HARNESSING HETEROGENEITY TO IMPROVE PATIENT OUTCOMES

Jonathan Hibbard

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics.

Chapel Hill
2017

Approved by:

Sonia Davis

Scott Hultman

Michael Kosorok

Yufeng Liu

Donglin Zeng

ABSTRACT

Jonathan Hibbard: Harnessing Heterogeneity To Improve Patient Outcomes
(Under the direction of Michael Kosorok)

We investigate methods of improving medical outcomes through exploiting heterogeneity, with focus on actual implementation.

Advances in data-mining and big-data methods have allowed new and exciting opportunities to alter the precise nature of statistical medical research. Whereas traditional science experimentation has attempted to eliminate causes of variability beyond a small set of variables of interest to be investigated, machine-learning techniques to extract weak and complex signals from noisy data now allow handling of heterogeneous experiments and subjects.

We propose that viewed through the lens of these modern machine-learning methods, heterogeneous and highly-variable data should be regarded as a boon not a nuisance. In particular such data allows for the investigation and construction of individualized treatment rules for patients, that is for the advance of precision medicine.

Two facets of this view are especially explored. Firstly the practical design and implementation of appropriate data collection experiments allowing for a machine-learning approach, whilst simultaneously permitting a traditional experimental view in order to satisfy investigators from both paradigms. We reference a particular example, the design for a clinical trial investigating the optimal treatment of burns patients (the LIBERTI trial).

This example highlights some particular challenges, statistical, philosophical and logistical, and hopefully some corresponding solutions, that arise when bridging traditional and modern paradigms. Whilst we present our design as an initial solution,

from the attempted implementation of this trial we discover, and then explore, particular aspects that are apt for further improvement.

Secondly we investigate methods to combine and make effective traditional clustering techniques in higher dimensional data with weak signals, where existing techniques may fail. Motivated by an example of COPD sufferers data (the SPIROMICS study), we attempt to develop ways combining more traditional methods with a machine-learning approach, and more fuzzy data-mining methods, with ones permitting better inference.

We illustrate our methods on Fisher's Iris data, and the Wisconsin Breast Cancer data set. We explore extensions of the traditional Gaussian mixture model to more general log-concave distributions and highlight what should be interesting theory for such approximations.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION.

Heterogeneity has long been the enemy of statisticians, and indeed much of statistics historically has been focused on accounting for and removing this heterogeneity. For a specific example, in clinical trials often a homogeneous population as possible has been used to reduce patient variance, and ensure a clean signal can be extracted with good inference.

However this approach is suboptimal, for numerous reasons. In particular, firstly for generalizability. The conclusions of such an experiment can only be confidently generalized to a similar population, and it is hypothesized that this has resulted in some of the failures of recent supposed medical advances to be replicated in further experiments. Secondly often such investigations and trials focus on the majority subpopulations such, as white males, for these being the largest subpopulations are both easiest to recruit and furthermore provide the largest market when there are commercial concerns. This leaves minority subpopulations unfairly underrepresented with respect to research and medical advances. Thirdly, and perhaps we might argue most crucially, it is a missed opportunity to exploit patient variability to tweak and thence optimize treatment for any given patient.

It is widely reported that we are entering an era of precision medicine. According to the National Institutes of Health, precision medicine 'an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person.' (NIH (2016)). To crudely paraphrase, people are highly heterogeneous, therefore treatments appropriate to each

one should be heterogeneous too. Clearly doctors have known and appreciated this since the modern era, if not before as Katsnelson (2013) points out. For instance that a febrile infant demands different considerations than a febrile geriatric provides a trite example; but through the science of collecting and providing much data on each patient, not just simple demographics, but also genetic and environmental factors amongst others, and then through the available computing power to make use of this big data, there is the current possibility to tailor to each patient a treatment truly deserving of the word 'precision'.

Whilst data collection abilities and computing power have exploded, statistics, including biostatistics, has not entirely kept pace, at least in the same direction. This has resulted in new terminology such as 'data science' to nomenclate the development and application of novel, nontraditional algorithms and techniques to cope with examining big data, and extracting noisy, complex and sparse signals. While there is much that is laudable in the new realm of 'data science', in particular the drawing together of ideas from computer science, pure mathematics, applied mathematics, physics, and statistics among other fields, the liberation from the more rigid traditional statistics viewpoint also brings disadvantages.

A prominent warning as to the misuse of big data is the reported problems with the Google Flu Trends project to predict epidemic outbreaks in the United States (Lazer et al. (2014)). Lack of sufficient modeling assumptions and most saliently the misapprehension of bias in the data sample, are reputed to have led to some major errors. This decidedly showed simply using bigger data does not always give a better result, and indeed Google retired the project. However, further research has reported that the addition of more modelling and statistical inference methods into such framework can reap real rewards (Schumacher et al. (2015), Preis and Moat (2014)).

Dealing with bias, modelling assumptions and inferences is the home territory of the biostatistician, and hence biostatistics is in prime position to propel precision medicine into a successful new era, particularly by fusing this familiarity with both the use and development of powerful data mining methods. In this script, we shall describe our own, rather minuscule, efforts to contribute to this fusion.

In the first chapter of this document, we define and focus on a particular type of precision medicine, the adaptive, or dynamic, treatment regime (DTR), and advances in calculating these. In particular we shall comment on the role of the Sequential Multiple Randomized Trial (SMART). We do so through presenting a particular example of a SMART, that we have designed and of which we have begun implementation. The design has the interesting features of being able to be regarded either from the modern precision medicine viewpoint, amenable to develop precision treatment rules, or from a more traditional perspective suited to making classic inference. We hope that this design, the first of its type (to our knowledge) outside of mental health, might serve as a template for launching other SMARTs.

For the second chapter, we shift our focus to a broad and more general type of data mining, namely clustering. One might argue that the goal of clustering is to strike a balance between parsimoniously describing the data, whilst recognizing the presence of appreciable heterogeneity, which is to say we aim to group the data into a relatively small number of groups, or clusters, such that the objects within each group are appreciably more heterogeneous than objects between two different clusters. This is a well researched and vast field, with little consensus on the optimal methods, or indeed precise definitions, and we explore some of the issues through the difficult analysis of high dimensional data on COPD patients, with the aim of extracting COPD subtypes. We propose a viewpoint to allow reasonable conclusions as to existence of subspecies/subtypes in a dataset, and illuminate the ideas through

application to Fisher's Iris data. We then examine what these methods tell us when applied to data collected from the SPIROMIC Study, which provides a large amount of data from COPD patients.

The third chapter examines extensions of our method, and explores the result of these, in particular the use of certain nonparametric mixture models. We illustrate this on a classical dataset, the Wisconsin Breast Cancer dataset. We tie our results theoretically to other recent developments in this area, and suggest that, at least in some instances, our methods offer, in some ways, benefits over competing ideas.

The concern of the fourth chapter is possible directions for further work on these themes. In particular we discuss limitations we found on implementing the trial design presented in the second chapter. We discovered that logistics, insurance issues, and patient perceptions caused implementation of the design to be problematic in this particular setting. Hence we briefly discuss new proposals, involving judicious and adaptive randomization, that we hope will result in both circumnavigating these issues and furthermore actually improve the statistical power of our machine learning techniques to calculate optimal decision rules. We also consider further how to continue the work we present on data mining the SPIROMIC Study, and ways to move it more fully into the realm of machine learning. Finally we outline intended further work on density approximation, in particular with regard to techniques which might be of use in nonparametric mixture models.

CHAPTER 2: LIBERTI, A MULTIPURPOSE SMART STUDY.

2.1 Summary.

Background/Aims:

Laser treatment of burns scars is considered by some to be standard of care, despite little evidence-based research. Evaluating the efficacy is made difficult by substantial heterogeneity in patient response, possible delayed effects from the laser treatment, a large number of treatment options/settings, and treatments can be provided in conjunction. We design a trial capable of coping with these issues that may be viewed as a classic randomized clinical trial, ensuring standard interpretations, but also viewed from a more modern paradigm to give extra insight regarding optimized treatment algorithms and precision medicine. It will be the first randomized trial to compare the effectiveness of laser treatments for burns scars.

Methods:

We propose using the Sequential Multiple Assignment Randomized Trial (SMART) framework to investigate the effect of various permutations of laser treatment on hypertrophic burn scar amelioration. We also examine the resulting trial design from a classical viewpoint, to provide sample size calculations, and assurance of good power for specified treatment group comparisons under a traditional Randomized Clinical Trial (RCT) viewpoint.

Results:

We demonstrate that viewed classically the trial has power to produce determinations of laser treatment effect. We also appeal to recent methodological research to predict, and describe, the many benefits of the new SMART paradigm coupled with machine learning.

Conclusions:

We show that some trials may be effectively viewed both as a SMART and a RCT simultaneously, allowing the enhanced interpretations and power of the former, while ensuring the well understood framework and analyses of the latter. We believe that SMARTs designs should be commonly used, but until their benefits have been frequently demonstrated, we suggest a chimera trial, such as proposed here, which allows use of the SMART paradigm with confidence that classic results may still be attained. Further, this is, to our knowledge, the first use of a SMART in surgery, and possibly the first application of AI techniques to SMARTs outside of psychiatry, and evinces the potential benefits of SMARTs throughout all fields of medicine.

2.2 Introduction and Background.

Overview:

The Laser Induced BioEngineered Remodeling of Thermally Injured skin (LIBERTI) Trial is a randomized clinical trial in burn scar surgery, which combines both conventional and precision medicine paradigms. The trial will investigate and evaluate the efficacy of certain sequences of laser treatment of hypertrophic burn scars, and is designed to both address the salient clinical questions in a standard fashion, while also allowing the calculation of precision medicine, patient specific, treatment

rules which optimize individual patient outcome. To achieve this dual purpose both standard Randomized Clinical Trial (RCT) and Sequential Multiple Assignment Randomized Trial (SMART) paradigms are employed.

Both RCT and SMART approaches are of interest, as they allow the reasonably powered comparison of a fairly large number of disparate treatment options, as well as the rudimentary mathematical estimation of the effect of no treatment in a trial where assignment to no treatment is impossible practically, yet should be considered.

The SMART approach is of further interest in this scenario, as it can make use of very appreciable patient heterogeneity and is intended to provide an individualized Dynamic Treatment Rule (DTR) that might allow patient specific precision medicine.

Indeed the SMART approach is of more interest still by virtue of being the first such trial (to our knowledge) in surgery certainly, and possibly the first to employ machine learning methods outside of psychiatry. As such, certain problems inherent in the implementation of SMARTs in a new area are identified and overcome.

Burn Injury:

Hypertrophic burn scars are the cause of significant ongoing morbidity in burn victims (Finnerty et al. (2016), Hultman et al. (2013)). Physical sequelae include itching, pain, stiffness and contracture (Finnerty et al. (2016), Hultman et al. (2013)), while psychological sequelae include depression, post traumatic stress disorder and great social anxiety (Arno et al. (2014), Bayat et al. (2003), Brown et al. (2008)). Up to 70% of burns victims develop such scars but, of yet, optimal therapy combinations, treatment timings, and indication remain unknown (Finnerty et al. (2016)). Further clinical trials are needed to address the efficacy of currently used treatments, in addition to elucidating the little known molecular pathways instrumental in the formation of these scars (Finnerty et al. (2016), Friedstat and Hultman (2014), Porter

et al. (2016)).

There are two laser treatments that are hypothesized to ameliorate hypertrophic scars: vascular-specific Pulsed-Dye Laser (PDL), and ablative fractional CO2 laser (CO2) (Hultman et al. (2013), Hultman et al. (2012)). These two types of therapies operate through completely different mechanisms. PDL essentially vaporizes capillaries by targeting haemoglobin thereby reducing vascularity. The CO2 laser targets water and essentially burns away multiple small vertical columns into the scar, promoting collagen formation and scar remodelling (Hultman et al. (2013), Hultman et al. (2012)). Some plastic surgeons familiar with these therapies consider them effective for remodelling burn scars, but the evidence for this is as yet unclear, a major gap in the research (Friedstat and Hultman (2014)). There is a considerable lack of large scale randomized trials regarding laser treatments.

Determining the effectiveness of a laser treatment is problematic for various reasons (Friedstat and Hultman (2014)). A major one of these is the heterogeneity across patients and types of burn scars. The burns mechanism could be an important indicator of outcome, with say electrical burns being fundamentally different from flame burns (Duke et al. (2012), Kidd et al. (2007)). The amount of melatonin in the skin is a factor in response to laser treatment, with for example hypopigmentation being an issue for darker skin types (Fontana et al. (2013)). Even genetic factors, particularly regarding the inflammatory response, (Barber et al. (2006), Schwacha et al. (2005)), will play a part in scar formation and hence treatment.

A further problem is that laser treatment is usually provided in combination with other treatments, such as standard medical care (MED) involving non surgical interventions such as silicone gels, compression garments and medical ointments, or indeed within a sequence of laser treatments of more than one type, (Bloemen et al. (2009), Kerwin et al. (2014)), causing the isolation of the effects of specific, individual

treatments to be problematic.

SMART Studies:

The classic gold-standard in evidence based medicine is the randomized clinical trial (RCT) (Bothwell et al. (2016)). Judicious use of randomization removes unmeasured confounders and permits unbiased estimation and precise statistical inference. However the classic paradigm of falsifying one simple hypothesis is no longer as appropriate in the age of precision medicine, big data and artificial intelligence (AI), when rather than have a few treatment options to choose between the opportunity exists to tailor treatment to the individual (Dawson and Lavori (2003)). As such the clinical statistician's task becomes less about comparing fixed treatments, and more about data mining for possibly complex individualized treatment rules (ITRs) (Murdoch and Detsky (2013), Alyass et al. (2015)).

Classical clinical trials may focus on a homogeneous subgroup (Akli et al. (2015)), in order to reduce the outcome's variance, and increase the statistical power. This is not such an appropriate paradigm when the subjects are naturally extremely heterogeneous, nor especially when an aim is to investigate ITRs that precisely depend on this heterogeneity (Collins and Varmus (2015), Hayward et al. (2006)).

When the goal of main interest is evaluating, or discovering, a dynamic treatment regime (DTR), that is optimized for an individual patient, classical clinical trials also fall short, as a possibly large number of treatment sequences must be compared, arising from permutations of different possible treatments at subsequent timepoints (Kidwell (2014)).

Moving away from the classic paradigm to address these issues and towards a modern paradigm of data mining and constructing DTRs, Murphy (Murphy (2005)) introduced the nomenclature of a Sequential Multiple Assignment Randomized Trial

(SMART). In a SMART, participants are randomized to different treatments at subsequent timepoints, with the randomization options and possibilities based on tailoring covariates, particularly previous treatments, and a patient's response to them (Almirall et al. (2014)).

Mechanically, a SMART might be viewed simply as a subclass of RCT, especially if the randomization probabilities remain constant between all arms at all timepoints. However, the aim, of looking for a personalized DTR, welcoming heterogeneity, and particularly the emphasis on use of newly developed, machine learning based analytical apparatus, as opposed to classical techniques, provides a clear philosophical separation between SMARTs and RCTs and their utilities (Liu et al. (2014)).

While extensive theoretical research into SMARTs blossoms (Chakraborty and Moodie (2013), Kosorok and Moodie (2015)), their practical use has so far been limited, and restricted mostly to psychiatry, despite expected benefits of wide application throughout the medical field (Zhao et al. (2009)). This is not of course unreasonable. While a SMART offers a promise of powerful analysis, its unfamiliar nature might rightly make clinicians, reviewers, IRBs and funding bodies uncomfortable or unwilling to approve such designs (Murphy (2005) ,Almirall et al. (2014)). This is a particular problem due to the fact that precise statistical inference remains very difficult (Laber et al. (2014)). This might even result in querying the ethics of a SMART, as a patient should not be in a trial unless it is well planned to have a feasible chance of success.

Authors have answered this issue by, among other ways, suggesting regarding SMARTs as simply pilot trials for feasibility (Almirall et al. (2012)), or else embedding a simple aim in the SMART such as a classic RCT comparison of first stage treatments (Murphy (2005)), allowing a sample size calculation to be performed for this secondary aim.We present a design that extends this idea, in that we propose

a SMART and a complimentary classical analysis that is capable of powerfully addressing the main research question of interest: the simultaneous comparison amongst multiple DTRs (albeit not individualized DTRs). We further propose a simple but innovative halfway analysis between the classic RCT and modern SMART paradigm that is well powered to find *better* and *worse* treatments, if not *the* best and worst treatments.

2.3 Methods/Trial Design.

Aims/Hypotheses:

This study concerns the evaluation of effects of two different types of laser treatment, (PDL and CO2), on hypertrophic burn scars in comparison to traditional non-surgical medical therapy (MED). There are multiple aims.

1: Address the gap in current knowledge regarding clinical evidence for the efficacy of laser treatment of hypertrophic burn scars.

2: Address the lack of knowledge as to the optimal timing, type and order of laser treatment.

3: Develop a personalized approach to the treatment of scars, acknowledging that each participant could possibly respond to treatment significantly differently depending on individual participant variables such as, for example, scar severity, skin pigment and burn type.

Addressing these aims will be accomplished by testing hypotheses as appropriate and also by using machine learning techniques to indicate the optimal sequence. Our main outcome will be the Vancouver Scar Score (VSS) well-validated quantitative burn scar assessment tool Fearmonti et al. (2010)). Specifically:

1: We will test the following hypothesis regarding the short-term change in VSS between entry to the study (measured at Visit 1) and the end of the first four-month treatment block (measured at Visit 2):

$H_{0\alpha}$: No significant difference in VSS change between PDL and MED.

$H_{0\beta}$: No significant difference in VSS change between CO2 and MED.

$H_{0\gamma}$: No significant difference in VSS change between PDL and CO2.

2: After 24 months of the study, we will determine the long-term effect of treatment on change in VSS (between Visits 1 and 5) by testing the following hypotheses:

H_{01} : No significant difference in VSS change between laser and medical treatment.

H_{02} : No significant difference in VSS change between CO2 and PDL.

We will also construct a confidence interval for the effect of a 'better' treatment combination over a 'worse' one.

3: We will analyze whether there is an optimal treatment sequence using Q-learning (Moodie et al. (2014), Lu et al. (2011), Zhang et al. (2012)) and Outcome Weighted Learning (OWL) (Zhao et al. (2012), Zhao et al. (2015a)), and investigate the difference between estimated participant specific treatment optimal outcomes with a general (that is patient non-specific) optimal treatment outcome.

Interventions/Measurements:

The study will dovetail with current treatment practice at the University of North Carolina (UNC) Burn Reconstruction & Aesthetic Center. Following initial intensive care and surgery, patients generally have a consultation with a plastic surgeon at the center. Typically one or two laser treatment cycles of either CO2 or PDL, as well

as concurrent MED, are offered over a period up to a year, with the combination choice being driven by the surgeon's (unevidenced) intuition. Each treatment block is four months long, during which time each patient receives multiple sessions of laser treatment. Only one of CO2 or PDL is offered during any specific four-month treatment block.

Our study design will offer all patients two four-month blocks of laser treatment. Initial patient consultation suggested that patients would be very unwilling to enter a trial if there was some chance of not receiving the maximum laser treatment they might otherwise receive outside of the trial. On presentation, patients will be screened for the study inclusion criteria. These are minimal, and essentially are suitability for laser treatment, and willingness to be in a study. If met, then the patient will be asked by the surgeon to consider consenting to the trial, which involves two four-month blocks of laser treatments over the period of one year, selected randomly, as well as concurrent standard medical care, and a follow-up visit at the two year mark, one year after the end of treatment.

Year 1 of the study is divided into three four-month blocks. Patients will be randomized so that laser treatment, augmented with medical therapy, is given in two of these blocks, and solely medical therapy is given in the other block. Patients will be block randomized equally between all allowed treatment sequences.

The patients will be examined at baseline, and at the end of each four-month block. Further, they will be examined at the end of Year 2, for a total of 5 visits (at 0, 4, 8, 12 and 24 months). The treatment path options, and the timings are shown in Figure 1. At the initial visit, and the subsequent two visits, the patient is randomized to either CO2, MED, or PDL, under the constraints each patient must receive exactly two laser treatments during the first year. There are 12 such treatment paths that a patient could be set on.

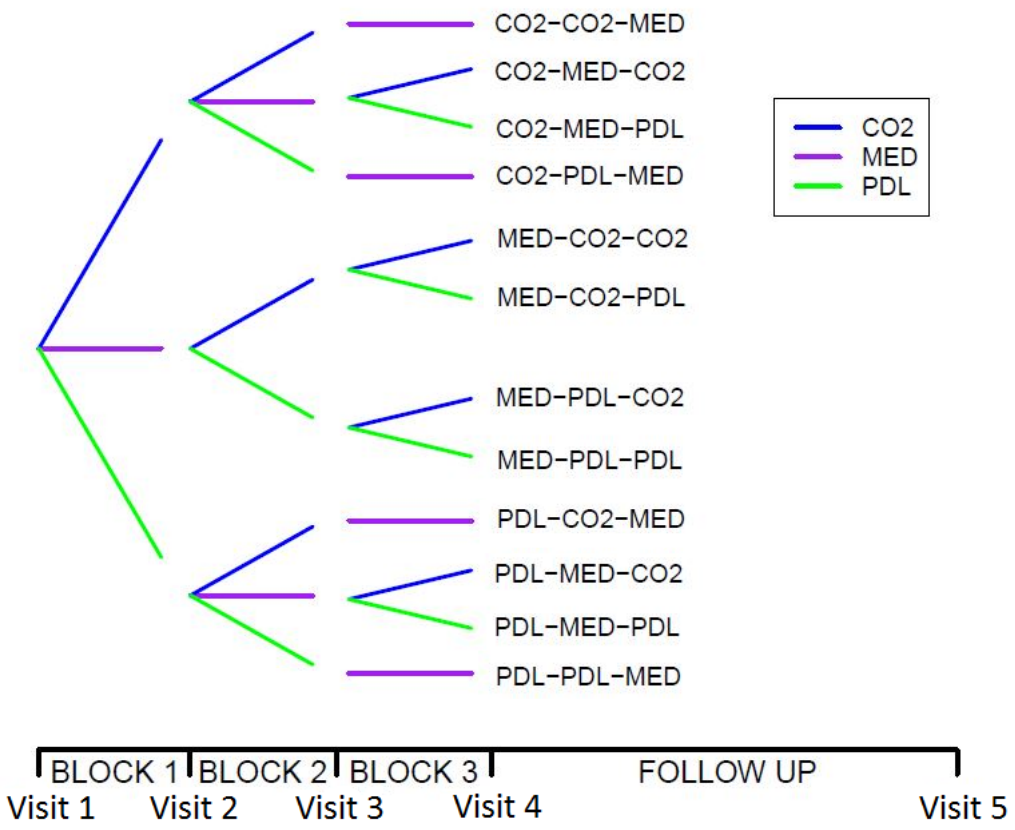

Figure 2.1: Schematic diagram illustrating patient treatment sequences.

In total upwards of 30 variables will be collected from each patient, over 10 of which are time varying and must be recorded at every visit.

The main outcome of interest will be the well validated VSS, which encapsulates many scar issues with one single score that should be ideally minimized Fearmonti et al. (2010). Standard demographics, injury/treatment factors and skin/burn measurements (Zhao and Laber (2014a), Robins (2004)) will be used as controlling variables in the analysis, and more importantly as tailoring variables to determine a DTR. Quality of life measures (Edwards et al. (2007)), and alternative scar scores (Fearmonti et al. (2010)) will be collected as secondary outcomes of interest.

Patients:

The study population will be all patients referred to UNC's burn center that satisfy the inclusion criteria. Extrapolating from previous patient numbers and types, we expect to recruit a fairly heterogeneous population, at an estimated rate of 10 patients a month. We propose a sample size of 180 patients.

Models/Analysis Methods:

We outline our main analyses for the main outcome change in VSS. We have stated three stated aims, which we attempt to fulfill using the following analyses:

1: For evaluation of short term treatment effect, we will test H0, H0 and H0, by comparing, using standard methods, the mean change in VSS from baseline to four months, between the patients receiving MED for the initial treatment block, those receiving CO2, and those receiving PDL. We will make pairwise comparisons using 2-sided t-tests and the Hochberg modification of the Bonferroni multiple comparison procedure, with an overall Type I error rate of 5% (Hochberg (1988)).

2: The comparison of 2-year change in VSS between sequences of three treatment blocks is less standard. The SMART nature of the design means we compare 12 treatment options at once, with possible major heterogeneity. Further we want to compare laser treatments versus no laser treatment, even though all patients will receive some laser treatment. With certain models and assumptions, for example additive treatment effects, this would not be an issue, however experience with laser treatments suggests these may not be realistic assumptions. Instead we propose the following innovative regression model.

We will test H_{01} and H_{02} using a one-sided test with Type I error of 5%. We first test H_{01} and only if this is rejected we test H_{02}. Step-down testing will preserve a

total Type I error rate of 5%.

We use a regression model applied to the treatments effect in each of the blocks. We denote a treatment sequence by a triplet (i, j, k) with each of i, j, k corresponding to choice of treatment in the first, second and third block, respectively, and being set to 0, 1, or 2, to represent MED, CO2, and PDL respectively. We let the change in VSS over the study of any given participant on treatment sequence (i, j, k) be denoted by $z_{i,j,k}$. We model the change in VSS by a normal variable, with a global constant variance, as follows:

$$z_{i,j,k} \sim N(\mu_{i,j,k}, \sigma^2), \quad E[z_{i,j,k}] = \mu_{i,j,k} = \mu + \alpha_{1,i} + \alpha_{2,j} + \alpha_{3,k} + \beta_{i,j} + \gamma_{i,j},$$

subject to the constraints:

$$\Sigma_i \alpha_{x,i} = 0, \quad \Sigma_x \alpha_{x,i} = 0,$$

$$\Sigma \beta_{p,j} = 0, \quad \Sigma_q \beta_{q,i} = 0, \quad \Sigma_a \beta_{a,a} = 0,$$

$$\Sigma_k \gamma_{q,k} = 0, \quad \Sigma_j \gamma_{j,r} = 0, \quad \Sigma_a \gamma_{a,a} = 0.$$

So we model participants on a given treatment path to have random normal responses, with a common variance. We thus assume a mean model which has 11 free parameters, and 28 total variables, that is $\mu, \alpha_{x,p}, \beta_{p,q}, \gamma_{p,q}$ for all $1 \leq x \leq 3$, $0 \leq p \leq 2$, $0 \leq q \leq 2$.

We note that the interpretation of these parameters are intuitive: μ is the grand mean, $\alpha_{x,p}$ represents the first order effect of having treatment p at timepoint x, $\beta_{p,q}$ represents the interactive effect of treatment p at the first timepoint and at treatment q at the second, and $\gamma_{p,q}$ represents the interactive effect of treatment p at the second timepoint and treatment q at the third.

The imposed conditions should force the model to immediately identify a 'better' option for treatment at both a given timepoint, as well as 'better' treatment combinations between first and second treatments and also second and third treatments. These will simply correspond to the signs of the $\alpha_{x,p}$, $\beta_{i,j}$, $\gamma_{j,k}$ respectively.

We note $\mu + \alpha_{x,w}$ corresponds to the expected outcome for participants assigned treatment w at timepoint x and randomized to treatments at other timepoints with equal probability. Other more elaborate interpretations may similarly be given. Thus if, for example, we have an $\alpha_{x,u} - \alpha_{x,0} \geq q$, we have an interpretation that there is a treatment with an average outcome different from the current standard by more than q. Testing the $\alpha_{x,w}$ to determine whether or not the outcome is different for various treatments is then reasonable.

Under the null hypothesis H_{01}, the model is true, and all parameters except μ are 0, in particular all the $\alpha_{x,i}$ must be 0. Further if there is no difference between the effect of CO2 and PDL, the standard regression estimators must give that $\alpha_{x,1} = \alpha_{x,2}$ for each timepoint x. Testing these relationships is then a valid test of H_{01} and H_{02}.

For testing H_{01}, we will use an $F(2, 157)$-test to determine whether $\alpha_{1,0} = \alpha_{2,0} = \alpha_{3,0} = 0$, and for testing H_{02}, use an $F(2, 157)$-test to determine whether $\alpha_{1,1} - \alpha_{1,2} = \alpha_{2,1} - \alpha_{2,2} = \alpha_{3,1} - \alpha_{3,2} = 0$. We will allow for a 5% type I error.

The predicted effects of unobserved treatments, as calculated by the model, will depend most certainly on our model assumptions. Nevertheless we hope these predicted values might provide insight and guide future study.

The model we propose seems clinically reasonable; for example it contains the 'usual' regression model of nontemporal effects and two way interactions between first and second treatments and also between second and third. It also seems statistically reasonable; for example we have 11 free parameters for 12 distinct observation points.

This would seem to be as flexible as we could hope, yet still have some hope that the model is robust and can extrapolate somewhat. Furthermore, it appears to have very decent power for rejecting our proposed hypotheses, despite having comparatively large degrees of freedom, as is needed to realistically fit data at 12 points.

We will attempt to answer the question of laser effect difference with a non-model based approach as well. To do this we will empirically estimate the mean population outcome for each treatment, and rely on standard asymptotic theory to assume that these estimates are approximately normally distributed with differing means $\mu_{\mathbf{i}} = \mu_{i,j,k}$ a constant variance of $\sigma^2_{\mathbf{i}} = \sigma^2$ for all treatment sequences. Instead of trying to identify the absolute best and worse treatment sequences, we simply search for a 'better' and 'worse' one. Our estimator for this is constructed by ordering the 12 treatment sequences as $\mathbf{i}_1, \ldots, \mathbf{i}_{12}$ corresponding to decreasing observed average outcome. We take the sequence, $\mathbf{i}_1$ with highest observed average as the 'better' sequence, and the sequence, $\mathbf{i_{12}}$ with lowest observed average as the 'worse' sequence, provided this difference of averages is too large to be observed by chance if $\mathbf{i}_1$ indeed is not better than $\mathbf{i}_{12}$ overall, more precisely provided $m_{\mathbf{i}_1} - m_{\mathbf{i}_12} > 1.15s$, with s being an estimate of σ. The chance of a type I error of a false conclusion is then about 10%, yet we have good discriminatory power to indeed find a better and worse sequence. Using a step-down procedure to preserve the error rate, we may continue, adding a second 'better' treatment if $m_{\mathbf{i}_2} - m_{\mathbf{i}_12} > 1.15s$, then a second worse if $m_{\mathbf{i}_2} - m_{\mathbf{i}_11} > 1.15s$, to get a set of better treatments $\mathbf{i}_1, \mathbf{i}_2, \ldots, \mathbf{i}_m$ compared with a set of worse treatments $\mathbf{i}_n, \mathbf{i}_{11}, \ldots, \mathbf{i}_{12}$ and so on until either the appropriate statistic says we cannot include another treatment in a particular group, or else all treatments are divvied up between the 'better' and 'worse' groups.

3: We then proceed according to aim 3, to analyze the data using the recently developed Q-learning framework (Schulte et al. (2014)), which attempts to find the

optimal treatment regime tailored to an individual based on their individual covariates and response to already given treatments, through regressing and essentially theoretically reassigning, recursively, and simultaneous outcome weighted learning (SOWL) framework (Zhao et al. (2015b)), which attempts to estimate the DTR directly without calculating theoretical reassignments. Q-learning is extremely powerful if each regression step is well modelled (Zhao and Laber (2014b)), whereas SOWL has excellence performance for higher dimensional covariate space (Zhao et al. (2015b)). We do not specify the precise analysis, for we fully expect significant theoretical developments whilst the study is running, and we will avail ourselves of the best ones available at the study's conclusion. One very promising new method, is the possible combining of Q-learning and OWL, so that the Q-learning is favoured if the associated models are estimated to be correct, but nonparametric OWL is favoured if they are not.

If successful in this, we will have estimated not just the optimal treatment regime for each individual, but also mean outcomes for the population if a population level decision rule was followed instead. Specifically this will allows us, to evaluate the result of varying the timing or order of the laser treatments at the general, non-personalized level, and to compare this with our results from non-machine learning methods.

Size Calculations and Power:

We estimate 180 participants will give us good power to ensure all three of our aims are satisfied.

To account for some expected participant dropout, we calculated power for 162 participants (estimating a worse scenario of 10% dropout of our original 180). We perform different sample size calculations for each of our different three differing aims,

using calculations associated with the methods we have detailed to achieve them. In particular we calculate:

1: We will reject either of $H_{0\alpha}$, $H_{0\beta}$, or $H_{0\gamma}$ with 80% power, at overall size 5%, adjusting for multiple comparisons, if there is an effect size of more than 0.63 standardized units. Pilot study data suggests that we then have an 80% power of observing a minimal clinically significant difference of 0.9 VSS units in treatment effects.

2: The sample size will give good power for rejecting H_{01} or rejecting H_{02}. The precise conditions are slightly difficult to interpret. But we may make the following observations:

a: the power of rejecting H_{01} is more than 80% if choosing MED at a particular timepoint, and assigning random treatment at the other timepoints results in an outcome difference of more than 0.35 standardized units in comparison to randomly assigning treatment at each timepoint.

b: the power of rejecting H_{02} (if H_{01} is rejected) is more than 80%, if choosing CO2 laser therapy at a particular timepoint, and assigning random treatment at the other timepoints results in an outcome difference of more than 0.37 standardized units in comparison to assigning PDL therapy at the given timepoint and assigning random treatment at other timepoints.

