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  
http://ro.uow.edu.au/cssmwp/16
Higher density difference region estimation with application to flow cytometric data

Tarn Duong, Inge Koch and Matt P. Wand
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.

![Scatter Plot Matrix](image)

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), Baillo, 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 \( \chi^2 \)-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 \( \chi^2 \) 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 \( \pi \) 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

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

so that \( g = g^+ - g^- \). For \( 0 \leq \tau \leq 1 \) the \( \tau \) highest positive density difference region is

\[
R^+_\tau \equiv \{x \in \mathbb{R}^d : g^+(x) \geq g^+_\tau\} \quad \text{where} \quad g^+_\tau \text{ is the greatest number for which } \int_{R^+_\tau} g^+(x) \, dx \geq 1 - \tau.
\]

The \( \tau \) highest negative density difference region \( R^-_\tau \) and corresponding threshold \( g^-_\tau \) are defined analogously. The \( \tau \) highest density difference region is then

\[
R_\tau \equiv R^+_\tau \cup R^-_\tau.
\]

Figure 2 provides a graphical description of \( R_\tau \) 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 \( R_\tau \) from random samples

\[
X^+_1, \ldots, X^+_n \sim f^+ \quad \text{and} \quad X^-_1, \ldots, X^-_n \sim f^-.
\]
Let \( \hat{R}_+^\tau \) and \( \hat{R}_-^\tau \) respectively be estimators of \( R_+^\tau \) and \( R_-^\tau \) so that \( \hat{R}_\tau = \hat{R}_+^\tau \cup \hat{R}_-^\tau \) is an estimator of \( R_\tau \). We quantify the error in \( \hat{R}_\tau \) via

\[
\text{err}(\hat{R}_\tau) = \mu(\hat{R}_+^\tau \Delta R_+^\tau) + \mu(\hat{R}_-^\tau \Delta R_-^\tau)
\]

for some measure \( \mu \) 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 \( R_-^\tau \). There are several options for \( \mu \). 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:

\[
\mu(A) = \frac{1}{2} \left\{ f^+(x) + f^-(x) \right\} \, dx
\]

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

\[
\mu(A) \approx \frac{1}{N} \sum_{i=1}^{N} I(X_i \in A)
\]

where \( X_1, \ldots, X_N \) 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 \( f^+ \): \( X_1^+, \ldots, X_n^+ \) 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 \( X_i^+ \)'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 \( X_i^+ \)'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 \( \mathbb{R}^2 \) 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
recursive binary partitioning have been performed to produce quite quickly. Friedman & Fisher (1999) devised the Patient Rule Induction Method (PRIM) to re-

Hastie, Tibshirani & Friedman, 2001). Binary partitioning means that the data are fragmented 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 re-

dress 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
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:

\[(X_1, Y_1), \ldots, (X_n, Y_n)\], where \(X_i \in \mathbb{R}^d\) and \(Y_i \in \mathbb{R}\). The output is a set of boxes in \(\mathbb{R}^d\) that estimate regions corresponding to maxima in \(E(Y|X = x)\).

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 \(\mathbb{R}\) package \texttt{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

\[H_0 : p_{1j} = \cdots = p_{rj}, \quad j = 1, \ldots, c \quad \text{versus} \quad H_1 : \text{not } H_0\]  

(4)

where \(p_{ij}\) is the probability of the \(j\)th category of the \(i\)th multinomial distribution. The Pearson \(X^2\) statistic is

\[X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.\]  

(5)

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

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

\[\delta = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(p_{ij} - \overline{p}_j)^2}{\overline{p}_j}\]

is the discrepancy between the \(p_{ij}\) and the \(\overline{p}_j \equiv \sum_{i=1}^{r} p_{ij} / r\). 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

\[
X^2 \approx \chi^2_{(r-1)(c-1)}(n\delta_0) \quad \text{under } H_0 \quad (6)
\]

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 \(\nu\). Patnaik (1949) is an early reference for derivation of (6). Generalised chi-squared tests thus involve setting the indifference parameter \(\delta_0\) 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 \(n^+_j \equiv O_{1j}, n^-_j \equiv O_{2j}, n^+ \equiv O_1,\) and \(n^- \equiv O_2\). Then the Pearson \(X^2\) statistic (5) can be written as

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

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

\[
X^2 > \chi^2_{c-1; 1-\alpha}(n\delta_0) \quad (8)
\]

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

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

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 \((X_i, Y_i)\) and \((X_i, -Y_i)\) where the \(X_i\)’s are the \(X^+_i\) and \(X^-_i\) data pooled together and the \(Y_i \in \{-1, 1\}\) are indicators of whether \(X_i\) is from the – or + sample. The first application of PRIM, using the \((X_i, Y_i)\), produces a partition of \(\mathbb{R}^d\), the elements of which are candidates for \(\hat{R}^+_\tau\). Similarly, candidates for \(\hat{R}^-\) are generated via application of PRIM to the \((X_i, -Y_i)\). 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 \(R_\tau\). 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.
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 \( R_\tau^+ \); an analogous approach applies to \( R_\tau^- \).

