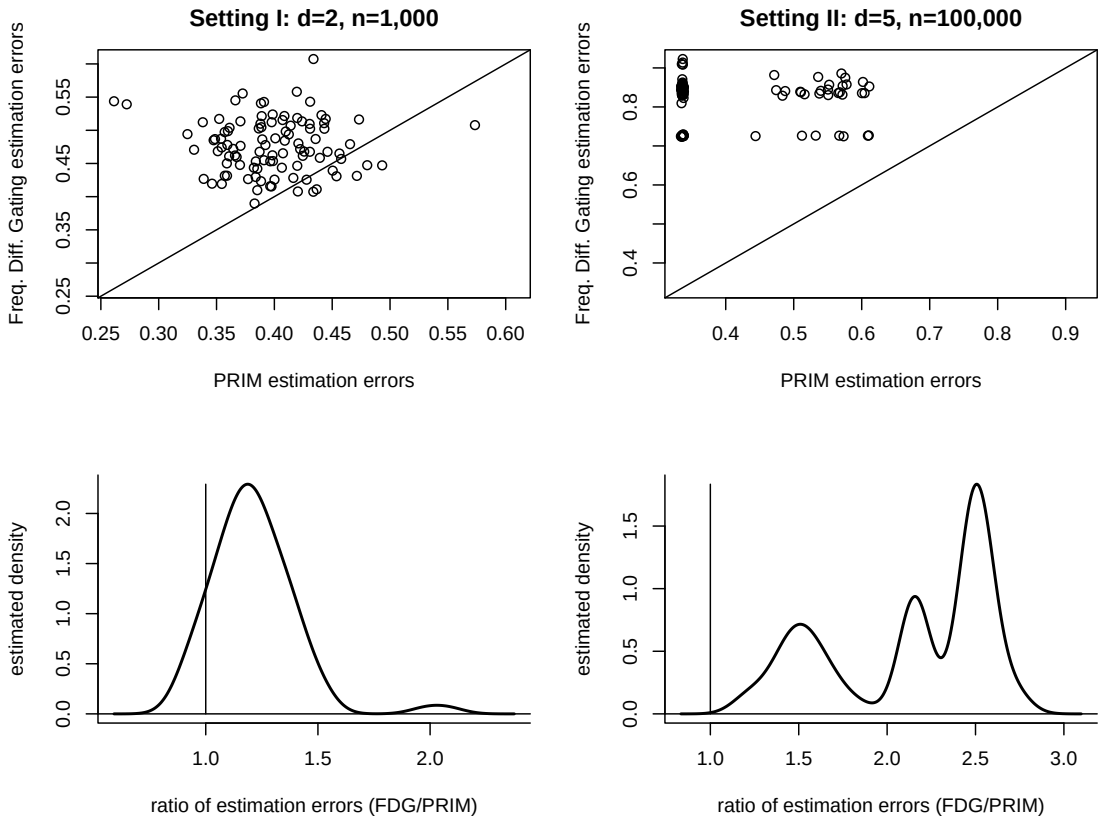

Figure 7: *Summary of simulation results. The upper panels are scatterplots of the* $\mathrm{err}(\widehat{R}_{0.5})$ *values for Frequency Difference Gating versus PRIM-based estimation, together with the 1:1 line. The lower panels are kernel density estimates of the error ratios (with Frequency Difference Gating error on the numerator).*

7 Conclusion

Highest density difference region 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 & 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 Frequency Difference Gating approach of Roederer & 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 & Scovel, 2005) and general classification methodology (e.g. Hastie, Tibshirani & Friedman, 2001) also await exploration. Finally, data-driven rules for the selection of smoothing and auxiliary parameters in