More involved power estimates and their corresponding clinical interpretations convince us that we have more than sufficient power.

For the better/worse treatment analysis we may show that we expect to identify a better and worse treatment 90% of the time if the true best and true worse treatment sequences have expected outcomes separated by $\mu_{\mathbf{i}_1} - \mu_{\mathbf{i}_{12}} > 1.15\sigma$. If

there is a laser effect we imagine this would be satisfied. We provide the worse scenario power, and under realistic assumptions, we expect to have 90% power with far weaker assumptions. Corresponding power calculations for longer sequences can also be detailed.

3: The sample size is expected to be enough to enable insights into both general optimal treatment path and a patient-specific optimal treatment path. More specifically we expect to find a superior personalized treatment (within approximately 90% of the true optimal treatment) with probably at least 80%, based on published simulation results (Zhao et al. (2011)). Further while the covariate dimension is fairly large with 30 covariates, the sample size should allow decent performance of the support vector machinery utilized by OWL (Zhao et al. (2015b)).

2.4 Expectations/Discussion.

We have provided a trial design and analysis that will allow comparisons of large number of treatment options in a classical manner without drastic loss of power. We have offered a simple and innovative analysis which allows calculation of sets of better and worst treatments rather than calculating precise treatment effects. We argue this different style of analysis has good power to provide most useful scientific information as to differences in many treatment options. We have proposed a design that while being acceptable and powered for these analyses alone, should allow application of machine learning techniques to mine precision medicine individualized treatment regimes. That is, we have proposed a SMART for a challenging situation, which has assurances of classical utility, even if the novel SMART analyses should fail. We believe that until the use of SMARTs are well understood and practiced, this is an ethical and effective way to implement such trials.

The LIBERTI trial design received approval from both UNC's IRB, and support from grant bodies to allow the trial to start recruiting.

Grants Acknowledged:
NIH grant P01 CA142538,
NIH grant UL1 TR001111.

CHAPTER 3: SURROGATE DRIVEN CLUSTERING.

3.1 Introduction and Background.

Clustering is a name given to the machine learning task of grouping together objects within a set so that any object within each group, or cluster, is more 'similar' to any other object in the same cluster than it is to any object in any other cluster. This task is clearly not well-defined, depending inherently on what 'similarity' means, and this will change with context and with analyst.

Though not well-defined, the idea is however a clearly natural one, and that may be easily intuitively grasped by most. For example, presented with an orange, an apple, a pear, a dog and a cat, one suspects almost everybody would immediately partition these objects in their mind to two subgroups or clusters, fruits and pets. Now this starts to become murkier with further examples, for instance if the dog and cat were ginger coloured, whilst the apple and pear were green, and the orange was, erm, orange, then one could argue that an equally good partition would be into green objects and orange objects. It seems possibly far less natural to cluster into colours than fruits and pets, but at some point this becomes the natural partition, if for example there are many such objects added, all of which are green or orange.

Being so natural, clustering has been used by various nomenclatures, or none, and with differing mathematical and statistical subtleties, for a long period in academic research. For example, its use in psychology goes back to at least 1938, when Zubin (1938) proposed creating clusters of 'likeminded' individuals, in anthropology

to at least 1932, when Driver and Kroeber (1932) proposed subgrouping polynesian cultures, and in sociopolitical theory to at least 1927, when Rice (1927) partitioned members of New Jersey's senate based on year-long voting history. Even at this intial stage, these authors note the computational complexity of the problems they consider, and with modern computing power, it is not difficult to imagine that uses of clustering has only exploded exponentially.

A most comprehensive starting paper on current state-of-the-art in clustering is provided by Jain (2010). The author begins with an observation that even one half-a century after its introduction, and despite thousands of newer competing algorithms, k-means is the prevalent clustering algorithm, and Jain claims this is evidence for both the ill-defined nature and also difficulty of clustering.

A different philosophy is to regard clustering in terms of mixture models, as described for example by Fraley and Raftery (1998) and McLachlan and Peel (2004), who propound modeling the data as a sample from a finite mixtures of distributions from some parametric families, most often some subset of gaussians. The mission then is to identify the parameters corresponding to the separate distributions. The usual approach to doing this is via the EM-algorithm, as explained by Dempster et al. (1977). Fraley and Raftery (1999) give a very efficient R package, Mclust, that assumes a normal mixture model, and estimates the associated parameters.

A related but separate way has been proposed by Huang et al. (2012). They tackle the problem of high dimensional data, and the problem of significance testing in this setting, especially with low sample size, introducing a method known as SigClust. This gives test for deciding if the mixture is better represented as two normal components or one. They then propose a divisive scheme, repeatedly dividing clusters showing evidence of mixing further. This major benefit of this approach is a high powered test despite the high dimensions, although there is some cost that a divisive

algorithm, that is one that proceeds dividing clusters, might not be the optimal way to find a clustering. Once a mistake is made and a cluster partitioned it can not become visible again. Furthermore investigation is needed to understand well effects missing data, for in higher dimensions, if some features have much missingness, a complete case analysis may indeed have very few subjects. Imputation would be a way to overcome this, but the resilience of the method to imputation and model misspecifications is not well studied.

Once the data's distribution is represented as a mixture model, then clustering simply becomes a matter of identifying which distribution a sample comes from. This would seem simple if the distributions have good separation, however when the separation is not complete, it becomes clearly more problematic practically and indeed philosophically.

Walther (2002) runs further with this idea, discussing how the evidence of mixture distributions may be viewed as evidence of subtypes, or subgroups. Walther considers cases when the (univariate) mixture distribution is unimodal but we still have hope of concluding they are in fact different subtypes, referencing blood pressure distributions, and radial star velocities as examples.

A most related issue is then how many subtypes are there? There are many methods to address this question, as summarized in for example by Fraley and Raftery (2002). This is naturally a big issue for unsupervised learning, when the answers given may not be checked against any gold-standard. Our experience in clustering has highlighted for us a separate, but similar issue, namely whether a subtype is a clinically meaningful one. As we shall discuss with respect to our SPIROMICS example below, in the age of big data, not only must we guard against false clusters, but real ones of no significance. For example, many biomarkers exhibit clustering artifacts between genders, or race. If we were clustering Alzheimer disease patients,

a cluster distinguishing males and females might not be of scientific interest. However, a similar example would be finding a cluster of Native Americans amongst study data for Osteogenesis Imperfecta (Brittle Bone disease), yet the study actually managed to interpret this cluster as a very different subtype. Hence determining whether a cluster is a true subtype or not most delve deeper than simple comparisons of clusters. We consider this problem further, and propose what we have termed surrogate driven clustering as a possibility to help tackle this problem.

3.2 Use of a Surrogate.

Goal and Problems:

We consider Fisher's Iris data, consisting of 50 samples of each of 3 species of iris, with petal width and length, and sepal width and length measurements. This is a classic supervised learning data set, but conventional wisdom says that it is unsuitable for unsupervised learning, as either according to some sources, the species cannot be well separated, or, according to other sources, they may separated by quite advanced clustering methods but there is no way to determine how many species there are.

We claim surrogate driven clustering will evidence that there are three subtypes of iris, and cluster them appropriately.

The idea behind surrogate driven clustering is that firstly clusters may not correspond to distinct (relevant) subtypes. For instance clustering testosterone levels in Alzheimers patients would likely give two pretty excellent clusters, but they would not be fantastically interesting. This is especially important to bear in mind when exploring big(gish) data without understanding the underlying biological mechanisms fully (e.g. shotgun clustering as we shall designate it!).

The second saliency inspiring surrogate driven clustering is that clusters may merge together, and need to be separated, yet determining the number of clusters (with or sans clinical relevance) is done by adhoc methods and little rigour, or justification. This is the situation for the iris example. As seen in Figure 3.1, there appear to the eye to be two clusters, yet this would seem to be from species 2 and 3 having distinct clusters that merge into one another, once one inspects the species labels. Indeed it would seem to human inspection that the fact there are three species might not be recoverable from this data, although it seems apparent there are at least 2 subtypes of iris. A method capable of suggesting there were 3 species would be therefore of interest.

What would it take to convince us that one of those two apparent to the eye clusters could actually be composed of two distinct species? Well if there are different species, some variables must differentiate these well (whether we have access to the variables or not). We shall call these variables surrogates (as in that they are surrogates for subtype).

Overview of Methods:

In general we shall be interested in considering biomedical data, where clinical variables are available and somewhat differentiate subtypes. For the iris data, we have no clinical variables, but we can take some of the main variables to be surrogates instead. The science would make this appear attractive, different species of iris are most likely to have different underlying relationships of these variables. And here the choice would appear obvious: petal width generally increases with petal length, yet the rate of increase is likely to be different across species. We could observe the same regarding sepal length and sepal width. Also the correlation of petal length with sepal length is likely to be high, but change over species. And similarly for sepal and petal widths.

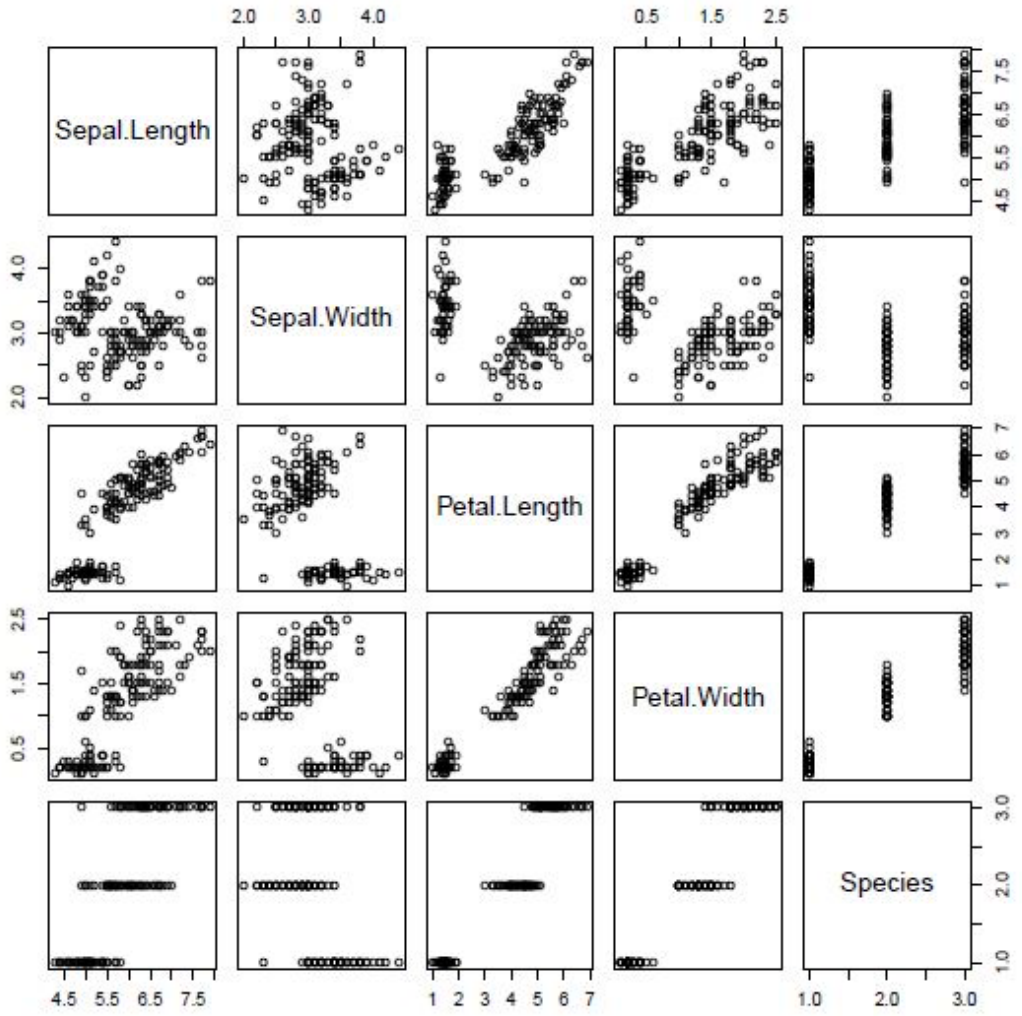

Figure 3.1: Iris dataset

To educe the different relationships across species we thus partition the variables naturally as: Main variables: petal length, sepal width, and Surrogate variables: petal width, and sepal length. And by pairing as uncorrelated variables as possible together, we expect to maximize the ability to cluster.

In summary we postulate if needing to divide variables into Main Variables and Surrogate Variables, variables should be partitioned to maximize in some way correlations between Main Variables and Surrogates to highlight changing relationships for

different subtypes well, and to minimize in some way correlations within the group of Main Variables to provide as much information to guide our clustering as possible.

Once we have Main Variables and Surrogates chosen, we need to decide on a clustering method. Our so called surrogate driven clustering is not a new clustering method itself, but rather a way to both select the correct number of clusters and also verify clinical relevance, for any analyst chosen clustering method.

Our general preferred clustering method is unconstrained normal mixtures. This is a flexible, robust method with superb implementation in R through Mclust. It goes without saying (but we will anyway) that it would be particularly appropriate when the underlying variables, for each (unknown!) subtype have a normal distribution. We would expect that maybe for the variables we have for the iris dataset (such measurement variables are well known to be normally distributed in a species to very good approximation).

K-means might be another general choice of method, however it seems too coarse given we often expect different populations to have different heterogeneity.

Mixture models utilizing t-distributions have been put forward for modelling of gene expression data (McLachlan et al. (2002)). A particular advantage of this is that outliers are much better tolerated, whereas they can make it very difficult fitting a normal mixture. A particular disadvantage is that there is no (we believe) implementation in available routines that approaches the ability, or flexibility, of the R package Mclust in modelling normal mixtures. (Further research to develop an expanded package would likely be most worthwhile).

The data may require transforming to better choose a model. As an example, for biomarkers in the blood, it is observed they often follow a log-normal distribution, thus a logarithmic transform followed by modelling as a normal mixture could be

appropriate.

One may always precheck the appropriateness of a certain clustering algorithm, at least from a mixture modelling paradigm. For example a reasonable check to see (well, technically, as per usual, to falsify) whether some multivariate samples might be from a normal mixture, would be to see if each univariate variable could be reasonably modelled with a normal mixture (following on from the knowledge that normal projections are themselves normal).

We begin by using our chosen clustering algorithm on our dataset. As we, and nobody else, knows (we pretend for the iris dataset) the correct number of clusters, we will use our algorithm to divide the patients (well flowers!) into first 2 clusters, then 3, then 4, and so on to a suitably large number which we believe is bigger than how many subtypes we could possibly be looking for. So far, so usual for an analyst. The crux of our method is that we now evaluate the quality of each of these clusters, not with an internal measure as per unsupervised learning (these simply comment on the cluster algorithms interaction with the data not the intrinsic relevance), nor an external measure as per supervised learning (we are assuming one is not available), but with an intrinsic measured found from using the surrogates.

Figure 3.2 shows how the iris species vary by the two main variables, petal width and sepal length. The dataset we are using is truncated to the nearest 0.1 for each variable, so to highlight multiple flowers with identical truncated measurements, in Figure 4.2 we have randomly altered all measurements by 0.05. The iris species are differentiated by colour, with each of cyan, blue and green corresponding to a different species.

As we had expected from the pairwise plots of Figure 3.1, there is clear evidence of two clusters, which may or may not correspond to differing subspecies. But it would

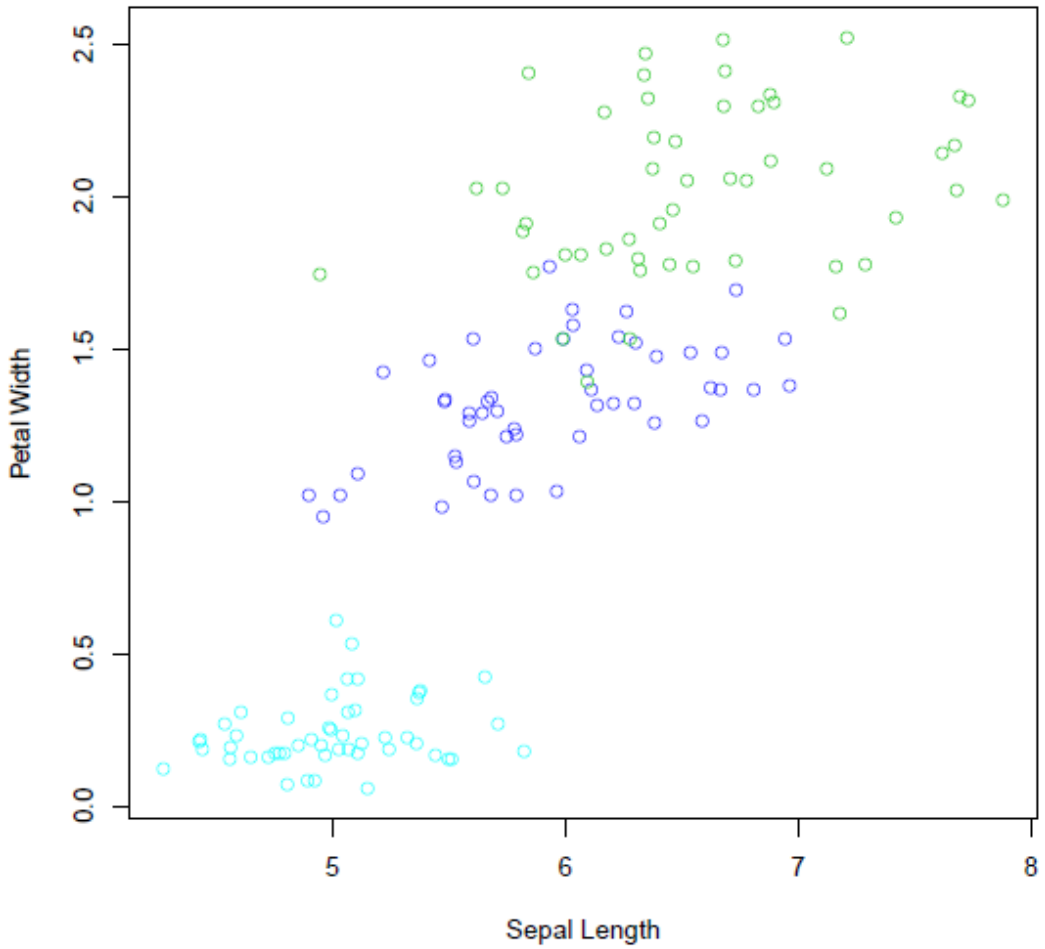

Figure 3.2: Main variables, with species distinguished by colour.

be hard to conclude there could be three subspecies in this data, although knowing the answer we might say that the green species can be identified from having different variance from the blue. We would need good precise evidence though to be confident of this.

Figure 3.3 shows the results of clustering using a normal mixture unconstrained (except for 9 clusters, which requires a more specified structure to fit). The points are, as in Figure 3.2, colored by species, whereas the cluster each point is assigned to

is shown by the numeric label of that point.

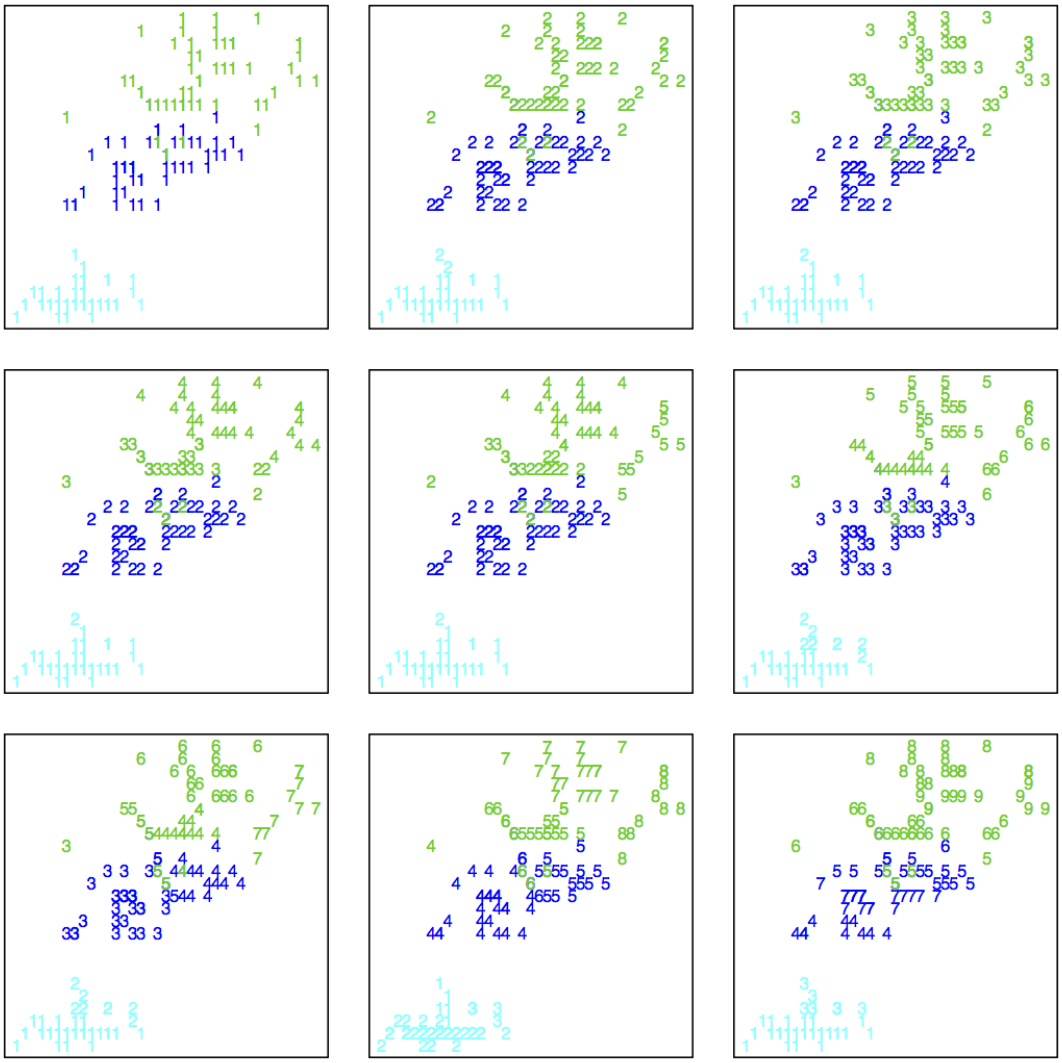

Figure 3.3: Main variables clustered by normal mixtures into 1-9 clusters.

We remark that this cluster algorithm performs excellently, given we know the labels. Clearly a mixture of normally distributed variables is a good model of this data. We also note that the cluster algorithm separates the species excellently when the correct number of clusters is chosen, that is for three clusters, the misclassification rate of assigning species to the wrong cluster is an extremely impressive 4.7%. The cluster algorithm performs very well in this example in that as the number of clusters

is increased beyond the true number of components, the algorithm tends to break up clusters into further subclusters, rather than as might happen provide a very different grouping to the clusterings with less clusters. But we do see that for some choices, such as 7 clusters, the cluster algorithm starts to perform less well, and as an example where the species were very well distinguished when classified into 6 clusters, for 7 clusters, cluster number 4 groups together both the green and the blue species.

A little more precisely, we proceed as follows. Having asked our algorithm for a total of M clusters, then for each of these outputted clusters we may see whether there might be any evidence of multiple subtypes within the cluster.

Of course we need to find a way to evaluate such evidence, and there are many choices. We consider a couple here, and see how they perform in the context of the iris dataset.

The underlying thought behind our choices of evidence, is that, as we hinted at above, for the correct choice of surrogates, the different subtypes will have different relationships between the main variables and the surrogates. There is certainly always an option of surrogate which will elucidate this and it is merely (or majorly more like) a matter of finding it.

As an illustration, consider clustering genetic cancer data. If our method of clustering gives you three clusters, and the survival times of each of the subjects for each cluster appeared to be the same, would one leap to the conclusion that these clusters are clinically relevant subtypes, for shouldn't we expect different subtypes to have different survival times for example, at least to a high resolution? Or would one more likely decide that either these were real clusters but not so clinically relevant (e.g. maybe we had just picked up clusters of ethnicity in the genetic data, which was independent of much cancer relevance), or would one conclude that there was no real

cluster and our algorithm for clustering (as such algorithms often are prone to do) artificially divided up one group (and internal validity methods give some indication as to which of these two options we should choose, but neither option is what we had hoped to find)?

Conversely if one should find that in each produced cluster, a highly different, 'non-overlapping' survival time distribution, then this would surely be worthy of further investigation, and perhaps the existence of discrete subtypes would be the natural working explanation.

It is clear that such evaluations and determinations are not going to be precise and unequivocal answers, just as there is no precise and unequivocal answer to the question, 'does this regression model fit well?'. But it seems that by considering the algorithms we are suggesting, an analyst would be able to make arguments for or against the existence of subtypes, just as an analyst has ways to make arguments for or against the appropriateness of a regression model.

Our first idea applies to a very idealized situation. Assume there are very well defined clusters, that our clustering algorithm will elucidate (if the correct number of clusters is specified). Also assume that for each subtype, there is a different distribution of surrogate. Then we should simply increase the number of clusters we ask the algorithm to produce until we see that no outputted cluster contains a mixture of distributions. Of course this might not get us far, for indeed how do we know if a distribution is best regarded as a mixture or not? One method, we do not investigate here, could be to assume and check that the surrogate variables have unimodal distributions within each cluster. This could be problematic as there is not yet a good way to estimate unimodal distributions, as the maximum likelihood estimator is undefined, although we believe we have a new method to skirt this issue. Moreover though a unimodal distribution can arise as a mix of (unimodal)

distributions, so this is possibly not the most sensitive investigation we may follow. Walther explores further this very question (Walther (2002)).

A little different track though could be to think that if a cluster contains a mix of subtypes, and those subtypes each have a differently distributed surrogate, then we should be able to tell (with enough data) that the distribution changes as we move in the cluster. That is, we may simply check whether the main variables and the surrogate variables are independent within each cluster. If not, we need further divisions. To perform this test, one could use, as we propose, BM Covariance (Székely et al. (2009)), a recently proposed powerful and widely applicable test of independence between variables.

Preliminary Results:

In Table 3.1 we display the results of our procedure applied to the Iris data. More specifically for differing total numbers of clusters $1 \leq m \leq 9$ wecluster the data through use gaussian mixture models (calculated with the EM-algorithm) with m components. For a specified m total number of clusters, we then examine each of the clusters, $1 \leq n \leq m$ in this clustering. For each of the clusters we evaluate the dependence of the surrogate variable restricted to members of this cluster on the spatial position within the cluster, using brownian distance correlation.

As interpretation of these results is somewhat dependent on cluster size, in Table 3.2 we display how many irises are in each corresponding cluster.

We observe that the normal mixture model divides the patients into fairly sizeable groups at each clustering, but that almost none of the clusters, for any total number of clusters, have a p-value which indicates that the surrogate and main variable are independent in that cluster. Simply put we do not see any structure in the data from our analyses.

Total Clusters:	1	2	3	4	5	6	7	8	9
Cluster: 1	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.450	0.025
2		0.005	0.005	0.005	0.005	0.070	0.045	0.005	0.305
3			0.005	0.005	0.065	0.005	0.005	0.805	0.685
4				0.005	0.160	0.025	0.005	0.005	0.005
5					0.160	0.200	0.150	0.005	0.005
6						0.165	0.230	0.150	0.005
7							0.165	0.255	0.040
8								0.105	0.355
9									0.320

Table 3.1: Within-cluster p-values for independence of main, surrogate.

Total Clusters:	1	2	3	4	5	6	7	8	9
Cluster: 1	150	48	49	49	49	40	40	7	19
2		102	53	56	63	10	10	38	12
3			48	19	7	52	29	5	12
4				26	21	17	28	29	12
5					10	21	14	28	13
6						10	19	14	38
7							10	19	17
8								10	18
9									9

Table 3.2: Total number of members within each cluster.

This, on reflection, of course is not surprising. One of our main variables was petal width, one of our surrogates petal length, and almost surely these will have a strong correlation no matter the subspecies. Thus we should not expect to find main and surrogate variable independence in this example.

Such a hiccup is not an unsurmountable objection to the proposed method however. For while we expect that the petal length and width should have high correlation no matter the species, we might perhaps reasonably expect the precise nature of this relationship to change. And therefore the nature of the relationship is what we need to examine within each cluster.

We make a brief pause for two observations. Firstly our methods might be adapted to, instead of finding clusters per se, regress the surrogates onto the main variables. The division of the main variables into clusters should occur to group together either

regions of high density in the main variables, or else very different classes of the main variables. We may keep dividing until the surrogate appears to be approximately constant statistically, to obtain a partition of the main variables, with little change of surrogate in each partition. In some sense, we would have a blobby random forest.

Secondly this has significant implications for standard ways of validation for cluster analysis. Often analysts will find a clustering (often using an internal, subjective, metric to decide how many clusters) and then check the relevance clinically by seeing if some clinical variable, essentially for us the surrogate, is significantly different between clusters. If so, then the cluster is validated both in spatial and clinical terms. The above example shows that this is very erroneous. A simple regression effect between the clustered variables and the clinical one, such as we have in this example with our main and surrogates will automatically result in the analyst falsely declaring that the clustering is valid and clinically relevant. In this example that would have been a conclusion no matter how many clusters were chosen, due to the very high correlation. This issue, and ameliorating it, was actually the motivation for much of our own work.

Enhanced Method:

Returning to our investigation of the iris dataset, we have remarked that we would like to examine the nature of the relationships between the surrogate and main variables, expecting these to change over species. Our question is then precisely how we should do this.

We have a few suggestions to answer this question. An obvious solution would be to require that the within-cluster relationships between the surrogates and the mains is either say monotonic or else unimodal. As earlier this could be difficult to assess in higher dimensions, though we could utilize say the grenander estimator to fit a model

within each cluster in low dimensions, or else model with log-concave distributions. We might also perhaps assess how well simple regressions explains the relationships between the main variables and the surrogates, or whether for instance changepoint models explains them better than simple regression.

However, while it would be worthwhile examining these ideas further, we shall for now move in a different direction. These ideas all involve an assessment of how appropriate our modelling choice is. For example, maybe petal length does grow for each species with petal width, but nonlinearly. With enough samples, a simple regression would then be rejected in each species, as would any posited relationship, unless we happened to guess the precise truth, and our method would not give the correct results that there are just three species.

Instead what we shall choose to focus more on, as a measure of whether the within cluster relationship changes, is whether the *residuals*, having regressed the surrogate onto the main variables, are independent of the main variables. This measure focuses far less on the total model accuracy, and much more on whether it is equally accurate, or otherwise, over a whole postulated cluster. We hypothesize that this then will be able to highlight precisely any significant changes in relationship with clusters as we wish to find. In particular we believe that examining the residuals for change over each cluster will exploit situations where the heterogeneity of the surrogate differs over subpopulations, not merely the average level.

We proceed with analyzing the iris data in this way. As before we use a normal mixture to cluster the data, and then within each cluster we calculate the residuals resulting from simply regressing the surrogates onto the main variables, for just patients in that cluster. It is the calculated residuals that are then tested for spatial independence in each particular cluster by brownian distance correlation tests.

Total Clusters:	1	2	3	4	5	6	7	8	9
Cluster: 1	**0.005**	0.820	0.855	0.815	0.845	0.770	0.785	1.000	0.950
2		**0.010**	0.155	0.225	0.180	0.990	0.995	0.870	0.950
3			0.310	0.570	0.495	0.560	0.955	1.000	0.995
4				0.310	0.920	0.415	0.640	0.955	0.085
5					0.950	0.895	0.700	0.620	0.470
6						0.945	0.905	0.765	0.965
7							0.970	0.910	1.000
8								0.955	0.890
9									1.000

Table 3.3: Within cluster p-values for independence of main, residuals.

Total Clusters:	1	2	3	4	5	6	7	8	9
Cluster: 1	150	48	49	49	49	40	40	7	19
2		102	53	56	63	10	10	38	12
3			48	19	7	52	29	5	12
4				26	21	17	28	29	12
5					10	21	14	28	13
6						10	19	14	38
7							10	19	17
8								10	18
9									9

Table 3.4: Total number of members within each cluster.

Table 3.3 shows the p-values obtained via this method. Table 3.4 is a duplication of Table 3.2 and provides the number of patients in each cluster.

We find a far more interesting picture than when we examined the within-cluster relationships between the unregressed surrogates and the main variables.

Firstly we see that there is very strong evidence ($p = 0.005$) that one cluster may be inappropriate, and strong evidence that two clusters is inappropriate (Bonferroni corrected $p = 0.02$). But the story in moving from one to two clusters overall is worth scrutinizing further. Not only do the p-values increase. as would be expected due to decreasing sample size in each cluster, but we also find that cluster 1 in the two cluster solution appears to have broken off some patients into a homogeneous cluster ($p = 0.820$), thus we may be confident that something is different about these patients, and that they indeed may be grouped together. Likewise when we

look at the results that correspond to the three cluster solution we find that the heterogeneity found in cluster 2 of the 2 cluster solution is entirely addressed by moving to three clusters. So it is not simply small (overal)l p-values we are interested in, but behavious where the p-values jump suddenly from indicative of a bad fit, to not contradicting a good fit. We are interested then in seeing a sigmasoidal curve in the p-values, showing that subspecies are being picked off and separated as the cluster becomes more refined.