Let \( Z_1, \ldots, Z_N \) 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 \( g^+_\tau \) is the \( \tau \)th sample quantile of \( g^+(Z_1), \ldots, g^+(Z_N) \). Since
\[
g^+ \leq |g| = |\frac{1}{2} f^+ - \frac{1}{2} f^-| \leq \frac{1}{2} f^+ + \frac{1}{2} f^-
\]
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:

Setting I: \( d = 2, \quad n = 1,000 \)
Setting II: \( d = 5, \quad n = 100,000 \).

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
\[
\frac{1}{2} N \left( \left[ \begin{array}{c} 0 \\ \frac{1}{8} \\ 0 \\ 0 \\ 0 \\
\end{array} \right], \left[ \begin{array}{c} \frac{1}{16} \\ 0 \\ \frac{1}{16} \\ 0 \\ 0 \\
\end{array} \right] \right) + \frac{1}{4} N \left( \left[ \begin{array}{c} 2 \\ 0 \\ 0 \\ 0 \\ 0 \\
\end{array} \right], \left[ \begin{array}{c} \frac{1}{64} \\ 0 \\ \frac{1}{64} \\ 0 \\ 0 \\
\end{array} \right] \right) + \frac{1}{8} N \left( \left[ \begin{array}{c} \frac{5}{2} \\ 0 \\ 0 \\ 0 \\ 0 \\
\end{array} \right], \left[ \begin{array}{c} \frac{1}{64} \\ 0 \\ \frac{1}{64} \\ 0 \\ 0 \\
\end{array} \right] \right)
\]
and \( f^- \) to be
\[
\frac{5}{8} N \left( \left[ \begin{array}{c} 0 \\ -\frac{3}{8} \\ -\frac{3}{8} \\ 0 \\ 0 \\
\end{array} \right], \left[ \begin{array}{c} -\frac{3}{8} \\ 0 \\ 0 \\ \frac{1}{16} \\ 0 \\
\end{array} \right] \right) + \frac{1}{4} N \left( \left[ \begin{array}{c} 2 \\ 0 \\ 0 \\ 0 \\ 0 \\
\end{array} \right], \left[ \begin{array}{c} \frac{1}{16} \\ 0 \\ \frac{1}{16} \\ 0 \\ 0 \\
\end{array} \right] \right) + \frac{1}{8} N \left( \left[ \begin{array}{c} \frac{5}{2} \\ 0 \\ 0 \\ 0 \\ 0 \\
\end{array} \right], \left[ \begin{array}{c} \frac{1}{64} \\ 0 \\ \frac{1}{64} \\ 0 \\ 0 \\
\end{array} \right] \right).
\]

For Setting II, both densities are of the form
\[
w_1 N(\mu_1, \Sigma_1) + w_2 N(\mu_2, \Sigma_2) + w_3 N(\mu_3, \Sigma_3) + w_4 N(\mu_4, \Sigma_4)
\]
where each \( \mu_k \) is a \( 5 \times 1 \) vector and each \( \Sigma_k \) is a \( 5 \times 5 \) covariance matrix (1 \( \leq k \leq 4 \)). We obtained the normal mixture parameters for \( f^+ \) and \( f^- \) by fitting normal mixtures to actual flow cytometry data so that for this application, Setting II has an element of realism with respect to the motivating application. Table I 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 \( \text{err}(\hat{R}_{0.5}) \) 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 \( \text{err}(\hat{R}_{0.5}) \) 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 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, CD8/3-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 $\tau$ 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 $\text{err}(\hat{R}_{0.5})$ 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 $k$th-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
<table>
<thead>
<tr>
<th>Control patient</th>
<th>GvHD patient</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Control patient" /></td>
<td><img src="image2" alt="GvHD patient" /></td>
</tr>
<tr>
<td><img src="image3" alt="Control patient" /></td>
<td><img src="image4" alt="GvHD patient" /></td>
</tr>
<tr>
<td><img src="image5" alt="Control patient" /></td>
<td><img src="image6" alt="GvHD patient" /></td>
</tr>
<tr>
<td><img src="image7" alt="Control patient" /></td>
<td><img src="image8" alt="GvHD patient" /></td>
</tr>
<tr>
<td>Scatter Plot Matrix</td>
<td></td>
</tr>
</tbody>
</table>

