UiT

THE ARCTIC UNIVERSITY OF NORWAY

Faculty of Science and Technology
Department of Computer Science

Small data: practical modeling issues in human-model -omic data

—

Einar Holsbø

A Dissertation for the degree of Philosophiae Doctor — 2018

This thesis document was typeset using the *UiT Thesis LaTeX Template*.

“Behind everything simple is a huge tail of complicated.”
–Terry Pratchett, *I Shall Wear Midnight*

“I’m no biologist.”
–Bjørn Fjukstad, multiple occasions

Abstract

Human-model data are very valuable and important in biomedical research. Ethical and economical constraints limit the access to such data, and consequently these datasets rarely comprise more than a few hundred observations. As measurements are comparatively cheap, the tendency is to measure as many things as possible for the few, valuable participants in a study. With -omics technologies it is cheap and simple to make hundreds of thousands of measurements simultaneously. This few observations–many measurements setting is a high-dimensional problem in the technical language. Most gene expression experiments measure the expression levels of 10 000–15 000 genes for fewer than 100 subjects. I refer to this as the **small data setting**.

This dissertation is an exercise in practical data analysis as it happens in a large epidemiological cohort study. It comprises three main projects: (i) predictive modeling of breast cancer metastasis from whole-blood transcriptomics measurements; (ii) standardizing a microarray data quality assessment in the Norwegian Women and Cancer (NOWAC) postgenome cohort; and (iii) shrinkage estimation of rates. These three are all small data analyses for various reasons.

Predictive modeling in the small data setting is very challenging. There are several modern methods built to tackle high-dimensional data, but there is a need to evaluate these methods against one another when analyzing data in practice. Through the metastasis prediction work we learned first-hand that common practices in machine learning can be inefficient or harmful, especially for small data. I will outline some of the more important issues.

In a large project such as NOWAC there is a need to centralize and disseminate knowledge and procedures. The standardization of NOWAC quality assessment was a project born of necessity. The standard operating procedure for outlier removal was developed so that preprocessing of the NOWAC microarray material should happen in the same way every time. We take this procedure from an archaic R-script that resided in peoples email inboxes to a well-documented, open-source R-package and present the NOWAC guidelines for microarray quality control. The procedure is built around the inherent high value of a single observation.

Small data are plagued by high variance. Working with small data it is usually profitable to bias models by shrinkage or borrowing of information from elsewhere. We present a pseudo-Bayesian estimator of rates in an informal crime rate study. We exhibit the value of such procedures in a small data setting and demonstrate some novel considerations about the coverage properties of such a procedure.

In short I gather some common practices in predictive modeling as applied to small data and assess their practical implications. I argue that with more focus on human-based datasets in biomedicine there is a need for particular consideration of these data in a small data paradigm to allow for reliable analysis. I will present what I believe to be sensible guidelines.

Acknowledgements

I have come a long way for someone whose most notable achievement to date is dropping out of upper secondary school. First of all thank you for reading this: the only part of this document interesting to anyone who isn't me, my advisers, or the committee. Make sure to also check out the clever quotes at the beginning of each chapter; I worked very hard on them.

Thank you to friends and family. Particularly to Martine who shares my considerable house mortgage and has put up with her fair share of mood swings. Thank you Bjørn, my constant moral support and office wife during almost all of my nine (!) years at university. This is truly the end of an era. Thanks Jan-Ove for proof-reading Chapters 2 and 5 after I did some revisions. I offered to buy you a beer for each chapter but you kept buying the next round and thus negating my efforts.

Thank you to academic friends and collaborators, especially Vittorio who is always willing to hear out a bad idea and has been very encouraging of my statistical work in spite of my limited formal education in the subject. I have learned so much from our talks and correspondence.

Thank you to my advisers Lars Ailo, Etienne, and Eiliv. You have all provided guidance while still allowing me to form my own ideas and make my own mistakes, of which there have been many.

For the SOP article we are thankful to and impressed by the women that donated blood for NOWAC. Bente Augdal, Merete Albertsen, and Knut Hansen were responsible for all infrastructure and administrative issues. Kajsa Møllersen has provided many useful comments.

For the shrinkage article we would like to thank our anonymous reviewer for the very thorough and very useful comments and suggestions.

The image on the front page is a detail from J. M. W. Turner's *Fishermen at Sea*. I feel that it evokes the feelings of helplessness and despair that go hand in hand with writing a Ph. D. dissertation.

Contents

List of Figures

List of Tables

List of Code Listings

1

Introduction

"[. . .] there's nothing like millions of years of really frustrating trial and error to give a species moral fibre [. . .] "

–Terry Pratchett, *Reaper Man*

1.1 The human model

There is a major concern in biomedical research about the extent to which results from model organisms such as mice apply to humans. There are eg. several known discrepancies between the immune systems of mice and humans.[1] Hence the need for human-derived data grows in both academic and commercial endeavors. Ethical, practical, and economical concerns limit the access to such data, and the most common sample-size calculation is not a power calculation, but the simple equation of

$$n = \frac{\text{money available}}{\text{cost per subject}}.$$

Therefore human-model biomedical datasets are often small-to-medium sized. Having once recruited a person to participate in a research project it is tempting to take as many measurements as is feasible, extra measurements being

[1] See the editorial *Of men, not mice*, Nature Medicine volume 19, page 379 (2013).

comparatively easy to come by.

As measurements go, a blood sample is a fairly quick, cheap, and unintrusive procedure with a large potential. Liew et al. (2006) argue that circulating blood cells could, since blood passes through and interacts with all other parts of the human body, act as "sentinels" that respond to the processes in other organs and hence that blood could act as a surrogate tissue to those that are harder to access.

The *central dogma of molecular biology* (Crick, 1958, 1970) describes the information flow from deoxyribonucleic acid (DNA) via ribonucleic acid (RNA) to fully-formed proteins, which perform the tasks in our bodies. In simplified form this information flows in the straight line

$$\text{DNA} \xrightarrow{transcription} \text{RNA} \xrightarrow{translation} \text{protein.}$$

This path (with nuances) describes the processes of life from the first building blocks of inherited genetic material to the incredibly complicated end-product of a human being. The blood transcriptome—ie. the complete set of transcribed RNA molecules and their abundance—as a representation of the functional elements of the genome—provides a large pool of potential sentinel biomarkers for a given process.

1.2 Measuring the transcriptome

There are two main technologies for charactering the transcriptome: DNA microarrays and RNA-Seq. I describe these in quite general terms below.

DNA microarrays contain short strings of DNA, probes, designed to complement and attach to different target messenger RNAs (mRNAs) corresponding to genes in the genome. A fluorescent label is attached to the mRNA extracted from biological samples and the mRNAs are allowed to hybridize and bind to the probes. Whatever material did not bind to probes is washed away and the microarray is analyzed with a scanner that detects fluorescence. The intensity of this fluorescence at the location of a certain probe family is then a measurement of the abundance of the particular mRNA this probe is designed to attach to.

RNA-Seq is a technology based on high-throughput sequencing of DNA. Complimentary DNAs (cDNAs) are generated from extracted RNAs by reverse transcription. This allows the use of high-throughput sequencing designed for DNA, which is a quite mature family of technologies. In shotgun sequencing, cDNAs are broken into shorter fragments so that they can be sequenced by the Sanger method (Sanger et al., 1977) in parallel. They fragmented reads are re-assembled by computers based on their partial overlap and, possibly, fit to a reference genome. A measurement of abundance for a given RNA is then provided by counting the occurrences of its cDNA.

At the time that I write this (late 2018) the DNA microarray is mostly considered passé. RNA-Seq has several advantages over DNA microarrays (Wang et al., 2009), including lower background noise, less RNA needed for analysis, and the ability to investigate novel RNAs as there is no need for targeted probes as with microarrays. With RNA-Seq we are also able to detect a wider range of genetic expression, as microarrays are limited at the low end by background noise and at the high end by probe saturation. Challenges in RNA-Seq include biases induced by fragmenting the RNA strands for sequencing and the fact that assembly of these fragments is both data and compute intensive.

1.3 Transcriptomics, small data, and statistics

With -omic-type technologies we have access to thousands of measurements from a single subject. As a result these data contain few observations in high dimensions. Analysis of such small, high-dimensional biomedical data comes with several challenges. It is reasonable to expect small effect sizes and high variance. An interest in building diagnosic tests for clinical use in -omics-type projects shifts the focus from estimating means and variances to finding linear (or even higher-order) combinations of potential predictors of the outcome of interest—among thousands of candidates. This search for a needle in a haystack stretches already small data sets even thinner because we often need a certain amount of data to specify the model and additional data to evaluate the final model. Common strategies and conventions from big data machine learning may not readily extend to this setting. It is certainly not a situation imagined by the giants of classical statistics (paraphrasing Efron (2012)).

Figure 1.1 describes what I am going to call the **small data** setting of biomedical -omics research. The figure shows a distribution over typical sample sizes of transcriptomics data sets derived from human subjects. The vast majority of these data sets comprise fewer than 100 observations. Compare with the around 20 000 known genes in the genome. I have been working with the human-model transcriptome in blood. Blood is among the most variable tissue types in the body. It consists of multiple cell types whose relative proportions vary both over time and from subject to subject. The most transcriptionally active cells, white blood cells, make up only 1% of the blood volume. The white blood cells themselves comprise several subtypes. It can be very hard to

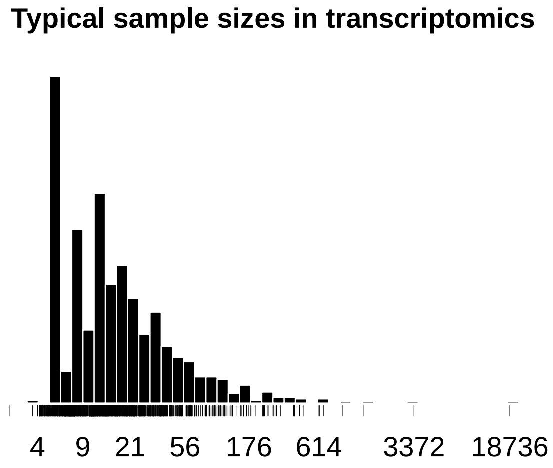


Figure 1.1: Sizes of human-derived transcriptomics data uploaded to the EMBL-EBI (n=1178). The figure is described in more detail in Chapter 5.

distinguish biological variation due to outcome—the good kind that we want to study—and biological variation due to the tissue itself—a nuisance (for all of this see Fan and Hegde (2005)). Additional variation comes from the processing of the blood samples themselves, decisions made in preprocessing, etc.

The 20 000 genes for < 100 observations setting falls within the field of high-dimensional statistics. The most basic type of question asked of such data in transcriptomics is *which genes are differentially expressed between outcome groups?* Such a question is meant to inform on the functional-genomic properties of a certain disease or phenotype. There are of course many ways to detect such differences from classical statistics and the statistical arguments made around this question tend to be gene-wise hypothesis tests of the kinds that we variously investigate in the appendix to Chapter 2.

Tailor-made methods tend to take advantage of the parallel nature of a transcriptomic experiment. As we discuss in more detail in Chapter 4, quantities estimated in parallel can always be improved, in a squared-errors-sense, by shrinking their estimates either toward zero or toward one another, effectively borrowing information between estimates. This is the thinking behind the LIMMA methodology, described by Smyth (2004) for microarrays, and later in edgeR, described by Robinson et al. (2010) as an adaptation of LIMMA-type methods to RNA-Seq data. These methods both moderate estimates of variation for their different test statistics in empirical Bayes procedures. Similarly SAM (for microarrays) by Tusher et al. (2001) adds some small data-dependent constant (calculated from all data, not just gene-by-gene) to estimates of spread to make them independent of expression level. This is extended to RNA-Seq data in the SAMSeq method (Li and Tibshirani, 2013). These are both nonparametric methods. Penalized likelihood methods as used in Chapter 2 also fall under this general shrinkage-type thinking but the shrinkage usually applies to coefficient estimates in the model instead of to their test statistics. Inference is tricky here, unless we use fully Bayesian approaches.

People adapt methods from microarray data to RNA-Seq data because the two technologies—although they ostensibly measure the same thing—produce different types of data. The numbers that come out of microarray platforms are continuous whereas those from RNA-Seq platforms are discrete counts. Hence much of the effort in developing methods for RNA-Seq has consisted of adapting the lessons learned from microarrays to work with counting distributions such as the Poisson or the negative binomial. Covering all different types of possible analyses and methods could fill several review articles, which indeed it already has, see eg. Kristensen et al. (2014) and Conesa et al. (2016).

There are also more general statistical issues to consider working with small transcriptomics data. For instance, if we want to preserve the error-statistical properties of an experiment while also performing thousands of hypothesis tests simultaneously we must take this multiplicity into consideration. Two well-established methods are either to control the family-wise error rate by Bonferroni correction (Dunn, 1961) or controlling the false discovery rate—a less stringent criterion—by the method of Benjamini and Hochberg (1995). If we want to use shrinkage, how do we determine the size of this shrinkage? I return to this in Chapters 2 and 5 (also indirectly in Chapter 4).

I touch on various methods and methodology throughout most of this dissertation, especially in Chapter 5 where I focus on predictive modeling. There are many models and methods for general high-dimensional problems. Often there is no agnostic a priori reason to prefer one model to another. The general approach is to hope that there is a lower-dimensional structure to be found in the data, and to use biased models to combat the high variance in the data themselves and in model estimation procedures. Theoretical results are usually asymptotic in n or p, but there is a long way from real-life data analysis to Asymptopia. No-one knows how to safely evaluate a prediction method except for use of data other than those used to fit the model. The gold standard is 100% independent validation data, if you can have them. Splitting data into train-, test-, and validation sets is a common practice in machine learning. Since epidemiological studies are small and are likely to remain so, keeping some of the data separate for the purpose of validation is a risky proposition. We shall see that such data splitting and other common practices can be harmful if applied haphazardly in the analysis of small, high-dimensional data.

1.4 Transcriptomics and cancer

There is a fair body of work on the use of transcriptomic measurements as markers of cancer subtype or as indicators of prognosis. The recent review by Kwa et al. (2017) enumerates six different gene-signature tests for early-stage breast cancers: OncotypeDX (Paik et al., 2004); Prosigna (Parker et al., 2009); MammaPrint (Van't Veer et al., 2002); Breast cancer index (Ma et al., 2008); EndoPredict (Filipits et al., 2011); and Genomic Grade Index (Sotiriou et al., 2006). These are all applied to tissue samples taken from the tumor, and are variously used for predicting prognosis, recurrence risk (including distant metastasis), benefit from extended therapy, etc. Most of these were first developed in smaller data sets (sub-100 to low 100s) with validation in larger sets (low-to-high 100s).