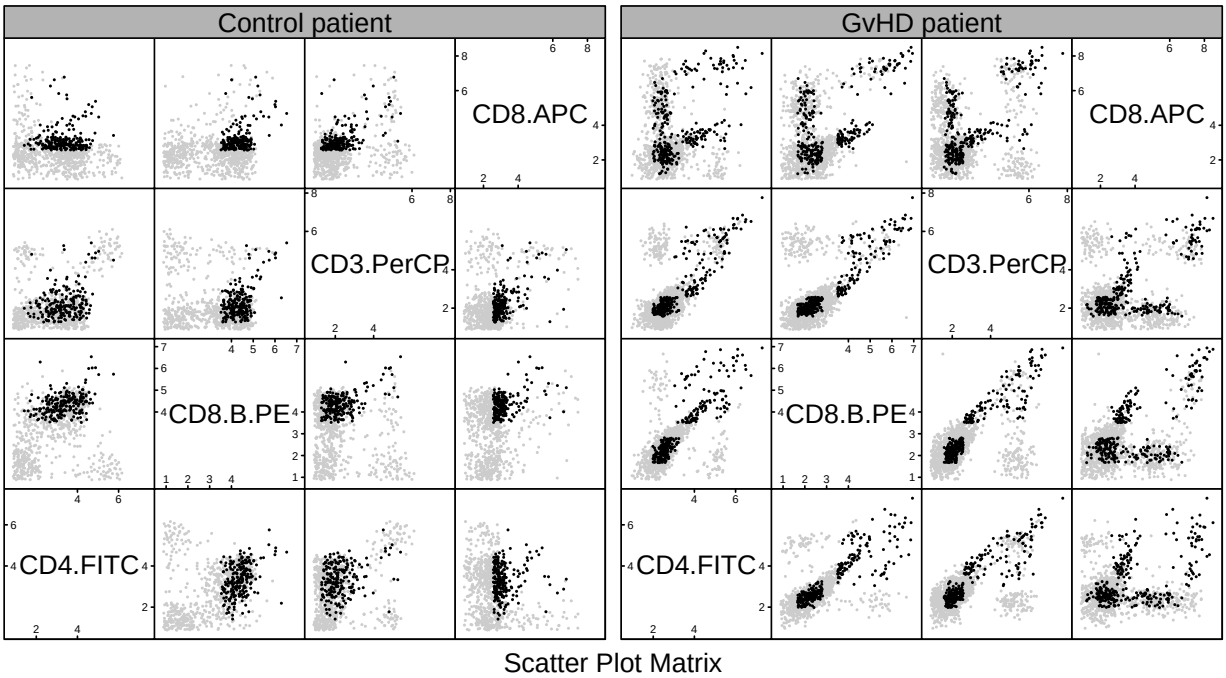

Figure 8: *Estimated positive highest difference density region ($\tau = 0.5$) for the GvHD data, after gating on lymphocytes. The dark points are those inside the region. A 10% sample has been taken to aid visualisation. All data have undergone the inverse sinh transformation before plotting and processing.*

the highest density and density difference region contexts remains an open topic.

Acknowledgments

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 Project grant (Project DP0556518).