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 \( R_\tau^+ \) and \( R_\tau^- \). This defines the target HDDR.
- \( X_1^+, \ldots, X_{n^+}^+ \) and \( X_1^-, \ldots, X_{n^-}^- \); two multivariate random samples in \( \mathbb{R}^d \); with underlying densities \( f^+ \) and \( f^- \) respectively.
- \( \alpha \); the significance level in the generalised chi-squared tests. The default is 0.05.
- \( \delta_0 \); 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.
- \( K_{\text{max}} \); the maximum number of boxes. The default is \( K_{\text{max}} = 20 \).
- \( \overline{\text{MBM}}^+ (\tau), \overline{\text{MBM}}^- (\tau) \); the minimum box means for estimation of \( R_\tau^+ \) and \( R_\tau^- \), respectively. We will suppose, for now, that they have been chosen from the data to be \( \overline{\text{MBM}}^+ (\tau) \) and \( \overline{\text{MBM}}^- (\tau) \). Appendix A.1 discusses a cursory data-driven rule for their choice.

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

1. Form the ‘regression’ data set \((X_i, Y_i), 1 \leq i \leq n \) \((n = n^+ + n^-)\), for input into PRIM. The \( X_i \)'s are the \( X_i^+ \) and \( X_i^- \) data pooled together. The \( Y_i \in \{-1, 1\} \) are indicators of whether \( X_i \) is from the \( - \) or \( + \) sample.
2. Feed the \((X_i, Y_i)\) data into PRIM to obtain \( K_{\text{max}} \) (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: \( P_1^+ = B_1; P_k^+ = B_k - \bigcup_{j=1}^{k-1} B_j, \) \( 2 \leq k \leq K^+ \) where \( K^+ = \max\{k : \text{ave}(Y; P_k^+) \geq \overline{\text{MBM}}^+ (\tau)\} \) and \( \text{ave}(Y; P_k^+) \) is the average of the \( Y_i \)'s in \( P_k^+ \).
3. Repeat Step 2, but with \((X_i, -Y_i)\), fed into PRIM to obtain \( P_1^- , \ldots, P_{K^-}^- \), with \( K^- = \max\{k : \text{ave}(Y; P_k^-) \leq \overline{\text{MBM}}^- (\tau)\} \).
4. Check that there is no overlap between the \( P_k^+ \) and \( P_k^- \) (for differing densities this is unlikely for reasonable choices of \( \overline{\text{MBM}}^+ (\tau) \) and \( \overline{\text{MBM}}^- (\tau) \)). If there is some overlap then increase \( \overline{\text{MBM}}^+ (\tau) \) and decrease \( \overline{\text{MBM}}^- (\tau) \) until the \( P_k^+ \) and \( P_k^- \) are disjoint. If the boxes become null then the estimate of \( R_\tau \) is null.
5. Form the following partition of \( \mathbb{R}^d \): \( P_k = P_k^+, 1 \leq k \leq K^+ \); \( P_{K^++k} = P_k^-, 1 \leq k \leq K^- \). \( P_{K+1} = \mathbb{R}^d - \bigcup_{j=1}^{K^+} P_j \) where \( K = K^+ + K^- \).
6. Obtain the counts among the \( X_i^+ \) and \( X_i^- \) 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 \( \alpha \) 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 \( \hat{R}_\tau^+ \) 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 \( \hat{p}_k^- < \hat{p}_k^+ \).
A.1. Data-driven choice of Minimum Box Mean

1. Obtain pilot estimates of $R^+_1$ and $R^-_r$ based on fitting normal mixture densities to the $X^+_i$ and $X^-_r$. Let these pilot estimates be denoted by $R^+_{\text{pilot}}$ and $R^-_{\text{pilot}}$. Details on the normal mixture fitting are deferred to Appendix A.2.

2. Apply PRIM to the $(X_i, Y_i)$ with $\text{MBM} = \Upsilon$ to give a sequence $P^+_1, \ldots, P^+_K$, with minimum box means $\text{ave}(Y; P^+_g), g = 1, \ldots, K^+_0$.

3. Apply PRIM to the $(X_i, Y_i)$ $K^+_0$ times each with $\text{MBM}$ equal to $\text{ave}(Y; P^+_g)$ to obtain estimates $\hat{R}^+_{\tau,g}$ for $1 \leq g \leq K^+_0$.

4. Take $\text{MBM}^+(\tau) = \arg\min_{1 \leq g \leq K^+_0} \text{err}(\hat{R}^+_{\tau,g}, R^+_{\text{pilot}})$.

5. Repeat Steps 2.-4.; but using the $(X_i, -Y_i)$ 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 $\text{MBM}^-_\tau = \arg\min_{1 \leq g \leq K^-_0} \text{err}(\hat{R}^-_{\tau,g}, R^-_{\text{pilot}})$.