Therefore we see, for the Irises, and looking at Table 4.3, evidence of three populations with different relationships between the main and surrogate variables. And indeed, as mentioned already, these 3 clusters divide superbly well the three species (misclassifying just 4.7%).

There is no evidence of more than three populations within the irises. At least we find no evidence given by this choice of main and surrogate variables. Perhaps with a large sample, more populations (subsubspecies?) would be educed, or perhaps another, more efficacious choice of surrogate would do so. However from this analysis it seems clear that it is worth hypothesizing there are 3 species and then investigating this claim further experimentally.

We may examine other diagnostics, such the BM covariance between surrogates in each cluster, to get a metric on the clusters, and the between cluster BM covariance of either the surrogate or the residuals. We merely remark that there is fecund ground for a plethora of diagnostics here, which we believe are generally interesting and worth developing, and in this example evidence further that there are three salient species.

When using p-values to determine whether a cluster shows evidence of containing multiple populations, as in this example one must consider the issues of multiple

hypothesis testing. Of course, one may simply correct for this via a bonferroni type correction. However we feel this is much too conservative. The data analyst would likely rather have a false positive than a false negative, thus while this is an open problem, we feel that either no correcting should be done, or else a test applied that focuses on small p-values and how many there are. For example if we had examined 20 clusters, each having a p-value of 0.005, the trivially corrected p-value is 0.1, non-significant. Yet surely noone would conclude all 20 p-values were just randomly this tiny. Other methods we have investigated include testing whether the whole set of values appears to emanate from a $U[0,1]$ uniform distribution, which gives results that are very interesting, but requires careful interpretation, as the p-values in actuality will not be uniformly distributed until large scale asymptotics may come into play. Another test we have investigated, is constructing a statistic based on how many p-values are below certain thresholds. This has given very promising results.

Another way to guard against false positives, is to view our process as a step-down hypothesis test procedure, where each cluster showing a significant p-value is split, but the ones which do not are left alone. This provides certain complications with our normal mixture model, as increasing the number of clusters does not simply result in an earlier one being split. However we believe this method may be adapted to non-divisive clustering algorithms such as mixture models.

It goes without saying that the failsafe method of guarding against overfitting and false positives may always be used, for example using both a training set to data mine and a test set to verify hypotheses of interest. It is not utterly clear how the hypotheses should be formulated however. Should they take into account and attempt a reproduction of the spatial clusters, or the between cluster relationships, or the within cluster relationships between main variables and surrogates? Likely the answer to this depends on the precise situation and question of interest.

Finally one diagnostic we should emphasize is the examination of the changes in p-values as a cluster is divided, or a clustering with more components is moved to. In the initial example, where the main variables and the surrogates were simply compared, we see that there was no dramatic change moving from one clustering solution to the next. However for the p-values found by comparing the main variables and the residuals, we note dramatic jumps from significant values to non-significant values, both moving from 1 to 2 clusters, and also for the division of cluster 2, moving from 2 to 3 clusters. We may interpret this as for the first example, showing us the choice of comparing surrogate and main variables is not appropriate, and for the second example showing us that the data is very well explained by clustering and comparing the residuals. In effect we should be looking for some elbow in these statistics, and in doing so, we should be asking whether the data is better explained by splitting one cluster into more. Taking inspiration from the gap statistic, perhaps we should be comparing the increase in p-values moving between clusters, with what the expected increase would be if the particular cluster should not be split and there was little spatial dependence of the surrogate or its residuals (as was applicable).

The aforementioned metrics as to whether the surrogate has a fundamental difference in distribution between clusters allows for an interesting possibility of boosting the clustering algorithm. As we noted mixture models, as well as many other clustering algorithms, do not just divide up an existing cluster when one more cluster is asked for, but indeed might reorder all the clusters in order to best fit in an extra one. More specifically often the situation arises where there are two clear well separated clusters, yet the clustering algorithm only divides up the patients respectful of this clear clustering when more than two clusters are requested, for it prefers to bisect one cluster before marking the other one as separate. This is a particular problem when one cluster is oversampled compared to the other, resulting in the value function some algorithms attempt to maximize benefitting more from cleaving an entire large

cluster up, rather than initially breaking apart the two clear clusters.

Thus to elucidate all the subtypes, it might be, indeed is probable that it is, necessary to split some single clusters into pieces before the algorithm finds all the clusters. This then gives us a method of creating a meta-algorithm. Whereas our proposed method was just designed to evaluate the results of a clustering algorithm, it may indeed be used to refine said algorithm. Specifically once the number of clusters is chosen so that we have confidence all subtypes have been found, the clusters can be compared, using our surrogates as explained, and decisions can be made to recombine these. This idea may of course be iterated.

Similarly our proposed method lends itself well to controlling algorithms that repeatedly divide or combine clusters, forming dendrograms. We have provided a statistical stopping point for such divisions, or agglomerations.

3.3 Binary Surrogate and Residuals.

For this iris dataset, understanding the nature of the variables, led to a natural and most efficacious choice of main and surrogates. Often the situation will not be so clear. One situation we are interested in is using as a surrogate a binary indicator, such as disease severity. This may possibly be a very good choice of indicator, as it combines numerous clinical variables into one and may be assigned by an expert diagnostician utilizing much experience and knowledge. Further when attempting to find disease subtypes, this is one variable we might expect to vary in distribution over the subtypes. Surely different subtypes would not have the same severity propensities? However do we lose too much information using a binary variable? We shall investigate this.

Our previous method requires no tweaks for comparing the main and surrogates

Species	Cyan	Blue	Green
$Y = 0$	0	15	42
$Y = 1$	50	35	8

Table 3.5: Distribution of Y within each species.

within clusters, even when the surrogate is dichotomous. To evaluate the effect of using a binary indicator, we create one for the iris example.

This we have done by randomly selecting two coefficients a and b and from the surrogate forming the variable

$$f(\text{flower}) = a(\text{sepal width}) + b(\text{petal length}).$$

Then a new variable Y was randomly sampled for each flower of the dataset from the distribution such that this variable for different flowers was independent, was binary, and took the value 1 with probability

$$p = e^{f(\text{flower})}/(1 + e^{f(\text{flower})}).$$

The χ^2 p-value for dependence between Y and the subspecies of iris was calculated, and this process was repeated until the found p-value was below 0.01. The parameters ultimately selected by this method were $a = 6.50$ and $b = -3.91$. We display the distribution of Y in Figure 3.4, which displays the flowers according to the main variables of petal width and sepal length and colours them according to species. The points are designated by crosses if the corresponding Y value is 0, and by exes if the corresponding Y value is 1. Table 3.5 gives the distribution of Y broken down by species.

We see that this new binary variable does encode well information from the species. Now we see if we can still conclude that there are the right number of subspecies using this new surrogate.

As earlier we may use BM covariance to calculate p-values for within cluster

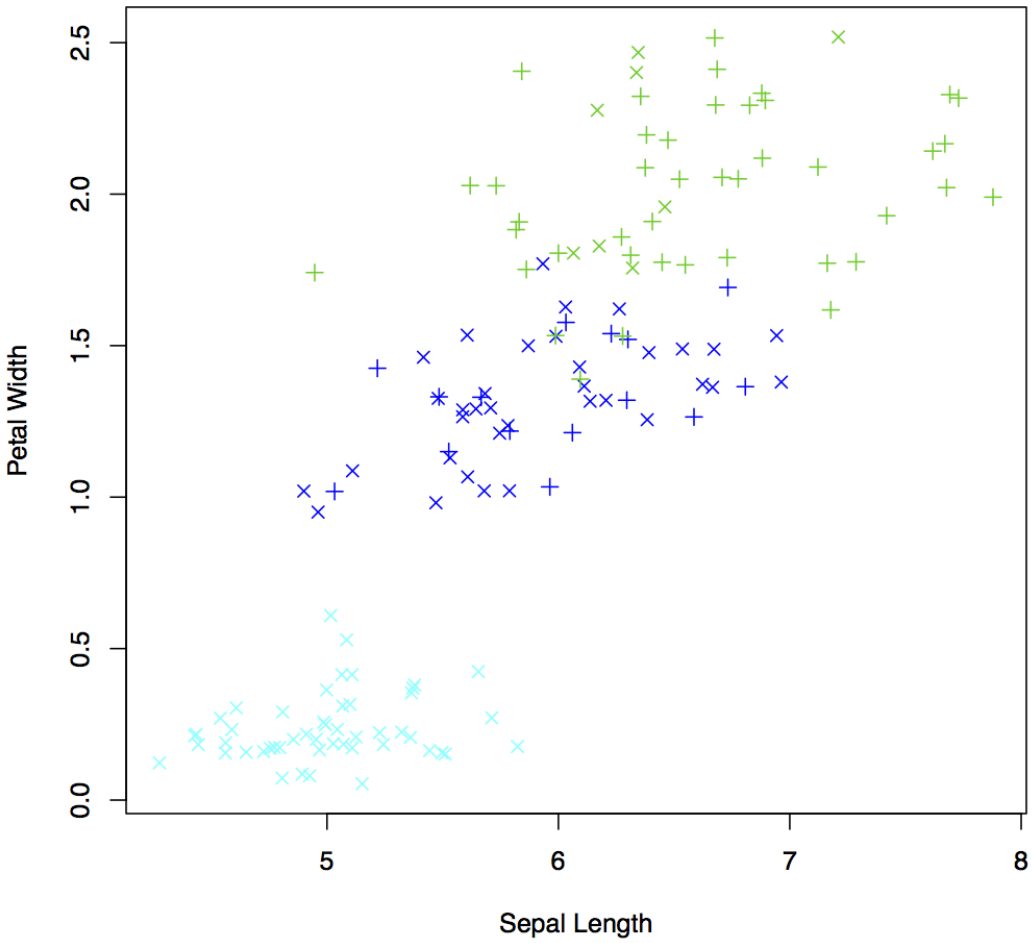

Figure 3.4: Main variables, coloured by species, marked by Y values.

independence of this surrogate and the main group of variables. The p-values are displayed in Table 3.6.

Evaluating these according to our new paradigm, we see that 3 species are once more indicated. But this is surprising, not only did the binary surrogate give evidence of 3 species, when the (non regressed) continuous did not, further the evidence is extremely strong that there are 3 clusters, with minuscule p-values jumping to large ones as soon as the patients are clustered into the right number of subtypes, that is

Total Clusters:	1	2	3	4	5	6	7	8	9
Cluster: 1	0.005	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2		0.005	0.485	0.090	0.165	1.000	1.000	1.000	1.000
3			0.135	0.605	0.495	0.850	0.750	1.000	1.000
4				0.105	0.125	0.450	0.035	0.715	0.080
5					1.000	0.450	0.350	0.715	0.080
6						1.000	0.075	0.040	0.445
7							1.000	0.120	0.750
8								1.000	0.040
9									1.000

Table 3.6: Within cluster p-values for independence of main, 0/1 surrogate.

the elbow is very extreme. We will not do so now, but we shall attempt to unravel why this binary surrogate should be so much more powerful than the continuous ones. Our aforementioned diagnostics would also confirm that there are 3 clusters of stark difference.

We could finish here, however what would have we done if we found no evidence of clusters from our binary surrogate? Would we be capable of applying our residual analysing method? That would of course be problematic. The corresponding regression would be a logistic one, but the usual considered 'residuals' are not identically, independently distributed for the correct model, so it would be unreasonable to look for independence of main variables and residuals. To circumnavigate this problem, we must define a new residual. We do this as follows. First given a binary surrogate, we regress logistically, and obtain the probability $f(\mathbf{X})$ that $Y(\mathbf{X}) = 1$. Now assuming the model is correct, we may define the error at each point. We expect that Y should be 1 more than 0 if $f(\mathbf{X}) > 1/2$ and 0 more than 1 if $f(\mathbf{X}) < 1/2$, so a natural error $Z(\mathbf{X})$ may be defined by, if $f(\mathbf{X}) > 1/2$ then $Z = Y$, and if $f(\mathbf{X}) < 1/2$ then $Z = 1 - Y$. Values of $Z = 1$ mean we observed the outcome we expected, and $Z = 0$ we observed a different outcome. However this error variable is biased in that, if we have a correct model, when for example $f(\mathbf{X}) \approx 0$ or 1, we should see very few errors, but the errors we do see should really count. We can make errors count equally no matter the spatial locale $\mathbf{X}$ by reweighting them. We do this

Total Clusters:	1	2	3	4	5	6	7	8	9
Cluster: 1	0.005	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2		0.059	0.560	0.319	0.337	1.000	1.000	1.000	1.000
3			0.505	0.527	0.549	0.646	0.527	1.000	1.000
4				0.477	0.465	0.473	0.511	0.523	0.417
5					0.493	0.488	0.408	0.494	0.579
6						0.492	0.474	0.433	0.527
7							0.500	0.476	0.641
8								0.504	0.431
9									0.509

Table 3.7: Average p-values for independence of main, 0/1 residuals.

by multiplying Z by idependent binary random variables V, defined by $V(\mathbf{X}) = 1$ with probability $(\max(f(\mathbf{X}), 1-f(\mathbf{X}))/2)$. This reweighting ensures that the random residual W that has definition $W(\mathbf{X}) = Z(\mathbf{X})V(\mathbf{X})$, if the model is correctly specified is a binary random variable that is 1 with probability 1/2. Thus we have created W, a random residual that is identically, independently distributed, if we have the true model, and looking for where the distribution of W departs from $W = 1$ with probability 1/2, $W = 0$ with probability 1/2, will highlight where the observations do not concur with the model, analogously to standard regression residual analysis. We have considered other ways to define W our residual, but are still examining the optimal method. Now our residual W is random, but we can perform our analyses with it, and then take an average over many simulations of V. Thus our techniques for residual analysis we have put forth already will now apply. Table 3.7 then shows the results of examining the within-cluster dependency of the residual for our example, with our simulated binary variable. The mean p-value over 700 simulations of V is displayed.

We see at a glance, adhering to our earlier logic for the analysis, that again we should predict 3 species. Our residual definition and analyses seems to work very adroitly. We notice there is now less evidence than given by using the original surrogate for 3 clusters, but the results show the underlying algorithm is rather sound. Further using the residual would, on consideration, seem preferable to using

the original variable, due to being both more interpretable and robust.

We note however that these mean values do not give the whole picture. Figure 3.5 gives a density estimate for the p-values of both clusters, when the algorithm is requested to give exactly 2 clusters.

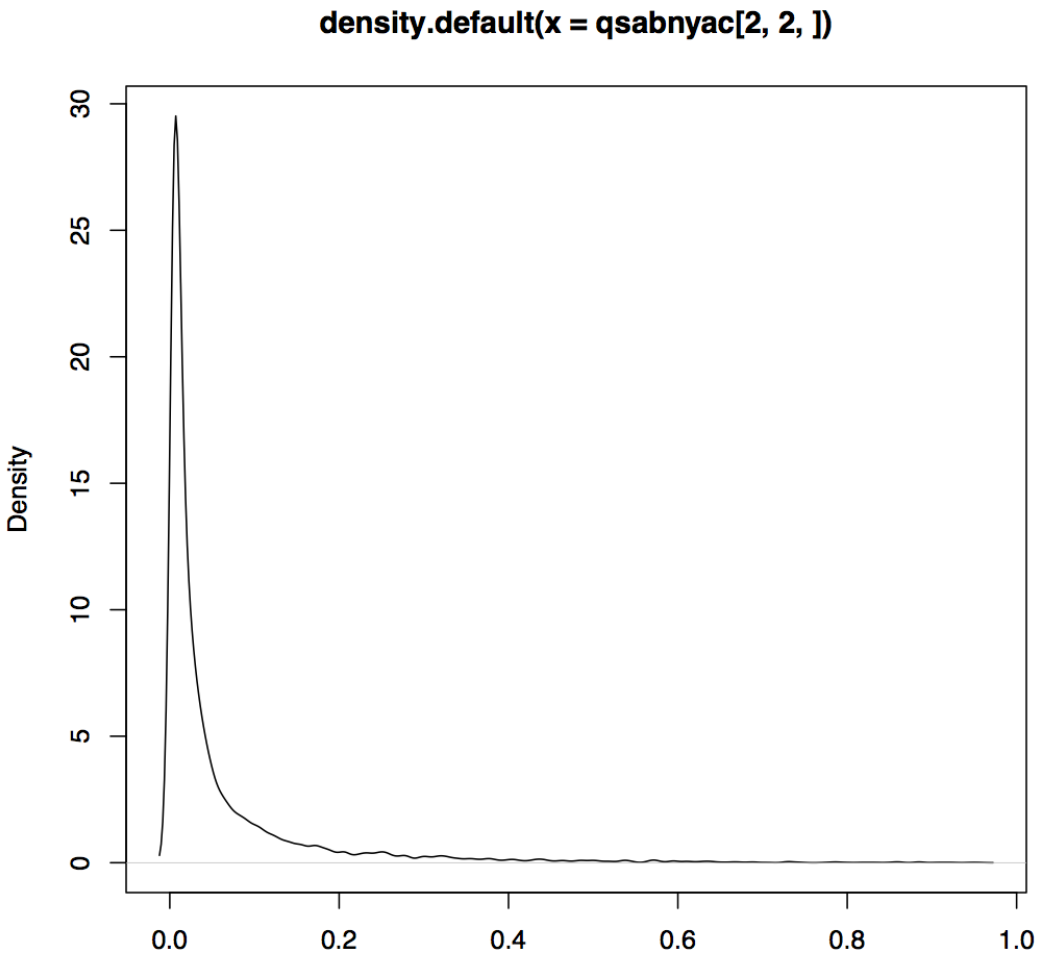

Figure 3.5: Estimation of density of p-value for second cluster of 2.

We notice that this is a distribution with a very large skew and long tail. The mean is really affected by outlying results, but most random p-values were discovered to be 0.005. It seems that simply averaging may not be optimal to calculate

the appropriate p-values, and maybe a bootstrap simulation, and comparison of the observed distribution with the expected one if we had independent within-cluster residuals, would provide a much more appropriate (and less liberal) p-value. We shall make further studies regarding the best calculation of residual p-values from the simulated random residuals.

3.4 Dimension Reduction Through a Surrogate.

We conclude this exploration with a discourse on how the surrogate variable not only may be utilized to estimate (with significance levels) the number of clusters, but also how it may be utilized in the other plague of clustering, namely reducing the dimension.

Clustering algorithms rapidly become infeasible in higher dimension, particularly ones that rely on fitting mixture models and density estimation. But in shotgun clustering, as we named it, we are given many variables, and only a fraction of these could be relevant. Screening of variables, particularly for gene data has been suggested to reduce dimension. Screening is oft performed by looking for evidence of clusters in each feature univariately, for example assessing the fit of a mixture model with more than one unimodal distribution. However this is likely to both overfit (if we have a large number of noise variables the true cluster signal could get crowded out), direct us to non clinically relevant clusters, and also miss a weaker signal requiring more than one feature to see. We hope that instead of fitting mixture models, the surrogate could be of use to sidestep these issues.

We ultimately will be interested in only those clusters that show evidence of difference in surrogate, so we should begin by screening for those features most likely to produce this. We make the suggestion that each individual feature should be

screened for showing dependency between the feature and the surrogate. We shall give the illustration of using our binary generated surrogate from the preceding examples. Here we only have two main variables, so it would not be too fascinating to perform screening on these by themselves. Thus we create 4998 further random features, with normal distribution, independent of the surrogate. We then screen using Kolmogorov-Smirnoff tests, and test for independence of those values of the feature corresponding to the surrogate being 0 with those values of the feature corresponding to the surrogate being 1.

The Sepal Length and Petal Width features give p-values with orders of -10 and -11. The next lowest p-value from the unassociated features is of order -4. If we correct (simply) for multiple tests, the p-values corresponding to the main variables are p=3x10^{-6}, and 7x10^{-8}. The other 4,998 p-values are all 1 following corrections. It is entirely obvious that screening for association between the feature and surrogate will proffer good outcomes, just as we see here.

We might also consider screening in couplets or triplets of variables. This can simply be done, but instead of splitting the features between where the surrogate is 0 and where it is 1 and testing if the distributions of each are the same, we need to use BM distance covariance. To do so, we test simply the dependence of the dimensional feature, and the surrogate. In a like fashion if we want to screen with a continuous, or multivariate surrogate, we can also take this route. We do not provide calculations for the iris dataset with a non-binary surrogate, as the methods and results should be trivial on comparison with the results for screening with Y.

The use of screening and dimension reduction should hopefully be immediately appealing. However as an interesting illustration we shall display how even reducing a small number of dimensions, of even very informative variables, might pay dividends. Assume we have been told there are three species of iris, and asked to deliminate them.

	Groups Calculated In 4D:			Groups Calculated In 2D:			
	I	II	III	I	II	III	Total
Species: 1	50	0	0	49	1	0	50
Species: 2	0	48	2	0	47	3	50
Species: 3	0	14	36	0	6	44	50
Total	50	62	38	49	54	57	150

Table 3.8: k-means constructed classifier performance.

Further assume that we decide on using k-means for this purpose. (For the iris data, normal mixture modelling will not exhibit this precise pathology). Table 3.8 shows the results of using 3-means (with 1000 repetitions of the algorithm of Hartigan–Wong) applied to all four dimensions to classify the flowers, and then the results if applied to just the two dimensions of Petal Length and Sepal Width. Immediately we see using all four dimensions for classification performs far more poorly than just using the selected two dimensions. This is a slightly shocking result as here we have only removed informative variables, that are no more noisy, we believe, than those we have kept.

Our overall misclassification rate when all four dimension are utilized is 10.7%. However using just two we immediately get a far smaller misclassification rate of 7.3%.

Thus we hope surrogate clustering (and surrogate screening) has much to offer. Certainly it is extremely effective for certain instances as we have shown.

3.5 Visualization Methods Involving Surrogate.

One issue of the use of clustering for data mining is that there often is no good visualization of the results, meaning we lack good diagnostics. Indeed this is somewhat due to the inherent nature of cluster analysis. For two-dimensional or three-dimensional data we have an instinctive idea of what constitutes a good clustering,

and plotting the data can show whether we have an intuitively appropriate clustering. Appropriate hypothesis testing can then confirm this, at least if we are proceeding through a model-centric viewpoint. However, for higher dimensional data, it is difficult to evaluate whether or not what we obtain is an appropriate meaningful clustering, and of course it is even more difficult to even know how to precisely define a meaningful cluster. A partial answer is provided by dendrograms, which may be produced if the clustering algorithm is a divisive one (though any algorithm naturally might be iterated to force it into a divisive-type algorithm). However, these are not a perfect, or indeed a particularly acceptable, solution, and to an untrained eye, the majority of dendrograms look pretty impressive and good evidence of clustering, even if the data really doesn't have meaningful clustering. We therefore put forth a very simple graphical representation designed for non divisive algorithms that allows visualization of our surrogate clustering technique.

Our representation, designed for non divisive algorithms expecting to find a reasonably low number of clusters, is to visualize the clustering as a graph, with edges having weightings reflecting both the spatial distance between corresponding clusters and the ability to tell the clusters apart using the surrogate variables. More precisely, given a clustering of a given number of clusters, we represent each of these clusters by points spread equidistant around a circle. To indicate the size of each cluster we mark its number with a symbol proportionate in size to how many members it has. To indicate the distance in the main feature space between clusters, the line connecting each cluster is given width inversely proportional to that distance, and to indicate the dissimilarity of clusters based on our surrogate measure, we shade the line in grayscale going from black to white, in proportion to the p-value that we calculate for whether if those two clusters are combined, then regressed onto the main features, the main features and residuals are dependent. Hence, clusters spatially close will have thick lines, but these will be not so visible obviously unless the surrogate evinces

a level of evidence for distinct subtypes - indeed they will be invisible against the background if the surrogate evinces no evidence whatsoever. Thus both surrogate and main feature distance may be seen on the diagram. While certainly not a solution to all cluster diagnostics, we feel this contributes a small amount to the ability of graphically representing clusters. Figure 3.6 displays the results of using normal mixture modeling on the iris dataset, for one, two, three, four and lastly five clusters. The salient absorption from the graph should be the contrasting feature and surrogate distances, as opposed to absolute magnitudes, and especially the results when shifting the number of clusters up or down by one.

Examining these plots provides a fair bit of information. Crossreferencing with Figure 3.3 as required to see the precise location of the clusters in main feature space, and we shall write cluster m/n to designate the mth cluster out of n in the clustering with that total number of clusters. We see that cluster 1/2 is quite distinct, both spatially and even more so in surrogate relation distance to cluster 2/2, and that this separation continues in subsequent clusterings with at first cluster 2/2 being split into two equal size clusters, of the same magnitude as 1/2, and a further of these 3/3 is split again and again in subsequent clusterings. It appears thus we have quite a stable clustering. Interestingly when three clusters are asked for we see that 2/3 is closer to 3/3 in absolute distance than it is to 1/3, but closer in surrogate relation distance to 1/3 than 3/3. Further 3/3 is very distinct in surrogate relation metric from 1/3, as well as in spatial absolute distance.It appears we have a true cluster effect here and that three clusters are most justified. This is further highlighted looking at the results of asking for four clusters, when cluster 3/3 is split into 3/4 and 4/4, with 4/4 a cluster very distinct from cluster 1/4, and with 3/4 being closer to 1/4 in surrogate relation distance, though not absolute distance, and much closer to 2/4 in absolute spatial distance, but more like 4/4 in surrogate relation metric. This then might be interpreted as one very distinct cluster, that is 1/3, and two clusters 2/3,

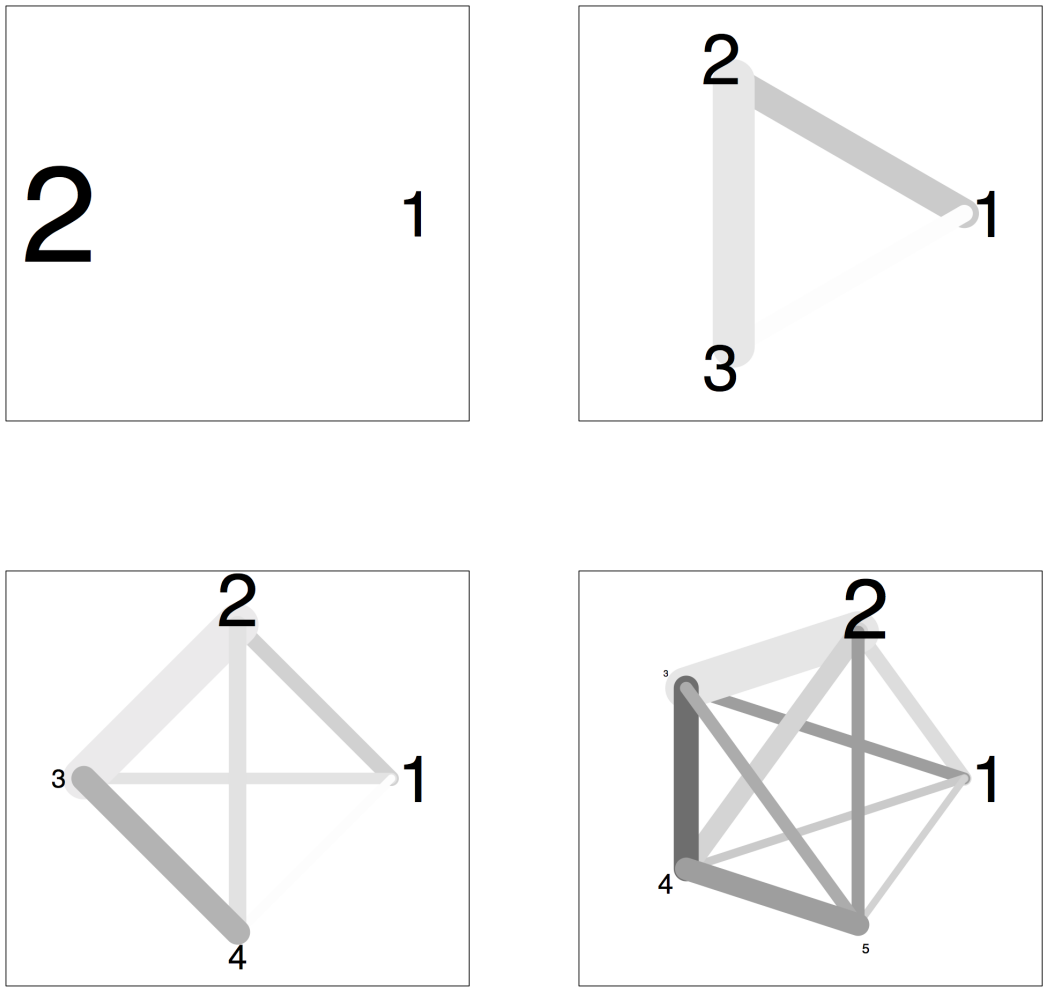

Figure 3.6: Cluster plots for iris dataset with 2-5 clusters.

3/3 that merge into each other spatially but contain at least two subtypes evinced by a switch in surrogate relationships. Moving to both four and five clusters shows that our clustering is attempting to discern where 2/3 ends and 3/3 starts, by splitting up points that are hard to identify. We conclude there is good evidence for three subtypes.

Naturally the effects of statistical estimation may be also included in the cluster

plots. For example markers, and lines may be multishaded to represent confidence intervals. But we have not pursued this here.

3.6 Example: SPIROMICS Data.

Overview:

We proceed by attempting to utilize surrogate clustering to identify subtypes of COPD using blood biomarkers. As part of the SPIROMICS study, data was collected from over 1,000 COPD sufferers and normal controls. The data includes both clinical variables, and 112 blood biomarker variables. We are interested in determining subtypes of COPD through clusters in these 112 biomarkers.

Chronic Obstructive Pulmonary Disease (COPD) is a serious progressive and debilitating lung condition that affects millions of sufferers, and is the third leading cause of death in the United States alone. It is a term that denotes loss of air flow through the lungs due to a number of possible causes, including loss of elasticity of the airways and air sacs, wall being destroyed between air sacs, inflamed walls of the airways, clogging from too much mucous. COPD includes two main conditions, emphysema and chronic bronchitis, that are often concurrent in COPD sufferers. There is no known cure or way to reverse damage, though progression can be slowed. While risk factors, such as smoking, are known, other factors to explain the propensity of a person to develop COPD are not. Furthermore, the precise pathophysiology of COPD, and in particular how many disparate pathways there are that give rise to symptoms denoted as COPD, are unknown. The latter query is of particular interest, for different pathways would beget different possible interventions. We intend to explore this question and identify different subtypes.

The SubPopulations and InteRmediate OutcoMes In COPD Study (SPIROMICS)

is a multicenter program funded by the National Heart, Lung and Blood Institute, coordinated by the University of North Carolina at Chapel Hill, to promote the collection and analysis of phenotypic, biomarker, genetic, genomic, and clinical data from COPD subjects and corresponding controls. The primary aims include identifying disease subpopulations, and discovering and validating surrogate makers of disease severity that may be used as intermediate outcomes to asses whether potential treatments are efficacious. Hence SPIROMICS provides a large, high dimensional data set, which quite possibly will evince subtypes of COPD with differing pathophysiologies, when subjected to the right analyses.

As noted, cluster analysis is one method to find subtypes of disease. For subtypes of disease with different pathophysiologies, we expect a different distribution among subtypes of some biomarkers related to these pathophysiologies. In particular we might expect the high density regions to be quite different for each subtype. Not knowing the subtypes and beginning with these biomarkers, clustering into unimodal, high density regions should then elucidate the underlying subtypes. This logic has been applied successfully in many areas of biomedical research, and for a particular example to examine cancer subtypes, where often the biomarkers are levels of particular gene expressions.

Our initial results, using standard methods were unable to reveal any meaningful, interpretable structure in the data. This motivated our development of surrogate clustering as laid out above. We apply this method to the SPIROMICS data set, and the results are detailed below.

Results Summary:

Using machine learning techniques, we isolate 10 of the 118 biomarkers to target for clustering: ICAM1, CXCL10, CXCL9, TNFRSF1B, CRP, MMP3, CHGA,

TIMP1, CCL16, TNFRSF10A. We then predict that we may find the most clinically meaningful clusters in just four of these features taken together simultaneously: ICAM1, CXCL10, MMP3, TIMP1 (and CXCL10 may be essentially interchanged with CXCL9).

We identify 4 possible clusters in the subspace of these features.

We identify a high association between these clusters and the 57 clinical features which may be divided into physical, demographic, lung capacity, response to bronchodilator, lung disease indication, and biomarker subsets (with the biomarker falling into four distinct categories).

Figure 3.7 displays a visualization of these clusters in a two dimensional representation of the four dimensional biomarker space, followed by a two dimensional visualization of the clinical variables (listed above, but excluding the physical height and gender variables), colored by these four clusters.