The literature is much sparser when it comes to the use of blood samples, especially in metastasis prediction, which has been my focus. A fairly-recent study by Aristizábal-Pachón et al. (2015) provides some evidence that the expression of Mammaglobin A in peripheral blood has potential as a marker for breast cancer metastasis. Some various examples of blood-based predictors that do not use transcriptomic measurements are mentioned in Chapter 2.

The Norwegian Women and Cancer (NOWAC) study—which has provided a general backdrop to my work—is fairly unique in that it has collected blood samples prospectively and buffered these to prevent the degradation of mRNAs. Hence their data enable the investigation of pre-diagnostic transcriptomic signals of carcinogenesis and metastasis. Such investigations could provide a valuable basis for early detection, or provide valuable system-epidemiological insights. The NOWAC postgenome cohort (Lund et al., 2008) is a prospective population-based cohort that contains blood samples from 50 000 women born between 1943 and 1957. Out of these in total about 1 600 case–control pairs (3 200 blood samples) have at various times been processed with DNA microarrays to measure mRNA abundance. These measurements are combined with questionnaires and disease/death status from The Norwegian Cancer Registry, and the The Cause of Death Registry in Statistics Norway. Lund et al. (2016) provides a recent study of a pre-diagnostic blood transcriptome signal of breast cancer presence. Another example of this focus on a pre-diagnostic signal in blood is that of Sandanger et al. (2018), who investigate both mRNAs and DNA methylation as potential biomarkers of lung cancer.

1.5 Contributions and outline

This dissertation is part case-study, part original research, and part methodological guideline. I argue that in human cohort studies it is useful and needful to set Big Data dreams aside and consider a paradigm of **small data** where observations are very valuable indeed and high dimensionality is the norm. In particular I investigate the dangers of some common machine learning practice as applied to data from an epidemiological cohort study. I also touch upon some of the analysis-adjacent problems of such a study, such as replicability and knowledge dissemination.

For the most part I have either worked directly with the NOWAC material, or worked with problems that I became aware of as a result of my work with the NOWAC material. NOWAC is an exploratory study, so the focus is on hypothesis generation.

I organize this dissertation around three main projects in the form of papers. Below I briefly describe these projects. For each I describe the contributions of the work, their small-data implications, and my personal contributions to the project:

Breast cancer metastasis prediction: I present this work in Chapter 2. It is the largest part of my Ph. D. work. We analyze gene expression measurements from blood samples in the NOWAC postgenome cohort study. We find indication that there is predictive information of metastasis in these blood samples, which were taken some time before cancer was detected. We provide a quantitative description of the genes that most strongly associate with metastasis in these data. Early detection of metastasis could potentially reduce mortality, and investigation of the functional-genomic aspects in blood could shed light on the systemic response to aggressive cancers. In addition to these main results I provide some methodological considerations in Chapter 5 that I consider key in a small-data setting and warn against some potentially harmful standard machine learning practices. I am first author on this work and provided all modeling, implementation, validation, and most writing.

In Chapter 6 I briefly outline two pieces of work that follow this one. First, we have done a small analysis of a data set collected in NOWAC to assess the **potential effect of psychological stress on blood transcriptomics.** Such an effect might well get in the way of any information predictive of eg. cancer and confound the analysis. We find little evidence of a stress effect. I am joint first authors with Dr. Karina Standahl Olsen on this work. It should be considered in-progress but very nearly ready for submission. I provide methodology and analysis.

Second, the analyses in the appendix to Chapter 2 indicate a certain reduction in variance when time-to-diagnosis is part of the variable selection process. This makes biological sense as there is little reason to think that a blood sample provided 10 years ago should show any systemic response to a cancer found today. I outline a **followup-weighted extension to the significance analysis of microarrays**. I have some promising early simulation results for this method, not presented in this document. Formalization and further development should be considered future work.

A NOWAC standard operating procedure for outlier assessment: It was realized in NOWAC that (i) there should be a standard procedure of detecting and removing technical outliers: observations that for lab-technical reasons cannot be used for analysis, and (ii) this procedure should remove as few observations as possible. The article, presented in Chapter 3, provides a formal description of the standard operating procedure and an R-package that implements the methods to carry the procedure out. I am joint first authors on this work with Dr. Hege Bøvelstad. We provide experiments that indicate that the approach should be applied carefully by a human rather than automatically based on standard thresholds. I developed the `nowaclean` R package from the canonical NOWAC R-script, passed from person to person by email. I also provided writing and the experimental evaluation.

Shrinkage estimation of rates: Techniques for introducing bias and borrowing "extra" information are very important in a small-data situation. Shrinkage estimation has seen a very successful application in gene expression analysis (Smyth, 2004). This work, presented in Chapter 4, provides a tutorial of a common Bayesian approach to shrinkage estimation of a high-dimensional vector of rates and method assessment by realistic simulations. We provide a new result in terms of coverage properties of the posterior credible intervals of such a procedure. I am first author on this work and provided most of the initial modeling and implementation, and most writing and experimental evaluation.

In a longer-term perspective, I use Chapter 5 to point out some common problems in the predictive modeling and management of small data. With more focus on human-based datasets in biomedicine, I believe that the analyses provided by considering a small data paradigm will be increasingly important for both reliable and reproducible results. Hence my thesis: **A small data paradigm is needed for reliable analysis of small, high-dimensional data.**

2 Metastasis prediction

"Ordinary fortune-tellers tell you what you want to happen; witches tell you what's going to happen whether you want it to or not. Strangely enough, witches tend to be more accurate but less popular."

–Terry Pratchett, *The Wee Free Men*

This chapter is an extension and improvement of our working paper on predicting breast cancer metastasis from blood gene expression measurements (Holsbø et al., 2018). Importantly I have fixed a subtle flaw in our methodology that introduced some serious overfitting problems for the penalized likelihood regression methods (lasso and ridge in the original manuscript). For discussion of this issue, see Chapter 5.

The text below will show that the fixing of this flaw leads to a superior model to any of the ad-hoc variable selection methods that we previously considered. There is no innovation of ours in the model itself, and with the sample size at hand I judge it unlikely that refinements will clearly improve the present results. The improved model lets us shift our focus from *can we reliably model the data at all?*—the main problem of the original text—to *what do the data tell us?* So instead of presenting an array of models with general observations about them as an ensemble, I choose to present a single model and go into more detail about its fit and the properties of the predictors chosen. I believe this makes a more valuable scientific contribution: though the results are necessarily uncertain and exploratory, they define a very concrete route for further investigation.

For completeness I have included the pertinent material on the models from Holsbø et al. (2018) as an appendix to this chapter (see page 24). The text there is identical to the original material, differing only from the cited manuscript in what I have omitted as no longer relevant. I recommend you, the reader, keep Tables 2.4, 2.5, and 2.6 from the appendix in mind to compare with the results in the main text below.

During the work on this problem, which has been central during my time as a Ph. D. student, we learned a lot about modeling small high-dimensional data, which eventually informed the bulk of my views presented in Chapter 5.

Abstract: We investigate whether blood gene expression measurements contain predictive information of breast cancer metastasis. Our data comes from the NOWAC epidemiological cohort study, which also provides nested controls. The women who contributed to these data provided a blood sample up to a year before receiving a breast cancer diagnosis. We estimate a penalized maximum likelihood logistic regression, which we evaluate by extensive resampling in terms of calibration, concordance, and stability. By this model we identify a set of 108 candidate predictor genes that exhibit clear fold change in average metstasized case–control pairs where there is none for the average non-metastasized pair. This set provides a promising starting point for future research.

2.1 Introduction

About one in ten women will at some point develop breast cancer. About 25% have an aggressive cancer at the time of diagnosis, with metastatic spread to axillary lymph nodes.[1] Spread is detected by a sentinel node biopsy: a surgical procedure to check the lymph nodes closest to the cancer site for metastasized cancer. A cancer that has developed to the point of metastasis is much more dangerous than a local one. The absence or presence of metastatic spread largely determines the patient's survival. Early detection is hence very important in terms of reducing cancer mortality. Were we able to detect signs of metastasis or metastatic potential by a blood sample, perhaps in a screening setting, we could conceivably start treatment earlier and treat the cancer before the onset of large, deadly metastasized tumors.

Several recent articles develop this idea of *liquid biopsies* (Chi, 2016). Dif-

[1] http://oncolex.org/Breast-cancer/

ferent relevant signals appear in blood for already-diagnosed breast cancer. For instance: circulating tumor cells (Giuliano et al., 2014), circulating tumor DNA (Cohen et al., 2018), serum microRNA (van Schooneveld et al., 2012), or tumor-educated platelets (Best et al., 2017). A recent review in *Cancer and Metastasis Reviews* (Lim and Hortobagyi, 2016) lists liquid biopsies and large data analysis tools as important challenges in metastatic breast cancer research.

The Norwegian Women and Cancer (NOWAC) postgenome cohort (Lund et al., 2008) is a prospective population-based cohort that contains blood samples from 50 000 women born between 1943 and 1957. Out of these in total about 1 600 case–control pairs (3 200 blood samples) have at various times been processed to provide transcriptomic measurements in the form of mRNA abundance. These measurements combine with questionnaires, disease status from The Norwegian Cancer Registry, and death status from the The Cause of Death Registry in Statistics Norway to provide a high-quality dataset. These data are used for exploration and hypothesis generation.

Transcriptomics data are challenging to model, especially for exploratory data analysis. Such data usually comprise fewer than 100 observations and tens of thousands of measurements for each observation. As traditional statistical methods do not lend themselves to such high-dimensional problems, the hope is to uncover lower-dimensional structures. For instance, we expect genes to work together in pathways and do not expect all genes to be relevant in all processes (or indeed any given process). The analysis of high-dimensional data is an active research area of statistics and machine learning (Frigessi et al., 2016). The common methods for discovering low-dimensional structure are projection approaches like PLS-methods (Liquet et al., 2015) and variable selection such as shrinkage (Tibshirani, 1996; Zou and Hastie, 2005).

We examine 88 prospective case–control pairs from the NOWAC study. The blood samples were provided 6–358 days before diagnosis. We fit a penalized likelihood logistic regression with the ElasticNet-type penalty (Zou and Hastie, 2005). This approach provides built-in variable selection in the estimation procedure.

We evaluate our model by extensive resampling (Efron and Gong, 1983). We demonstrate that there is a signal that predicts metastasis in blood transcriptomics but there is work to be done before such a model could have practical utility. Our model uncovers 108 predictor genes that form a promising direction for further research.

2.2 Material and methods

2.2.1 Data

We analyze 88 pairs of cases with breast cancer diagnoses and age-matched healthy controls from the NOWAC Post-genome cohort. Dumeaux et al. (2008) describe the NOWAC study in detail. In brief, women in a certain age group received an invitation to participate by random draw from the Norwegian National Registry. The women who chose to participate filled out a questionnaire and provided a blood sample. Over the years the Cancer Registry of Norway provided followup information on cancer diagnoses and lymph node status. The women in this particular data set received a breast cancer diagnosis at most one year after providing a blood sample.

The NTNU genomics core facility processed the blood samples with Illumina microarray chips of either the HumanWG-6 v. 3 or the HumanHT-12 v. 4 type. To keep case–control pair as comparable as possible, the pair is intact throughout processing pipeline. This means that they are processed on the same day by the same person and lie next to one another physically on the microarray chip. All NOWAC data sets go through a standardized technical quality control (Bøvelstad et al., 2017).

From a starting-point of some 30 000 microarrays probes per observation, we have removed those of low quality and those that showed no signal in a certain fraction of the observations. We next map probes to known genes, where, when several probes map to the same gene, we choose the probe with the largest inter-quartile range. Finally we have normalized the data by quantile normalization before analysis. All of this is NOWAC standard, and the details of preprocessing for these particular data is described in detail by Lund et al. (2016).

The above reduces the dimensionality considerably. The end-result is a 88×12404 **fold change matrix**, X, on the $\log_2$ scale. For each gene, g, and each case–control pair, i, we have the measurement $\log_2 x_{ig} - \log_2 x'_{ig}$. Here x_{ig} is the g expression level for the ith case, and x'_{ig} is the corresponding control. The response variable, **metastasis** ($\in \{0, 1\}$), indicates whether a sentinel node biopsy showed evidence of metastasis. We sometimes refer to this as spread. The cancers in these data were detected in one of three settings: (i) *screening* cancers are detected in the regular screening program; (ii) *interval* cancers are detected in the two-year interval between screenings in a woman who participates in the screening program; finally *clinical* cancers are detected at a clinic in women who either never took part in the screening program, or did not attend a screening in at least two years. Table 2.1 shows the incidence of

	Screening	Interval	Clinical
No spread	43	10	13
Spread	6	6	10

Table 2.1: Incidence of metastasis across detection strata. There is noticeable between-stratum variation. The incidence is much lower in the screening stratum.

metastasis stratified by the three different detection settings. There is some heterogeneity, but since strata are quite small we choose to pool them.

2.2.2 Predictive model

We model the probability of metastasis, $p(m)$, given gene expression, x, by a penalized likelihood logistic regression with an ElasticNet-type penalty (Zou and Hastie, 2005). This takes the usual log–linear form

$$\log \frac{p(m)}{1 - p(m)} = \beta_0 + \beta_1 x_1 + \ldots + x_p$$

but likelihood is maximized under the constraint that

$$(1 - \alpha) \sum |\beta_j| + \alpha \sum \beta_j^2 \leq t \tag{2.1}$$

for some user-specified penalty size t (or, in its alternative formulation, λ) and penalty mixing parameter, α. In a penalized likelihood procedure we essentially set a budget for the size of the fitted regression coefficients. The penalty expressed in 2.1 provides a trade-off between the lasso penalty (Tibshirani, 1996), which provides a variable selection that tends to choose haphazardly between correlated variables, and the ridge penalty (Hoerl and Kennard, 1970), which provides no variable selection and tends to shrink correlated variables toward one another. The idea is that the combined penalty should provide grouped variable selection: some variables get shrunk out of the model entirely and correlated variables get shrunk toward one another.