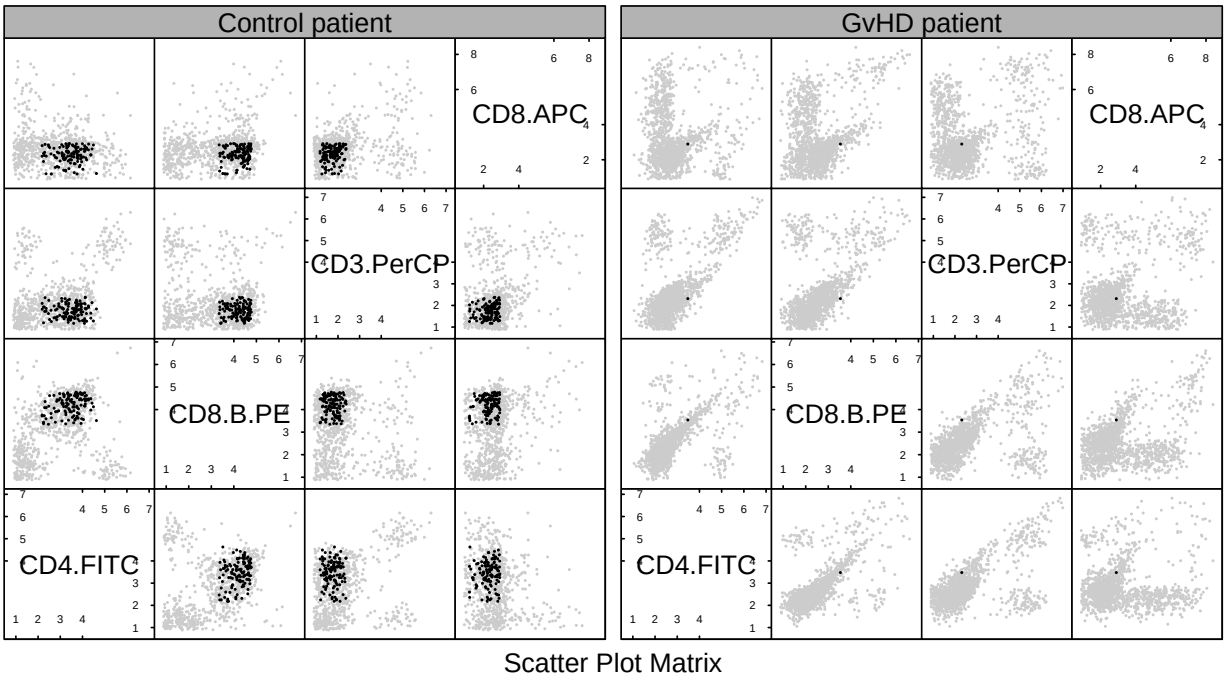

Figure 9: *Estimated negative highest difference density region ($\tau = 0.5$) for the GvHD data, after gating on lymphocytes. The dark points are those inside the region. A 10% sample has been taken to aid visualisation. All data have undergone the inverse sinh transformation before plotting and processing.*

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.
- $\boldsymbol{X}_1^+, \ldots, \boldsymbol{X}_{n^+}^+$ and $\boldsymbol{X}_1^-, \ldots, \boldsymbol{X}_{n^-}^-$; 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 $K_{\text{max}} = 20$.
- $\text{MBM}^+(\tau)$, $\text{MBM}^-(\tau)$; the minimum box means for estimation of R_τ^+ and R_τ^-, respectively. We will suppose, for now, that they have been chosen from the data to be $\widehat{\text{MBM}}^+(\tau)$ and $\widehat{\text{MBM}}^-(\tau)$. Appendix A.1 discusses a cursory data-driven rule for their choice.

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

1. Form the 'regression' data set $(\boldsymbol{X}_i, Y_i),\ 1 \le i \le n$ ($n = n^+ + n^-$), for input into PRIM. The $\boldsymbol{X}_i$'s are the $\boldsymbol{X}_i^+$ and $\boldsymbol{X}_i^-$ data pooled together. The $Y_i \in \{-1, 1\}$ are indicators of whether $\boldsymbol{X}_i$ is from the $-$ or $+$ sample.
2. Feed the $(\boldsymbol{X}_i, Y_i)$ data into PRIM to obtain K_{max} (ordered) boxes $B_1, \ldots, B_{K_{\text{max}}}$ 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^+.
3. Repeat Step 2, but with $(\boldsymbol{X}_i, -Y_i)$, fed into PRIM to obtain $P_1^-, \ldots, P_{K^-}^-$, with $K^- = \max\{k : \text{ave}(-Y; P_k^-) \le \widehat{\text{MBM}}^-(\tau)\}$.
4. 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.
5. 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 $K = K^+ + K^-$.
6. Obtain the counts among the $\boldsymbol{X}_i^+$ and $\boldsymbol{X}_i^-$ samples, respectively, over the partition $P_1, \ldots, P_{K+1}$. Combine the counts into a $2 \times (K+1)$ contingency table.
7. 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 $K+1$.
8. 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 $\widehat{R}_\tau^+$ to be the union of all P_k for which significance is achieved and $\hat{p}_k^+ > \hat{p}_k^-$.

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

A.1. Data-driven choice of Minimum Box Mean

1. Obtain pilot estimates of R_τ^+ and R_τ^- based on fitting normal mixture densities to the $\boldsymbol{X}_i^+$ and $\boldsymbol{X}_i^-$. 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.
2. Apply PRIM to the $(\boldsymbol{X}_i, Y_i)$ with MBM $= \overline{Y}$ to give a sequence $P_1^+, \ldots, P_{K_0^+}^+$, with minimum box means $\text{ave}(Y; P_g^+), g = 1, \ldots, K_0^+$.
3. Apply PRIM to the $(\boldsymbol{X}_i, Y_i)$ K_0^+ times each with MBM equal to $\text{ave}(Y; P_g^+)$ to obtain estimates $\widehat{R}_{\tau,g}^+$ for $1 \le g \le K_0^+$.
4. Take $\widehat{\text{MBM}}^+(\tau)$ to be $\underset{1 \le g \le K_0^+}{\text{argmin}}\ \text{err}(\widehat{R}_{\tau,g}^+, R_{\tau,\text{pilot}}^+)$.
5. Repeat Steps 2.-4.; but using the $(\boldsymbol{X}_i, -Y_i)$ data in PRIM to obtain $P_1^-, \ldots, P_{K_0^-}^-$, with minimum box means $\text{ave}(-Y; P_g^-), g = 1, \ldots, K_0^-$ and take $\widehat{\text{MBM}}^-(\tau) = \underset{1 \le g \le K_0^-}{\text{argmin}}\ \text{err}(\widehat{R}_{\tau,g}^-, R_{\tau,\text{pilot}}^-)$.