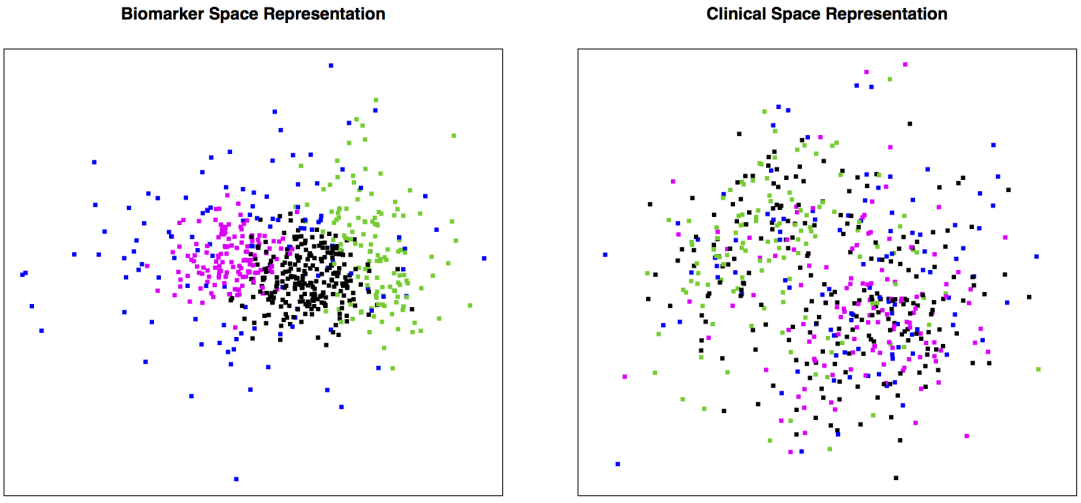

Figure 3.7: Two dimensional visualizations of the 4 clusters, both in biomarker space (in which they are defined) and clinical space.

We may observe that the clusters found from the biomarkers appear to have clinical significance. Indeed, when viewed in clinical feature space, there is evidence of clustering among the clinical variables, and this clustering calculated from the biomarkers suggests greater evidence of subtypes than when viewed in biomarker space. We explore and comment further below.

Data:

The SPIROMICS data available for our analysis is 439 clinical and 118 biomarker variables taken from 1,697 patients.

Demographic (gender, race, etc.), physical (weight, BMI, etc.), physiological (lung capacity, sleep quality) and medical (drug usage information) constitute the clinical variables, which are a mixture of continuous and categorical variables. The biomarkers are measurements of a number of different proteins obtained from the plasma of each patient. The 112 different biomarkers were proposed by the investigators as being judged to be the biomarkers most likely to elucidate further information about COPD.

The patients are stratified into four strata: healthy non-smoker, healthy smoker, mild/moderate COPD, severe COPD.

The data has appreciable missingness. Certain biomarker variables are over 90% missing. Missingness is of multiple types. We have frequent left censoring due to receiving reported measurements below the accuracy level of the assays, and less frequently right censoring due to the measurement being above the upper quantifiable limit of the assay. Further there is occasionally general, presumably non-ignorable, missingness.

We also have appreciable missingness within certain clinical variables. We have

439 variables, but only 270 with less than 25% missing, and only 162 with less than 5% missing.

Method:

The fairly high dimensionality of the dataset, combined with the amount of missing data, and need for clinical interpretability make this a difficult analyses. Standard analyses did not reveal interesting structure in the data. We needed to develop specific machine learning techniques for this task, as described below.

The biomarker measurements are all positive numbers, of differing ranges. For statistical, scientific, and philosophical reasons, we work with the logarithmic transforms of these. Below, all referenced biomarkers are referencing the log transform of the original data biomarkers.

We shall use a training and test set split. As much has to be estimated from the training set (that is appropriate dimension reduction, clustering, clinical relevance) we propose that a 3:2 ratio is appropriate, stratified by COPD severity. We divide our data randomly into a training set of 636 patients from strata 3 and 4 (with a breakdown of 439 in stratum 3, 205 in stratum 4), and a test set of 423 patients from strata 3 and 4 (with a breakdown of 287 in stratum 3, 136 in stratum 4).

Our hope is that different pathways may be educed by extracting different signatures in the patients measured biomarkers, more precisely by finding clusters, each of which corresponds to a different pathway or pathophysiology. Examining the clinical measurements is an alternative attack, but our hope is pathway differences will have a more immediate and clearer signal in the biomarkers than the clinical variables. Furthermore all patients have similar presentations of the disease, suggesting finding subtypes from presentation may be difficult. However we will crossreference our findings with clinical variables to determine if our results have clinical meaning. Simple

standard explorations appear to be unable to find meaningful clusters that are well separated in biomarker space, or else with good interpretation in terms of clinical relevance, requiring us to extend existing methods.

We utilize subspace clustering, with clustering for each subspace begin performed by fitting normal mixture models, using the Mclust algorithm of Fraley et al.

To reduce the number of subspaces to examine to manageable numbers, we screen each biomarker first univariately, and then in pairs bivariately, for suggestions of clinically different subtype. This then flags the ten biomarkers: ICAM1, CXCL10, CXCL9, TNFRSF1B, CRP, MMP3, CHGA, TIMP1, CCL16, TNFRSF10A.

We exhaustively search through and rank all 1024 possible subspaces of these to identify those with possible clinical meaningful subtypes. We select the 4-dimensional subspace of ICAM1, CXCL10, MMP3, TIMP1, as that which is most promising for the identification of subtypes. Within this subspace we infer there are 4 meaningful clusters.

The presence of clinical subtypes is sought for by examining and seeking sharp changes in regressed residuals of the clinical variable LLN FEV1/FVC onto the biomarkers over different clusters. This clinical variable is the lower limit of normal ratio of volume of air exhaled in 1 second to total volume that may be exhaled. The lower limit of normal for an individual is defined as the lower fifth percentile of this variable for a group similar in age, height, gender and race. It is inherently bound up with the definition and diagnosis of COPD, and we might hope different subtypes of COPD have this variable distributed somewhat differently. It is thus an acceptable surrogate.

Biomarker	p-value
ICAM	1.4×10^{-12}
CXCL10	1.1×10^{-8}
CXCL9	6.2×10^{-7}
TNFRSF1B	3.8×10^{-4}
CRP	5.0×10^{-4}
MMP3	5.2×10^{-4}
CHGA	7.1×10^{-4}
TIMP1	3.9×10^{-3}
CCL16	7.6×10^{-3}

Table 3.9: Biomarker rankings for possible indication of disparate subtypes.

Preliminary Conclusions:

Scrutinising changes in relationships between the biomarkers and the surrogate LLN FEV1/FVC variable, we propose that the 9 biomarkers, as ranked in Table 3.9, show strong evidence of possible disparate clusters or subtypes. The p-values shown are for those calculated by brownian distance correlation t-tests, for the null hypothesis that regressing the surrogate results in residuals independent from the biomarker, using the training set.

Of subspaces spanned by features chosen from these nine, the 10 we rank most interesting and likely to display clusters of disparate subtypes are indicated by Table 3.10. The estimated statistic is calculated from a regression of log p-values onto number of clusters, ranging from 1 to 8, where the p-values are for the hypothesis that normal mixture modelling into the requisite number of clusters results in clusters where the residuals of the regressed surrogate are independent of spatial position within the cluster. We conjecture high values of this statistic are likely to indicate a significant number of clusters representing disparate subtypes. It appears that ICAM1, MMP3, TIMP1 and one of CXCL10 or CXCL9 are the most important biomarkers according to this statistic. Interestingly it seems that the noise to non-mutual signal ratio of CXCL10 and CXCL9 is not high enough to justify including the additional dimension of having both these biomarkers in the clustering.

Combination	1	2	3	4	5	6	7	8	9	10
ICAM1	X		X	X	X	X	X	X	X	
CXCL10	X					X	X		X	
CXCL9		X	X	X	X			X	X	X
TNFRSF1B					X				X	X
CRP										X
MMP3	X	X	X	X	X	X		X	X	
CHGA						X				
TIMP1	X	X	X		X	X		X	X	
CCL16						X		X		X
Statistic Estimate	19.6	18.2	17.9	17.6	17.4	17.1	16.8	16.8	16.4	16.3

Table 3.10: Subspaces giving highest statistics.

Total Clusters:	1	2	3	4	5	6	7	8
Cluster: 1	1×10^{-12}	3×10^{-4}	2×10^{-3}	0.65	0.88	0.84	0.81	0.88
2		0.79	0.52	0.81	0.68	0.84	0.59	0.80
3			0.78	0.90	0.67	0.72	0.78	0.64
4				0.80	0.88	0.91	0.91	0.87
5					0.74	0.69	0.72	0.83
6						0.71	0.81	0.64
7							0.88	0.89
8								0.92
Overall:	1×10^{-12}	6×10^{-4}	6×10^{-3}	1.00	1.00	1.00	1.00	1.00

Table 3.11: Training p-values for ICAM1, CXCL10, MMP3, TIMP1.

Examining the p-values whose logarithms were regressed to give the proposed statistic, we see that for the subspace of ICAM1, MMP3, TIMP1 and CXCL10, we choose 4 populations as the best estimate for the number of clusters in this 4d subspace. These p-values are shown in Table 3.11. More specifically the table displays the within cluster p-values calculated by brownian distance correlation for the null hypothesis that the regressed surrogate residual with each individual cluster, if the subspace is clustered into differing number of total clusters, is independent of spatial location within the cluster. The simply Bonferroni corrected overall p-value is also given for each choice of the overall number of clusters. Table 3.12 shows the breakdown of how many training subjects are in each cluster, if four clusters are sought. The clusters are designated partly by the colors they will be represented by in later illustrations.

The results of the clustering calculated on the training data may be shown visually

Cluster (of 4):	1 (Green)	2 (Magenta)	3(Black)	4 (Blue)
Number of subjects:	132	139	247	114

Table 3.12: Number of training subjects assigned to each cluster.

through either pairwise feature plots, as in Figure 3.8, or a two dimensional classical multidimensional scaling (MDS) of a data matrix, performed using cmdscale in R, as shown in Figure 3.9. In both cases the 4 clusters are color coded. Corresponding to the ordering in Table 3.12 for a total number of 4 clusters, Green represents cluster 1, Magenta cluster 2, Black cluster 3, Blue cluster 4.

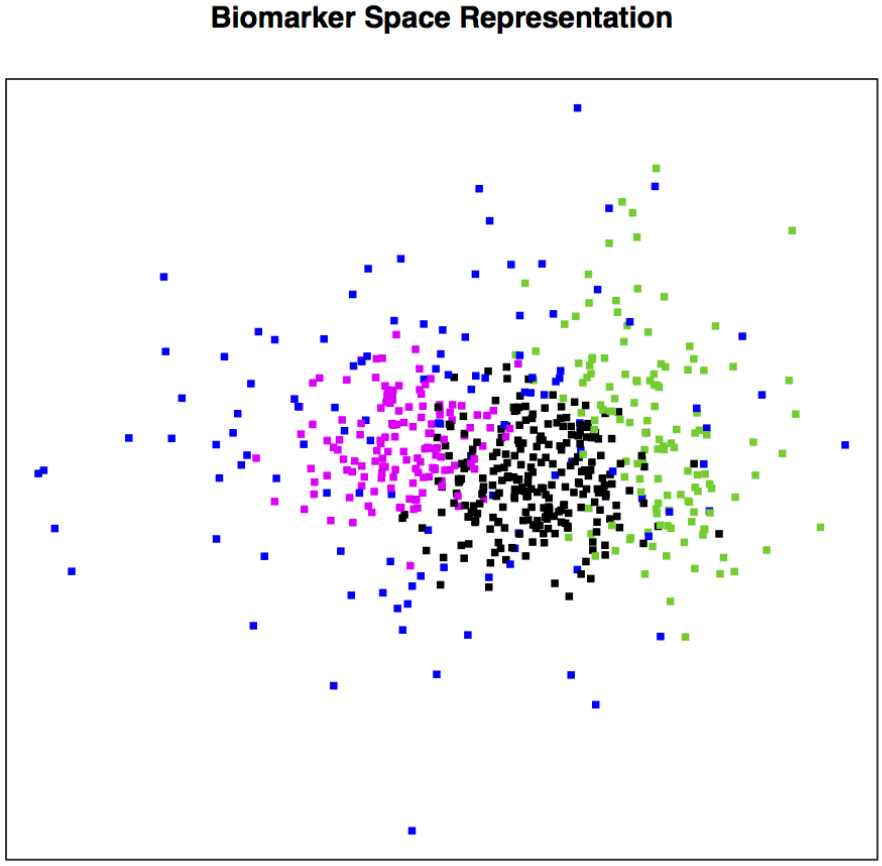

Figure 3.8: 2D visualization of the 4D biomarker space, colored by 4 clusters.

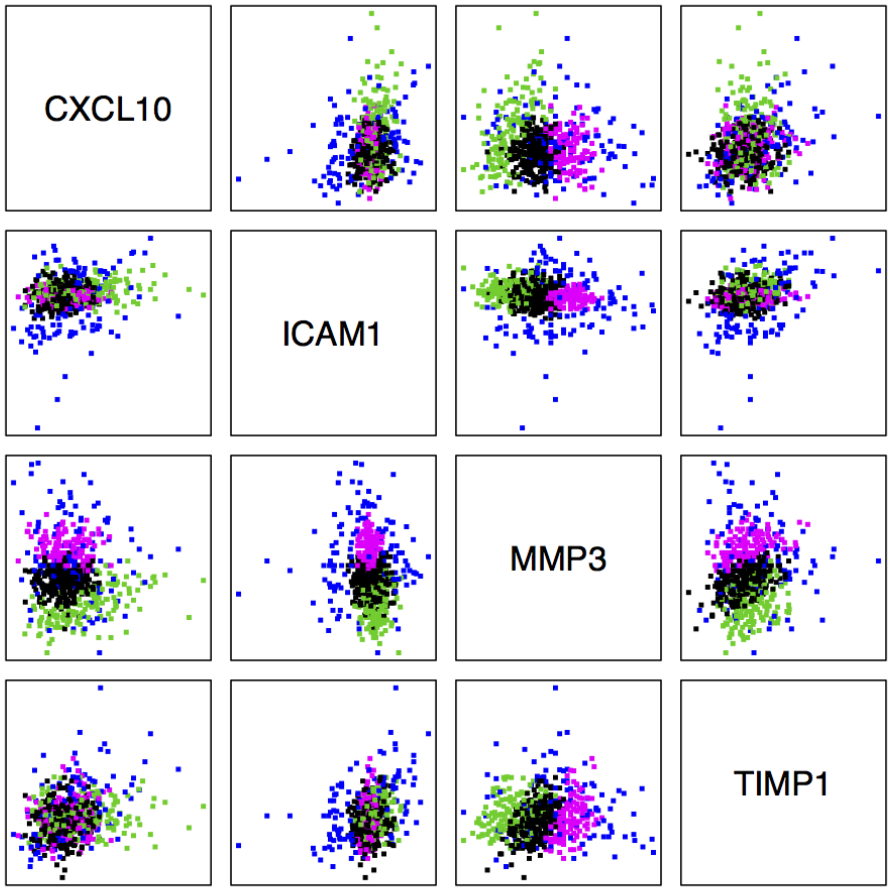

Figure 3.9: Pairwise feature plots, colored by 4 clusters, for training data.

The univariate marginal distribution of the four biomarkers within each cluster is summarized in Figure 3.10.

We find that in the two dimensional multidimensional scaling representation as given in Figure 3.8, clusters 2, Magenta and 3, Black, seem to be compact somewhat spherical clusters, with not unreasonable delineation, whereas cluster 1, Green is ellipsoidal and cluster 4, Blue is also ellipsoidal, although somewhat orthogonally oriented to cluster 1, Green. Indeed cluster 4, Blue seemingly is a cluster of outliers,

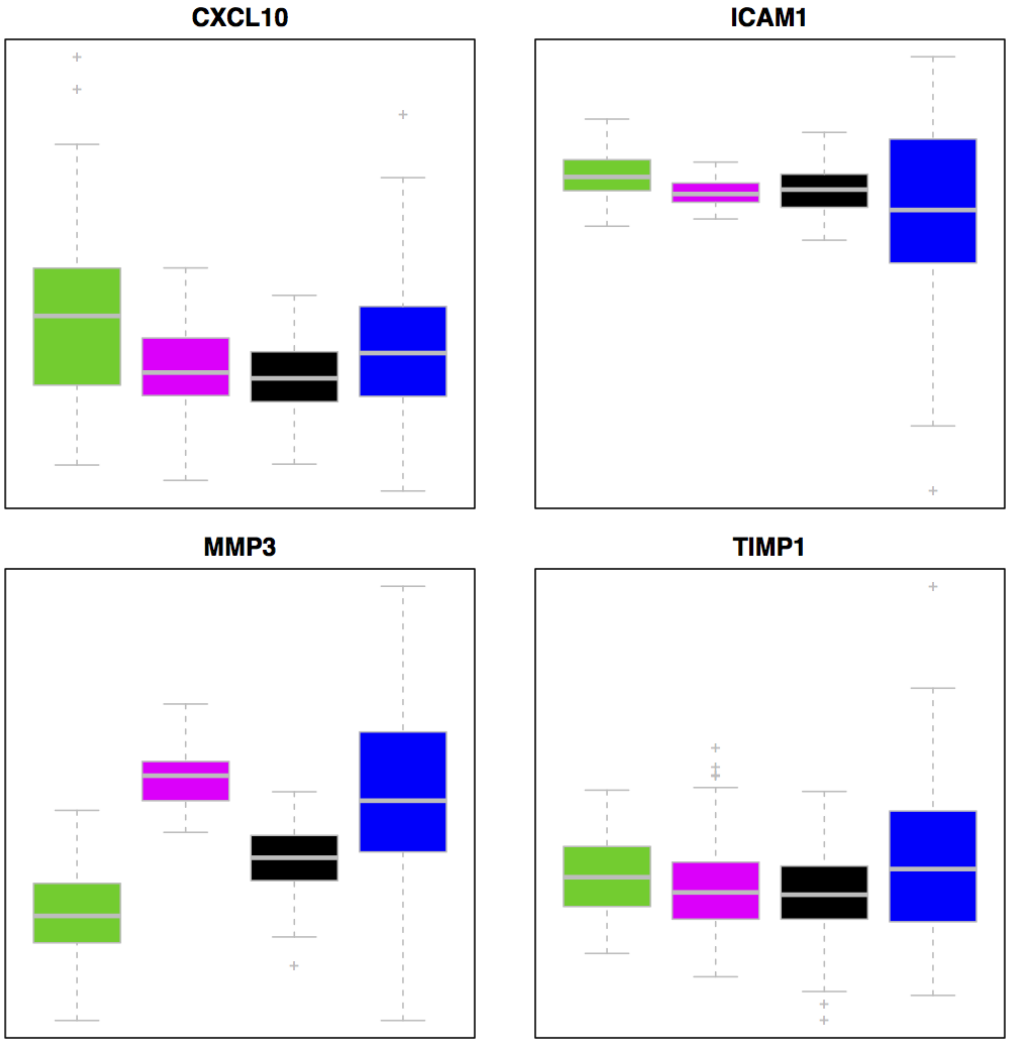

Figure 3.10: Marginal distribution of biomarkers by cluster.

or extreme points in the 4 dimensional space. From the pairwise plots of Figure 3.9, we may obtain the same observations, and indeed expand them noting that cluster 1, Green is fairly compact and spherical except mainly in the CXCL10 feature, whereas cluster 4, Blue is diffuse in all four variables. Cluster 4, Blue in particular might possibly be characterized as differing from the other clusters due to strong positive correlation between CXCL10 and ICAM1, whereas cluster 1, Green shows strong positive correlation between MMP3 and CXCL10. From examining the univariate distributions by cluster, as displayed in figure 3.10, we reinforce our diffuseness and

Cluster (of 4):	1 (Green)	2 (Magenta)	3(Black)	4 (Blue)
Number of subjects:	84	81	164	92

Table 3.13: Number of test subjects assigned to each cluster.

compactness observations, and note that a fairly sharp boundary in MMP3 seems to be the discriminating factor between assignment into clusters 2, Magenta, and 3, Black, and these clusters otherwise are hard to distinguish. The fact that the medians (and other similar quantities) of all clusters may not be simply arranged into any monotone increasing pattern across all variables simultaneously suggests that we have not simply divided up one unimodal elliptical distribution.

Having fitted a normal mixture model, we may then use the parameters estimated to provide a classifier for new subjects. This essentially classifies subjects by evaluating four different quadratic forms of the four features ICAM1, CXCL10, MMP3 and TIMP1, and the minimum of these indicates which cluster the subject should be placed in. Thus we may classify our test data to examine performance. We note that as the sampling mechanism is identical for the test and training data, we may take into account the relative size of the clusters in constructing the classifier, as opposed to if we had a different sampling mechanism when it would be preferable to use a pure maximum likelihood classifier. Table 3.13 shows how many subjects in the test set are classified into each of the clusters.

As normal mixture modelling gives consistent estimates, and as the sampling mechanism is the same, we should expect to see our clustering repeatedly exactly in the test set once it is classified, and clearly it is, as shown in Figure 3.11. Similarly the marginal distributions of the feature among the clusters in the test set should be a repeat of that for the training data, up to statistical sampling variability.

What are not consistent estimates however are the p-values we calculated to determine the correct number of clusters. These should be validated in the test set.

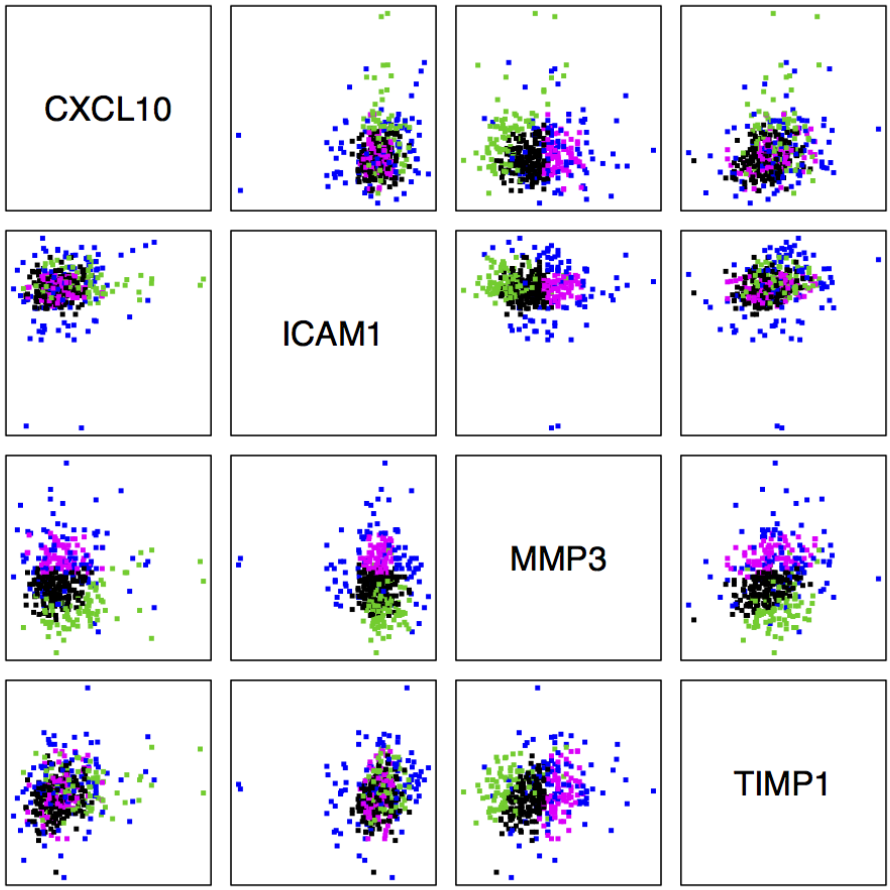

Figure 3.11: Pairwise feature plots, colored by 4 clusters, for test data.

Tables 3.14 provides the corresponding p-values in the test set.

These p-values are, by our above reasoning, indicative of at least three clusters, and are not obviously contradictory to the hypothesis of four clusters. Hence we are willing to believe that our results from the training data were not unduly biased by our feature and subset selection process, and thus there indeed are four clinically meaningful clusters in these four features.

Total Clusters:	1	2	3	4	5	6	7	8
Cluster: 1	3×10^{-6}	0.82	0.12	0.39	0.30	0.88	0.38	0.33
2		0.02	0.34	0.18	0.64	0.49	0.74	0.88
3			0.49	0.81	0.81	0.65	0.86	0.30
4				0.82	0.49	0.89	0.65	0.74
5					0.75	0.84	0.81	0.75
6						0.75	0.80	0.62
7							0.77	0.90
8								0.87
Overall:	3×10^{-6}	0.04	0.36	0.73	1.00	1.00	1.00	1.00

Table 3.14: Test p-values for ICAM1, CXCL10, MMP3, TIMP1.

Clinical Interpretations:

It does appear that our clustering approach has given interesting and appropriate results for the biomarker space. We now explore how the corresponding clusters appear in clinical space. A simple visualization is a multiple dimensional scaling of the clinical variables. We have some missingness in these clinical variables, and to deal with these, we consider those variables with less than 10 missing measurements amongst COPD sufferers in the training set, and then those subjects with no missing observations for these variables. This allows us to have complete data on 562 COPD sufferers for 113 demographic and clinical variables. Of these features 63 features are non-discrete and amenable to multidimensional scaling techniques. We separate the demographic variables, namely Age, BMI, Height, Weight, Follow Up Period, Pack Years of Smoking History, from the 57 clinical variables. We perform multidimensional scaling to two dimensions separately for these 6 demographic variables and then for the 57 clinical variables, due to the disparate natures of these two categories. We color the points according to their cluster assignments in biomarker space. The results for the demographic variables is shown in Figure 3.12, and that of the clinical variables in Figure 3.13.

Both representations display structure, though it is hard to interpret. The representation of demographic variables is surprisingly discretized, and the representation

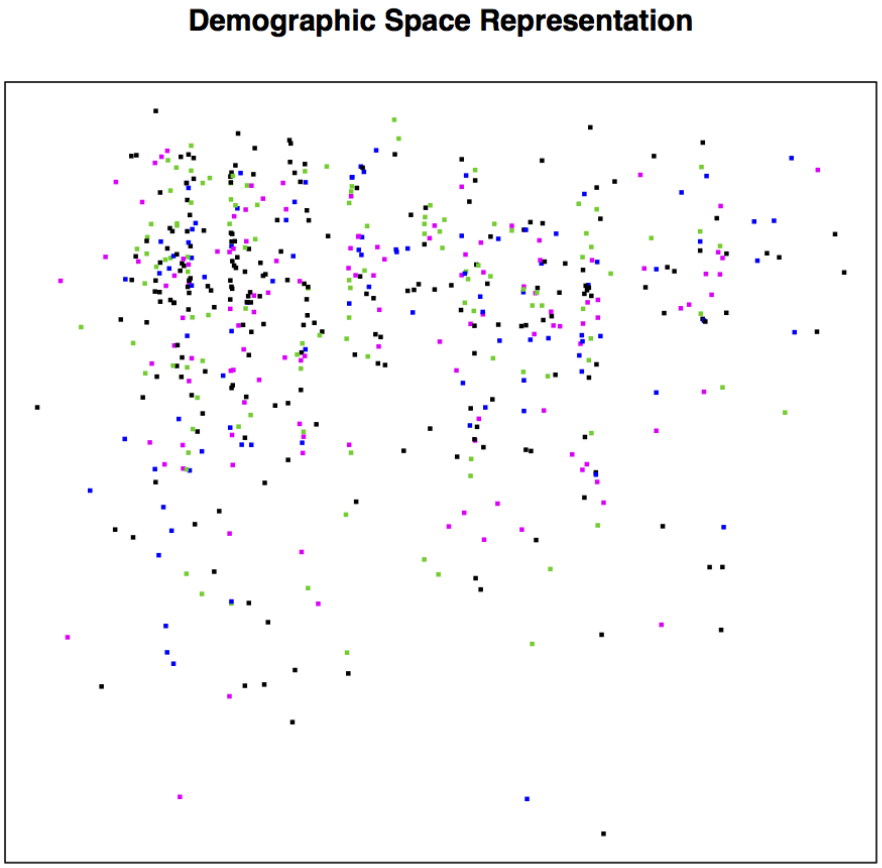

Figure 3.12: Two dimensional visualization of the 4 clusters, both in demographic space of six chosen continuous variables.

of clinical variables might be interpreted to hint at a number of clusters. However it is not easy to see how our biomarker based clusters relate to the representations in clinical or demographic space. To try to determine any correspondence we attempt to denoise, and extract only relevant signal in the clinical feature space. To do so we examine which variables differ significantly between the biomarker clusters. To measure this we use brownian correlation p-values that test the dependence of variable on the cluster number, and a correction for multiple testing is made. A density

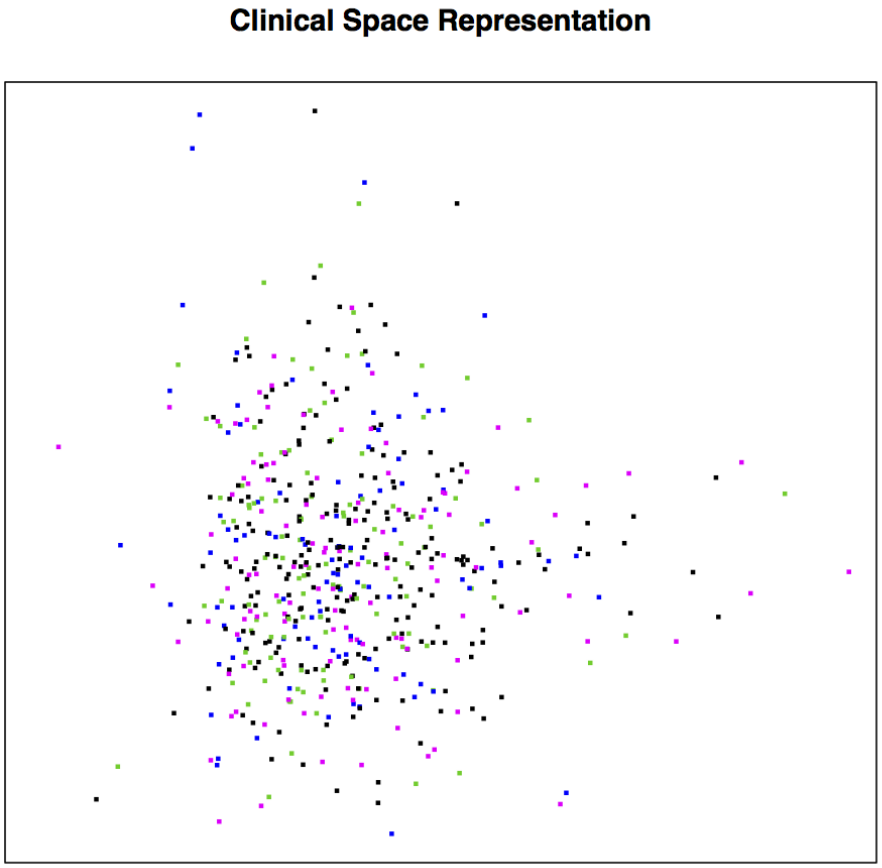

Figure 3.13: Two dimensional visualization of the 4 clusters, both in space of 57 chosen continuous clinical variables.

estimate of the resulting p-values is shown in Figure 3.14.

We make an assumption that the density of p-values for the non-informative features is unimodal. From examining the inflection points of the estimated density of the p-values, we suggest using10^{-5} as a threshold to indicate likely informative features. This results in identifying 79 variables of interest, which are displayed in Tables 3.15, 3.16 and 3.17 with the corresponding p-values calculated both in the training set and also the confirmatory p-values found from the test set. We find that

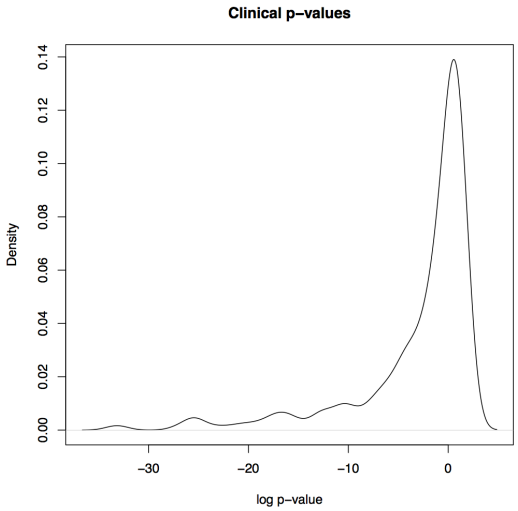

Figure 3.14: Density estimate of clinical p-values of training data.

the 10^{-5} cut off is appropriate., and many p-values reproduce very well in the test set. Those features which failed to reproduce are greyed out.

It is clear that the found clusters differ markedly amongst the variables highlighted. There is certainly a race and gender effect, with cluster 1, Green, isolating many white males (best lung function), cluster 2, Magenta isolating many white females, and cluster 4, Blue isolating many blacks (worst lung function). However there are also clear purely clinical feature differences, so the race and gender effect, while interesting, does not appear to fully explain the separation of these clusters in clinical space. Furthermore, if we are indeed looking at clusters indicative of subtypes, it would be reasonable to expect the subtypes to affect gender and race at different rates.