Tuning parameters

We set the penalty trade-off, α, in expression 2.1 to .5 a priori: half ridge, half lasso. We do this because we have no strong opinion about what α should be and would rather avoid optimizing tuning parameters over a two-dimensional surface.

We choose the penalty size λ by optimizing for a modified version of Akaike's Information Criterion (Akaike, 1973; Verweij and Van Houwelingen, 1994),

$$AIC' = LR\chi^2 - 2k,$$

where k is the effective degrees of freedom of the model and $LR\chi^2$ is the likelihood ratio χ^2 for the model (Wilks, 1938), ignoring the penalty. This has the advantage that we do not need to split the data in a cross-validation and should often result in a sensible model according to Harrell (2013, p. 211).

An estimate for the effective degrees of freedom (EDF) for an additive error lasso model is simply the number of non-zero coefficients (Zou et al., 2007). Scrupulous estimation of EDF for our model is made more complicated because we fit a logistic regression model and because we have added another term to the penalty that introduces additional penalty (and hence lowers the EDF). Hastie and Tibshirani (1990), among others, define EDF for a model, $\hat{\eta}$, as

$$\mathrm{df}(\hat{\eta}) = n - \mathrm{E}[\mathrm{dev}(\eta_s, \hat{\eta})],$$

n being the number of observations, and $\mathrm{dev}(\eta_s, \hat{\eta})$ the deviance between the saturated model and the model under consideration. Our experiments with this quantity has shown no material change in the general properties of the (AIC', λ)-curve in our data, and hence we use the much quicker lasso shorthand of counting non-zero coefficients.

2.2.3 Validation

Metrics

We evaluate models by several criteria. **Brier score** (Brier, 1950) is the mean squared error,

$$\bar{B} = n^{-1} \sum (\hat{y}_i - y_i)^2,$$

between the probability that was predicted by the model, $\hat{y}$, and the known outcomes, y. This is a one-number summary of the calibration of predicted probabilities; we also assess calibration by means of a **calibration curve**. This is an estimate of proportion of true successes as a function of predicted probability, which we calculate by smoothing the true zero/one outcome as a function of predicted probability (lowess with a span of $\frac{2}{3}$). If n observations receive a prediction of $\hat{p}$, $n\hat{p}$ of them should have the predicted condition for a well-calibrated model.

Concordance probability is the probability of ranking (in terms of predicted probability) a randomly chosen positive higher than a randomly chosen negative. A concordance probability of unity means that all positives have a higher predicted probability than all the negatives, one of .5 is equivalent to random guess, and between .5 and zero means that somehow negatives are ranked higher than positives.This is equivalent to the area under the receiver operating characteristic curve (AUC), and is proportional to the Mann-Whitney-Wilcoxon U statistic (Hanley and McNeil, 1982).

Finally, **stability** is the proportion of overlap between predictor genes chosen during different realizations of the modeling procedure. We follow Haury et al. (2011) and measure this by the Jaccard index, $\frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$, where S_1 and S_2 are two sets of predictor genes.

Estimation: Optimism bootstrap

For estimation of concordance probability, Brier score, and calibration curve we take the optimism-corrected bootstrap approach described, eg., in Efron and Gong (1983). This has the advantage of using all of the data in estimating model performance opposed to data splitting procedures such as out-of-bootstrap or k-fold cross-validation where only a portion of the data is used to fit the model.

The apparent score (or training score), $sc(x, \hat{F})$, is the score of a model fit to the sample, x, w.r.t. the empirical distribution of this sample, $\hat{F}$. This is necessarily an overoptimistic estimate. To correct for this, we estimate the expected overoptimism, ω, by the bootstrap:

$$\hat{\omega}_{\text{boot}} = B^{-1} \sum_{b=1}^{B} \left(sc(x_b^*, \hat{F}) - sc(x_b^*, \hat{F}_b^*) \right),$$

where B is the total number of bootstrap resamples, x_b^* is the bth bootstrap sample, and $\hat{F}_b^*$ is the empirical distribution function of the same. Hence $sc(x_b^*, \hat{F}_b^*)$ is the apparent score of the bth bootstrapped model, and $sc(x_b^*, \hat{F})$ is the score of the same model w.r.t. the empirical distribution of the original sample. The optimism-corrected expected score of our model becomes

$$\widehat{sc}_{\text{boot}} = sc(x, \hat{F}) + \hat{\omega}_{\text{boot}},$$

which is a bias correction of the apparent score.In the case of the calibration curve we do this as a pointwise procedure along the curve.

Since there is no notion of “apparent stability,” we take a slightly different approach for this score. Let $S(x)$ be the gene set selected in the original data and $S(x_b^*)$ be the gene set selected in the bth bootstrap sample. The bootstrap estimate of expected stability is then

$$\widehat{st}_{\text{boot}} = B^{-1} \sum_{b=1}^{B} \frac{|S(x) \cap S(x_b^*)|}{|S(x) \cup S(x_b^*)|}.$$

We keep track of the bootstrap gene sets $\{S(x_b^*)\}$. This allows us to make secondary calculations about selected genes such as how often a given gene is selected and which genes that tend to be selected together under resampling.

2.3 Results

We begin this section by describing the fitted model and its predictive performance, and go on to describe the predictor genes selected in the fitting.

2.3.1 Model fit

Choice of penalty size

The top panel of Figure 2.1 shows AIC' as a function of penalty size. Instead of showing a clear maximum, AIC' keeps improving with higher penalty. This is in part because of overestimated effective degrees of freedom, but as we mention above, more sophisticated (and slower to compute) estimates do not materially improve the situation for these data. Instead we observe that improvement slows down considerably with higher shrinkage and choose the point at which improvement slows down as our λ. This is the point of maximum curvature, indicated by a dotted line. Detection of this point can easily be automated by applying a smoother—lowess with a span of $\frac{2}{3}$ in our case—to the (AIC, λ) points and finding the λ-point with the largest absolute second derivative along this smooth line.

To the extent that it is possible to speak of an optimal penalty, the above procedure may not find it. Some degree of undershrinkage should be expected, which may contribute somewhat to the poor calibration outlined below. However, for the purpose of selecting variables, the impact of undershrinkage is not too severe: The bottom panel of Figure 2.1 shows the coefficients of the 108 selected predictor genes (described in Section 2.3.2 below). The ticks along the lambda-axis show the points at which a gene from this set gets shrunk out of the model. A larger penalty would naturally shrink out more genes, but the chosen set will be a subset of the one we present here. We do not show the coefficient for the remaining thousands of genes not chosen, but at no point do they re-enter the model after dropping out.

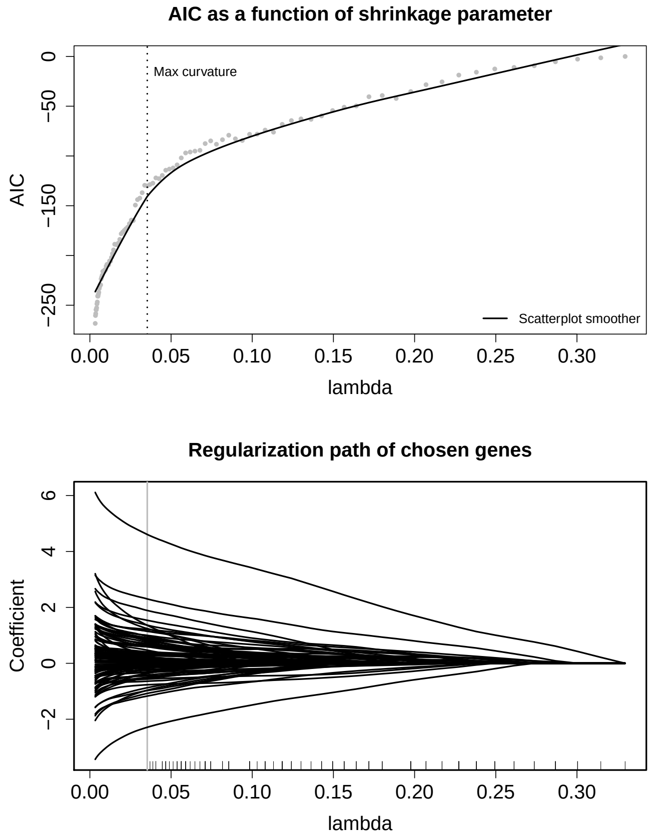


Figure 2.1: Top: Selecting shrinkage size by AIC'. After the automatically detected elbow by maximum curvature at roughly $\lambda = .04$ the gains in AIC' slow considerably. Bottom: Regularization path (coefficients) of selected gene set as function of shrinkage size. The ticks show the point at which a variable drops out of the model.

Evaluation metrics

Figure 2.2 shows the bootstrap for our estimates of Brier score, concordance probability, and stability. The solid lines show point estimates and the dotted lines indicate the middle .8 of each distribution. The Brier score for our model is roughly .1, while that of an intercept-only null model is roughly .18. Since Brier score is the mean square error of predicted probabilities we can take its root to get an average error on the probability scale; $\sqrt{.1} \approx .32$, which suggests that the predicted probabilities are not very accurate. Figure 2.3 corroborates this. The figure shows the pointwise calibration of predicted probabilities, ie., for a given predicted metastasis probability, how great a proportion observations turned out to have metastases. For a predicted metastasis probability $< .4$ the true proportion is $\approx .1$, while for a predicted metastasis probability $> .8$ the

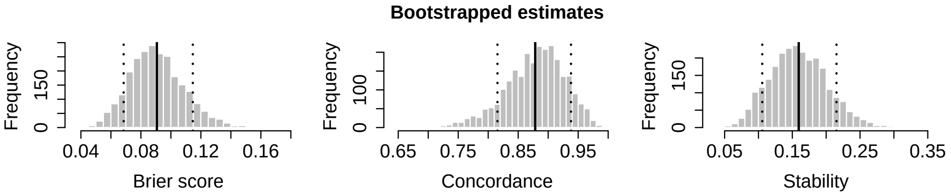


Figure 2.2: Bootstrap distribution of optimism-corrected estimates for Brier score, concordance/AUC, and stability for the Elasticnet model. The solid vertical lines show point estimates, and the dotted vertical lines show the middle .8 of each distribution.

true proportion is ≈ .7. In other words we overestimate low probabilities and underestimate high ones. The model is somewhat more calibrated for higher probabilities but far from perfect.

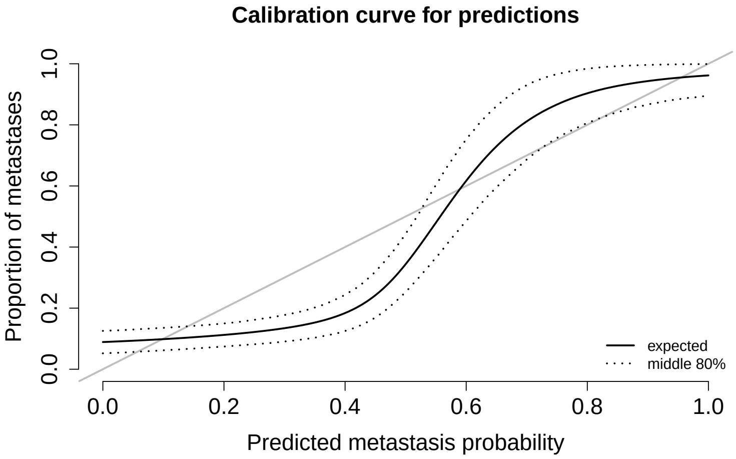


Figure 2.3: Expected calibration of predicted probabilities shown in solid black. The dotted line shows middle .8 of the bootstrap distribution. Ideally, .8 of the observations for which .8 metastasis probability was predicted should turn out to show metastasis. In other words the ideal calibration is a diagonal line (shown in grey). Our model tends to overestimate lower probabilities and underestimate higher ones.

Returning to Figure 2.2, the concordance probability (or AUC) is quite high at roughly .88, with a lower bound for the middle .8 of the distribution at .81. Contrast this with random guess at .5. This suggests that the model consistently selects gene sets that separate metastases from non-metastases in their expression levels in spite of the fact that the predicted probabilities are poorly calibrated. The stability of these chosen gene sets is around .16, which suggests the likely scenario that there are many correlated genes to choose from. With a stability of .16 for 108 genes you might expect a 17-gene overlap when fitting a similar model to similar data.

Assessment of bootstrap procedure

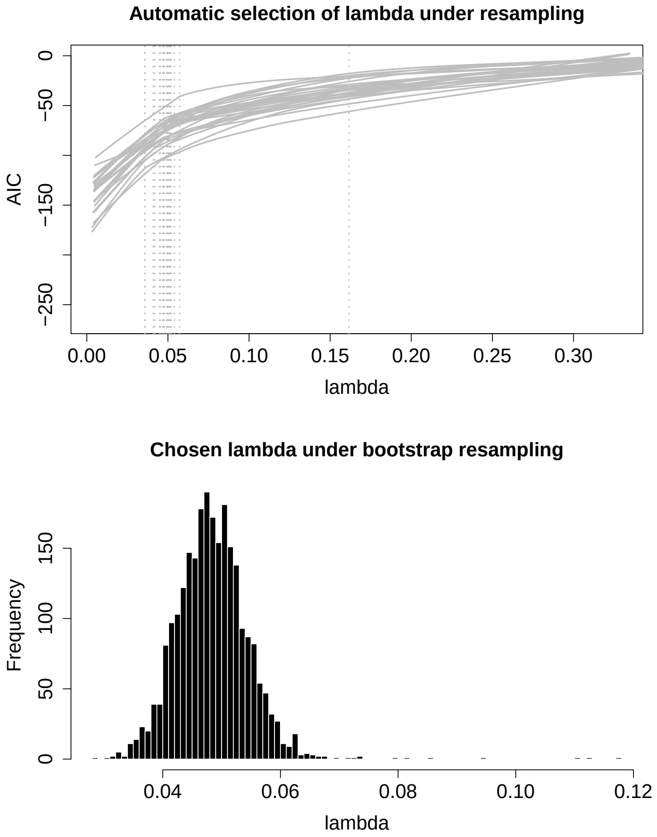


Figure 2.4: Selection of shrinkage size by automatically detecting the “elbow” in the AIC–λ curve under resampling. Top shows 25 such detections under resampling, while bottom shows the selected λ for all the 2500 bootstrap resamples in the main estimation procedure. The automated procedure shows no marked inconsistency.