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.

1. For each $k = 1, \ldots, k_{\text{max}}$ 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 $W(1), \ldots, W(k_{\text{max}})$ 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 $\overline{\mathbf{c}}_{\cdot\cdot} = \sum_{i=1}^{k} \overline{\mathbf{c}}_{i\cdot}/k$.
2. Choose the number of clusters $k^* = \min\{k_{\text{CH}}, k_{\text{KL}}, k_{\text{WV}}\}$ where
$$k_{\text{CH}} = \underset{1 \le k \le k_{\text{max}}}{\text{argmax}}\ \text{CH}(k), \quad k_{\text{KL}} = \underset{1 \le k \le k_{\text{max}}}{\text{argmax}}\ \text{KL}(k) \quad \text{and} \quad k_{\text{WV}} = \max\{k : \text{WV}(k) > 1.2\},$$
$$\text{CH}(k) = \frac{B(k)/(k-1)}{W(k)/(n-k)}, \quad \text{KL}(k) = \frac{D(k)}{D(k+1)}, \quad \text{WV}(k) = \frac{W(k)}{W(k+1)}$$
and $D(k) = (k-1)^{2/d} W(k-1) - k^{2/d} W(k)$. $\text{CH}(k)$ is due to Calinski and Harabasz (1974), $\text{KL}(k)$ to Krzanowski and Lai (1985) and $\text{WV}(k)$ to Bumgarner (2007). This results in a partition of the data of size k^*.

3. 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 $\mathbf{\Sigma}$, vech($\mathbf{\Sigma}$) is the vector of entries of $\mathbf{\Sigma}$ on and below the diagonal; stacked in order from left to right.

Normal mixture parameters for f^+.											
w_1	$\boldsymbol{\mu}_1$	vech($\mathbf{\Sigma}_1$)	w_2	$\boldsymbol{\mu}_2$	vech($\mathbf{\Sigma}_2$)	w_3	$\boldsymbol{\mu}_3$	vech($\mathbf{\Sigma}_3$)	w_4	$\boldsymbol{\mu}_4$	vech($\mathbf{\Sigma}_4$)
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^-.											
w_1	$\boldsymbol{\mu}_1$	vech($\mathbf{\Sigma}_1$)	w_2	$\boldsymbol{\mu}_2$	vech($\mathbf{\Sigma}_2$)	w_3	$\boldsymbol{\mu}_3$	vech($\mathbf{\Sigma}_3$)	w_4	$\boldsymbol{\mu}_4$	vech($\mathbf{\Sigma}_4$)
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

Table 1: *Normal mixture parameters for simulation Setting II.*

References

Baïllo, A., Cuesta-Albertos, J. & 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. & Stone, C.J. (1984). *Classification and Regression Trees*, Belmont, California: Wadsworth Publishing.

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.

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

Calinski, R.B. & 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. & Oosterhoff, J. (1989). Power approximations to multinomial tests of fit. *Journal of the American Statistical Association*, **84**, 130–141.

Duong, T. (2007). `prim 1.0.1`. R package. `http://cran.r-project.org`.

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

Hall, P. & 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. & Friedman, J. (2001). *The Elements of Statistical Learning.* New York: Springer-Verlag.

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. & 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., & 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. & 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. & 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, Second Edition*. Belmont, California: Duxbury Press.

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

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

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

Shapiro, H.M. (2003). *Practical Flow Cytometry, 4th Edition.* New York: John Wiley & Sons.

Steinwart, I., Hush, D. & 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. *Statistical Papers in Honor of George W. Snedecor.* (T.A. Bancroft, ed.) 293–316. Ames: Iowa State University Press.