Overall 56 variables were still flagged as significant after validating replication in the training set. We remark that numerous PEX features (discrete measures related to particulates) were flagged in the training set, but this failed to be reproduced in the test set. It is possible this is an artifact due to the discrete nature of these variables,

Feature	Train p-value	Test p-value
HT_CM01	0	0
GENDER	0	0
RACE	0	0
AGECAT_BY10	1×10^{-7}	3×10^{-12}
LLN_FEV1_FVC	0	0
LLN_FEV1	0	0
LLN_FVC	0	0
PEX_TOT0101	0	0.5
PEX_HCUTOT0101	0	1
PEX_ANTITOT0101	9×10^{-10}	0.4
PEX_STEROIDTOT0101	0	2
PEX_DRUGTOT0101	3×10^{-8}	0.6
PEX_SEVERETOT0101	0	4
INFECTIOUS_DISEASE_CONDITION01	1×10^{-9}	3
EMPHYSEMA_DIAGNOSED01	3×10^{-7}	0.3
FVC_BDRESPONSE_VOL01	6×10^{-11}	4
BRONCH_DIAGNOSED01	9×10^{-8}	0.2
CBC_LYMPHOCYTE_CNT01	0	0.6
CBC_LYMPHOCYTE_PCT01	0	0
LONGESTJOBDURATION01	0	0.3
EMPH_ULN01	0	0
ANT05	0	0
ANT06C	0	0
PRH09A	2×10^{-12}	2
PRH09C	2×10^{-8}	4
PRH09D	7×10^{-8}	1
CBC07	0	4×10^{-11}
CBC08	4×10^{-6}	0

Table 3.15: Selected p-values for demographic, clinical features (initial).

which can sometimes cause problems with the interpretation of p-values calculated with brownian distance correlation, however it is more plausible that missing values in these data dropped the available data below the corresponding detection thresholds.

For ease of displaying an overview of the result, we group them (a little arbitrarily) together. We create a demographic features subset (Gender, Race, Height, Age), a Physical features subset (Sleep quality measurements, Physical well being measurements), a lung function subset (involving all the LLN measurements), a Bronchodilator Response subset (pre- and post- bronchodilator FEV1, FVC and SVC), a Lung Condition subset (Indicators for the presence of Emphysema, Infectious Disease, Bronchitis), and subsets grouping all the SSV variables together, the PFV variables

Feature	Train p-value	Test p-value
SSV33_DERV	0	0
SSV34_DERV	0	0
SFV31_DERV	0	0
SFV40_DERV	0	0
SFV30_DERV	0	0
SFV45_DERV	2×10^{-8}	0.7
SFV33_DERV	0	0
SFV32_DERV	0	0
SFV34_DERV	0	0
SFV62_DERV	2×10^{-11}	0.01
PSV33_DERV	0	0
PSV34_DERV	0	0
PFV31_DERV	0	0
PFV40_DERV	0	0
PFV30_DERV	0	0
PFV45_DERV	9×10^{-12}	0.4
PFV33_DERV	0	0
PFV32_DERV	0	3×10^{-15}
PFV34_DERV	0	0
PFV62_DERV	0	0.02
SDF03A	0	0
SDF03B	0	0
SDF03C	3×10^{-6}	0.2
SDF03D	0	0
SDF05A	0	0
SDF05B	0	0
SDF05C	9×10^{-12}	0.1
SDF05D	0	0

Table 3.16: Selected p-values for demographic, clinical features (continued).

together, the SDF variables together, and a sundry biomarker subset (including ANT, PRH, CBC, SFH, DEM). As we have a large number of clinically relevant variables, subgrouping them for simplification might be valuable in determining what precisely these subgroups represent.

We may visualize the relationship between these subgroups and biomarker clusters by using multidimensional scaling maps to embed the variations in each subgroup with two dimensions. Representations are shown in Figures 3.15 to Figures 3.23.

We see much structure in these representations. The bronchodilator response and physical subsets in particular, seem to be in the shape of interesting manifolds. There is a clear partition into two sets in LLN variables, that is predominantly respected

Feature	Train p-value	Test p-value
SFH_RP01	0	7×10^{-16}
SFH_SF01	4×10^{-6}	6×10^{-7}
SGR_ACTIVITYSCORE01	5×10^{-9}	0.01
SGR_IMPACTSCORE01	4×10^{-6}	1×10^{-4}
SGR_TOTALSCORE01	5×10^{-9}	4×10^{-4}
PSQ_TOTALSCORE01	3×10^{-15}	0.1
PSQ_SLEEPDISTURBANCE01	2×10^{-6}	2×10^{-7}
PSQ_SLEEPLATENCY01	8×10^{-8}	0.004
PSQ_SLEEPEFFICIENCY01	6×10^{-15}	1
FACIT_PHYSICALWELLBEINGSCORE01	5×10^{-12}	4×10^{-10}
VSASCORE01	2×10^{-7}	6×10^{-2}
DEM02	2×10^{-10}	0.5
DEM04	0	3
RDS10A2	2×10^{-6}	3
RMU02	0	1×10^{-5}
BMH08A	3×10^{-8}	1
SYMBICORT	5×10^{-6}	2×10^{-5}
PREBD_FVC	0	0
PREBD_FEV1	7×10^{-7}	0.5
POSBD_FVC	0	0
POSBD_FEV1	9×10^{-12}	0.1
PREBD_SVC	0	0
POSBD_SVC	0	0

Table 3.17: Selected p-values for demographic, clinical features (final).

by the biomarker clustering with cluster 1, Green and cluster 2, Magenta being split between the two clusters - though of course the choice of a LLN variable as a surrogate likely influences this. Other subspaces hint at more clusters, for example three would be a reasonable extraction from the CBC, ANT, PRH and DEM subspace.

Conclusions: We are confident in a meaningful division of COPD patients in the SPIROMICS study into four clusters determined by biomarkers. The structure and signal is messy, but nevertheless there is structure and signal present. Clinical interpretation and whether these could indeed be indicative of disparate subtypes is now necessary.

Our lessons from this analysis include that clustering noisy data in high dimensions is hard! We have proposed and partly validated a screening method that seems to give clinically meaningful clusters in projected dimensions. As a result of this,

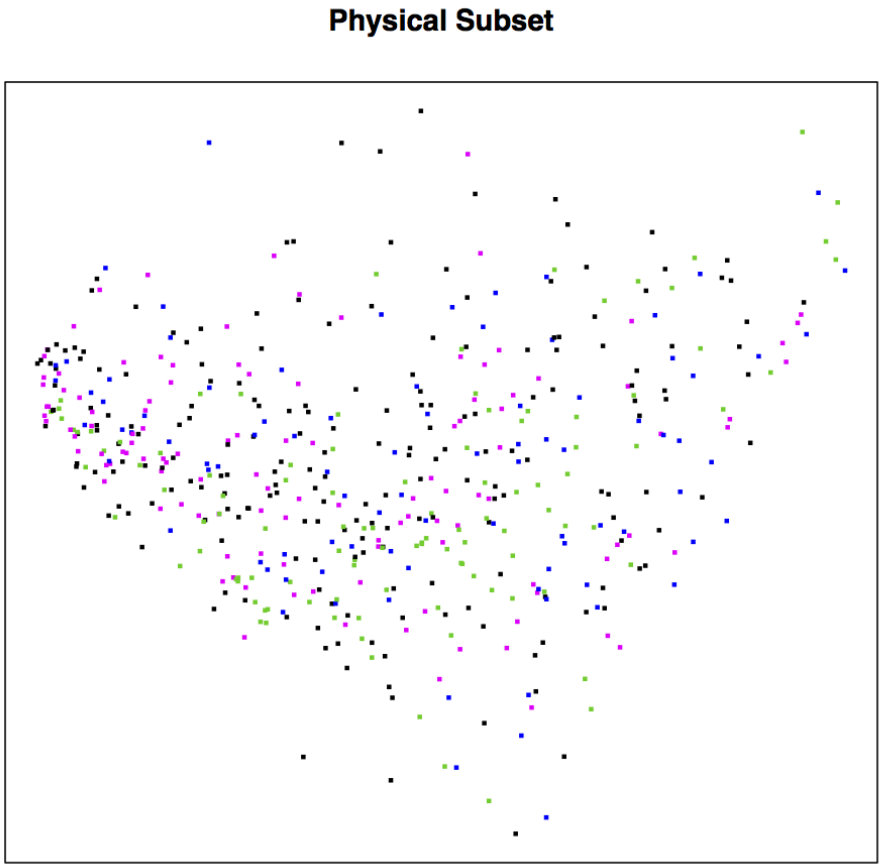

Figure 3.15: Cluster visualization in the space of 11 physical variables.

we have also introduced a new method for selecting the number of clusters in a meaningful data driven way.

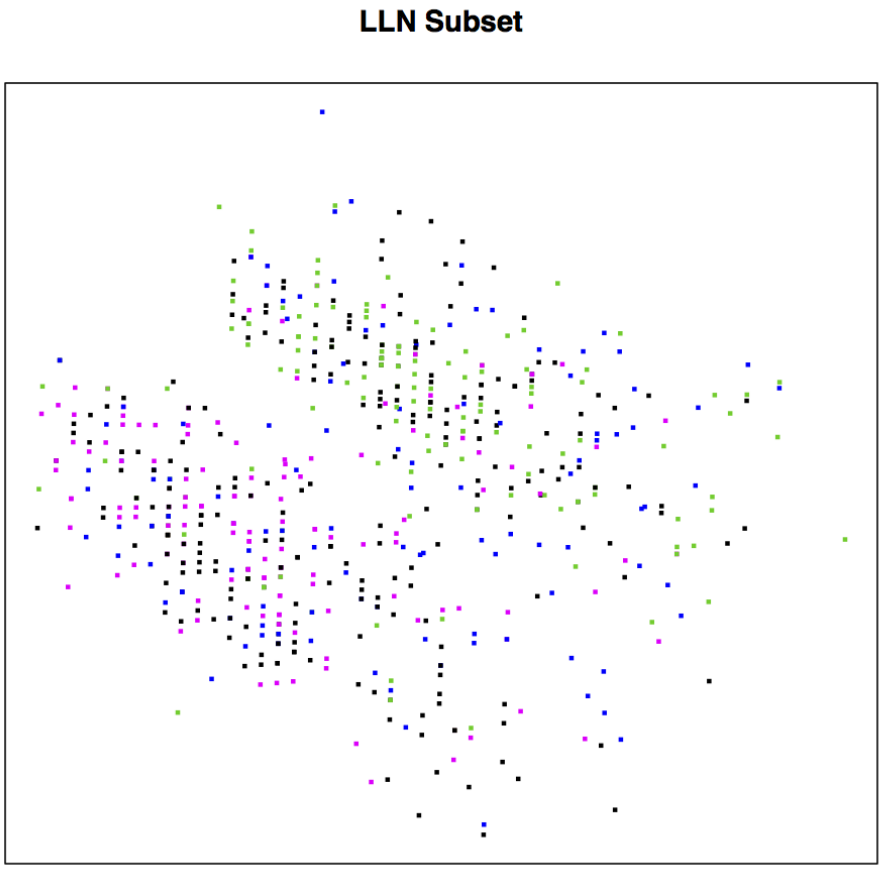

Figure 3.16: Cluster visualization in the space of 3 LLN variables.

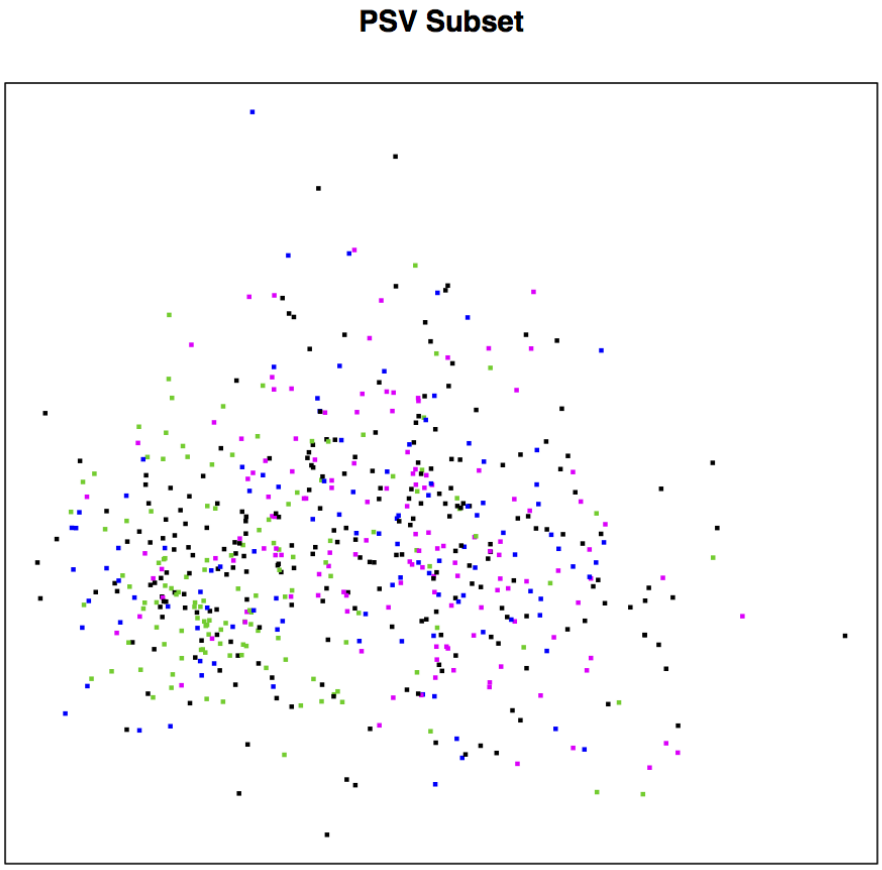

Figure 3.17: Cluster visualization in the space of 2 PSV variables.

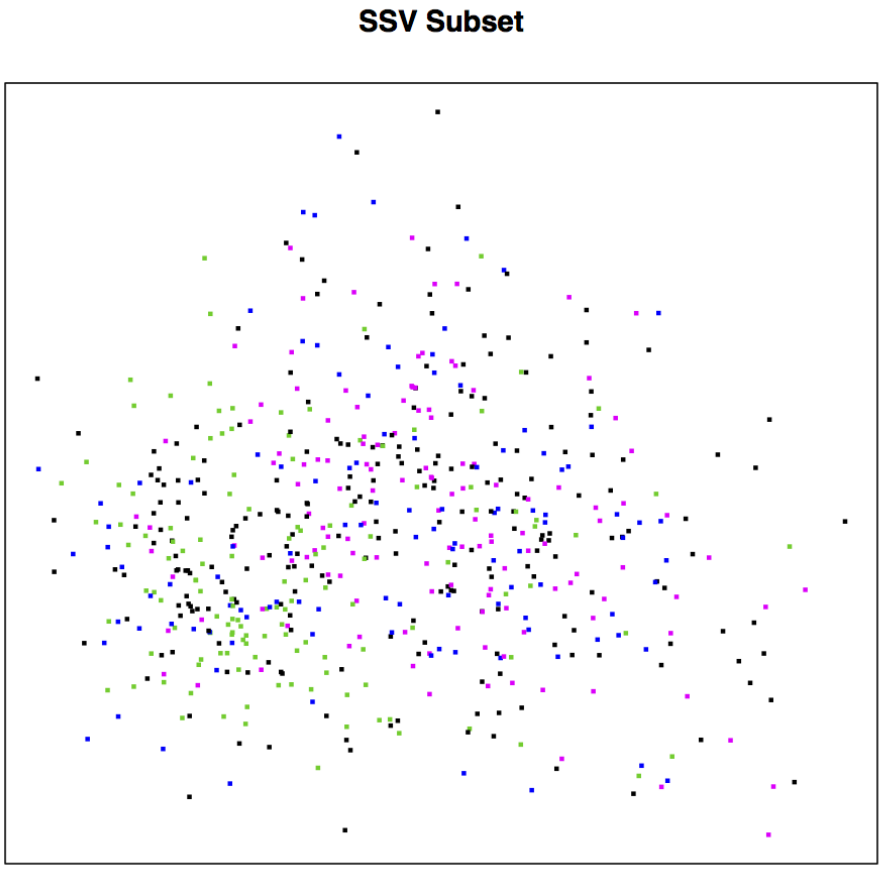

Figure 3.18: Cluster visualization in the space of 2 SSV variables.

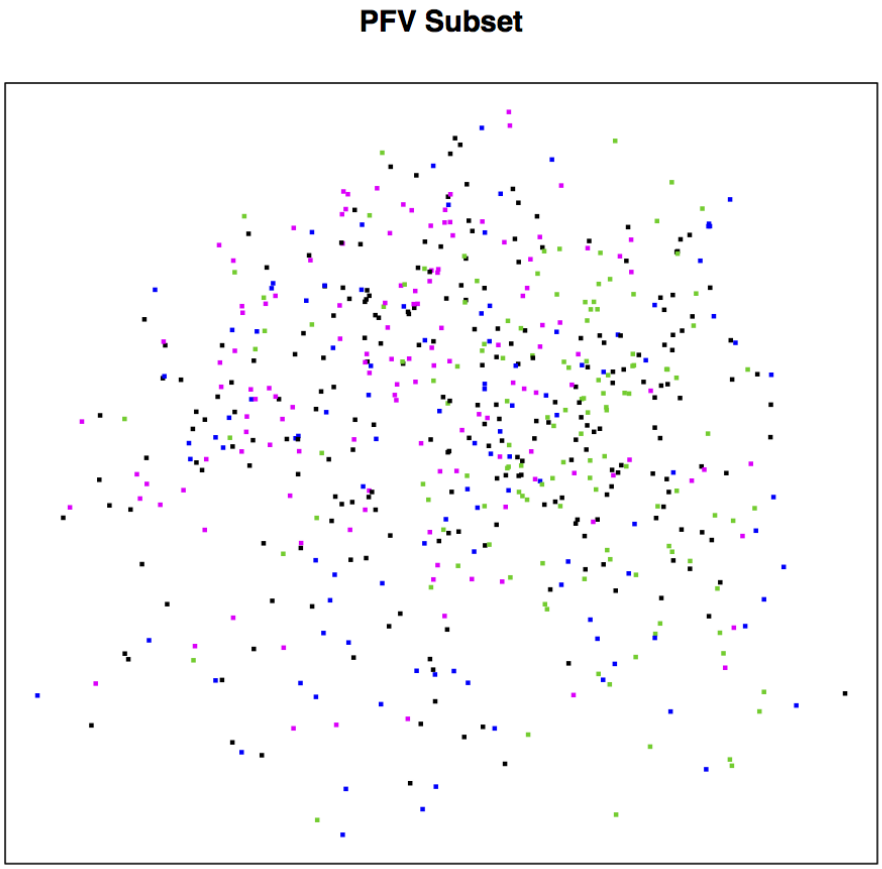

Figure 3.19: Cluster visualization in the space of 9 PFV variables.

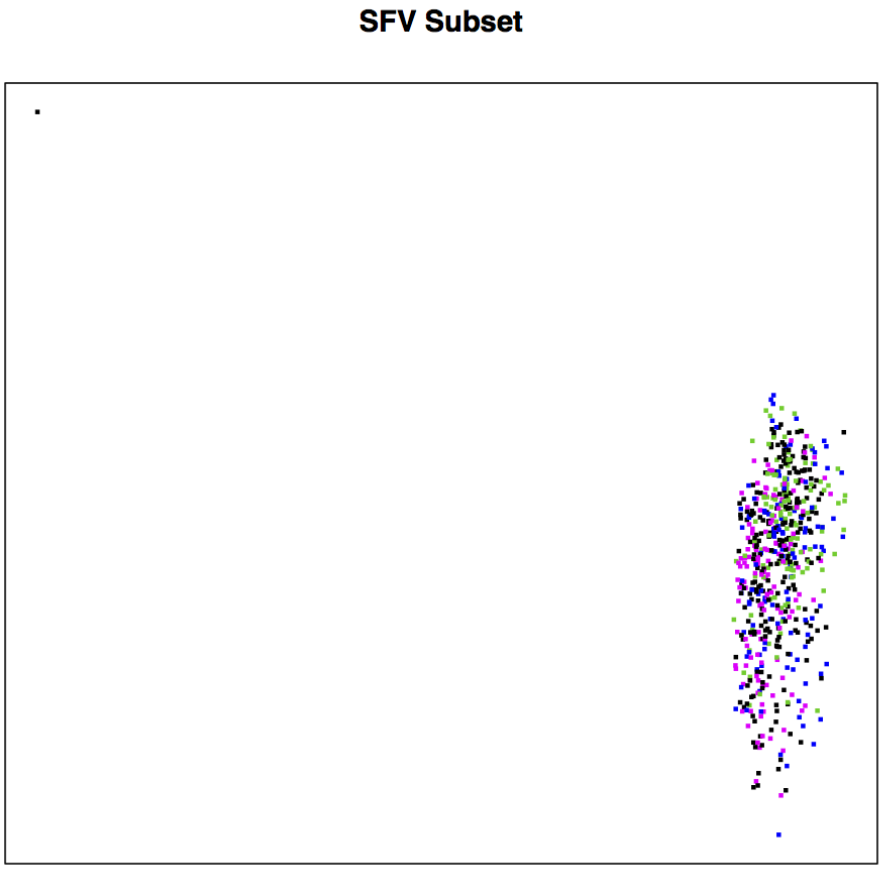

Figure 3.20: Cluster visualization, in the space of 8 SFV variables.

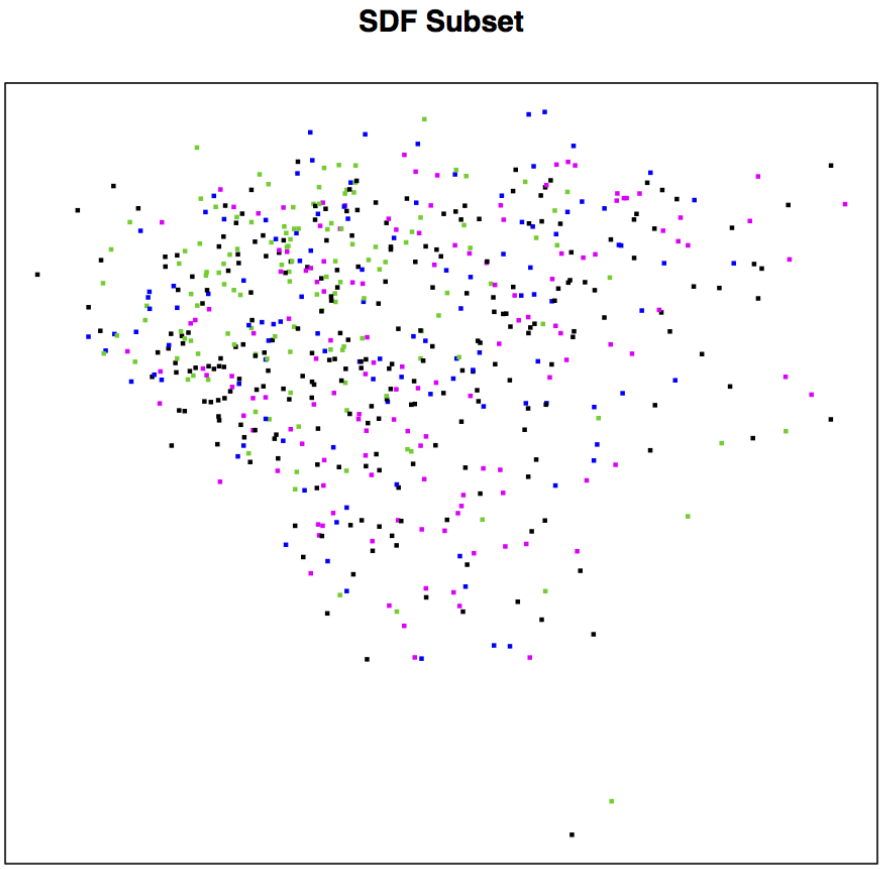

Figure 3.21: Cluster visualization in the space of 7 SDF variables.

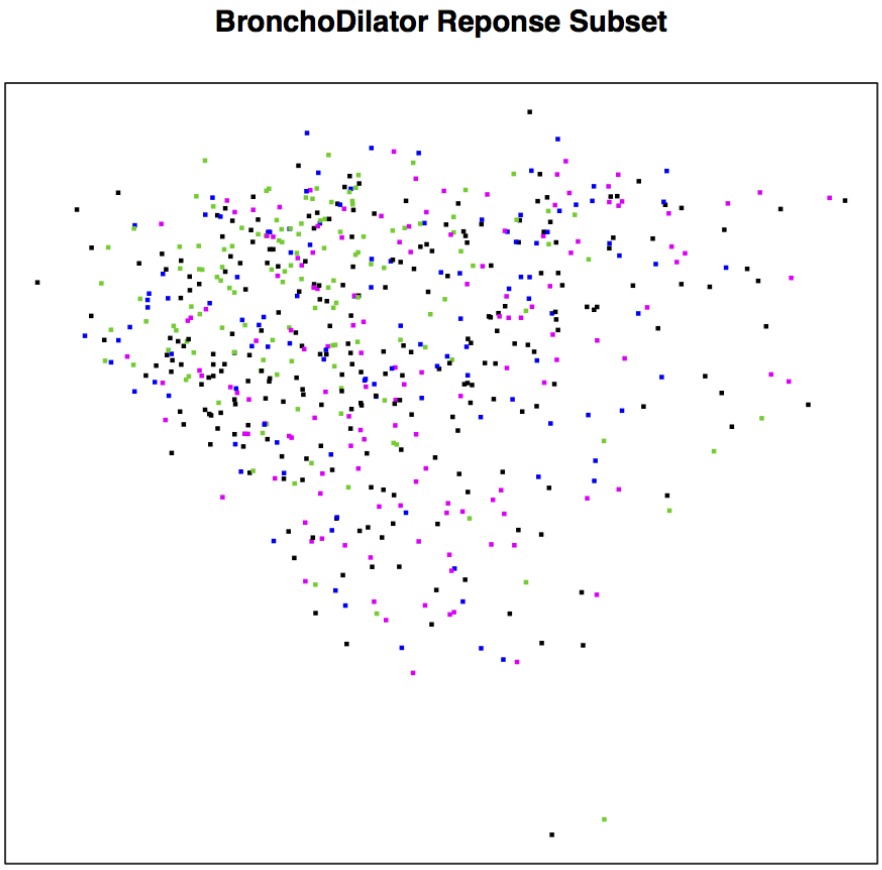

Figure 3.22: Cluster visualization in space of 6 bronchodilation effect variables.

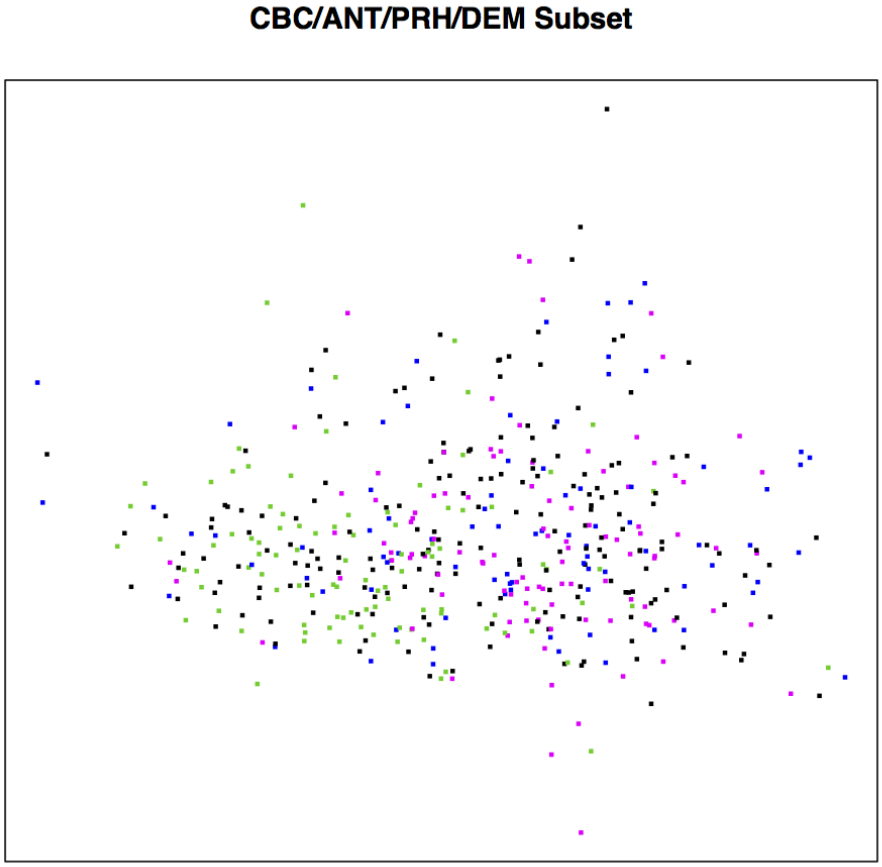

Figure 3.23: Cluster visualization in space of 10 chosen features from the set of CBC, ANT, PHR, DEM variables.

CHAPTER 4: JUMBLED KDES AND MIXTURE MODELS.

4.1 Introduction and Abnormal Distributions.

We detailed numerous possible extensions that might be useful for our surrogate clustering method. We now consider a class in more detail.

The decision to use mixture modelling to attempt to unscramble subtypes seems a very justifiable and defendable one, as we have noted (Walther (2002)). A simple, and well-established, distillation of this idea is to estimate the number of modes of a distribution, and conclude that each of these corresponds to a separate univariate population (Silverman (1981)). However this is problematic in higher dimensions even in simple situations, for as Ray and Lindsay (2005) point out, the property of normal mixture models in one dimension that the modality equaling the number of components is no longer true in higher dimensions. Thus different subpopulations may not be as simply identifiable. However, they consider other functionals that allow recovery of how many components there are, under the strong assumption the mixture truly is a normal mixture. It is thus clear we must be careful in higher dimensions, and our intuition from lower dimensions does not transfer. The same phenomenon precisely in the one dimensional case when the distributions are not normal is the motivation for Walther (2002) to model with log-concave distributions instead of simply mode hunting. Walther also points out that this approach may provide far more power than simple mode hunting would.

It is then clear that we should perhaps be flexible in our choice of distributions.

A different reason to be flexible is pointed out by McLachlan et al. (2002), who note that the heavy tails of distributions of gene expression levels might not be at all suitable for modelling as a gaussian mixture. Indeed the authors propose examining whether t-distributions are a better modelling assumption than simple normals for these situations.

In our work we have often observed, as in the case of the SPIROMICS data, biomarkers may not be assumed to follow gaussian distributions, and in the situations we have considered heavy tails have not been much of an issue. As we noted earlier, one workaround was to transform variables that they might be approximated by normal distributions, such as taking a logarithm when the variable was strictly positive, and hence certainly non-normal, yet specific work-arounds are undesirable for machine learning situations where as much of the decision process as possible should be automated. It seems therefore we should maybe continue with Walther's paradigm, and consider modelling with mixtures of log-concave distributions.

The use of log-concave distributions has many attractive features. If we are modelling mixtures with a class of distributions in higher dimensions (with the features being possibly fairly convoluted or interrelated) then we should hope for some closure properties for certain operations on this set. For example, reasonable properties to require would be that marginal densities are in the class if joint densities are, and addition and multiplication of random variables with densities in the class results in a random variable with density also in this class. These properties are met by the class of log-concave distributions (see for example Bagnoli and Bergstrom (2005)).

The log-concave distributions also form a rather wide class, from normals and chi-squared distributions, to exponential and compactly supported uniform distributions, enabling them to model appropriately many situations. Further the log-concave distributions all necessarily have the restriction that they are unimodal, vital for use in

extracting subpopulations. We thus concur with Walther that log-concave mixtures offer much promise.

4.2 Shape Constrained Density Estimates and Jumbled Kernel Density Estimators.

Modern clustering is usually performed on non-univariate features, which makes modelling with a non-parametric class of distributions fairly problematic. Recently Dümbgen, Lutz and Rufibach (Dümbgen et al. (2009)) introduced and proved uniqueness of the MLE for a univariate log-concave density. Cule, Samworth and Stewart introduced a MLE of a multivariate a log-concave distribution (Cule et al. (2010b)), and studied its theoretical properties (Cule et al. (2010a)). This was a major step forward as it allowed them to use log-concave densities within an EM-algorithm for fitting mixtures. As Cule and Samworth note, the MLE of a unimodal distribution must be undefined, but placing the log-concave restriction allows estimation and identifiability. However we have observed that significant improvement, especially for higher dimensions or low sample size may be obtained simply through another method.

Shape constrained density estimation is a well considered problem, going certainly as far back as Grenander's celebrated MLE for estimation of a monotone density over a finite interval (Grenander (1956a), Grenander (1956b)). What appears to have received less attention however is the possibility of simply tweaking existing, well-understood estimators, such as kernel density estimators (KDEs) to fit the relevant constraints. Suggestions of doing so appear more or less explicitly in a number of papers, such as in Kosorok's consideration of bootstrapping the grenander estimator (Kosorok (2008)), as well as more expansively elsewhere (Birke (2009), Wolters (2012)). However, there seems to be neither much acknowledgement as to the nu-

merical advantages of this technique, nor the theory that can be inherited from KDE theory. We shall highlight the possibility of both of these.

Kernel density estimation is well-established, having been first introduced separately by Rosenblatt and Parzen (Rosenblatt et al. (1956), Parzen (1962)). One feature that makes it a particularly attractive method is the fact that it has optimal convergence properties amongst non-parametric density estimators (Roberts et al. (2001)), as well as the well-developed theory of KDEs in general.