A resampling procedure rests on our ability to automate the choices made during modelling. The top panel of Figure 2.4 shows 25 resampling realizations of the automated selection of λ described above. This looks sensible compared with Figure 2.1 (top). The bottom panel of Figure 2.4 shows the distribution of all λs selected in the main bootstrap procedure. The distribution is fairly concentrated, though under resampling a slightly larger λ tends to be chosen.

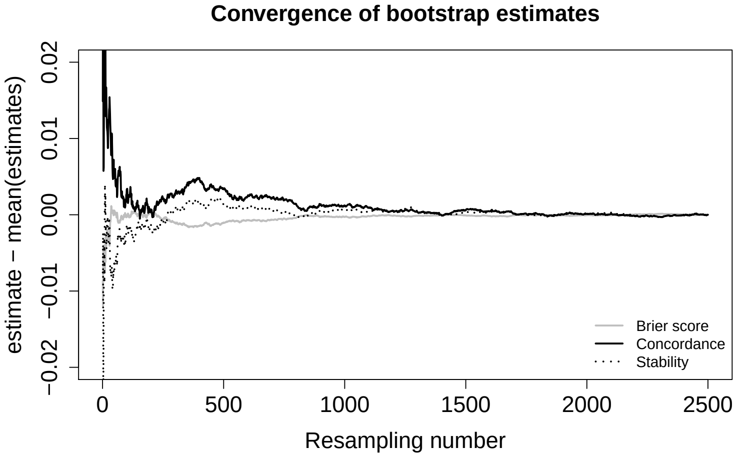


Figure 2.5: Convergence of bootstrap estimates. All estimates look stable after about 1700 resamples, which suggests that the Monte-Carlo error due to resampling is small.

Figure 2.5 shows the quantity

$$b^{-1}\sum_{i=1}^{b} sc(x_i^*, \hat{F}_i^*) - B^{-1}\sum_{j=1}^{B} sc(x_j^*, \hat{F}_j^*), \qquad b \in \{1, 2, \dots, B\},$$

for the three different single-number score estimates. This is the difference between the score after b resamples and the final score, so the figure shows the convergence of these scores toward their final estimates. After around 1700 resamples the change in these curves looks negligible, which indicates that the Monte-Carlo error from resampling should also be negligible.

2.3.2 Selected genes

In this section we list the 108 genes selected by penalized likelihood and describe them in general quantitative terms. As mentioned above, we keep track of the selected gene sets under resampling and can hence calculate statistics for how often a given gene is selected and for how often a given gene is co-selected with any other gene. Table 2.2 shows the 108 selected genes ordered by their individual selection probabilities. Apart from the first few genes, the selection probabilities are not very high. It is quite likely that (i) a larger set of genes correlate with the ones we select and get selected in their place some of the time, and (ii) our selected genes correlate with one another and the selection of one some times makes the selection of another less likely. This is a natural consequence of doing variable selection: “redundant” information may shrink out of the model.

Table 2.2: Resampling selection probability for the 108 elasticnet-selected genes.

GRK5[a]	0.853	C1orf115	0.290	ANO8	0.221	FBLN5	0.157
GPATCH4	0.682	LOC654055	0.287	PTTG1IP	0.219	BLMH	0.156
GNGT2	0.474	RNF214	0.280	3NDg8gVCd. . .[b]	0.218	FCRL3	0.149
PDGFD[c]	0.467	SULT1A1	0.278	USF1	0.216	TDRD9	0.143
FAM24B	0.457	ZNF365	0.271	BCCIP	0.210	ACY1	0.142
PTPRN2	0.442	USE1	0.267	MGC29506	0.209	ZFP57	0.142
CBLB	0.440	DNMT3A	0.267	GRK5[a]	0.207	SLIC1	0.138
PDCL	0.410	LOC649210	0.266	WTIP	0.205	PICK1	0.135
RASA2	0.380	CNTNAP2	0.265	BCL10	0.204	RTN4IP1	0.134
C11orf48	0.376	IL2RA	0.265	DLGAP2	0.200	CDCA7L	0.132
TCEB1	0.374	CCT5	0.264	HRAS	0.199	BEX4	0.131
CAPN3	0.354	R3HDM1	0.263	RAD1	0.189	FCAR	0.130
STK19	0.351	MRPL43	0.260	PRKCE	0.187	ANKRD35	0.111
GUCY1A3	0.348	SLC38A1	0.256	UBAP2L	0.186	USP39	0.109
ZDHHC11	0.345	GNG8	0.255	BPI	0.186	KIAA0495	0.106
SULT1A3	0.336	PLA2G4C	0.251	DTX1	0.184	BRI3BP	0.106
Z6FIQGkeo. . .[d]	0.335	TCF4	0.248	LASS5	0.182	TUBA4A	0.105
FAM89A	0.328	uX15cu4f_. . .[e]	0.247	GSTT1	0.182	IDH1	0.102
rh13dQXo4. . .[f]	0.324	C20orf107	0.245	SPATA20	0.182	DDX52	0.100
LANCL2	0.323	VCL	0.242	IGLL1	0.172	ANKRD57	0.094
SERPINE2	0.318	EZH2	0.242	SPG3A	0.172	TFG	0.087
ADIPOR2	0.314	PRPSAP2	0.237	PPAP2A	0.172	LILRA6	0.080
GPR177	0.312	ISY1	0.235	NOTCH2NL	0.172	C6orf47	0.078
PDGFD[c]	0.299	UGDH	0.234	TAF6	0.168	WDR60	0.075
LOC647460	0.294	ABCF2	0.230	CCDC90B	0.166	AHCYL2	0.068
WEE1	0.293	C16orf5	0.229	LOC731486	0.158	HAUS4	0.068
ITM2C	0.291	VAV3	0.225	CDH2	0.157	MAD2L2	0.053

a. Two probes map to the same gene GRK5. Combined selection probability is 1.06, implying that both get selected together at least some of the time.
b. Illumina probe id 3NDg8gVCdQkNdcg.Ko, missing annotation.
c. Two probes map to the same gene PDGFD. Combined selection probability is 0.766.
d. Ilummina probe id Z6FIQGkeoCSiVAoKeg, missing annotation.
e. Illumina probe id uX15cu4f_VUIuXoSTo, missing annotation.
f. Illumina probe id rh13dQXo4hUS7uOpRQ, missing annotation.

Figure 2.6 shows the (log fold change) expression levels in each of the 108 selected genes for the metastasized and non-metastasized observations. The shaded area shows the middle .8 of the bootstrap distribution for difference in medians between the two groups; the white notch shows the expectation of this distribution, by which the genes are ordered. The black snake-shaped line marks the two group medians. The non-metastasized median is usually around zero, so the difference in medians is mostly dominated by the median fold change of the metastasized observations. In other words, for these genes the average case–control pair is similar in the non-metastasized group, while the average pair is dissimilar in the metastasized group.

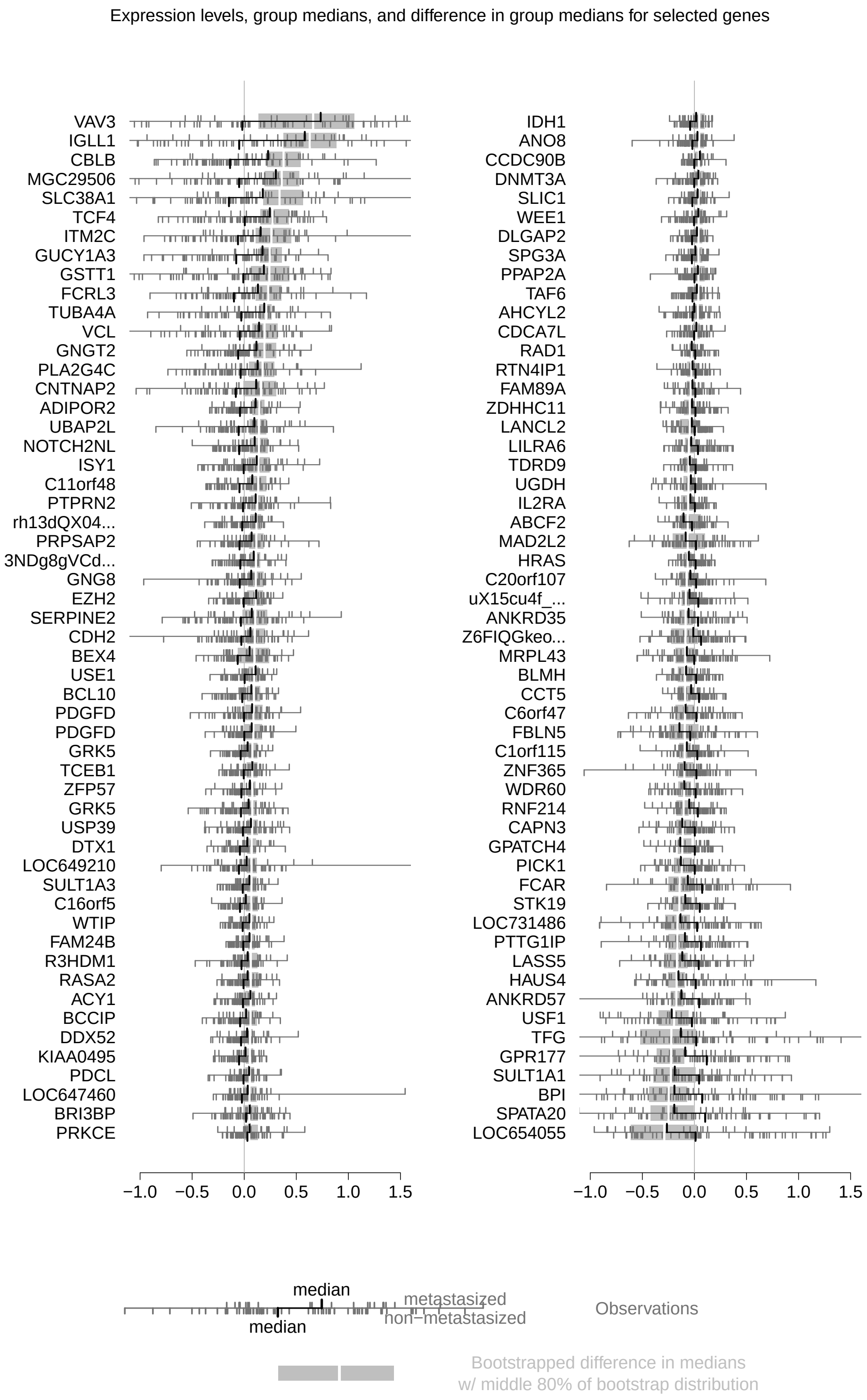


Figure 2.6: Expression levels of selected genes ordered by difference in medians between metastasized and non-metastasized observations.

By collecting pairwise counts of co-selection during the bootstrap we can form an idea about which genes tend to get selected together. We form a probability from these counts by dividing the count by the number of times either gene was selected, so a co-selection probability of unity tells us that every time one of the genes was selected, the other was also. In Figure 2.7 we show a heatmap of these pairwise co-selection probabilities. The rows and columns—it is a symmetrical matrix—are clustered so that similar co-selection patterns are closer to one another. The plot shows a group in the bottom right that tends to get selected together.

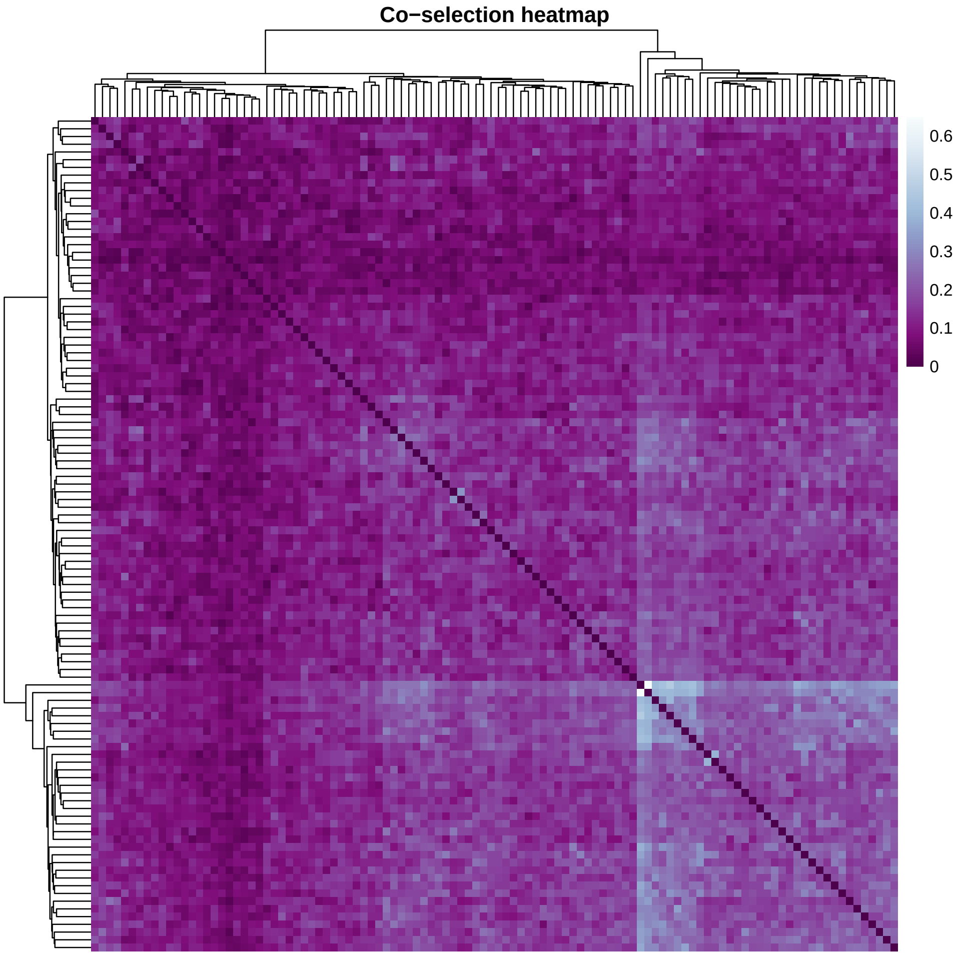


Figure 2.7: Pairwise probabilities for co-selection of genes, ordered by euclidean-distance McQuitty hierarchical clustering. The bottom-right group stands out as co-selected.

Table 2.3 shows the co-selected genes indicated in Figure 2.7. Many of these also have high individual selection probabilities in Table 2.2, so there is some indication that they reliably separate metastases from non-metastases in these data even under resampling.