A.2 Details of Normal Mixture Fitting

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

- A continuous sample in $\mathbb{R}^d$.
- $k_{\text{max}}$: 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 $i$th cluster ($1 \leq i \leq k$) and $c_{ij} = c_{ij}(k, \ell)$ be the $j$th point in the $i$th cluster ($1 \leq j \leq n_i; 1 \leq i \leq k$). Compute the within-cluster variabilities

\[
W_\ell(k) = \frac{1}{k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} ||c_{ij} - \bar{c}_i||^2
\]

where $\bar{c}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} c_{ij}$ 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 $B(k) = \sum_{i=1}^{k} ||c_i - \bar{c}_i||^2$ with $\bar{c}_\tau = \sum_{i=1}^{k} \bar{c}_i / k$.

2. Choose the number of clusters $k^* = \min\{k_{\text{CH}}, k_{\text{KL}}, k_{\text{WV}}\}$ where

\[
k_{\text{CH}} = \arg\max_{1 \leq k \leq k_{\text{max}}} \text{CH}(k), \quad k_{\text{KL}} = \arg\max_{1 \leq k \leq k_{\text{max}}} \text{KL}(k) \quad \text{and} \quad k_{\text{WV}} = \max\{k : W(k) > 1.2\},
\]

\[
\text{CH}(k) = \frac{B(k)/(k-1)}{W(k)/(n-k)}, \quad \text{KL}(k) = \frac{D(k)}{D(k+1)}, \quad \text{WV}(k) = \frac{W(k)}{W(k+1)}
\]

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 $W(k)$ to Bumgarner (2007). This results in a partition of the data of size $k^*$. 

16
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 $\Sigma$, $\text{vech}(\Sigma)$ is the vector of entries of $\Sigma$ on and below the diagonal; stacked in order from left to right.

<table>
<thead>
<tr>
<th></th>
<th>$w_1$</th>
<th>$\mu_1$</th>
<th>$\text{vech}(\Sigma_1)$</th>
<th>$w_2$</th>
<th>$\mu_2$</th>
<th>$\text{vech}(\Sigma_2)$</th>
<th>$w_3$</th>
<th>$\mu_3$</th>
<th>$\text{vech}(\Sigma_3)$</th>
<th>$w_4$</th>
<th>$\mu_4$</th>
<th>$\text{vech}(\Sigma_4)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f^+$</td>
<td>0.08</td>
<td>420</td>
<td>6609</td>
<td>0.31</td>
<td>194</td>
<td>888</td>
<td>0.36</td>
<td>223</td>
<td>2244</td>
<td>0.25</td>
<td>203</td>
<td>2003</td>
</tr>
<tr>
<td></td>
<td>367</td>
<td>1312</td>
<td>327</td>
<td>334</td>
<td>808</td>
<td>337</td>
<td>433</td>
<td>122</td>
<td>248</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>122</td>
<td>301</td>
<td>-223</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>303</td>
<td>652</td>
<td>124</td>
<td>447</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>222</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>292</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>179</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>626</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$f^-$</td>
<td>0.17</td>
<td>193</td>
<td>671</td>
<td>0.33</td>
<td>189</td>
<td>605</td>
<td>0.27</td>
<td>183</td>
<td>1001</td>
<td>0.23</td>
<td>300</td>
<td>6658</td>
</tr>
<tr>
<td></td>
<td>270</td>
<td>226</td>
<td>605</td>
<td>179</td>
<td>273</td>
<td>277</td>
<td>343</td>
<td>1241</td>
<td>2216</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>319</td>
<td>80</td>
<td>277</td>
<td>163</td>
<td>429</td>
<td>256</td>
<td>144</td>
<td>1264</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>316</td>
<td>36</td>
<td>273</td>
<td>116</td>
<td>118</td>
<td>300</td>
<td>1701</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>116</td>
<td>228</td>
<td>118</td>
<td>158</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>321</td>
<td></td>
<td></td>
<td></td>
<td>676</td>
<td>623</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>72</td>
<td></td>
<td></td>
<td></td>
<td>371</td>
<td>696</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td>91</td>
<td>330</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>268</td>
<td></td>
<td></td>
<td></td>
<td>112</td>
<td>431</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>621</td>
<td></td>
<td></td>
<td></td>
<td>2669</td>
<td>2477</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>52</td>
<td></td>
<td></td>
<td></td>
<td>62</td>
<td>1054</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>123</td>
<td></td>
<td></td>
<td></td>
<td>82</td>
<td>1354</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1340</td>
<td></td>
<td></td>
<td></td>
<td>2167</td>
<td>2731</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>279</td>
<td></td>
<td></td>
<td></td>
<td>-440</td>
<td>495</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3503</td>
<td></td>
<td></td>
<td></td>
<td>4131</td>
<td>2460</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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