We let $\{\mathbf{X_i}\}$, $1 \leq i \leq n$, be an i.i.d. sample from the random variable $\mathbf{X}$ which has density function $f_{\mathbf{X}}(\mathbf{x})$ with respect to lebesgue measure on $\mathbb{R}^k$. The kernel density estimator with bandwidth b is a map

$$K_{n,\,b(n)} : (\mathbb{R}^k)^n \to \mathcal{D}$$

where $\mathcal{D} = \{f \in L^1(\mathbb{R}^k) : f \geq 0,\ \int f\,dx = 1\}$. Depending on the choice of bandwidth (and actually kernel), which could be data dependent, this might or might not be a linear map. For a non data dependent smart choice of bandwidth it is a linear map, and for appropriate bandwidths (such as those minimizing expected mean integrated squared error (MISE)) it is known that

$$\sqrt{n}\int g(K_{n,b(n)}(\mathbf{X_n}) - f)\,dx \rightsquigarrow_{l^\infty(\mathcal{G})} \mathbb{G}$$

where $\mathcal{G}$ are certain classes of functions and $\mathbb{G}$ is a generalized brownian bridge process indexed by $\mathcal{G}$. Proving such results requires smoothed empirical processes (Giné and Nickl (2008)).

Let $\mathcal{C} \subset \mathcal{D}$ represent the class of densities in which that of X is known (or assumed) to be. For our procedure, we propose choosing some $T : \mathcal{D} \to \mathcal{C}$, then on application of the functional delta method (see for example Kosorok (2007)), we have that

$$\sqrt{n}\int g(T(K_{n,b(n)}(\mathbf{X_n})) - f)\,dx \rightsquigarrow_{l^\infty(\mathcal{G})} T'(f)(\mathbb{G})$$

provided T is hadamard differentiable tangentially to $\mathcal{C}$ at f, and where we have used the fact (assuming a sane choice of T!) that $T(f) = f$. We will name T the correction map.

Therefore we have provided an estimator $J_n : (R^k)^n \to \mathcal{C}$ with precise definition

$$J(\mathbf{X}) = T(K_{n,b(n)}(\mathbf{X_n}))$$

for which we may obtain both consistency and inference results, given the proviso that we may find a good choice for T.

As the new estimator is constructed through simply taking a KDE and deforming it as needed, we refer to this type of estimator as a Jumbled Kernel Density Estimator, that is a JKDE.

4.3 A Monotone Jumbled Kernel Density Estimator.

Example, Grenander estimator: The grenander estimator is a well studied estimator for a monotone univariate density. It is a MLE estimator, resulting in a piecewise constant density function. Inference is challenging, e.g. standard bootstrapping is inconsistent (Kosorok (2008)). We investigate the proposed method in comparison to the grenander estimator. We assume a monotone decreasing density on, without loss of generality, $[0, 1]$. Our choice of kernel density estimator is one that is respectful of this finite support. In order to give a density on $[0, 1]$, we use a kernel density estimator on the transformed data when the mapping $U : \mathbb{R} \to \mathbb{R}$ defined by $U(x) = \log(x/(1 - x))$ is applied, the estimated density is then transformed back to $[0, 1]$. Here we use a kernel density estimator which selects bandwidth to correspond with the sample standard deviation. Specifically we use the density() function in R. There are numerous options for T. We consider two.

Error	N=10		N=100		N=1000		N=10000	
	Grenander	JKDE	Grenander	JKDE	Grenander	JKDE	Grenander	JKDE
L1	0.44(0.17)	**0.31(0.16)**	0.17(0.09)	**0.12(0.040)**	0.08(0.04)	**0.054(0.015)**	0.035(0.009)	**0.025(0.005)**
L2	0.85(0.85)	**0.57(0.47)**	0.65(3.03)	**0.29(0.31)**	0.50(1.48)	**0.12(0.09)**	0.15(0.30)	**0.05(0.02)**
Pointwise	0.27(0.19)	**0.17(0.12)**	0.12(0.09)	**0.08(0.06)**	0.063(0.046)	**0.042(0.030)**	0.031(0.024)	**0.022(0.017)**
Probability	0.15(0.09)	**0.12(0.084)**	0.034(0.027)	**0.028(0.020)**	0.013(0.009)	**0.012(0.008)**	**0.004(0.003)**	0.005(0.003)

Table 4.1: Grenander and JKDE empirical errors for a truncated exponential on [0,1].

1: The first choice of T we consider is a decreasing rearrangement, $T = T_d$, defined by the fact that $T_d f(x) \geq T_d f(y)$ if $0 \leq x \leq y$, and $|\{x : T_d f(x) \geq c\}| = |\{x : f(x) \geq c\}|$ $\forall c$. This is a standard real analysis tool, and maps a given function to a nonincreasing one of same L^p norm for $p \geq 1$. This is simple to calculate, and the complexity is not high when programmed smartly. The main advantage of this choice though is that T is differentiable (in the hadamard sense), with the derivative being simply $\mathrm{id}(x) = x$ provided the true density f has no piecewise constant sections, which is to say $\nexists a, b, c$ so that $f(x) = c$ $\forall x \in [a, b]$. This seems a rather trivial requirement for real life densities.

In Figure 4.1, a comparison of the decreasing rearrangement kernel density estimator and the grenander estimator is given for a truncated exponential on [0,1], for sizes of sample, N=10, 100, 1000 and 10000. Visually, the rearranged density estimator is much more attractive. It seems the improvement is extreme for lower sample sizes, when the grenander is very coarse. Table 4.1 shows the findings of a simulation of 100 samples from this density regarding the performance of the two estimators measured by expected errors in L^1 and L^2 norms, the error in pointwise evaluation at the (arbitrary point) $x = 0.4$, and also the error of estimating the probability $P(X \leq 0.4)$ (again with an arbitrary choice being made for this interval), and corresponding standard deviations. It is immediate that the proposed estimator outperforms the grenander considerably. Although we have not detailed this here, similar improvements are seen when considering median and quantiles of errors instead of expected error and variances.

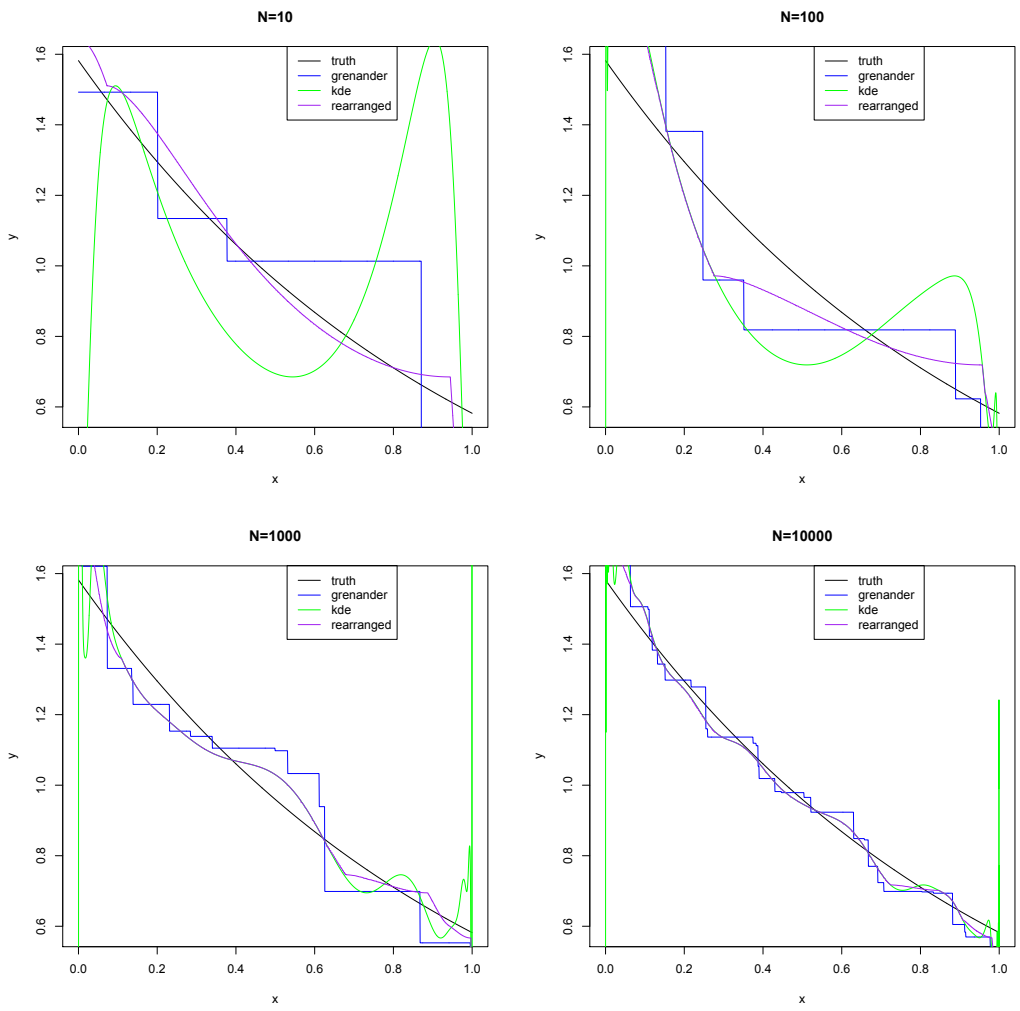

Figure 4.1: Grenander and JKDE density estimates for a truncated exponential on [0,1].

Now we consider the effect of choosing different distributions to generate the data. We examine the distributions that corresponding to a truncated normal of variance 1, a piecewise uniform with jump at 0.4 and 0.7, and a piecewise uniform with 10 jumps (that were chosen at random). We consider having $N = 100$ samples. Simulations of the estimated densities are shown in Figure 4.2 for visual comparison.

2: A second possible choice of T we could consider is simply the map which sends a function to *the* nearest (using the L^1 metric) decreasing positive function of integral 1, $T = T_o$.

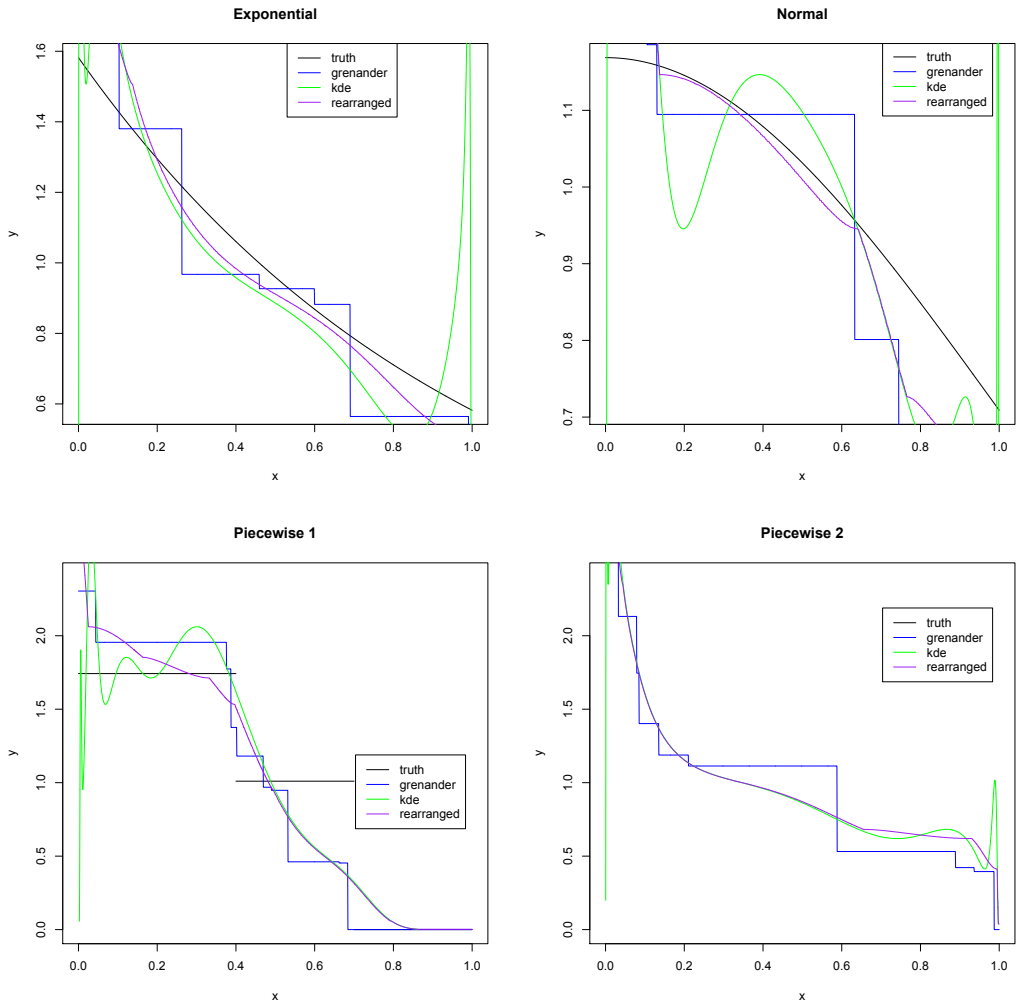

Figure 4.2: Grenander and JKDE estimates for diverse densities.

This may be calculated by using linear programming. For example, we choose a (large) number of quadrature points $u_1, \ldots, u_n$ in the unit interval, at which we shall determine the value of $J(\mathbf{X})$. First we evaluate $v_i = K_{n,b(n)}(\mathbf{X})(u_i)$ for $1 \leq i \leq n$. Denote $w_i = J(\mathbf{X})(u_i)$. All we need do is simply solve the convex optimization:

$$\begin{aligned} \text{Minimize:} \quad & \Sigma_{i=1}^{n} |w_i - v_i|(u_i - u_{i-1}) \\ \text{such that:} \quad & w_i \geq 0,\ w_{i+1} \geq w_i \text{ for each } 0 \leq i \leq (n-1) \\ & \Sigma_{i=1}^{n} w_i (u_i - u_{i-1}) = 1. \end{aligned}$$

Obviously choosing the nearest monotone function using a different norm (such as

L^2) provides a number of other choices. Furthermore a choice of using the L^2 norm has the additional advantage that the resulting optimization is not just convex, but quadratic. We have not pursued the performance of this correction map further in this work.

4.4 A Log-Concave Jumbled Kernel Density Estimator.

There are available estimators for a log-concave density in high dimension (Cule et al. (2010b)). Such estimators are found using MLE methods. There are several issues however with using these. One complication using the available estimators is that the running time for computations becomes impractical in higher dimension, with a possibility of a week running time in say $\mathbb{R}^{10}$, if not earlier. Clearly this hampers the use of log-concave densities, especially in mixture models, and even more so in EM algorithms. Additionally, such estimators do not compare extremely well with kernel density estimators regarding errors, which might indicate that a JKDE as we propose would be superior. Furthermore, as would be expected from an estimator that maximizes likelihood, it is a rather greedy estimator, often underestimating the extent of support, particularly when the signal is sparse, either due to low sample size, or simply the result of the curse of dimensionality.

For us it is an open question as to the optimal map from a KDE to a log-concave JKDE, and likely the optimal map will depend on the metric studied. However a most simple method would be to estimate the logarithmic density as usual, then we can calculate a concave majorant simply from calculating the appropriate convex hull of the interpolation points. Then we apply exponentials and finally normalize to a norm of 1 in L^1. This can be done rather fast: quickhull is one of many algorithms that can do this, and the complexity of quickhull is conjectured to be $n \log(n)$ independent of the space dimensionality, subject to certain so-called 'balance' conditions (Barber

et al. (1996)). It is reported that it is efficient certainly for $\mathbb{R}^8$ and lesser dimensions.

The normalization is not as simple as might be thought in higher dimensions, as the curse of dimensionality requires that a partition for numerically evaluating an integral must contain a very large, and computationally awkward, number of points to produce good accuracy in higher dimension. However this difficulty may be offset by the application of a Monte Carlo integration technique (Morokoff and Caflisch (1995)).

If there is an issue with a spiky density estimate due to a discontinuous density, then it is likely worth considering down-weighting spiky areas by weightings according to the confidence we have in our KDE at that point.

We now compare through simulations the performance of our JKDE for log-concave functions, which we are most interested to apply to the SPIROMICS project, with its competitor the MLE for log-concave functions. We shall perform the comparison in two-dimensions, as firstly the MLE becomes prohibitively expensive computationally in higher dimensions, and secondly we may visualize the performances and obvious pathologies of each method simply in two dimensions. To simulate the sparsity of higher dimensions, and to really highlight the features of the estimators, we will use a small sample size. We also compare these with the uncorrected KDE.

Figures 4.3 and 4.4 show the results of estimating a pdf of a standard two dimensional gaussian and a standard two dimensional uniform on the unit square respectively. We note that the JKDE seems to be clearly the best performing, smoothing out the standard KDE, whereas the simple MLE estimates the support too sparsely. Table 4.2 shows the average results of 335 experiments generating the gaussian distribution with 8 samples, and calculating the L^1, L^2, L^∞, and hellinger norms of the difference between the estimate and the true distribution. The monte carlo standard

deviations over the runs is given in brackets.

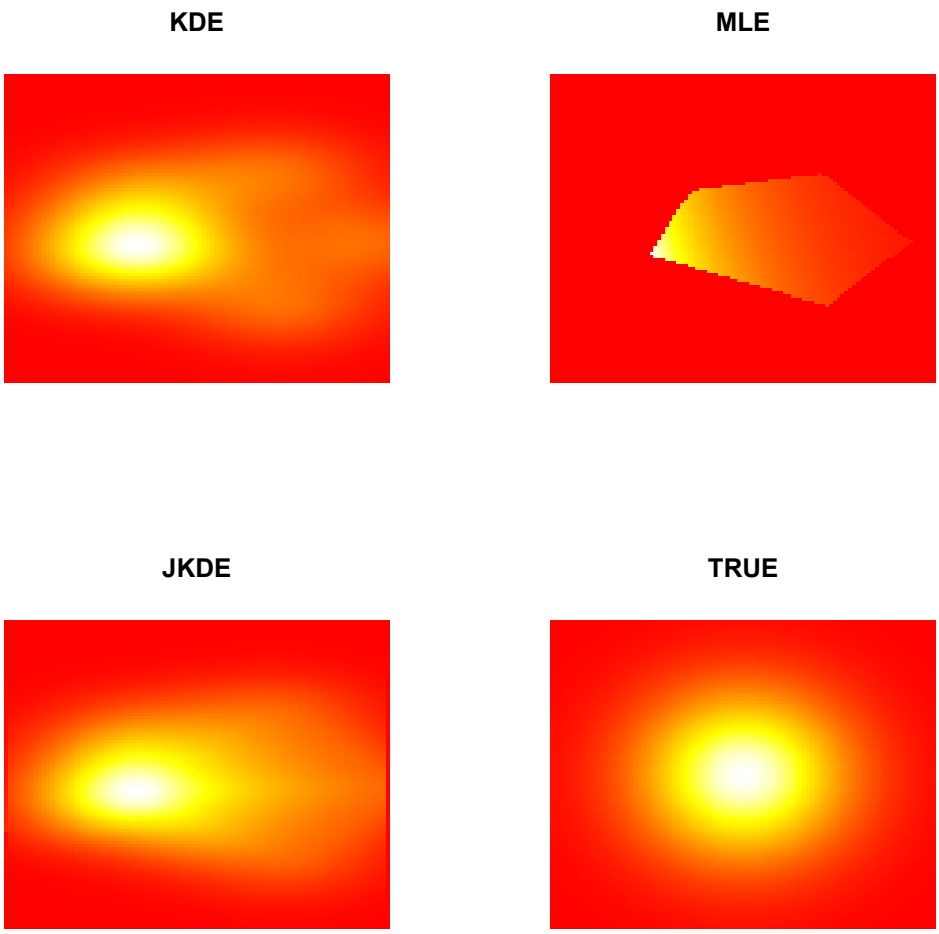

Figure 4.3: Estimators of a 2D gaussian.

We observe that indeed the JKDE performs far better than the MLE in simulation, giving savings in all error norms considered of between 50% and 80% over the errors of the MLE, a very appreciable benefit.

The JKDE also clearly beats the uncorrected KDE, saving between 5% and 10% over the KDE error in the L^1, L^2, L^∞ norms, and a large 45% in the hellinger norm. Quite why the JKDE is so much better than both KDE and MLE in the hellinger

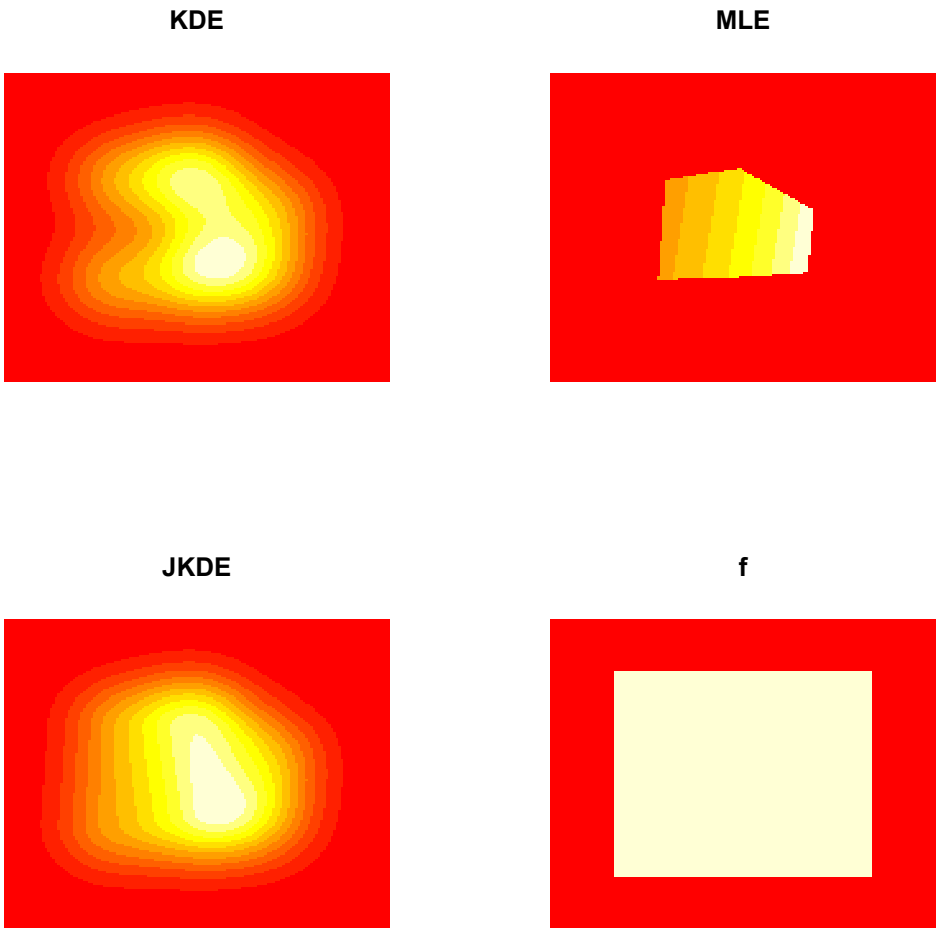

Figure 4.4: Estimators of a 2D uniform.

norm, which we did observe consistently in other experiments, we have not yet understood, even though the explanation might be quite simple.

Even though the KDE and JKDE have fairly similar performances in the L^1, L^2, L^∞ norms, albeit with a definite advantage to the JKDE, other considerations show that the JKDE is a far better and more appropriate estimator in many uses. For example when using the EM-algorithm to fit some mixture of log-concave densities, clearly the KDE estimate simply cannot help resolve which samples belong to which mixture.

Error Metric	MLE	KDE	JKDE
L^1	1.0 (0.21)	0.47 (0.11)	**0.45** (0.11)
L^2	0.33 (0.096)	0.13 (0.031)	**0.12** (0.030)
L^∞	0.35 (0.19)	0.084 (0.023)	**0.075** (0.021)
Hellinger	0.32 (0.080)	0.11 (0.040)	**0.063** (0.026)

Table 4.2: Error estimates for MLE, KDE, JKDE, across 335 simulations.

4.5 Theory.

We now summarize our main theorems that provide the backbone of our theory for JKDEs. We then will show how these may be applied in particular circumstances, in particular with respect to the monotone density operator of section 4.3, and more importantly its higher dimensional analogues.

We begin with a trivial theorem, that exploits asymptotic optimality of KDEs, and that this will not change if one is corrected with a contraction, or projection, map for the corresponding L^p. The JKDE immediately inherits (or perhaps betters) the L^p approximation properties of the KDE, when the underlying density is indeed preserved by the map. Hence, given (under weak conditions) that the KDE has optimal error in both mean integrated squared error (MISE) and also L_1 error, for appropriate bandwidths, the JKDE also inherits this convergence rate. We restate that result as follows:

Theorem 4.5.1. *Let $b(n)$ be a sequence of bandwidths such that $K_{n,b(n)}(\mathbf{x})$ achieves optimal asymptotic convergence in the MISE (or L_1) norms, for some class of densities $\mathcal{F}$. Further let $T|_{\mathcal{F}} = id$ for some restricted contraction map T. Then trivially $J_{n,h_n}(\mathbf{x}) = T(K_{n,h_n}(\mathbf{x}))$ achieves the same, or better, asymptotic convergence rate in the MISE (or L_1) norm.*

While consistency and the optimal asymptotic rate in certain error norms, over appropriate classes of densities come almost for free, we can say more without too

much further work, if we may exploit the differentiability of the correction map. Giné and Nickl, who provide the theoretical framework for proving some uniform central limit theorems and plug-in properties of the KDE with optimal (MISE or L_1) bandwidth. We may extend their theorem without to much effort.

We basically will work from two key theorems, one specific for KDEs (as proved by Giné and Nickl), and the second the standard functional delta method from empirical process theory. Following these authors we define a kernel $K : \mathbb{R}^d \to \mathbb{R}$ of real order $r > 0$ to be a integrable function (with respect to lebesgue measure), such that:

$$K(-\mathbf{x}) = K(\mathbf{x}), \qquad \int_{\mathbb{R}^d} K(\mathbf{x})\, dV(\mathbf{x}) = 1,$$

$$\int_{\mathbb{R}^d} \|\mathbf{x}\|^r |K(\mathbf{x})|\, dV(\mathbf{x}) < \infty \qquad \int_{\mathbb{R}} x_m^j K(\mathbf{x})\, dx_m = 0,$$

where $1 \leq m \leq d$ and $0 \leq 1 \leq [r]$, with $[x]$ denoting the smallest integer below x.

Then the standard kernel density estimator is given by

$$K_{n,h_n}(\mathbf{x}) = \mathbb{P}_n * K_{h_n}(\mathbf{x})$$

where $K_{h_n}(\mathbf{x}) = h_n^{-d} K(\mathbf{x}/h_n)$ and $\mathbb{P}_n = \Sigma_{i=1}^n \delta_{\mathbf{X}_i}$ is the standard empirical measure derived from n i.i.d. samples of $\mathbf{X}$.

The theorem of most interest to us from Giné and Nickl (2008) is proposition 4, which we state (a slightly weaker version of) below:

Theorem 4.5.2. *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be i.i.d. according to the (continuous) law $\mathbb{P}$ on $\mathbb{R}^d$, where $d\mathbb{P}(\mathbf{x}) = p_0(\mathbf{x})\, dV((\mathbf{x}))$, with $\|p_0\|_{L^\infty(\mathbb{R}^d)} < \infty$, then for all translation-invariant $\mathbb{P}$-Donsker class $\mathcal{C}$ of convex sets with uniformly bounded diameter, we have:*

$$\sqrt{n}(K_{n,h_n}(\mathbf{x}) - \mathbb{P}) \rightsquigarrow_{l^\infty(\mathcal{C})} \mathcal{G}$$

where $\mathcal{G}$ is the $\mathbb{P}$-Brownian bridge indexed by $\mathcal{C}$.

Giné and Nickl actually prove this for a wider range of kernels and classes of sets, however these are defined through their extension of the space of functions of bounded variation to higher dimensional space, and their wider framework is unnecessary here.

We now quote the functional delta theorem from Kosorok (2007).

Theorem 4.5.3. *For normed spaces $\mathcal{D}, \mathcal{E}$ let $\phi : \mathcal{D}_\phi \subset \mathcal{D} \to \mathcal{E}$ be hadamard differentiable at θ tangentially to $\mathcal{D}_0 \subset \mathcal{D}$. Assume that $h_n(X_n - \theta) \rightsquigarrow X$ for some sequence of constants $r_n \to \infty$, with X_n taking values in $\mathcal{D}_\phi$, and X being a tight process taking values in $\mathcal{D}_0$. Then*

$$h_n(\phi(X_n) - \phi(\theta)) \rightsquigarrow \phi'_\theta(X)$$

where ϕ'_θ is the hadamard derivative of ϕ evaluated at θ.

Placing these theorems together, and using the fact that our correction map is the identity at the true density, we then inherit convergence of our JKDEs whenever our correction map ϕ is hadamard differentiable. That we have simply deduced:

Theorem 4.5.4. *Let $J_n(\mathbf{X}) = T(K_{n,b(n)}(\mathbf{X}_n))$ be a JKDE, with $\{\mathbf{X}_i\}$ being i.i.d. samples from a continuous bounded density on $\mathbb{R}^d$ which has continuous density $f_{\mathbf{X}}(\mathbf{x})$ with respect to lebesgue measure, then*

$$\sqrt{n}(J_n(\mathbf{x}) - \mathbb{P}) \rightsquigarrow_{l^\infty(\mathcal{C})} T'(X)\mathcal{G}.$$

Thus to show convergence of any proposed JKDE it is simply necessary to examine the derivative of the correction map. With a reasonable choice of correction map though, this should be simple. For example consider our monotone density estimator

of section 4.3. Here the correction map is the decreasing rearrangement operator on $L^1([0,\infty))$. We may prove:

Proposition 4.5.1. *Let $T : L^1([0,1)) \to L^1([0,1))$ be the decreasing rearrangement operator, and $f \in L^1([0,\infty))$ be such that it is strictly monotone decreasing, then T is hadamard differentiable at f, and $T' = id$.*

To prove this proposition, we must more carefully define the decreasing rearrangement operator. The rearrangement operator was introduced by Hardy and Littlewood (1930) in relation to maximal inequalities. As some of their presentation, such as the explanations by appeal to examples from the game of cricket, do not immediately lend themselves to our needs, we redevelop the theory we require.

Definition 4.5.1. We let $\mathcal{P} \subset L^\infty([0,1))$ be the set of functions having the form

$$\mathcal{P} = \left\{ f(x) = \sum_{i=1}^{2^n} a_i 1_{[(i-1)2^{-n}, i2^{-n})}(x) | a_i \geq 0 \right\}$$

and define the operator $T : \mathcal{P} \to \mathcal{P} \subset L^\infty([0,1])$ by

$$T\left(\sum_{i=1}^{2^n} a_i 1_{[(i-1)2^{-n}, i2^{-n})}(x) \right) = \sum_{i=1}^{2^n} \tilde{a}_i 1_{[(i-1)2^{-n}, i2^{-n})}(x)$$

such that the sequence $\tilde{a}_1, \ldots, \tilde{a}_n$ is a reordering of $a_1, \ldots, a_n$ with $\tilde{a}_1 \geq \ldots \geq \tilde{a}_n$.

We immediately may note:

1. $T(f)(x) \geq T(f)(y)$ whenever when $y \geq x$.
2. $\mu(x : T(f)(x) \geq \lambda) = \mu(x : f(x) \geq \lambda)$.

   hence T is a decreasing rearrangement operator on $\mathcal{P}$. Further we have :
3. $\|T(f)\|_\infty = \|f\|_\infty$.
4. $\|T(f)\|_1 = \|f\|_1$.

Proposition 4.5.2. *T is a contraction map operator on $L^p(\mathbb{R}^+)$, that is*

$$\|T(f) - T(g)\|_p \leq \|f - g\|_p$$

for all $f, g \in \mathcal{P}$, and for $1 \leq p \leq \infty$.

Proof: We need only show this for simple functions and appeal to the resulting continuity. Without loss of generality we consider when $p = \infty$. We may proceed by induction. Given $f(x) = \sum_{i=1}^{2^n} a_i 1_{[(i-1)2^{-n}, i2^{-n})}(x)$, $g(x) = \sum_{i=1}^{2^n} b_i 1_{[(i-1)2^{-n}, i2^{-n})}(x)$, by considering $m \vee n$, and a consistent reordering of sequences $a_1, \ldots, a_m$ and $b_1, \ldots, b_n$ we may assume without loss of generality that $m = n$ and $a_1 \geq \ldots \geq a_n$. Then assume $\|f - g\|_\infty = \epsilon > 0$.

The L^∞-norm condition gives us that $|a_i - b_i| \leq \epsilon$ for each $1 \leq i \leq n$. By choice $T(f) = f$. Let $T(g) = \sum_{i=1}^{2^n} \tilde{b}_i 1_{[(i-1)2^{-n}, i2^{-n})}(x)$ and let $k = \arg\max_i b_i$, so then $b_k = \tilde{b}_1$.