Table 2.3: Genes that tend to be selected together, ordered alphabetically.

ADIPOR2
C11orf48
C1orf115
CAPN3
CBLB
DNMT3A
FAM24B
FAM89A
GNG8
GNGT2
GPATCH4
GRK5
GUCY1A3
ITM2C
LANCL2
LOC647460
LOC654055
PDCL
PDGFD
PDGFD
PRPSAP2
PTPRN2
R3HDM1
RASA2
rh13dQX0...
SERPINE2
STK19
SULT1A1
SULT1A3
TCEB1
TCF4
WEE1
Z6FIQGke0...
ZDHHC11
ZNF365

2.4 Conclusion

We have demonstrated predictability of metastasis in these data. We can, with a high probability, rank case–control pairs in terms of predicted metastasis probability. However we should not count the model itself as a reliable tool due to poor calibration and stability, and since these results stem from exploratory modeling we should be moderate in our expectations; further investigation is needed to establish reliable results.

We provide 108 candidate predictor genes as an avenue for future research. We are currently investigating their biological properties. An interesting statistical investigation may be to review the importance of the stratification and how to build this into a shrinkage model, as the results in the appendix below indicate that this may lead to improvements. We believe however that it is necessary to obtain independent data to be able to make any inference stronger than general indication.

2.A Appendix: variable selection methods

In addition to the main results presented above we previously explored various ad-hoc variable selection schemes. The results of these explorations are not competitive compared with the above penalized likelihood model, but I present them here for completeness and comparison. To make the next sections complete we must define the **followup time** of a case. This is the number of days between provision of the blood sample and the eventual diagnosis of cancer. Although followup introduces a time aspect, these are not time series data in the strictest technical sense. Each observation stems from a different woman, so there should be no autocorrelation to speak of, and followup time is random.

What follows is the "historical" text describing these variable selection methods. Take note that the flaw in methodology only affected the baseline shrinkage methods; the other numbers should be reliable.

2.A.1 Variable selection

We investigate four ways to rank genes, which we describe briefly in this section. The methods all assess differential expression between groups in some way. We propose the first—ANOVA—to take into account a hypothetical functional relationship between gene expression and time. The other three—SAM, t-tests, and LIMMA moderated t-tests—are well-established methods for ranking genes and assessing differential expression.

ANOVA

We hypothesize that the expression of genes that are relevant to the cancer process diverges over time. To detect this behavior we regress fold change, e, on time, t, and metastasis, M, in the following model:

$$e = \beta_0 + \beta_1 t + \beta_2 M + \beta_3 tM + \epsilon, \tag{2.2}$$

where ϵ is iid noise. We refer to this as **ANOVA-f** below.

We suspect that different genes may be relevant in different detection strata. We model this by expanding equation 2.2 to include an interaction with stratum, S:

$$e = \beta_0 + \ldots + \beta_4 S + \beta_5 tS + \beta_6 SM + \beta_7 tSM + \epsilon. \tag{2.3}$$

We refer to this as **ANOVA-fs** below.

Finally we entertain the possibility that followup is not important and that stratum alone is of interest. This yields the model

$$e = \beta_0 + \beta_1 S + \beta_2 M + \beta_3 SM + \epsilon, \tag{2.4}$$

which we refer to as **ANOVA-s** below. Note the abuse of notation in Equations 2.3 and 2.4; S has three levels and will be coded as a dummy variable.

We rank genes by the F-statistic obtained under the null hypothesis that the model in Equation 2.2, 2.3, or 2.4 is no better than the intercept-only model, $e = \beta_0 + \epsilon$. Ignoring both stratum and followup is equivalent to a regular t-test for metastasized vs not as in Section 2.A.1 below.

t-test

We rank genes by Welch's two-sample t-statistic (Welch, 1947) between metastasized and non-metastasized cases (**t-test** below). This is complementary to the three methods above, as regressing on a single binary grouping variable can be used as a test for difference in means.

SAM

The Significance Analysis of Microarrays (**SAM**) procedure of Tusher et al. (2001) defines the relative difference in gene expression for the ith gene as:

$$d(i) = \frac{\bar{x}_I(i) - \bar{x}_U(i)}{s(i) + s_0}. \tag{2.5}$$

Here $\bar{x}_I(i)$ and $\bar{X}_U(i)$ are the average expression levels of gene i in the two states I and U (metastasized or not), $s(i)$ is the pooled standard deviation estimate in the two states, and finally s_0 is a small positive constant added to all genes to make the variance of d_i independent of gene expression level. We rank genes by $d(i)$.

LIMMA t-test

Smyth's Linear Models for Microarray Data (LIMMA) is a general empirical Bayes framework for assessing differential expression (Smyth, 2004). The LIMMA moderated t-statistic, $\tilde{t}_i$, is similar to the SAM $d(i)$ in that it modifies the denominator of a regular t-statistic. In this case we have that for the ith gene,

$$\tilde{t}_i = \frac{\bar{x}_I(i) - \bar{x}_U(i)}{\tilde{s}_i \sqrt{v_i}},$$

where v_i is a factor that has to do with the variance of $\bar{x}_I(i) - \bar{x}_U(i)$. The standard deviation estimate $\tilde{s}_i$ has been shrunk by empirical Bayes methods toward the average standard deviation across all genes. We refer to this as **LIMMA-t** below.

2.A.2 Prediction

Having ranked genes and chosen the top k as predictors, we use these in the following logistic regression model for the probability, p, of metastasis:

$$\text{logit}(p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k. \tag{2.6}$$

This model uses only gene expression levels regardless of whether stratum (or time) was used in selecting the predictors. This model can be used in a screening setting (where the cancer has not yet happened and hence we do not have information about its detection). Since followup time is a result of these data coming from a cohort study, it could not be used in a realistic predictive model.

Considering Table 2.1 it is likely that detection type is informative of the probability of metastasis. Conceivably a predictive model could be used at time of diagnosis where this information is available. Our model for such a setting is simply Equation 2.6 with an extra interaction with stratum, much like in the variable selection in Section 2.A.1:

$$\begin{aligned}\text{logit}(p) =&\beta_0 + \ldots + \beta_{k+1}S + \beta_{k+2}Sx_1 + \\ &\ldots + \beta_{(2k+1)}Sx_k.\end{aligned} \tag{2.7}$$

We estimate models 2.6 and 2.7 by Bayesian generalized linear models with a weakly informative prior from Gelman et al. (2008). This is more for convenience than from a particular wish to do Bayesian modeling: when selecting the k “best” predictors out of thousands of candidates it is quite likely to find some where the metastasis and non-metastasis points are linearly separable (ie. their respective convex hulls are disjoint). In such a setting, the classical iteratively reweighted least squares optimization does not converge. Predictors selected for some function of their effect size would likely regress toward the mean in new data and some amount of shrinkage is prudent. The standard prior of Gelman et al. provides a sensible and convenient regularization without a need for parameter tuning.

2.A.3 Baseline

We compare predictive performance against two naive and two more sophisticated baselines. The more sophisticated baselines use all genes without prior ranking and selection.

The first naive model we consider is the “random guess” intercept-only model $\text{logit}(p) = \beta_0$. Second we compare against using the stratum information, $\text{logit}(p) = \beta_0 + \beta_1 S$, corresponding to making a recommendation based only on the manner in which the cancer was detected.

The other two baselines are penalized logistic regression models. These models take the same form as Equation 2.6, using all predictors rather than the top k, but maximize the likelihood subject to the constraint that $c(\hat{\beta}) \leq t$. Ie. the magnitude $c(\cdot)$ of the coefficients β_i must not exceed some threshold t. We investigate ridge penalty, $c_r(\hat{\beta}) = \sum \hat{\beta}_i^2$, and the lasso penalty, $c_l(\hat{\beta}) = \sum |\hat{\beta}_i|$ (Tibshirani, 1996). These are well-known models. The lasso provides an end-to-end solution that does variable selection and model fitting in one go. Using a ridge penalty simply uses all predictors but shrinks coefficients toward zero quadratically in their magnitude. Both of these methods require the selection of t; we choose this by cross validation.

Standard errors

We measure uncertainty in the bootstrap estimates by the jackknife-after-bootstrap procedure. The jackknife estimate of standard error for any statistic $\hat{\theta}$ is

$$\hat{\sigma}_J = \sqrt{\frac{n-1}{n} \sum (\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)})^2}, \tag{2.8}$$

where $\hat{\theta}_{(i)}$ is the statistic computed with the i-th sample removed, and $\theta_{(\cdot)} = \frac{1}{n} \sum \hat{\theta}_{(i)}$. In principle the bootstrap procedure has to be repeated for each $\hat{\theta}_{(i)}$. But there is a computational shortcut here due to the fact that a bootstrap sample drawn with replacement from $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots x_n$ has the same distribution as a bootstrap sample drawn from $x_1, \ldots, x_n$ in which x_i does not appear (the *jackknife-after-bootstrap lemma* in Efron and Tibshirani (1994)).

2.A.4 Results

Below all bootstrapped results are based on 2 500 resamples. For all ranking methods we choose the top ten genes and use them as predictors for the models described in Equations 2.6 and 2.7. This is based on the folk wisdom to have around ten observations per estimated parameter in the regression model. In practice we end up with fewer observations per parameter, especially for model 2.7, so the models are slightly over-parameterized. This likely contributes to uncertainty in our results.

Tables 2.4 and 2.5 show the Brier score and AUC for predictions based on the different selection schemes we investigate. Models 2.6 and 2.7 refer to Equations 2.6 and 2.7 in Section 2.A.1. That is, respectively, the “screening” prediction using only gene expression in the model, and the “at diagnosis” prediction that uses the additional information of detection stratum.

Brier score is a measure of error: the lower the better. Table 2.4 shows Brier score as point estimate plus/minus two standard errors in decreasing order by Model 2.7 point estimate. The results do not suggest any simple interpretation,

	model 2.6	**model 2.7**
t-test	.17 ± .45	.17 ± .33
ANOVA-fs	.27 ± .13	.18 ± .10
SAM	.34 ± .11	.20 ± .15
ANOVA-s	.33 ± .22	.20 ± .25
ANOVA-f	.31 ± .084	.21 ± .11
LIMMA-t	.35 ± .14	.20 ± .17
intercept	.19 ± .010	
stratum	.22 ± .029	
lasso	.27 ± .19	
ridge	.23 ± .30	

Table 2.4: Brier scores presented as point estimate plus/minus two standard errors. Measures error in forecast probability: lower is better. Model number refers to the equations in Section 2.A.2. Model 2.7 includes stratum as a predictor. Below the break are the four baseline models.

but it is noteworthy that the intercept-only model is among the best-calibrated. The uncertainty is large enough that is difficult to say that any selection method is better than any other. It is clear that the interaction with detection method in model 2.7 improves calibration for all models. There is also lower uncertainty in the ANOVA-f/fs models.

AUC or concordance probability is a measure of a model's ability to discriminate between outcomes: the higher the better. Brier score alone does not provide full information about predictive performance; the intercept-only model is well-calibrated but cannot be used for prediction at all. Random guess (or forecasting a constant for every observation) yields AUC of .5; perfect discrimination yields AUC of unity. Table 2.5 shows AUC as point estimate plus/minus two standard errors in decreasing order by model 2.7. Again the clearest signal is that the added information from detection method is very important. Point estimates improve markedly and standard errors generally decrease. Also here does use of stratification and followup time in preselection reduce uncertainty.

The ridge regression baseline performance has a very good AUC point estimate, but the standard error is very large. Too large: it is a theorem that the upper bound on standard deviation in a variable $\in [0,1]$ is $\frac{1}{2}$. This says something about the imperfection of the jackknife as an estimator of standard error. The blame lies at least in part with the correctional factor $\frac{n-1}{n}$ in Equation 2.8, which was originally defined heuristically. Since it is difficult to suggest a sensible alternative, we choose to live with this.[2]

[2] This was really the result of nesting a cross-validation in the bootstrap: the methodology

	model 2.6	model 2.7
LIMMA-t	.44 ± .30	.76 ± .20
SAM	.46 ± .26	.75 ± .24
ANOVA-fs	.51 ± .29	.75 ± .16
ANOVA-s	.41 ± .57	.75 ± .38
t-test	.65 ± 1.5	.74 ± .71
ANOVA-f	.44 ± .25	.72 ± .21
intercept	.5	
stratum	.49 ± .055	
lasso	.36 ± 1.4	
ridge	.81 ± 3.3	

Table 2.5: AUC presented as point estimate plus/minus two standard errors. Measures the probability of forecasting a higher probability of metastasis for a randomly chosen metastasis case than for a randomly chosen non-metastasis case: higher is better. Model number refers to the equations in Section 2.A.2. Model 2.7 includes stratum as a predictor. Below the break are the four baseline models.

The collected results for model 2.7 suggest some reason for optimism. Due to the size of the standard errors we must necessarily be uncertain about even the first significant digit of our point estimates. But even accounting for uncertainty there seems to be predictive information better than random guess. As in the simulations, there is not too much difference between the different methods, perhaps apart from the simple t-test, for which we observe much variance. Note that both SAM and LIMMA are flexible frameworks and we could have accounted for stratum and followup in either. Our comparison is between using this information and various ways of not using it, and there is no reason to believe that either framework should perform poorly if we were to use more refined models there.

Table 2.6 shows the predictor set stability as point estimate plus/minus two standard errors. Stability is in general very low, and the standard errors suggest that there is even some uncertainty to the order of magnitude of the point estimates. A possible interpretation is that the correlation between genes is such that many different genes hold similar information. It is at least clear that we need much more data if we want to find a stable set of predictor genes. If we take the point estimates at face value, Table 2.6 reflects the fact that we see lower uncertainty using ANOVA-f/fs in Tables 2.4 and 2.5.

issue mentioned in the preamble to this chapter. For details see Section 5.2.2 and Section 5.3.

ANOVA-f	.095 ± .15
SAM	.070 ± .10
ANOVA-fs	.067 ± .16
LIMMA-t	.061 ± .10
ANOVA-s	.055 ± .086
t-test	.00036 ± .0039
lasso	0 ± .26

Table 2.6: Stability as point estimate plus/minus two standard errors. Stability is an estimate of the probability of recovering the same gene set with different realizations of a modeling procedure. A larger stability provides more certain biological interpretation. The lasso is the only baseline method included here as it is the only one that does variable selection.

3 Standardized data cleaning

“Nanny Ogg never did any housework herself, but she was the cause of housework in other people.”

–Terry Pratchett, *Lords and Ladies*

This chapter describes the NOWAC standard operating procedure (SOP) for quality assessment of our microarray material and the text is mostly the same as in Bøvelstad et al. (2017). The work has led me to think a lot about how process is managed and communicated in a large academic project, something I get back to in Section 5.4. Although we use the term “large-sample” below this really means large in the epidemiological human-model transcriptomics scale. For an idea of this scale, see Section 5.1.

Abstract: Transcriptome measurements and other -omics type data are increasingly more used in epidemiological studies. Most of omics studies to date are small with samples sizes in the tens, or sometimes low hundreds, but this is changing. Our Norwegian Woman and Cancer (NOWAC) datasets are to date one or two orders of magnitude larger. The NOWAC biobank contains about 50000 blood samples from a prospective study. Around 125 breast cancer cases occur in this cohort each year. The large biological variation in gene expression means that many observations are needed to draw scientific conclusions. This is true for both microarray and RNA-seq type data. Hence, larger datasets are likely to become more common soon. Technical outliers are observations that somehow were distorted at the lab or during sampling. If not

removed these observations add bias and variance in later statistical analyses, and may skew the results. Hence, quality assessment and data cleaning are important. We find common quality assessment libraries difficult to work with for large datasets for two reasons: slow execution speed and unsuitable visualizations. In this paper, we present our standard operating procedure (SOP) for large-sample transcriptomics datasets. Our SOP combines automatic outlier detection with manual evaluation to avoid removing valuable observations. We use laboratory quality measures and statistical measures of deviation to aid the analyst. These are available in the nowaclean R package, currently available on GitHub.[1] Finally, we evaluate our SOP on one of our larger datasets with 832 observations.

3.1 Introduction

The use of –omics data in epidemiological studies is now common. Typical studies comprise sample sizes in the tens or low hundreds, but sizes in the order of thousands will soon be common. The NOWAC postgenome cohort (Lund et al., 2008) contains blood samples from 50000 women. In this cohort there are approximately 125 new breast cancer cases per year, and we have thus far extracted and processed blood samples from 1660 case—control pairs, or 3320 blood samples in total. Omics experiments are elaborate procedures with several steps. In the case of microarrays, these include mRNA isolation, hybridization, washing, and scanning. Each step may add random or systematic errors. Technical errors may also come from supplies or instruments. Mishaps may occur in the lab. The samples themselves can get contaminated in various ways. In whole-blood samples, there is also the added challenge of mRNA degradation due to high RNase activity. All this may be detrimental to the quality of the data and hence affect downstream analyses.

The goal of gene expression experiments is to detect differences in gene expression levels between groups. This is usually evaluated gene-by-gene or for sets of related genes. The methods for such analyses depend on accurate estimation of the sample variance. If there are technical outliers contributing unnecessary variance, removing these should increase power. However, removing biological outliers will result in underestimation of the natural biological variance. This in turn will increase the risk of spurious conclusions. There is a fine line to tread, and the accurate identification of technical outliers is important for later analysis.

Many publications guide the identification of outliers in gene expression

[1] `https://github.com/3inar/nowaclean`

data (Cohen Freue et al., 2007; Kauffmann et al., 2008; Kauffmann and Huber, 2010; Shieh and Hung, 2009). Yet, there is no real consensus on the best approach. For example, some authors such as Marczyk et al. (2014) propose automated procedures for outlier removal. Others such as Kauffmann and Huber (2010) warn against automation and instead recommend careful investigation.

Outlier removal is particularly challenging for studies based on blood samples, since there is larger biological variation in gene expression data from blood than in tumor tissue (Yang et al., 2006). The strength of the signal in tumor tissue makes it much more robust to variance than the signal in blood samples, which is weak and variable. This makes it more difficult to distinguish outliers from non-outliers and signal from noise. It is not well known whether lifestyle factors like medication use affect blood gene expression. All this complicates outlier identification, and it's inadvisable to remove outliers in a systematic, automated way.

R-packages such as arrayQualityMetrics (AQM) (Kauffmann et al., 2008) and lumi (Du et al., 2008) implement the most popular outlier detection methods for gene expression data. Important to these approaches is the combination of computational methods with interactive visualization. However, when dealing with several hundreds of observations, these methods are cumbersome for two reasons. First, some methods are slow and thus inefficient for interactive use. Second, their visualizations do not work well for larger sample sizes due to overplotting. The latter is in our opinion the most important aspect, as the decision to remove an outlier often rests on visual inspection by the analyst.

We also wish to provide numerical measures and standardized guidelines to help the user. The measures are statistics of deviation, derived from the data, and laboratory quality metrics. We believe that standardization removes some of the subjectivity from the task. Standardizing the outlier removal procedure as much as possible will enhance reproducibility and consistency.

Below we describe our standard operating procedure (SOP) for outlier removal in large-sample transcriptomics datasets. We believe our SOP will strengthen the reporting of observational studies in epidemiology (Von Elm et al., 2007). We have implemented the SOP as an open source R package that combines automated outlier detection with expert evaluation. The automated part consists of ranking observations by deviation metrics. We base these metrics on standard methods for outlier removal in data from microarrays. Our improvements are faster execution and easier-to-read visualizations. We provide a unified, interactive interface, saving computations for tinkering with thresholds and using standard R methods where available. We use data from the

NOWAC study (Dumeaux et al., 2008) for demonstration and evaluation.

3.2 Methods

3.2.1 Data

The NOWAC study is a nation-wide, population-based cancer study (Lund et al., 2008). A thorough description of the NOWAC postgenome cohort can be found in Dumeaux et al. (2008). To summarize: 97.2% of the women in the NOWAC cohort consented to donate a blood sample to research. Out of these, about two thirds ended up providing an actual blood sample. Blood sampling kits were sent out in batches of 500. These kits included a two-page questionnaire and a PAXgene tube (PreAnalytiX GmbH, Hembrechtikon, Switzerland). For the most part, the family general practitioner drew the actual blood sample. The sample was then mailed overnight to Tromsø. Between 2003 and 2006 the NOWAC biobank grew to comprise 48,692 blood samples. These make up the NOWAC postgenome cohort. The Norwegian Cancer Registry provides yearly updates about cancer cases. Statistics Norway provides yearly updates about emigrations and deaths. A control sample is assigned to each breast cancer case in the cohort yielding a nested case-control design. These are matched on mailing batch, time of blood sampling and year of birth. We keep each case—control pair together through every step in the laboratory. The statistical analysis of microarray data is described in Lund et al. (2016).

In this paper, we use a subset of 832 observations from the NOWAC cohort. The Genomics Core Facility at the Norwegian University of Science and Technology provided the laboratory work. They processed the samples on Illumina Whole-Genome Gene Expression Bead Chips,[2] HumanHT-12 v4. The raw microarray images are processed in GenomeStudio.[3] This is Illumina's own software for processing data from their platforms. The result is a table of 47323 probes for 832 observations on the summary level: one number per probe per observation.

[2] `http://technology.illumina.com/technology/beadarray-technology.html`

[3] `http://bioinformatics.illumina.com/informatics/sequencing-microarray-data-analysis/genomestudio.html`

3.2.2 The NOWAC pipeline

The outlier SOP is part of our data processing pipeline in NOWAC. The pipeline (Figure 3.1) contains three major steps where outlier removal is Step 2. We briefly describe the data preparation (Step 1) and the preprocessing (Step 3) to provide context for the SOP.[4]

STEP 1.1: Described in the "Data" section above.

STEP 1.2: Microarray gene expression measurements from the lab are merged into an R LumiBatch-object (Du et al., 2008) based on a unique lab number, along with external information from questionnaires, the Norwegian Cancer Registry, and Statistics Norway.

STEP 1.3: Yearly updates from the Cancer Registry can reveal that controls have become cases, or that cases have received a second cancer diagnosis. We considered these individuals non-eligible, and remove them along with their matching case/control.

For multivariate analyses, we remove 38 probes related to blood type, specifically the human leukocyte antigen (HLA) system. These are usually expressed strongly and have high variance, which will affect multivariate analyses. We have seen that they can dominate the variance-covariance pattern in the principal component analysis (PCA) transformation of the data (will be described in detail in the next section), and as such other patterns might be obscured. This is relevant for our SOP as we do PCA, so we recommend to take these out before outlier detection. It is possible to put these probes back after outlier detection. The decision will depend on whether the genes are interesting for subsequent analyses.

STEP 2: Described in detail below in the "Outlier Removal SOP" section.

STEPS 3.1 and 3.2: We apply the normal-exponential background adjustment method to make signals comparable across individuals (Plancade et al., 2012; Xie et al., 2009). We also use quantile normalization (Bolstad et al., 2003) and log2-transform the data to stabilize the variance.

STEP 3.3: Batch effects are systematic errors introduced when processing blood samples in multiple batches in the laboratory. Examples of a batch are all chips that are processed at the same day (named plate), laboratory technician, or the batch of laboratory regents used. It's important to adjust

[4] The pipeline has been developed further since this was written and may have changed in some particulars.

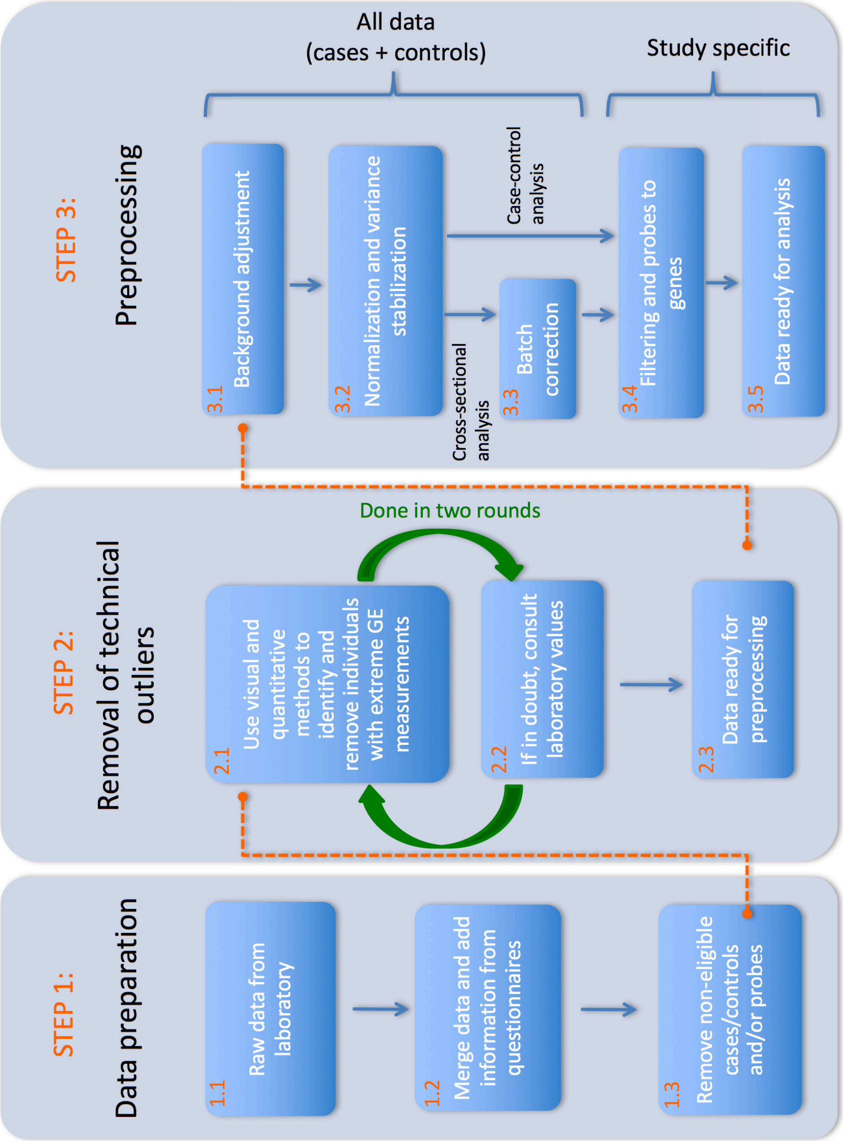

STEP 1:
Data preparation
1.1 Raw data from laboratory
1.2 Merge data and add information from questionnaires
1.3 Remove non-eligible cases/controls and/or probes
STEP 2:
Removal of technical outliers
Done in two rounds
2.1 Use visual and quantitative methods to identify and remove individuals with extreme GE measurements
2.2 If in doubt, consult laboratory values
2.3 Data ready for preprocessing
STEP 3:
Preprocessing
All data (cases + controls)
Study specific
3.1 Background adjustment
3.2 Normalization and variance stabilization
Cross-sectional analysis
Case-control analysis
3.3 Batch correction
3.4 Filtering and probes to genes
3.5 Data ready for analysis

Figure 3.1: NOWAC (10) standardized data analysis pipeline for cleaning and preprocessing the data. The pipeline is split into three steps, where all steps up to and including Step 3.3 are performed using all cases and controls that are eligible and not considered as outliers. Step 3.4 needs to be performed for each specific study/question at hand, fine-tuning the data to optimize the power of the analysis. Abbreviation: GE=Gene expression.

for batch effects with methods like e.g. ComBat (Johnson et al., 2007).

STEP 3.4: We filter out probes that are likely to be below the level of detection, probes that are expressed in only a few arrays, and probes that are known to have unreliable annotation. This reduces the number of probes and the risk of false positives in subsequent analysis.

3.2.3 Outlier removal SOP

Outlier removal is a subjective task. Our SOP combines guidelines, visualizations, and quantitative measures to help. An outlier should only be excluded if it is of technical origin, since biological outliers are valuable. Technical measures from the microarray lab describe the quality of the blood sample. They include information on RNA abundance, mRNA contamination, etc. They may provide hints to why an array might look wrong and help make the distinction between technical outliers and biological outliers. We describe the lab measures in detail further below. As extreme outliers have a strong influence on many of the plots and measures we use, we do outlier detection/removal in two rounds. In case—control designs, when an outlying observation is removed, the matching case/control will also be removed. An overview of the SOP is provided below:

1. Log2-transform your data to ameliorate heteroscedasticity. This is because you can expect higher variance for signals with higher intensity.

2. Find outlier candidates by looking at different views of the data with the methods detailed below: MA-plots, PCA-plots, and boxplots.