We have

$$a_1 + \epsilon \geq a_k + \epsilon \geq b_k \geq b_1 \geq a_1 - \epsilon.$$

Further let $\tilde{f}(x) = \sum_{i=1}^{2^n - 1} a_{i+1} 1_{[(i-1)2^{-n}, i2^{-n})}(x)$, $\tilde{g}(x) = \sum_{i=1}^{2^n - 1} b'_{i+1} 1_{[(i-1)2^{-n}, i2^{-n})}(x)$ where $b'_1, \ldots, b'_n$ is the extant sequence $b_1, \ldots, b_n$ with b_k omitted (that is we have removed the largest terms (a_1 and b_k respectively) in the series expansion definition of $f(x)$ and $g(x)$). Then we have that $T(f) = a_1 1_{[(0,2^{-n})}(x) + T(\tilde{f})(x + 2^{-n})$ and that $T(g) = a_1 1_{[(0,2^{-n})}(x) + T(\tilde{g})(x + 2^{-n})$.

But we have for $1 \leq j < k$ that

$$a_{j+1} + \epsilon \geq a_k + \epsilon \geq b_k \geq b_j \geq a_j - \epsilon \geq a_{j+1} - \epsilon$$

and of course $a_j + \epsilon \geq b_j \geq a_j - \epsilon$ for $j > k$. We may then deduce $\|\tilde{f} - \tilde{g}\|_\infty < \epsilon$. By an appeal to the inductive hypothesis, we have $\|T(\tilde{f}) - T(\tilde{g})\|_\infty \leq \epsilon$, and thus

combining our observations, we have proven $\|T(f) - T(g)\|_\infty \leq \epsilon$, and indeed the entire proposition.

Proposition 4.5.2 implies continuity of T on $\mathcal{P}$, hence T extends to a continuous operator on the completion of $\mathcal{P}$ in L^p. Thus, noting this completion in $L^p([0,1])$ is the entire space whenever $p < \infty$, we have an operator defined $T : L^p([0,1]) \to L^p([0,1])$, for $p < \infty$, which, by appeal to its continuity, also satisfies the properties following Definition 4.5.1 and furthermore Proposition 4.5.2. This may be regarded then as our decreasing rearrangement operator.

The situation for $p = \infty$ is not so straightforward, as $L^\infty([0,1])$ is not separable. We proceed by utilizing the established behaviour of T on $L^p([0,1])$ for large p.

Lemma 4.5.1. $\|f\|_\infty = \lim_{p\to\infty} \|f\|_p \quad \forall\, f \in L^\infty([0,1])$.

Proof: Let $f(x) = \Sigma_{i=1}^N a_i 1_{\alpha_i}$ be a simple function. Then $\|f\|_\infty = \max|a_i| = |a_m|$ for a certain m. We have

$$\|f\|_p = \left(\sum_i^N |a_i|^p \mu(\alpha_i)\right)^{1/p} = (\mu(\alpha_m))^{1/p}|a_m| \left(\sum_i^N \left(\frac{|a_i|}{|a_m|}\right)^p \frac{\mu(\alpha_i)}{\mu(\alpha_m)}\right)^{1/p}.$$

Now $|a_i|/|a_m| < (1-\eta)$ for some $0 < \eta < 1$ and $i \neq m$, hence

$$(\mu(\alpha_m))^{1/p}|a_m| \leq \|f\|_p \leq (\mu(\alpha_m))^{1/p}|a_m|(1 + Q(1-\eta)^p)^{1/p}.$$

Upon letting $p \to \infty$ the result follows.

Hence we simply extend the operator T to simple functions after noting

$$\|T(f) - T(g)\|_\infty = lim_{p\to\infty}\|T(f) - T(g)\|_p \leq \lim_{p\to\infty} |f - g\|_p = \|f - g\|_\infty.$$

Thus T not only extends to $L^\infty([0,1])$ by an appeal to continuity, but the properties of defintion 4.5.1 extend similarly.

Now we consider the differentiability of our T.

Proposition 4.5.3. *The decreasing rearrangement operator that we have defined, T, is hadamard differentiable at f with derivative being $T' = id$ the identity map, provided f is strictly decreasing.*

Proof: Given $h_n \in \mathbb{R}$ and $g_n \in L^\infty$ such that $h_n \to 0$ and $g_n \to g$, first we observe that

$$\frac{T(f + h_n g_n) - T(f)}{h_n} = \frac{T(f + h_n g) - T(f)}{h_n} + \frac{T(f + h_n g_n) - T(f - h_n g)}{h_n}.$$

An application of Proposition 4.5.2, as well as noting the multiplicative property of the definition of T, shows

$$\begin{aligned}
\left\| \frac{T(f + h_n g_n) - T(f - h_n g)}{h_n} \right\| &\leq \frac{\|T(f + h_n g_n) - T(f - h_n g)\|}{|h_n|} \\
&\leq \frac{\|(f + h_n g_n) - (f - h_n g)\|}{|h_n|} \\
&\leq \frac{|h_n| \|g_n - g\|}{|h_n|} \\
&\leq \|g_n - g\| \\
&\to 0
\end{aligned}$$

and therefore $\lim_n \frac{T(f+h_n g_n)-T(f)}{h_n} = \lim_n \frac{T(f+h_n g)-T(f)}{h_n}$, so it suffices to just consider $\frac{T(f+h_n g)-T(f)}{h_n}$. Now given N such that $|h_n| < \epsilon/2$ for all $n \geq N$, then there exists $f_\epsilon(x) = \sum_{i=1}^{2^m} a_i 1_{[(i-1)2^{-m}, i2^{-m})}(x)$ with $a_1 \geq \ldots \geq a_m$ and $|a_i - a_j| > 2\|g\|\epsilon$ if $i \neq j$, such that $\|f_\epsilon - f\| \leq \|g\|\epsilon/2$. Then note that if $x \in [(i-1)2^{-m}, i2^{-m})$ we have $a_m + \|g\|\epsilon/2 < f(x) < a_m - \|g\|\epsilon/2$, and then $a_m + \|g\|\epsilon < f(x) + h_n g(x) < a_m - \|g\|\epsilon$.

As $|a_i - a_j| > 2\|g\|\epsilon$ if $i \neq j$, we then have $f(x) > f(y)$ for any $x \in [(i-1)2^{-m}, i2^{-m})$, $y \in [(j-1)2^{-m}, j2^{-m})$ with $j > i$. Hence if $x \in [(i-1)2^{-m}, i2^{-m})$ then $a_m + \|g\|\epsilon > T(f)(x) > a_m - \|g\|\epsilon$. As a result we certainly have $\|T(f) - f\| \leq \|g\|\epsilon$ for all $n \geq N$, thus this then says that $T(f + h_n g)$ is calculated by taking a rearrangement

on each dyadic interval. We thence obtain $T(f + h_n g)(x) - T(f)(x) = h_n g(y)$ for some y in the same interval $[(i-1)2^{-m}, i2^{-m})$ as x. Letting $h_n \to \infty$, thus $m \to \infty$ and $\epsilon \to 0$ will prove the result.

Finally Propositions 4.5.3 and 4.5.2 together prove Proposition 4.5.1.

Combining Proposition 4.5.1 and Theorem 4.5.3, we then straightaway find for the monotone density JKDE of section 4.3 simply that

Theorem 4.5.5.

$$\sqrt{n}(J_n(\mathbf{x}) - \mathbb{P}) \rightsquigarrow_{l^\infty(\mathcal{C})} \mathcal{G}$$

for all translation-invariant $\mathbb{P}$-Donsker class $\mathcal{C}$ of convex sets with uniformly bounded diameter.

and thus we establish asymptotic properties.

This result is of course univariate, and does not make full use of the theory of Giné and Nickl and their extensions to higher dimensions, but we similar results may be obtained for all well chosen JKDEs. As an example we propose an estimator of a density on $[0, \infty)^d$ for which we expect the density to monotonically decrease in the direction of the axes, that is $f(x_1, \ldots, x_k, \ldots, x_n) \geq f(x_1, \ldots, \tilde{x}_k, \ldots, x_n)$ for all $\tilde{x}_k \geq x_k$. Let T_{x_j} be the decreasing rearrangement operator in the jth coordinate. Then

Proposition 4.5.4. *The JKDE on $[0, \infty)^d$ given by applying the correction map $T(\mathbf{x}) = T_{x_1} \circ \ldots \circ T_{x_n}$ is an estimator for monotonically decreasing densities on $[0, \infty)^d$.*

Proof: We proceed by simple induction on the dimension and liberal use of the fubini tonelli theorem. This immediately gives that $T : L_1([0, \infty)^d) \to L_1([0, \infty)^d)$,

and we have $\|T(f)\|_{L^1([0,\infty))} = \|f\|_{L^1([0,\infty))}$ (indeed it is also trivial to see for L^∞, and standard riesz-thorin interpolation results then give the same for L^p), and also that $T(f)$ is monotonically decreasing as needed.

The simple composition nature of this operator also allows differentiability results to be simply obtained. In particular we shall appeal to the chain rule for hadamard differentiability (see for example Van and der Vaart (1996)) that states

Theorem 4.5.6. *Assume $\phi : \mathbb{D}_\phi \subseteq \mathbb{D} \to \mathbb{E}_\phi \subseteq \mathbb{E}$ is hadamard differentiable at $\theta \in \mathbb{D}_\phi$ tangentially to $\mathbb{D}_0 \subseteq \mathbb{D}$ and that $\psi : \mathbb{E}_\psi \subseteq \mathbb{E} \to \mathbb{F}$ is hadamard differentiable at $\phi(\theta)$ tangentially to $\phi'_\theta(\mathbb{D}_0)$, then $\chi = \psi \circ \phi : \mathbb{D}_\phi \to \mathbb{F}$ is hadamard differentiable at θ tangentially to $\mathbb{D}_0$ with derivative $\chi'_\theta = \psi'_{\phi(\theta)} \circ \phi'_\theta$.*

We now may deduce asymptotic properties for this multivariate JKDE.

Theorem 4.5.7. *Let X be i.i.d. distributed according to a strictly monotone decreasing (coordinatewise) density f on $[0,\infty)^d$. Let $T : L_1([0,\infty)^d) \to L_1([0,\infty)^d)$ be the multidimensional rearrangement map of Proposition 4.5.2, and $J_n(\mathbf{X}) = T(K_{n,b(n)}(\mathbf{X}_n))$ be the corresponding JKDE. Then*

$$\sqrt{n}(J_n(\mathbf{x}) - \mathbb{P}) \rightsquigarrow_{l^\infty(\mathcal{C})} \mathcal{G}$$

for all translation-invariant $\mathbb{P}$-Donsker class $\mathcal{C}$ of convex sets with uniformly bounded diameter.

Proof: We simply have to exploit the sequential composition nature of T in terms of coordinate rearrangement maps T_{x_j}, which are all hadamard differentiable at the true density f. The chain rule of hadamard differentiability thus gives that T will also be and then that $T'(f) = id$.

While as noted, the idea of correcting other density estimates to take account of

shape constraints has been floated before, these are the first kind of such asymptotic results, to the best of our knowledge.

As noted we may prove similar theorems for numerous correction maps. It would be nice though to have unifying theory giving this automatically for entire classes and we continue to study this.

4.6 Example: Wisconsin Breast Cancer Dataset.

Our main interest and motivation in developing the JKDE, especially for log-concave densities, is to investigate the application of multivariate log-concave mixture models to complicated clustering situations, which might certainly be a promising method for the SPIROMICS data from perusal of the provided graphics. We finish by considering the application to the Wisconsin Breast Cancer Dataset, available from the UC Irvine machine learning repository. This dataset was analyzed by Cule and Samworth in their research into the log-concave MLE. These authors perform two experiments between their online preprints and published papers. Firstly they attempt to classify the samples, which are either benign or malignant, using two features relating to cell area (the standard deviation and worst measurements), that is features 13 and 14. The authors examine how unsupervised division into two clusters respects the benign or malignant designation, and compare the resulting performance with the clustering given by using a gaussian mixture. Secondly the authors repeat this but using the first two principle components of all 30 features. We compare the results for normal mixture models with those from the MLE EM-algorithm, and the same algorithm using instead the JKDE.

Figure 4.5 shows the distribution of the samples with respect to the selected area features, denoted by malignancy (pink denoting malignant, cyan denoting benign).

Figure 4.6 shows the same distribution of the samples classified into two clusters using a gaussian EM-algorithm (blue denoting mixture 1, green denoting mixture 2). Figure 4.7 shows the samples classified into two clusters using the JKDE EM-algorithm, initiated from the gaussian mixture (blue denoting mixture 1, green denoting mixture 2).

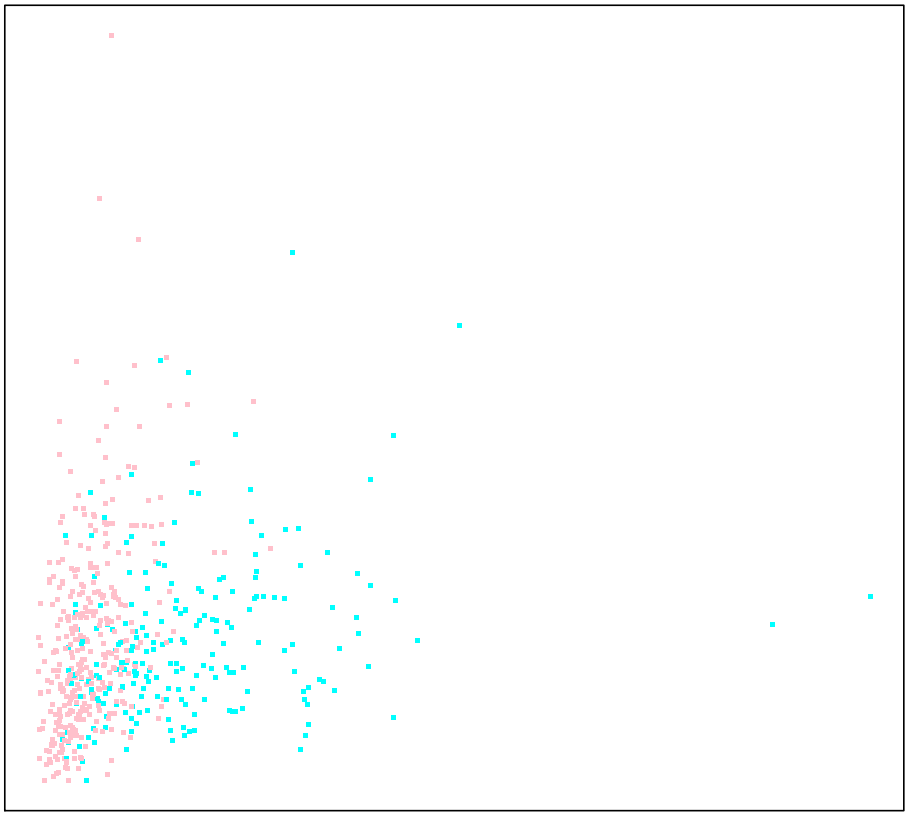

Figure 4.5: Samples displayed according to chosen area features classified by malignancy.

The JKDE EM-algorithm performs significantly better in distinguishing malignancy than the gaussian EM-algorithm, with only 120 misclassifications, as opposed

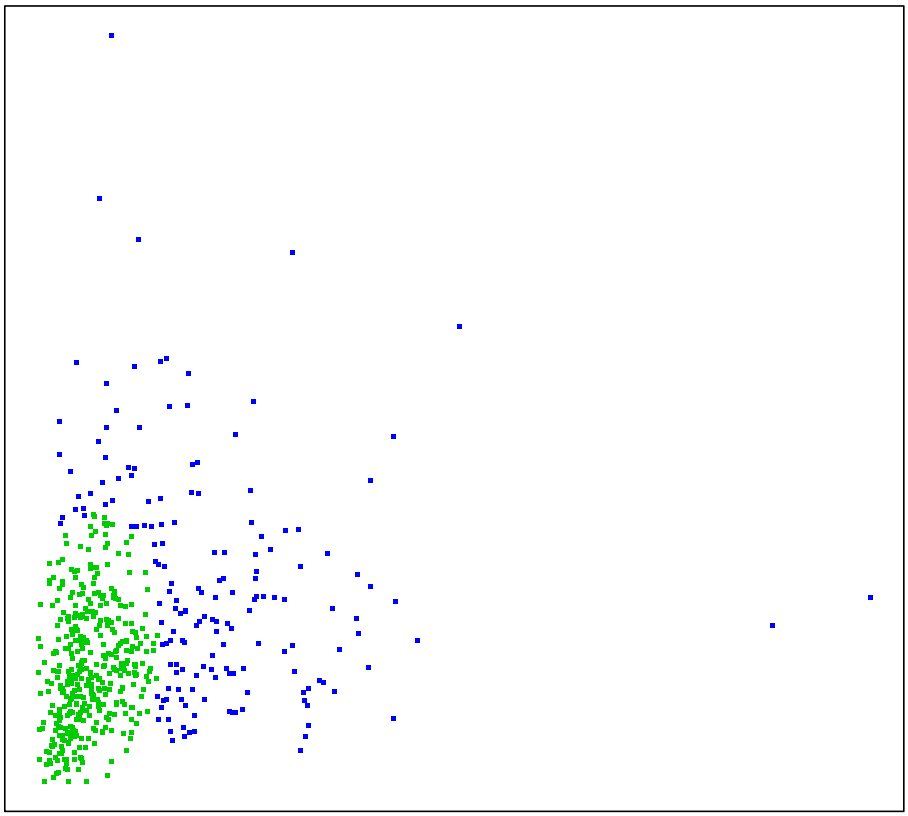

Figure 4.6: Samples displayed according to chosen area features classified by gaussian mixture cluster.

to 144. Samworth and Cule's MLE EM-algorithm has 121 misclassifications, so here there is a small benefit in using the JKDE with respect to misclassifications. How the classification breaks down precisely when normal mixtures are used is given in Table 4.3, and how it breaks down for log-concave mixtures is shown in Table 4.4.

We note that identfying malignancy using normal mixtures has 56% specificity and 86% sensitivity. These rise to 65% and 87% respectively when using log-concave

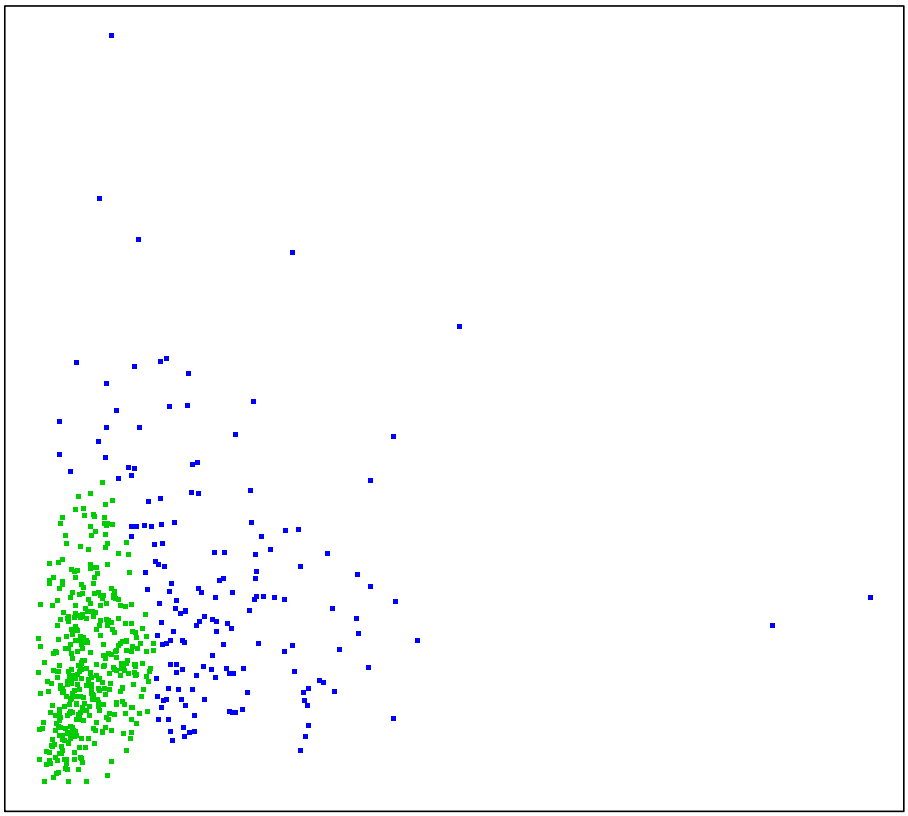

Figure 4.7: Samples displayed according to area features classified by log-concave (JKDE) mixture cluster.

mixtures. There is thus an apprecialbe improvement in specificity.

There may be considerable speed difference in using the JKDE. Secondly we note, as Samworth and Cule did, that indeed a mixture of log-concave happens to be far more appropriate in this scenario than a mixture of gaussians. Certainly we should bear this possibility in mind in clustering work.

Figure 4.8(a) shows the distribution of the samples with respect to the highest

	Malignant	Benign
Cluster I	118	50
Cluster II	94	307

Table 4.3: Breakdown for normal mixture.

	Malignant	Benign
Cluster I	138	46
Cluster II	74	311

Table 4.4: Breakdown for log-concave mixture.

two principle components, after centering and scaling, denoted by malignancy (pink denoting malignant, cyan denoting benign). Figure 4.8(b) shows the same distribution of the samples classified by the JKDE EM-algorithm, initiated from the gaussian mixture (blue denoting mixture 1, green denoting mixture 2).

Table 4.5 summarizes the breakdown of malignancy category between the two estimated mixtures from the JKDE EM-algorithm on the first two principal components. We observe that here the use of principal components do actually improve clustering performance (the use of principal components in clustering being a matter of some debate, see for example Ding and He (2004), Chang (1983)). The JKDE EM-algorithm gives 2 fewer misclassifications than the MLE EM-algorithm (as reported by Samworth and Cule), which is a change in the right direction.

We conclude firstly, with the comment that although the improvement for the JKDE over the MLE EM-algorithm is not huge for these examples, we expect that in higher dimensions, due to the greedy nature of the MLE, this improvement would be

Malignancy	Mixture 1	Mixture 2
Benign	350	7
Malignant	30	182

Table 4.5: Division of malignancy category between 2 clusters from JKDE EM-algorithm on the first 2 principal components.

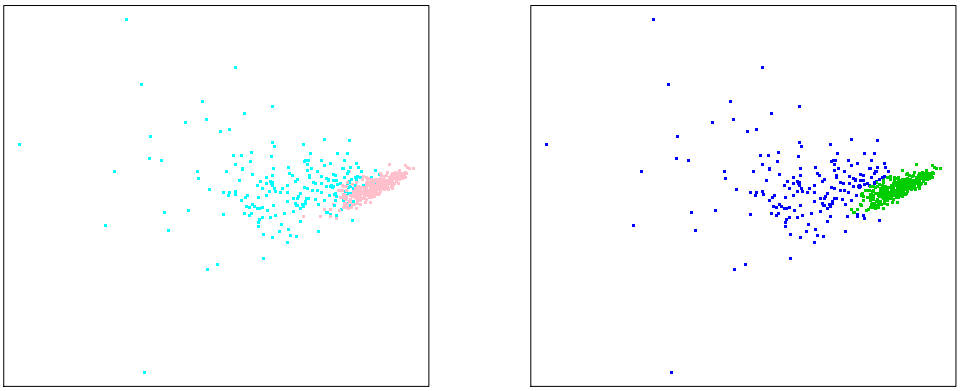

Figure 4.8: Samples displayed according to highest 2 PCs, classified according to (a) malignancy and (b) log-concave (JKDE) cluster.

much more significant. Further the computational saving of the JKDE may be very significant in high dimensions, and thus may allow for the JKDE to be used when the MLE is infeasible.

Secondly we conclude by observing that even though the samples were not well separated according to malignancy in either the two area coordinates, or the two highest principal components, the unsupervised clustering does result in much clinical significance. This example then gives hope that the clustering we obtained for the SPIROMICS data is also relevant, despite the clusters overlapping with respect to their defining biomarkers.

CHAPTER 5: FURTHER WORK.

5.1 LIBERTI.

In chapter 2, we presented a design for a trial that could be viewed as both a SMART and a RCT, that allows powerful comparisons of the effects of numerous different laser sequences on burns scars, the LIBERTI trial. We had been confident, that due to the powerful design, and possible standard interpretations that made IRB approval simple, we would achieve impressive results. Unfortunately, upon starting the trial, a major snag was hit, namely insufficient patient enrollment.

Enrollment for the trial began in January 2016, and efforts were made to get patients into the study through March 2016. At UNC Jaycee Burns Center during this period, 100 new burn reconstruction patients were seen, who might have been candidates for LIBERTI. Of these 68 patients satisfied the initial screening criteria, and were approached to enter the trial. However only 24 patients met the inclusion criteria of speaking english, committing to one year of laser therapy and one year of followup, and had insurance that would pay for the treatments involved. Of this last set of patients, 0 consented, rendering the trial infeasible.

The main problem for satisfying the screening criteria was the fact that burns injuries mainly occur as a result of workplace accidents, and hence often the main provider is Worker's Comp. However, Worker's Comp decided they would not pay for treatments in an experimental trial, even though they would pay for identical treatments not in a trial.

The main issues preventing patients from consenting, could be summarized by three reasons: 'I don't want to be experimented on, or be a guinea pig anymore,' 'I have been poked and prodded enough, I'm done waiting, and want to heal,' and 'I trust you to decide on the treatment Doc, just do what you think best'.

These logistical issues are of course perennial banes of all clinical trials. And even trials that do get off the ground often have low enrollments. Shortreed et al. (2011) reveal that the current average of people enrolled in schizophrenia trials, a prime area for trialing DTR discovery and SMARTs, is only 62 patients per trial. Thus clearly such trials could run into sample size problems, and be under powered, especially when missingness and drop out occurs.

This problem, with a particular reference to SMART trials has been investigate. Ogbagaber et al. (2016) we may data mine electronic health records, but we have then seen how the estimation of DTRs becomes far more complicated if we do not have the unmeasured confounders assumption. Liu et al. (2016) consider a scenario where power can be increased by enrolling patients at a the later stage after they had received out-of-study unrandomized and possibly doctor or patient decided treatment at a corresponding earlier stage. Liu et al. show this may be done in a way not to cause bias. This seems a very promising idea, although of course a seed of patients is still needed to begin the trial no matter the later stages. They call this technique SMARTEr (SMART with EnRichment).

It is of course disappointing these issues were also so apparent when recruiting for the LIBERTI trial, however given the amount of effort gone into designing the LIBERTI trial, and the fact that much set-up has been already done, and IRB approval obtained, we would like to save it, close to its current form if possible. Further a design that addresses these issues for the LIBERTI trial would be translational to many other types of trials, and hopefully serve as a future template.

Repackaging A Trial Through Adaptive Assignments.

We propose as a further research area, determining a design that preserves the aims and abilities of the LIBERTI trial, while addressing either the patient or insurers concerns to ensure sufficient enrollment. The essence of our idea is to repackage the trial, so that it may be also regarded as AI augmented treatment, instead of a trial, and be not just acceptable but indeed attractive to patients, and insurers alike. While considering this idea, we are also led to consider a similar idea which might not make the trial more acceptable to patients or insurers, but would make it more powerful, and hence also more ethical (Altman (1980)).

The classic way to ensure participation is to provide monetary incentives. However sometimes this is insufficient, and sometimes impossible due to cost constraints, as well as arguably unethical (Grady (2005)). Yet for a SMART, the fact that the goal is to find a precision medicine DTR, allows us to proffer a chance for the patients to benefit from this rule while actually participating within the trial.

We proceed considering constructing an ITR from the first stage of a clinical trial, using OWL (Zhao et al. (2012)), in order to elucidate our ideas. As we noted, we believe the superior higher dimensional performance of OWL (Zhao et al. (2012)) means it should be a leading contender to find ITRs in trials such as the LIBERTI trial, where we expect to have many covariates of interest.

Consider a basic trial to find an ITR deciding between two treatments, 0 and 1. As a simulation we assume there are 4 covariates, excitingly named X_1, X_2, X_3 and X_4, and the treatment effects $Y(1), Y(2)$ is given by:

$$Y(0) = N \times X_1 - E \times X_2 + U$$

$$Y(1) = Y(0) + 12X_1 - 9(1 + X_2)$$

with

$$N \sim N(0,1)$$

$$E \sim Exp(1)$$

$$U \sim U[-3,3]$$

being normal, exponential and uniform variables as indicated. Thus we observe only 2 covariates actually have any relevance, and X_3 and X_4 simply represent noise variables. The treatment effects are random, yet linear in X_1 and X_2. The contrast effect is fixed, and the ITR of optimal treatment (assuming positive outcomes are desirable) is given by $T = I(X_2 \leq \frac{4}{3}X_1 - 1)$ where $I(j)$ is an indicator function. We shall assume all four covariates are distributed as independent standard normals.

The basic paradigm proposed by Zhao et al. (2012) is to simply assign treatment at random with equal probability to subjects in a trial, and use this to estimate the ITR. We pretend we have 100 patients in the study, and then 10,000 patients who are not but we must decide the optimal treatment for. Figure 5.1 shows a plot of both trial and test subjects X_1 and X_2 coordinates, and labels these by what the optimal treatment should be: a green '+' for treatment 0, and a purple 'X' for treatment 0. Above each plot the mean Y outcome from following the shown treatment is given, along with the monte carlo experimental standard deviation in parentheses.

Figure 5.2 shows a plot of the effect of assigning random treatment to the test set, and using OWL to calculate the best treatment rule for the training set. We see, random treatment in the test set in effect costs a test patient an average decrease of $1.61 - (-4.97) = 6.58$ in outcome from the optimal. However the treatment set see benefits from this, and on average gain $0.29 - (-4.97) = 5.26$ over random assignments in outcome.

It does not seem entirely ethical the test set see no benefits, but the training set

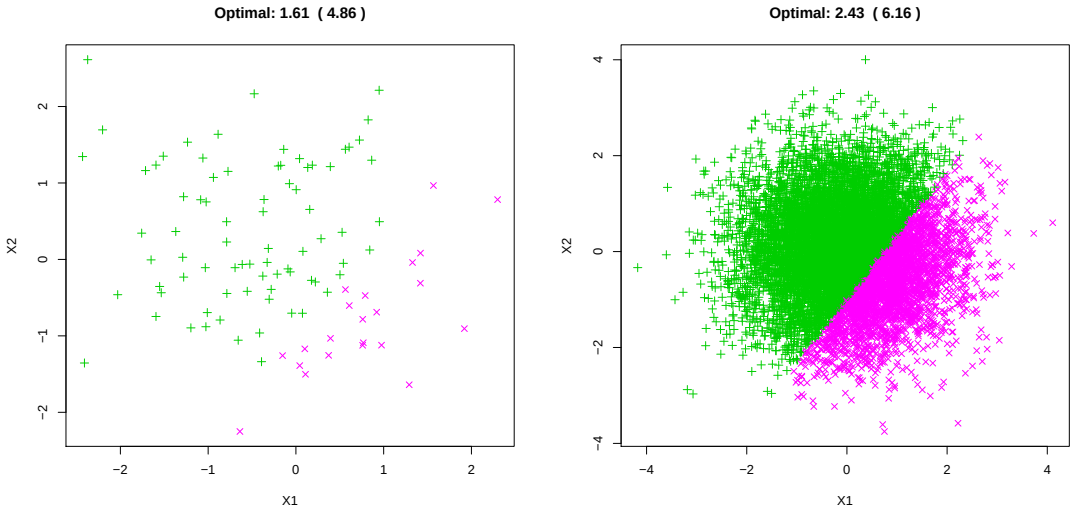

Figure 5.1: Training, test sets of patients, along with optimal treatment.

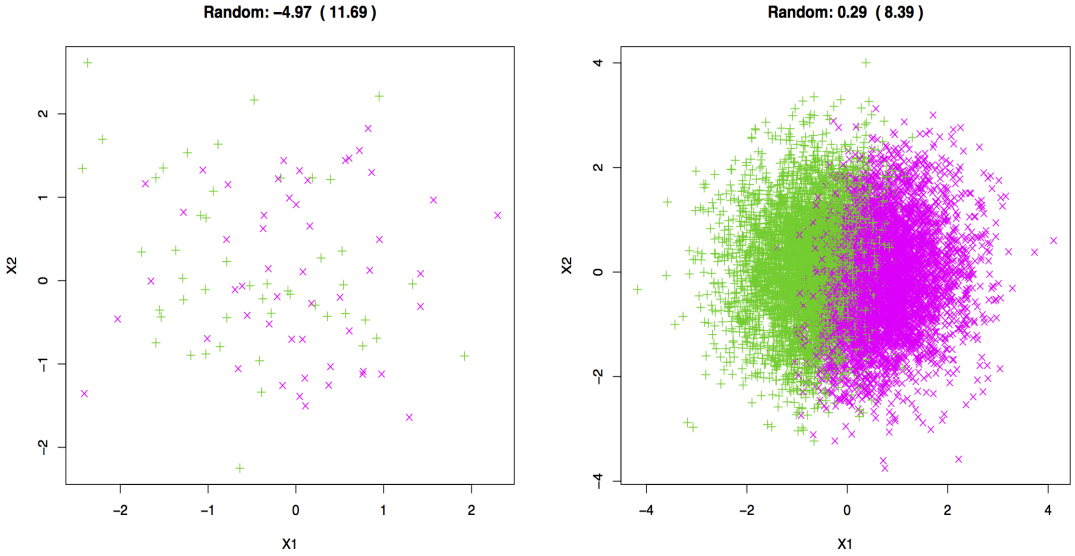

Figure 5.2: Training set with random trt and test sets with est optimal.

will. (Of course as evidenced by Miller and Brody (2003), there is much to debate over what is or is not entirely ethical.) A different assignment procedure we consider could be to attempt calculating the ITR after every patient, then assign the next patient to the predicted best treatment. We call this greedy assignment, and Figure 5.3 displays this. We see that this benefits the training set greatly, who gain on average

$-3.37-(-4.97)=1.60$ more than when treatments are assigned randomly. However the test set, while receiving treatment gains well above both random treatment and the greedy test set, do not now see as much of a gain as the random training set.