3. Investigate each candidate outlier by examining density plots and lab quality assessment measures, as described below.

4. Exclude observations that look irreparably strange. When in doubt, the standard cutoffs for lab measures may provide insight and help take a decision. Repeat steps 2–4 once more to be sure that no outliers are left in the data.

We evaluate individual array quality with array-wise MA-plots (Dudoit et al., 2002) where we compare each array with the median array. An MA-plot is a mean—difference plot that compares two assays on the $\log_2$ scale. Specifically, let A1 be a given array, and A2 the median array constructed by taking gene-wise medians over all arrays. Then, compute the two statistics: $M = \log_2 \frac{A_1}{A_2}$, and $A = \frac{1}{2} \log_2 A_1 A_2$. You should expect M to be constant as a function of A

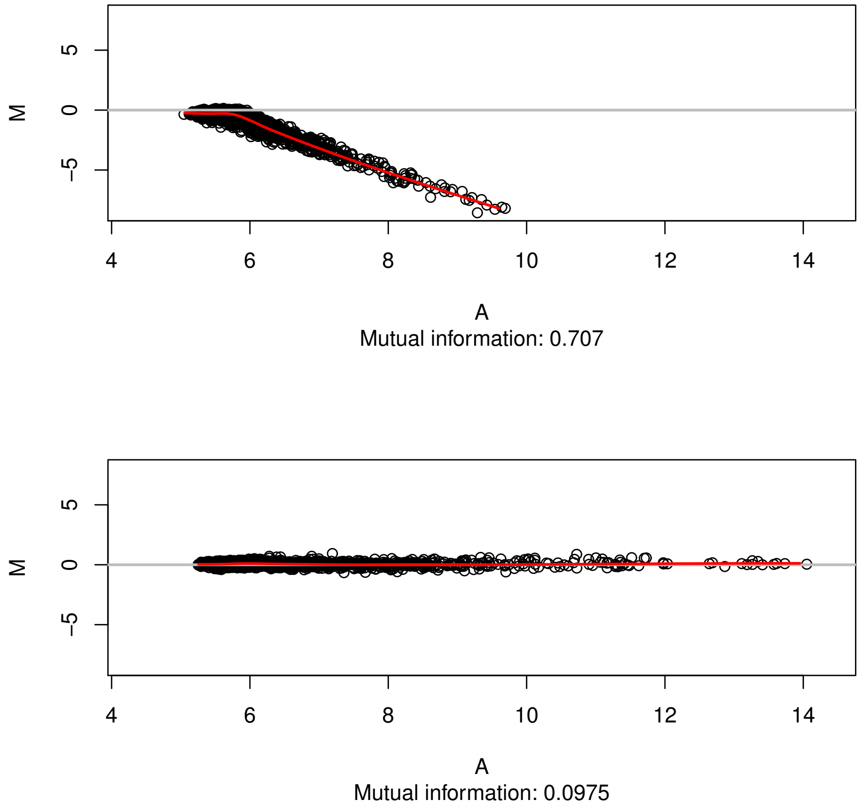


Figure 3.2: Illustrations of MA plots. On top: a potential outlier array. On bottom: a well-behaved array.

in well-behaved arrays (Figure 3.2, bottom panel). A trend in M as a function of A would indicate that gene expression values are somehow systematically skewed away from the median array (Figure 3.2, top panel).

You can measure independence of M and A in several ways, AQM uses Hoeffding's D statistic, which measures squared deviance from independence (difference between joint density and product of marginal densities). We use the similar measure of mutual information (MI), defined as

$$I(A, M) = \sum_{m \in M} \sum_{a \in A} p(a, m) \log \frac{p(a, m)}{p(a)p(m)},$$

simply because the R-implementation is considerably faster (Hausser and Strimmer, 2009). The joint density of the two statistics must be discretized, so some information may be lost. For this and many other reasons it's important to inspect the outliers yourself.

We evaluate homogeneity between arrays by inspecting boxplots. As we have hundreds to thousands of arrays, doing regular boxplots will result in overplotting. Hence, we use "compressed" boxplots where each quantile is represented by a single point, and the points for corresponding quantiles are connected by lines. The same is implemented for the lower and upper whiskers of the boxplot, given by the most extreme data point within 1.5 times the interquartile range from the median. This results in a plot with five continuous horizontal lines (Figure 3.3). We measure deviation from normal data by comparing the empirical cumulative distribution function (ECDF) of the expression intensities for each array with the ECDF of all arrays pooled. Distance from the pooled ECDF is measured by the Kolmogorov-Smirnov (KS) statistic (Wasserman, 2010) , which measures the largest distance between two distribution functions. By default, we order the boxplots by their respective KS statistic, but it may also be interesting to order by other things such as plate number to look for batch effects.

We define outliers as those observations that fall outside m standard deviations from the mean observed KS statistic. The value of m will depend on how conservative the analyst is in its search for outliers. The higher value of m, the fewer individuals will be marked as outliers. For the example shown in Figure 3.3 we used m = 3 (indicated by a red line).

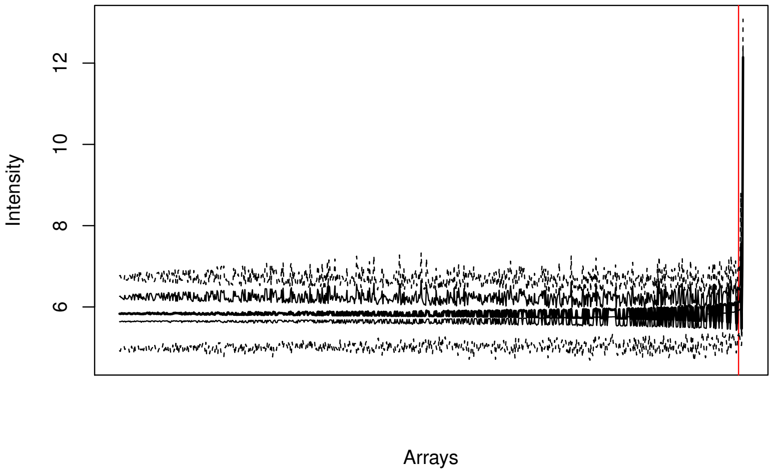


Figure 3.3: Boxplots, ordered by the KS-statistics. The individuals with a boxplot to the right of the red line has a KS-statistics above m=3 standard deviations from the mean observed KS-statistics. Points in this plot represent the components of a boxplot: median, lower and upper quartiles, lower and upper whiskers. Points of the same type (e.g. median) are joined by a line.

For between-array comparison we apply principal component analysis (PCA) (Hastie et al., 2009) to the data, and display the first two principal components in a scatterplot. As a quantitative measure to guide the outlier identification, we compute the Mahalanobis distance of all arrays to the mean array. In PCA-transformed data the Mahalanobis distance and Euclidean distance are necessarily equal. As shown in the top panel of Figure 3.4 there can be a distinct shape and rotation to the data that's not captured due to outliers. Hence, we define a "central cluster" that's used to compute distances. We obtain this central cluster by ignoring points that are less likely than 99% (adjustable) by Chebyshev's inequality. This leads to distances that fit the shape and rotation of the data better, see the bottom panel of Figure 3.4. Outliers are then defined as those more than n standard deviations from the data center in Mahalanobis distance. Once again, the value of n must be determined by the analyst; We used n = 3 in the analysis below.

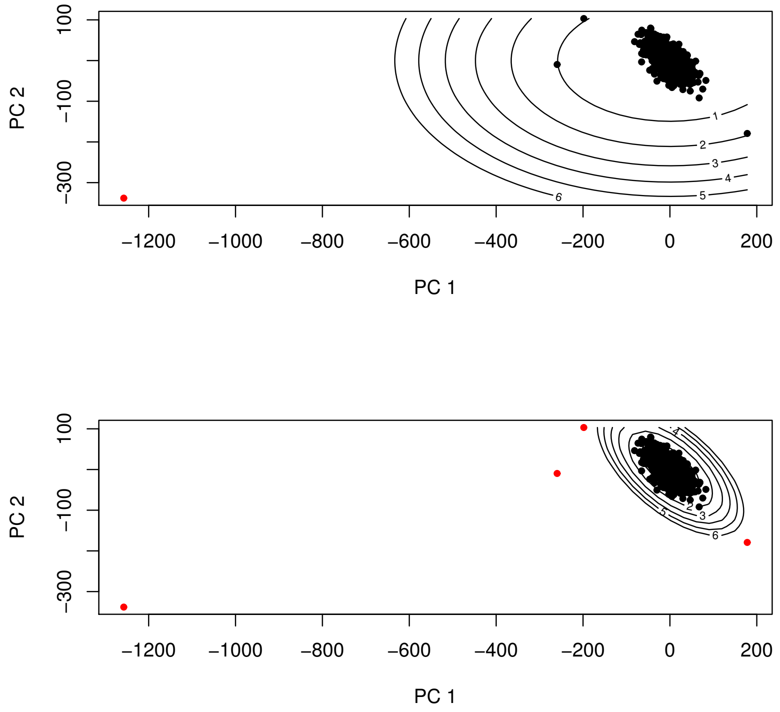


Figure 3.4: PCA plots. The lines show Mahalanobis distance to the center of the data (in standard deviations). The red points are considered potential outliers as they are farther away than two standard deviations. The top panel shows the distances computed from all data points. The bottom panel shows the same when leaving out the least likely points by Chebyshev's inequality.

Finally, we use density plots to inspect observations we suspect to be technical outliers based on the methods described above. These plots show distribution properties that are hidden in the other plots like severely skewed modes or several modes, neither of which you should expect to see in well-behaved data. Figure 3.5 below in the Results section shows an example of a density plot.

After exploring different outlier detection methods, the analyst is left with a selection of outliers and must decide which are technical outliers that should be discarded. Several technical measures from the laboratory may help guide this decision. These measures include information on RNA abundance; the quality of the blood sample in terms of mRNA degradation, quantified by RNA Integrity Number (RIN); and the level of contamination in the blood sample, quantified by NanoDrop 260/230 and 280/230 ratios. These values may help the analyst understand why some observations have outlying values. For borderline outliers, where the analyst is uncertain, we provide standard exclusion thresholds for each lab measure:

- RIN value < 7,
- 260/280 ratio < 2,
- 260/230 ratio < 1.7,
- and, RNA abundance outside the range of (50, 500)

If the observation is suspect and it falls outside of any of these thresholds, it may be regarded as an outlier and thus discarded. It is entirely possible for observations to look perfectly sensible despite bad lab measures, hence we don't exclude observations based purely on these numbers. We consult the lab measures only once we suspect an array to be a technical outlier based on the plots.

3.2.4 The nowaclean R package

The nowaclean R package (`https://github.com/3inar/nowaclean`) implements our standard operating procedure for detecting and removing technical outliers in the NOWAC microarray data. As mentioned above, the functionality we provide already exists elsewhere. The novelty of this R-package is the improved speed and visualizations for data sets with a large number of arrays. We have through this work identified four design principles that we believe improve the user experience: i) save computations so that users can tune thresholds; ii) force the use of names instead of indices into a matrix, in case several representations of the data are in use; iii) have a unified interface to the different methods: always use R's standard predict and plot methods, and provide the same set of arguments to these as far as possible; and finally iv) decouple the methods from special types of objects such as the Bioconductor standard `esets` and work on built-in matrices instead. This last point is to provide functionality to a broader user base.

3.2.5 Evaluation

To study how our SOP affects downstream analysis we need to quantify the effect of the outlier removal. As removing individuals may reduce power, removing technical outliers identified in our SOP should on the contrary increase the power and make sure that the downstream analysis leads to more sound and biologically reliable results. One way to quantify the effect of the SOP outlier removal is to count the number of genes that are differentially expressed between cases and controls before and after outlier removal, as described in Marczyk et al. (2014). A gene G is differentially expressed between cases and controls if $\mu(G_{\text{cases}} \neq \mu(G_{\text{controls}}$, i.e. the average expression $\mu(G)$ of gene G in one group is different from that of the other, as determined by the limma moderated t-test Smyth (2004).

We will examine the number of significant findings in two situations. First in a situation where we know there is no difference between groups. We create a pseudosample by assigning observations to groups randomly to ensure no relationship between group and gene expression levels (i.e. permutation). In this pseudosample we compute the statistic $\theta = \frac{\#\text{null rejections}}{\#\text{genes}}$, i.e. the proportion of significant genes. To get a distribution over θ, we will repeat the procedure for 1500 random pseudosamples. In this situation there is no difference between groups, and thus the null hypothesis should be rejected for about 5% of all genes tested when using a significance level of $\alpha = 0.05$. Hence θ is in effect an estimate of type-I error rate, the effective size of the hypothesis test.

In the second situation, we count the number of rejected null hypotheses when we expect a difference between groups. We will use an anonymized group variable from the NOWAC questionnaires, and generate data where we draw (observation, group) pairs with replacement from the real data, replicating the original dataset size (i.e. bootstrapping). We then compute the statistic defined above, and repeat the procedure for 1000 bootstrap samples. By doing this we get an estimate of the variance of θ, and not just a point estimate. In this situation θ says something about statistical power, but is not a direct estimate. We are primarily interested in a comparison of outlier removal strategies, so change in θ is what is most important.

The fraction of outliers to non-outliers is likely to be small, and their effect might be subtle in large samples. For this reason, we will inspect three dataset sizes: all of our data, half the data, and 10% of the data. We do this by, for each new pseudosample, removing the correct fraction of observations from the full data set but making sure that the identified outliers are kept in the pseudosample. This is done for both the permutation and the bootstrapping experiment. We then perform the preprocessing described in STEP 3 in Section 3.2.2, and finally compute for the data with and without outliers.

As for any procedure involving hypothesis testing you ideally want as high a statistical power as possible, and it's nice if you get the correct test size. That is, you want as many type-I errors as you'd expect so that $\alpha = \theta$ at the α level.

3.3 Results

We demonstrate our methods on a typical NOWAC raw data set comprising 47323 probes for 832 observations. After applying our SOP, we have identified four observations as technical outliers. We describe this process in detail in Appendix A. We used version 0.2.8 of nowaclean for these computations.

Figure 3.5 shows the expression densities of the four outliers (red lines) along with all the other observations (black lines). If by some chance the three right-skewed observations are e.g. all cases and the left-skewed observation is a control, the result would almost certainly be overestimation of differential expression. These same four observations are the ones highlighted in red above in the PCA plot of Figure 3.4.