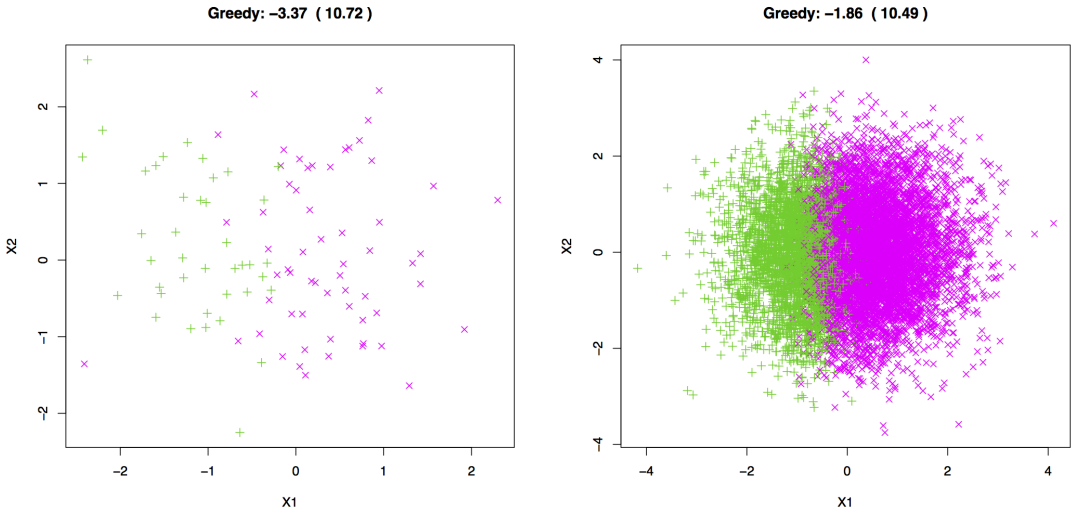

Figure 5.3: Training set with greedy trt and test set with est optimal.

It could be considered whether the greedy assignment is indeed too greedy. A compromise algorithm would be to assign test patients to treatments we knew were going to be beneficial, if we had confidence in their advantage, and if we had little confidence, then instead simply randomize. In the setting of OWL, one way of doing this is given a new patient, check whether or not they are within the soft margin of the classifier found with support vectors. If they are, we say we are not confident as to their treatment and they are assigned randomly, but if they lie outside, we say we are confident in their best treatment choice and act accordingly. We call this the trade assignment. The results of this process are shown in Figure 5.4. We see that surprisingly the test set has a better result on average under this assignment algorithm than it does for the greedy algorithm, gaining $-2.78-(-3.37)=0.79$, although this difference may not have significance. Less surprising, the estimated optimal ITR benefits the training patients, over what would have resulted from using the greedy

algorithm, with a patient gaining on average $-1.54-(-1.86)$=0.29, although once more this might not be of significance.

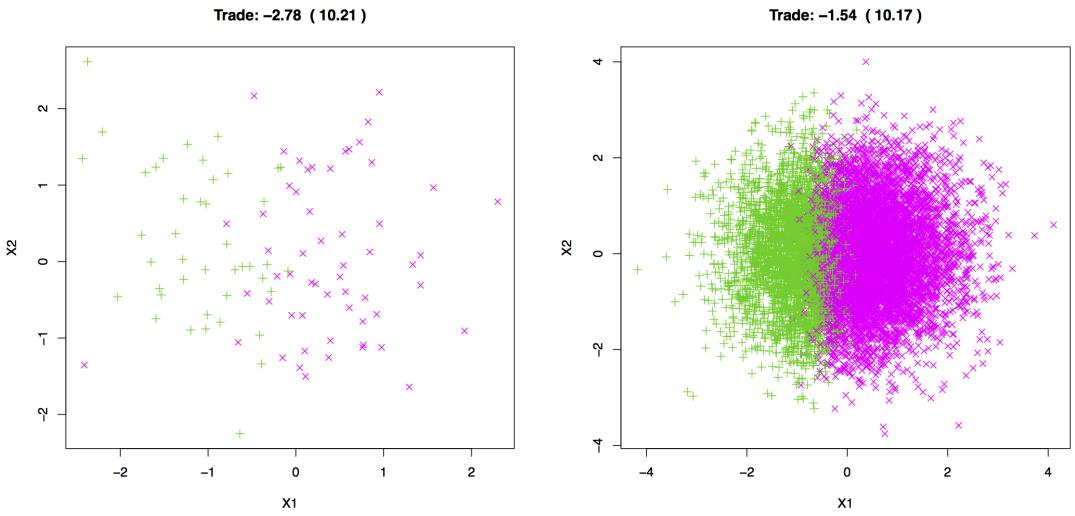

Figure 5.4: Training set with trade trt and test set with est optimal.

We therefore see that by altering the randomization probabilities through the trial, the test subjects can also see significant benefits from being in the trial. We propose to research this to find the optimal, or how one might define optimal, compromise between maintaining power of the trial, and giving benefit to those involved. We intend to prove consistency results, and asymptotic convergence rates.

Ultimately, it seems an aim would be one on-going trial, so that the optimal treatment rule is constantly updated. The mechanics of how to do this most efficiently are not clear, and the theory would have to be developed. The main issue naturally is that the randomization is no longer independent of patient covariates, or indeed previous subjects and their outcomes. It would seem a propensity score solution, to model how likely a patient was to have had been given a certain treatment, would be achievable. While we are most concerned about applying this for OWL, we hope the framework should be transferable to other methods.

It is of interest to investigate how propensity modeling affects the OWL performance anyway, independent of randomizing patients to a likely more beneficial treatment. Indeed van der Laan and Petersen (2007) have observed even for a totally randomized trial, superior performances may be obtained by estimating the propensity for the analysis. Extending this to OWL, which is of course different from many regression based methods, would be of great interest.

Another interesting investigation, would be ongoing dimension reduction throughout the trial. As an example in our LIBERTI trial, we start with a very large number of covariates that might, or might not be influential on the final optimal ITR. By varying the randomization to provide coarse information with regard to some of these dimensions sequentially, it may be possible to make good decisions about dropping these dimensions, reducing the collection of possible expensive to measure variables. Once more, it would be vital to give a consistency proof.

Alternatively, rather than attempting to assign test subjects to a treatment they may benefit from, the experiment could assign them simply to the treatment which gives most information to the investigator. For simple examples, the treatment assignment could be made to ensure an even distribution throughout the covariate space of all treatments. This might be very useful in higher dimensions in which we hope to use OWL.

We also consider methods of interpolating to see whether a new patient would give more information to the ITR estimation from being assigned to one treatment rather than the other. As a simple exploration of this idea, we return to the previous simulated example. For every new patient, we approximate using very basic nearest neighbor interpolation (see for example @articlealtman1992introduction) the effects of treatment, and assuming that effect, we see how much the ITR changes when they are assigned to one treatment than the other. We will call the method of selecting

the assignment we predict to alter the ITR most, (that is essentially gives most information), Info Assignment. The outcome of this is shown in Figure 5.5.

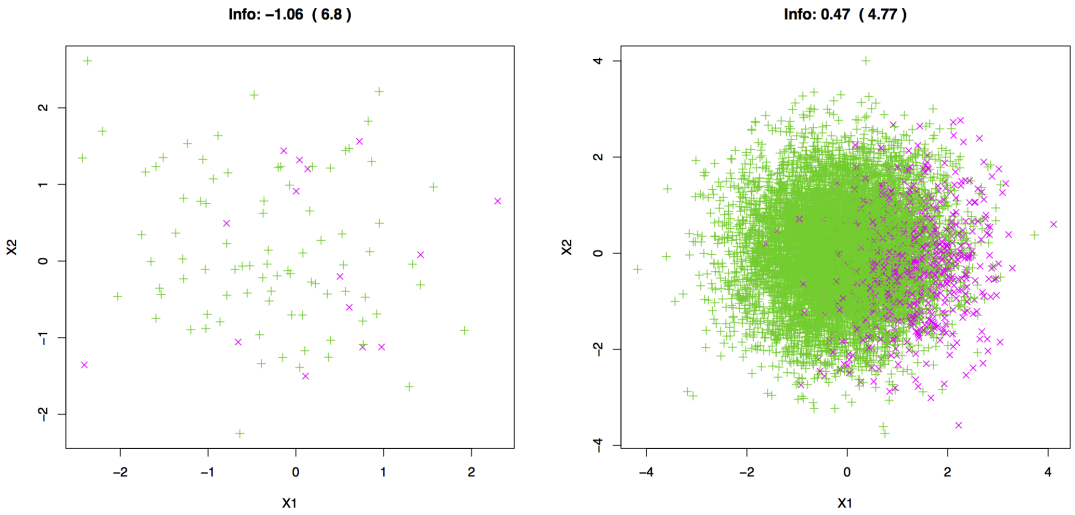

Figure 5.5: Training set with info trt and test set with est optimal.

We see that for Info Assignment, the final calculated ITR is superior to the random assignment final ITR, as we had hoped, by on average an increase of $0.47-0.29 = 0.18$, although this might not have much significance. However the startling observation is that the test set outcomes are far better for the Info Assignment than the Greedy or Trade Assignments and definitely the random assignment, which patients on average achieving benefits of $-1.06 - (-4.97) = 3.91$. We will investigate further much more advanced and sophisticated ways of finding this information measure, which necessitates considering what metrics should be used to quantify the possible change in ITR. We hope to see what then the properties of using this assignment regime are on the test set, and how much improvement could be possible.

In summary we are most interested in a somewhat classical question of amending using previous patient information to judge possible treatment benefits for the next patients, and also, unique to the SMART framework, using previous stage information to judge possible treatment benefits for the same patient at subsequent stages, to

firstly maximize benefit to each patient participating in the trial, while secondly maximizing the information gained from each patient by judicious assignment. These aims of course might be contradictory, and we shall investigate the tensions between them.

The idea of such information sampling as we term it, that is under some working model, assigning the treatment to constantly update and maximize the analysts knowledge of the underlying mechanism, needs to be solidified. For example, we could proceed under a proposed fixed model with the aim of, given firm belief in this model, maximizing patient benefit under it. Alternatively we could proceed by exploring an appropriate covariate space so that the covariants involved in the model are continuously selected for effect size and importance during the trial, in order to streamline both the model and also the logistics and data collection of the trial. A further idea would be to allow the complexity of the model to grow with patient accrual, at a pertinent rate until appropriate information and modelling is achieved.

Possible practical benefits of a successful project are the improved speed and costs of a trial, as well as better ethics and patient participation and easier accrual and enrollment. We aim to investigate the advantages in terms of each of these.

We expect the work to have connections to, amongst other areas, adaptive learning (Minsker et al. (2016)), reinforcement learning (Zhao et al. (2009)), thompson sampling (Chapelle and Li (2011), Agrawal and Goyal (2012)), multi-armed bandit questions (Vermorel and Mohri (2005)).

Relevance To LIBERTI.

Returning to consider LIBERTI, we are not exactly in the set up we have described with a single stage ITR to be found. Nevertheless we have confidence we can extend

these ideas to improve this trial and offer patients the chance to have an AI method calculate a better treatment than they would have received from the current standard of treatment. The notions need to be tweaked a little. For example the fact that there are multiple stages offers even the initial patients a chance to benefit, as under simple, reasonable assumptions of treatment effect additivity, we could model the effects of treatments two and three by that of treatment one, so that when the initial patients move through the trial, enough patients have complete the initial treatment, allowing us to predict a better treatment for the initial patients at the later stages, benefitting all participants.

Of course this is a rather lofty goal, and in reality it would be logistically challenging to collect the data, perform the appropriate analysis and implement the new assignment procedure. This would all take considerable time, and as our LIBERTI trial showed, even the 'simpler' trials may fail logistically. However it would be very interesting to contemplate what may be done theoretically and how this may be practically implemented to varying degrees.

5.2 SPIROMICS.

The work we present in SPIROMICS is just the tip of the iceberg as to what we believe should and indeed may be done for this clustering and subtype detection paradigm. In particular this was an exploration to familiarize ourselves with the thought processes and logical deduction framework that go into such analyses. The ultimate goal would be to fully implement this in a pure machine learning fashion. There are numerous ways we can proceed to bolster the results of this work and explore further.

At the expense of a direct clustering, Latent Supervised Learning (Wei and

Kosorok (2013)) might give a better division in high-dimensional covariate space. The issue of high-dimensional noise would need to be tackled, perhaps by adding a lasso penalty or other standard method (Meinshausen and Yu (2009)).

Biclustering is a term for finding multiple possibly overlapping clusters sparse in both patients and features (Cheng and Church (2000)). This topic is very closely related to clustering, yet has also a separate literature of its own (Prelić et al. (2006)). We calculated a clustering driven by one surrogate variable. It is possible to do this, and evaluate and rank the results, for varying clinical variables. This would give us a biclustering which might yield more information.

The biclustering approach alluded to might not give distinct biclusters, rather using different surrogates might identify repeated clusters with varying accuracy. Application of a super-learner method (Van der Laan et al. (2007)) could take advantage of this to firm up cluster definitions and combine the results from many surrogates to denoise.

Whether or not successful in fully adapting our process to a machine learning paradigm, certainly using machine learning classifiers, having found the clusterings in low dimensional biomarker space, we could extend to find corresponding rules back in higher dimensional space. It would be interesting whether an approach through lower dimensional subspace to obtain the signal, and then cross-referencing in high dimensional subspace to improve the result is possible.

If it does appear, after consultation with clinicians, that race and gender effects are too prevalent, and are non-informatively driving the definition of the clusters, we could regress these out from the biomarkers and surrogate, before performing similar analyses (Zhang (2003)). Similarly we could simply perform subgroup analyses, although we would likely have power issues (Vollmer et al. (2001)).

We have the option to bring data back in from the non COPD patients to see if the COPD clusters represent significantly different pathologies. This could be linked to any subgroup analyses.

We should investigate the stability of our results when e.g. subsampling techniques are applied. Indeed stability measures have been proposed as metrics of cluster validity, although their use is somewhat debatable (Lange et al. (2004), Wang (2010)).

We did not use missing data as much as we might due to inference problems, yet this might provide much information if it could be brought into our analysis (Zhang and Chen (2003)). Multiple imputation might be fruitful, although it is known that clustering becomes fairly unstable in the presence of even small amounts of missing data.

We only calculated the differences of the clinical variables between clusters. By our stated logic, it might be more germane to look at differences of residuals when the clinical variables are regressed over clusters. This runs the risk of being too cautious - at this point the signal might be rather weak - but would be worthwhile to check.

5.3 Jumbled Kernel Density Estimators.

The JKDE technique is clearly generalizable. Other maps of the estimated KDE may be considered, to achieve, for example, unimodal or more generalized unimodal densities as explored by Wolters (2012). There is much room for expanded methods, especially in higher dimensions. Estimators of unimodal densities in high dimension are not readily available. The methods we have provided work very well for the construction of a unimodal in any dimension. Use of the kernel density estimator bypasses the issue of no MLE existing, and the need of restrictions to cope with this

if we should attempt to use methods founded on likelihoods.

For ease of unimodal estimation though, it should be assumed the mode is at a sample point, as has been considered in other methods, for example in using the grenander estimator to find a unimodal estimator (Meyer (2001)). Naturally this would make theoretical results a little more difficult. The order of the algorithm would be increased by a factor of the resolution if the simplest algorithm was applied.

We could further consider a local unimodality restriction. For example a distribution constant on a curved manifold, may be considered unimodal if an appropriate transformation of $\mathbb{R}^k$ is used. Another generalizing distribution would be star shaped, so that radiating from the center on any vector gives a unimodal density. Such constraints are also discussed by Wolters (2012). Local unimodality conditions are simple to include by only requiring pertinent unimodality conditions for only the nearest quadrature points, and would allow construction of curved unimodal and star shaped densities.

This method does not solely need to be linked to KDEs, but a choice of any initial estimator of the density could be made, followed by application of our T to generate a density of the appropriate restriction. One upshot of this, is for example when working in $\mathbb{R}^k$ for sizeable k. We remark that obviously kernel density estimators perform increasingly badly in higher dimension. This, though, is not a failing of KDEs alone. MLEs equally have this problem, if not more so due to greedy behaviour and sparsity. However there are some ways to estimate densities (making for example the assumption it is elliptic, or through wavelet theory) which have (certain) errors independent of dimension (Liu and Wong (2014), Battey and Linton (2014)). Maybe these could be the initial estimate.

All of these should provide interesting applications to machine learning clustering

in higher dimensions.

BIBLIOGRAPHY

Agrawal, S. and Goyal, N. (2012). Analysis of thompson sampling for the multi-armed bandit problem. In *COLT*, pages 39–1.

Akli, E., Marinaki, L., and Halazonetis, D. J. (2015). Selecting subjects with high craniofacial shape homogeneity for clinical trials. *American Journal of Orthodontics and Dentofacial Orthopedics*, 148(6):1026–1035.

Almirall, D., Compton, S. N., Gunlicks-Stoessel, M., Duan, N., and Murphy, S. A. (2012). Designing a pilot sequential multiple assignment randomized trial for developing an adaptive treatment strategy. *Statistics in medicine*, 31(17):1887–1902.

Almirall, D., Nahum-Shani, I., Sherwood, N. E., and Murphy, S. A. (2014). Introduction to smart designs for the development of adaptive interventions: with application to weight loss research. *Translational behavioral medicine*, 4(3):260–274.

Altman, D. G. (1980). Statistics and ethics in medical research: Iii how large a sample? *British Medical Journal*, 281(6251):1336.

Alyass, A., Turcotte, M., and Meyre, D. (2015). From big data analysis to personalized medicine for all: challenges and opportunities. *BMC medical genomics*, 8(1):33.

Arno, A. I., Gauglitz, G. G., Barret, J. P., and Jeschke, M. G. (2014). Up-to-date approach to manage keloids and hypertrophic scars: a useful guide. *Burns*, 40(7):1255–1266.

Bagnoli, M. and Bergstrom, T. (2005). Log-concave probability and its applications. *Economic theory*, 26(2):445–469.

Barber, C. B., Dobkin, D. P., and Huhdanpaa, H. (1996). The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software (TOMS)*, 22(4):469–483.

Barber, R. C., Chang, L., Purdue, G., Hunt, J., Arnoldo, B. D., Aragaki, C., and Horton, J. (2006). Detecting genetic predisposition for complicated clinical outcomes after burn injury. *Burns*, 32(7):821–827.

Battey, H. and Linton, O. (2014). Nonparametric estimation of multivariate elliptic densities via finite mixture sieves. *Journal of Multivariate Analysis*, 123:43–67.

Bayat, A., McGrouther, D., and Ferguson, M. (2003). Skin scarring. *BMJ: British Medical Journal*, 326(7380):88.

Birke, M. (2009). Shape constrained kernel density estimation. *Journal of Statistical Planning and Inference*, 139(8):2851–2862.

Bloemen, M. C., van der Veer, W. M., Ulrich, M. M., van Zuijlen, P. P., Niessen, F. B., and Middelkoop, E. (2009). Prevention and curative management of hypertrophic scar formation. *Burns*, 35(4):463–475.

Bothwell, L. E., Greene, J. A., Podolsky, S. H., and Jones, D. S. (2016). Assessing the gold standardlessons from the history of rcts.

Brown, B., McKenna, S., Siddhi, K., McGrouther, D., and Bayat, A. (2008). The hidden cost of skin scars: quality of life after skin scarring. *Journal of Plastic, Reconstructive & Aesthetic Surgery*, 61(9):1049–1058.

Chakraborty, B. and Moodie, E. E. (2013). *Statistical methods for dynamic treatment regimes*. Springer.

Chang, W.-C. (1983). On using principal components before separating a mixture of two multivariate normal distributions. *Applied Statistics*, pages 267–275.

Chapelle, O. and Li, L. (2011). An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pages 2249–2257.

Cheng, Y. and Church, G. M. (2000). Biclustering of expression data. In *Ismb*, volume 8, pages 93–103.

Collins, F. S. and Varmus, H. (2015). A new initiative on precision medicine. *New England Journal of Medicine*, 372(9):793–795.

Cule, M., Samworth, R., et al. (2010a). Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density. *Electronic Journal of Statistics*, 4:254–270.

Cule, M., Samworth, R., and Stewart, M. (2010b). Maximum likelihood estimation of a multi-dimensional log-concave density. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(5):545–607.

Dawson, R. and Lavori, P. W. (2003). Comparison of designs for adaptive treatment strategies: baseline vs. adaptive randomization. *Journal of statistical planning and inference*, 117(2):365–385.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

Ding, C. and He, X. (2004). K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on Machine learning*, page 29. ACM.

Driver, H. E. and Kroeber, A. L. (1932). *Quantitative expression of cultural relationships*. University of California Press.

Duke, J., Wood, F., Semmens, J., Edgar, D., Spilsbury, K., and Rea, S. (2012). An assessment of burn injury hospitalisations of adolescents and young adults in western australia, 1983–2008. *Burns*, 38(1):128–135.

Dümbgen, L., Rufibach, K., et al. (2009). Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency. *Bernoulli*, 15(1):40–68.

Edwards, R. R., Smith, M. T., Klick, B., Magyar-Russell, G., Haythornthwaite, J. A., Holavanahalli, R., Patterson, D. R., Blakeney, P., Lezotte, D., McKibben, J., et al. (2007). Symptoms of depression and anxiety as unique predictors of pain-related outcomes following burn injury. *Annals of Behavioral Medicine*, 34(3):313–322.

Fearmonti, R., Bond, J., Erdmann, D., and Levinson, H. (2010). A review of scar scales and scar measuring devices. *Eplasty*, 10(e43).

Finnerty, C. C., Jeschke, M. G., Branski, L. K., Barret, J. P., Dziewulski, P., and Herndon, D. N. (2016). Hypertrophic scarring: the greatest unmet challenge after burn injury. *The Lancet*, 388(10052):1427–1436.

Fontana, C. R., Bonini, D., and Bagnato, V. S. (2013). A 12-month follow-up of hypopigmentation after laser hair removal. *Journal of Cosmetic and Laser Therapy*, 15(2):80–84.

Fraley, C. and Raftery, A. E. (1998). How many clusters? which clustering method? answers via model-based cluster analysis. *The computer journal*, 41(8):578–588.

Fraley, C. and Raftery, A. E. (1999). Mclust: Software for model-based cluster analysis. *Journal of Classification*, 16(2):297–306.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, 97(458):611–631.

Friedstat, J. S. and Hultman, C. S. (2014). Hypertrophic burn scar management: what does the evidence show? a systematic review of randomized controlled trials. *Annals of plastic surgery*, 72(6):S198–S201.

Giné, E. and Nickl, R. (2008). Uniform central limit theorems for kernel density estimators. *Probability Theory and Related Fields*, 141(3-4):333–387.

Grady, C. (2005). Payment of clinical research subjects. *The Journal of Clinical Investigation*, 115(7):1681–1687.

Grenander, U. (1956a). On the theory of mortality measurement. *Scandinavian Actuarial Journal*, 1956(1):70–96.

Grenander, U. (1956b). On the theory of mortality measurement: part ii. *Scandinavian Actuarial Journal*, 1956(2):125–153.

Hardy, G. H. and Littlewood, J. E. (1930). A maximal theorem with function-theoretic applications. *Acta Mathematica*, 54(1):81–116.

Hayward, R. A., Kent, D. M., Vijan, S., and Hofer, T. P. (2006). Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. *BMC medical research methodology*, 6(1):18.

Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802.

Huang, H., Liu, Y., and Marron, J. (2012). sigclust: Statistical significance of clustering. r package version: 1.0. 0.

Hultman, C. S., Edkins, R. E., Lee, C. N., Calvert, C. T., and Cairns, B. A. (2012). Shine on: review of laser-and light-based therapies for the treatment of burn scars. *Dermatology research and practice*, 2012.

Hultman, C. S., Edkins, R. E., Wu, C., Calvert, C. T., and Cairns, B. A. (2013). Prospective, before-after cohort study to assess the efficacy of laser therapy on hypertrophic burn scars. *Annals of plastic surgery*, 70(5):521–526.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666.

Katsnelson, A. (2013). Momentum grows to make'personalized'medicine more'precise'. *Nature medicine*, 19(3):249–249.

Kerwin, L. Y., Tal, E., Kader, A., Stiff, M. A., and Fakhouri, T. M. (2014). Scar prevention and remodeling: a review of the medical, surgical, topical and light treatment approaches. *International journal of dermatology*, 53(8):922–936.

Kidd, M., Hultman, C. S., Van Aalst, J., Calvert, C., Peck, M. D., and Cairns, B. A. (2007). The contemporary management of electrical injuries: resuscitation, reconstruction, rehabilitation. *Annals of plastic surgery*, 58(3):273–278.

Kidwell, K. M. (2014). Smart designs in cancer research: Past, present, and future. *Clinical Trials*, page 1740774514525691.

Kosorok, M. R. (2007). *Introduction to empirical processes and semiparametric inference*. Springer Science & Business Media.

Kosorok, M. R. (2008). Bootstrapping the grenander estimator. In *Beyond parametrics in interdisciplinary research: Festschrift in honor of Professor Pranab K. Sen*, pages 282–292. Institute of Mathematical Statistics.

Kosorok, M. R. and Moodie, E. E. (2015). *Adaptive treatment strategies in practice: planning trials and analyzing data for personalized medicine*. SIAM.

Laber, E. B., Lizotte, D. J., Qian, M., Pelham, W. E., and Murphy, S. A. (2014). Dynamic treatment regimes: Technical challenges and applications. *Electronic journal of statistics*, 8(1):1225.

Lange, T., Roth, V., Braun, M. L., and Buhmann, J. M. (2004). Stability-based validation of clustering solutions. *Neural computation*, 16(6):1299–1323.

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The parable of google flu: traps in big data analysis. *Science*, 343(6176):1203–1205.

Liu, L. and Wong, W. H. (2014). Multivariate density estimation via adaptive partitioning (i): Sieve mle. *arXiv preprint arXiv:1401.2597*.

Liu, Y., Wang, Y., and Zeng, D. (2016). Sequential multiple assignment randomization trials with enrichment design. *Biometrics*.

Liu, Y., Zeng, D., and Wang, Y. (2014). Use of personalized dynamic treatment regimes (dtrs) and sequential multiple assignment randomized trials (smarts) in mental health studies. *Shanghai archives of psychiatry*, 26(6):376.

Lu, W., Zhang, H. H., and Zeng, D. (2011). Variable selection for optimal treatment decision. *Statistical methods in medical research*, page 0962280211428383.

McLachlan, G. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

McLachlan, G. J., Bean, R., and Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, 18(3):413–422.

Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, pages 246–270.

Meyer, M. C. (2001). An alternative unimodal density estimator with a consistent estimate of the mode. *Statistica Sinica*, pages 1159–1174.

Miller, F. G. and Brody, H. (2003). A critique of clinical equipoise: therapeutic misconception in the ethics of clinical trials. *Hastings Center Report*, 33(3):19–28.

Minsker, S., Zhao, Y.-Q., and Cheng, G. (2016). Active clinical trials for personalized medicine. *Journal of the American Statistical Association*, 111(514):875–887.

Moodie, E. E., Dean, N., and Sun, Y. R. (2014). Q-learning: Flexible learning about useful utilities. *Statistics in Biosciences*, 6(2):223–243.

Morokoff, W. J. and Caflisch, R. E. (1995). Quasi-monte carlo integration. *Journal of computational physics*, 122(2):218–230.

Murdoch, T. B. and Detsky, A. S. (2013). The inevitable application of big data to health care. *Jama*, 309(13):1351–1352.

Murphy, S. A. (2005). An experimental design for the development of adaptive treatment strategies. *Statistics in medicine*, 24(10):1455–1481.

NIH (2016). Nih u.s. national library of medicine. `https://ghr.nlm.nih.gov/primer/precisionmedicine/definition`. Accessed: 2016-11-03.

Ogbagaber, S. B., Karp, J., and Wahed, A. S. (2016). Design of sequentially randomized trials for testing adaptive treatment strategies. *Statistics in medicine*, 35(6):840–858.

Parzen, E. (1962). On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076.

Porter, C., Tompkins, R. G., Finnerty, C. C., Sidossis, L. S., Suman, O. E., and Herndon, D. N. (2016). The metabolic stress response to burn trauma: current understanding and therapies. *The Lancet*, 388(10052):1417–1426.

Preis, T. and Moat, H. S. (2014). Adaptive nowcasting of influenza outbreaks using google searches. *Royal Society open science*, 1(2):140095.

Prelić, A., Bleuler, S., Zimmermann, P., Wille, A., Bühlmann, P., Gruissem, W., Hennig, L., Thiele, L., and Zitzler, E. (2006). A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9):1122–1129.

Ray, S. and Lindsay, B. G. (2005). The topography of multivariate normal mixtures. *Annals of Statistics*, pages 2042–2065.

Rice, S. A. (1927). The identification of blocs in small political bodies. *American Political Science Review*, 21(03):619–627.

Roberts, S. J., Holmes, C., and Denison, D. (2001). Minimum-entropy data partitioning using reversible jump markov chain monte carlo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(8):909–914.

Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the second seattle Symposium in Biostatistics*, pages 189–326. Springer.

Rosenblatt, M. et al. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837.

Schulte, P. J., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2014). Q-and a-learning methods for estimating optimal dynamic treatment regimes. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 29(4):640.

Schumacher, J., Wunderle, T., Fries, P., Jäkel, F., and Pipa, G. (2015). A statistical framework to infer delay and direction of information flow from measurements of complex systems. *Neural computation*.

Schwacha, M. G., Holland, L. T., Chaudry, I. H., and Messina, J. L. (2005). Genetic variability in the immune-inflammatory response after major burn injury. *Shock*, 23(2):123–128.

Shortreed, S. M., Laber, E., Lizotte, D. J., Stroup, T. S., Pineau, J., and Murphy, S. A. (2011). Informing sequential clinical decision-making through reinforcement learning: an empirical study. *Machine learning*, 84(1-2):109–136.

Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 97–99.

Székely, G. J., Rizzo, M. L., et al. (2009). Brownian distance covariance. *The annals of applied statistics*, 3(4):1236–1265.

Van, A. and der Vaart, J. (1996). Wellner, weak convergence and empirical processes. *Springer-Verlag*, 23:24.

van der Laan, M. J. and Petersen, M. L. (2007). Causal effect models for realistic individualized treatment and intention to treat rules. *The International Journal of Biostatistics*, 3(1).

Van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Statistical applications in genetics and molecular biology*, 6(1).

Vermorel, J. and Mohri, M. (2005). Multi-armed bandit algorithms and empirical evaluation. In *European conference on machine learning*, pages 437–448. Springer.

Vollmer, W. M., Sacks, F. M., Ard, J., Appel, L. J., Bray, G. A., Simons-Morton, D. G., Conlin, P. R., Svetkey, L. P., Erlinger, T. P., Moore, T. J., et al. (2001). Effects of diet and sodium intake on blood pressure: subgroup analysis of the dash-sodium trial. *Annals of internal medicine*, 135(12):1019–1028.

Walther, G. (2002). Detecting the presence of mixing with multiscale maximum likelihood. *Journal of the American Statistical Association*, 97(458):508–513.

Wang, J. (2010). Consistent selection of the number of clusters via crossvalidation. *Biometrika*, pages 893–904.

Wei, S. and Kosorok, M. R. (2013). Latent supervised learning. *Journal of the American Statistical Association*, 108(503):957–970.

Wolters, M. A. (2012). *Methods for Shape-Constrained Kernel Density Estimation*. PhD thesis, The University of Western Ontario.

Zhang, B. (2003). Regression clustering. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 451–458. IEEE.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2012). A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018.

Zhang, D.-Q. and Chen, S.-C. (2003). Clustering incomplete data using kernel-based fuzzy c-means algorithm. *Neural Processing Letters*, 18(3):155–162.

Zhao, Y., Kosorok, M. R., and Zeng, D. (2009). Reinforcement learning design for cancer clinical trials. *Statistics in medicine*, 28(26):3294–3315.

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118.

Zhao, Y., Zeng, D., Socinski, M. A., and Kosorok, M. R. (2011). Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics*, 67(4):1422–1433.

Zhao, Y.-Q. and Laber, E. B. (2014a). Estimation of optimal dynamic treatment regimes. *Clinical Trials*, 11(4):400–407.

Zhao, Y.-Q. and Laber, E. B. (2014b). Estimation of optimal dynamic treatment regimes. *Clinical Trials*, 11(4):400–407.

Zhao, Y.-Q., Zeng, D., Laber, E. B., and Kosorok, M. R. (2015a). New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association*, 110(510):583–598.

Zhao, Y.-Q., Zeng, D., Laber, E. B., and Kosorok, M. R. (2015b). New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association*, 110(510):583–598.

Zubin, J. (1938). A technique for measuring like-mindedness. *The Journal of Abnormal and Social Psychology*, 33(4):508.