We remove the four observations we consider technical outliers and compare with a fully-automated approach where we remove all suggested outliers without looking at them. This is done in one round with a cutoff of two standard deviations for all three methods of boxplots, PCA, and MA-plots. Accepting all outliers results in the removal of 59 observations. We also compare against removing no outliers. We refer to these three approaches as manual-, automatic-, and no outlier removal below.

Figure 3.6 below shows our results. There are 18 experiments in total. There are three outlier removal strategies: manual removal, no removal, and automatic removal. There are three dataset sizes: all data, half of the data, and 10% of the data. Finally, there are two hypothesis testing situations: one where we

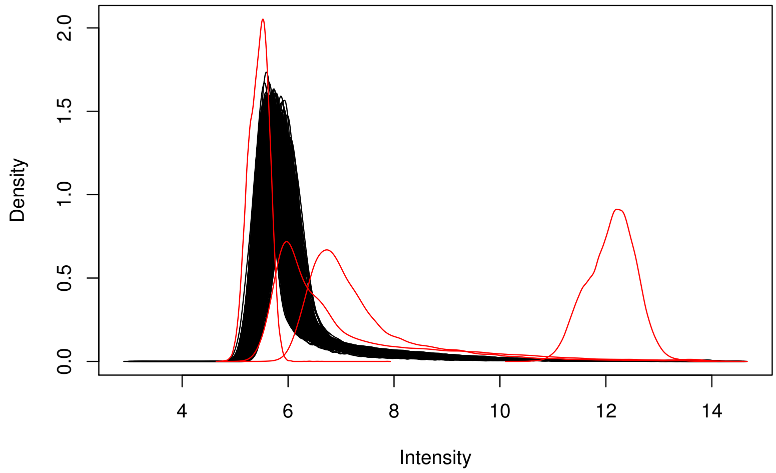


Figure 3.5: Densities of gene expression intensity across the arrays of the four SOP outliers (in red) along with the densities of the rest of the data (black lines).

expect no difference between groups (denoted as the permutation situation), and one where we do expect differences (the bootstrap situation) for some genes. We show the usual type of boxplot: median, quartiles, and whiskers that extend to the outmost points no farther away from the median than 1.5 times the interquartile range.

In the permutationsituation there is no clear advantage to any method, though the no outlier removal strategy has a slightly lower θ than expected, which is especially clear for the smallest dataset. Both manual and automatic outlier removal push this value closer to the 0.05 error rate you would expect. We suspect we still get a slightly lower error rate than 0.05 due to dependence between genes.

In the bootstrap situation careful, manual outlier removal improves power over no outlier removal for all dataset sizes. This effect goes away for the automatic removal as you start removing useful information. For the smallest dataset, any removal is better than none at all.

All in all, there is some evidence that manual outlier removal increases power and that it calibrates your error rate under the null.

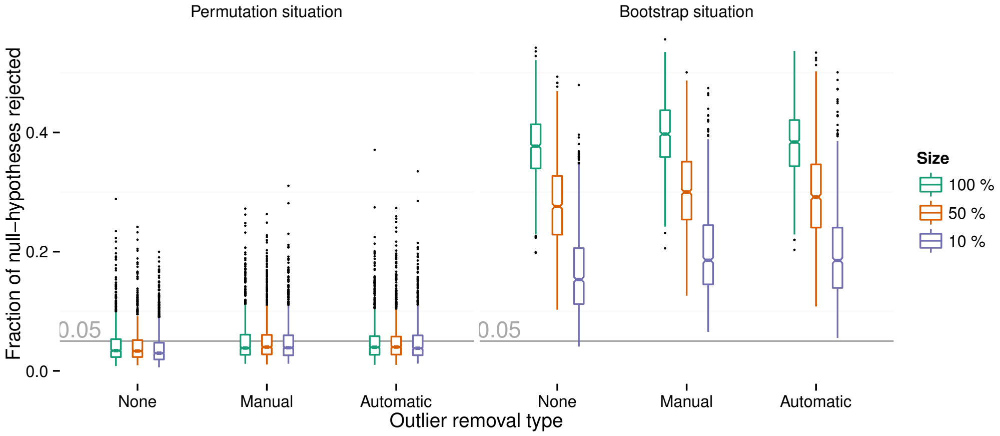


Figure 3.6: Fraction of null-hypotheses rejected at a 5% significance level. The boxplots are the standard kind, with whiskers extending to the most extreme points within 1.5 times the interquartile range from the median. We have examined three different data sizes: all 832 observations of our data (green boxes), half of these data (orange), and 10% of them (violet). There are three different approaches to outlier removal: no outlier removal, manual removal, and automatic removal. We have examined rejection rates in two situations, one where the null hypothesis of no difference between genes is true, and one where we expect to observe a difference between the two groups for some genes. We see slightly improved error rate calibration and increased power for manual outlier removal in all cases.

3.4 Conclusion

This paper describes the NOWAC standard operating procedure for the removal of technical outliers. We have described the methods we use and provide an R-package implementation. By defining a common set of methods and lab measure cutoffs to detect and evaluate technical outliers, we believe we ensure greater consistency in the preprocessing of large sample microarray data sets. Further, by providing a detailed stand-alone documentation of how we do this, we believe we make it easier to understand and reproduce the research conducted.

4

Shrinkage estimation

"You had to make choices. You never got told which ones were right. Oh, some of the priests said you got given marks afterwards but what was the point of that?"

–Terry Pratchett, *Carpe Jugulum*

This article is a case-study I wrote together with Dr. Vittorio Perduca (Holsbø and Perduca, 2018). It is written with a pedagogical purpose in mind targeted at advanced undergraduate and beginning graduate students in statistics as a tutorial around shrinkage estimation and Bayesian methods. Shrinkage is a central technique in a small-n-large-p setting such as in microarray data. We have framed the problem in a setting that is easier to reason about: the estimation of rates. Data and code for all our analyses, figures, and simulations are available at `https://github.com/3inar/crime_rates`

Abstract: This paper presents a simple shrinkage estimator of rates based on Bayesian methods. Our focus is on crime rates as a motivating example. The estimator shrinks each town's observed crime rate toward the country-wide average crime rate according to town size. By realistic simulations we confirm that the proposed estimator outperforms the maximum likelihood estimator in terms of global risk. We also show that it has better coverage properties.

4.1 Introduction

4.1.1 Two counterintuitive random phenomena

It is a classic result in statistics that the smaller the sample, the more variable the sample mean. The result is due to Abraham de Moivre and it tells us that the standard deviation of the mean is $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, where n is the sample size and σ the standard deviation of the random variable of interest. Although the equation is very simple, its practical implications are not intuitive. *People have erroneous intuitions about the laws of chance,* argue Tversky and Kahneman in their famous paper about the law of small numbers (Tversky and Kahneman, 1971).

Serious consequences can follow from small-sample inference ignoring deMoivre's equation. Wainer (2007) provides a notorious example: in the late 1990s and early 2000s private and public institutions provided massive funding to small schools. This was due to the observation that most of the best schools—according to a variety of performance measures—were small. As it turns out, there is nothing special about small schools except that they are small: their over-representation among the best schools is a consequence of their more variable performance, which is counterbalanced by their over-representation among the worst schools. The observed superiority of small schools was simply a statistical fluke.

Galton (1886) first described another stochastic mechanism that is dangerous to ignore. Galton observed that children of tall (or short) parents usually grow up to be not quite as tall (or short), i.e. closer to average height. Today we know this phenomenon as regression to the mean, and we will find it wherever we find variation. Imagine a coach who berates a runner who had an unusually slow lap time and finds that, indeed, the next lap is faster. The coach, who always berates slow runners, has not had the opportunity to realize that the next lap is very likely to be faster no matter what. As long as there is variability in lap time we will some times see unusually slow laps that we can do nothing about and make no inference from. In this case too do our intuitions about the laws of chance fail us. People, including scientists, make the mistake of ignoring regression all the time. Mathematically regression to the mean is as simple as imperfect correlation between instances.

4.1.2 These phenomena in official statistics

The small-schools example is egregious because it led to wasteful public spending. The statistics themselves were probably fine, but their interpretation was not careful enough. Such summary statistics are often presented without regard for uncertainty. For instance, every year Statistics Norway (`ssb.no`), the central bureau of statistics in Norway, presents crime report counts. The media usually reports these numbers as rates and inform us that some small town that few people know about is the *most criminal* in the country. Often the focus is on violent crimes. Figure 4.1 below shows these rates for 2016. Not knowing de Moivre's result it might be striking to observe that many of the towns with the highest rates are small towns. Similarly, not knowing regression it might be striking to observe that, on average, towns with a high rate in one year will have a lower one in any other year, see Figure 4.2 below. These are unavoidable stochastic phenomena. Thus there is reason to believe that we should somehow adjust our expectations about these numbers. We will see below that such an adjustment also makes statistical sense.

4.1.3 Shrinkage estimation

There is an astonishing decision-theoretic result due to Charles Stein: suppose that we wish to estimate $k \geq 3$ parameters $\theta_1, \ldots, \theta_k$ and observe k independent measurements, $x_1 \ldots x_k$, such that $x_i \sim N(\theta_i, 1)$. There is an estimator of θ_i that has uniformly lower risk, in terms of total quadratic loss, than the obvious candidate x_i (Stein, 1956). In other words, the maximum likelihood estimate is inadmissible. Stein showed this by introducing a lower-risk estimator that biases or *shrinks*, the x_is toward zero. James and Stein (1961) introduced an improved shrinkage estimator, which we will see below. Efron and Morris (1973) show a similar result and a similar estimator for shrinking toward the pooled mean. There are many successful applications of shrinkage estimation, see for instance the examples from Morris (1983). The common theme is a setting where the statistician wants to estimate many similar variable quantities.

4.1.4 An almost-Bayesian estimator

In this case study we consider the official Norwegian crime report counts. We assume that in a given year the number of crimes reported in town i, denoted k_i, corresponds to the number of criminal events in this town. We further assume that each inhabitant can at most be reported for one crime a year. Our goal is to estimate the *crime probability* θ_i: probability that a person will commit a crime in this town. The obvious estimator is the maximum likelihood estimate (MLE) for a binomial proportion $\hat{\theta}_i = \frac{k_i}{n_i}$, where n_i is the population of town i.

The MLE binomial model rests on an assumption that inhabitants commit crimes independently according to an identical crime probability. There are reasons to believe that this is not the case. The desperately poor might be more prone to stealing than the middle class professional. There is a weaker assumption called *exchangeability* that says that individuals are similar but not identical. More precisely we assume that their *joint* criminal behavior (some number of zeros and ones) does not depend on knowing who the individuals are (the order of the zeros and ones). It is an important theorem in Bayesian inference, due to De Finetti, that a sequence of exchangeable variables are independent and identically distributed conditional on an unknown parameter θ_i that is distributed according to an a priori (or prior) distribution $f(\theta_i)$ (Spiegelhalter et al., 2004). In the binomial sense, θ_i has the remarkable property that it is the long-run frequency with which crimes occur regardless of the i.i.d. assumption; the prior precisely reflects our opinion about this limit. By virtue of De Finetti's theorem, the exchangeability assumption justifies the introduction of the unknown parameter θ_i in a binomial model for k_i, so long as we take the prior into account.

To make an argument with priors is to make a Bayesian argument. Shrinkage is implicit in Bayesian inference: observed data gets pulled toward the prior (and indeed the prior is pulled toward the data likelihood). We propose an almost Bayesian shrinkage estimator, $\hat{\theta}_i^s$, that accounts for the variability due to population size. Our estimator is *almost* Bayesian because we do not treat the prior very formally, as will be clear below.

In a Bayesian argument we treat θ_i as random. The statistician specifies a prior distribution $f(\theta_i)$ for the parameter that reflects her knowledge (and uncertainty) about θ_i. As in the frequentist setting, she then selects a parametric model for the data given the parameters, which allows her to compute the likelihood $f(x|\theta_i)$. Inference about θ_i consists of computing its posterior distribution by Bayes' theorem:

$$f(\theta_i|x) = \frac{f(x|\theta_i)f(\theta_i)}{\int f(x|\theta_i)f(\theta_i)\,\mathrm{d}\theta_i}.$$

There are various assessments we could make about the collection of θ_i. If we assume they are identical we can pool them and use a single prior. If we assume they are independent we specify one prior for each and keep them separate. If we assume they are exchangeable—similar but not identical—it follows from De Finetti that there is a common prior distribution conditional on which the $\theta_1, \ldots, \theta_m$ are i.i.d. (Spiegelhalter et al., 2004).

We make this latter judgment and take a beta distribution common to all crime probabilities as prior. Our likelihood for an observed number of crime reports follows a binomial distribution. It is a classic exercise to show that the posterior distribution of θ_i is then also a beta distribution. The problem remains how to choose the parameters for the prior. On the idea that a given town is probably not that different from all the other towns, we will simply pool the observed crime rates for all towns and fit a beta distribution to this ensemble by the method of moments.

Under squared error loss, the posterior mean as point estimate minimizes Bayes risk. The posterior mean serves as our shrinkage estimate, $\hat{\theta}_i^s$, for θ_i. We will see that $\hat{\theta}_i^s$ in effect shrinks the observed crime rate $\hat{\theta}_i$ toward the country-wide mean $\bar{\theta} = \sum \frac{1}{m}\hat{\theta}_i$ by taking into account the size of town i.

Bayesian inference allows for intuitive uncertainty intervals. In contrast to a classical frequentist confidence interval, which can be tricky to interpret, we can say that θ_i lies within the Bayesian credible interval with a certain probability. This probability is necessarily subjective, as the prior distribution is subjective. We will conduct simulations to compare the coverage properties of our estimator to the classical asymptotic confidence interval.

4.2 Data

We will work with the official crime report statistics released by Statistics Norway (SSB) every year. These data contain the number of crime reports in a given Norwegian town in a given year. The counts are stratified by crime type, e.g. violent crimes, traffic violations, etc. We will focus on violent crimes. SSB separately provides yearly population statistics for each town. Figure 4.1 shows the 2016 crime rates (i.e. counts per population) for all towns in Norway against their respective populations. This is some times called a funnel plot for the funnel-like tapering along the horizontal axis: a shape that signals higher variance among the smaller towns.

Figure 4.2 compares the crime rates in 2015 with those in 2016 and shows that the more (or less) violent towns in 2015 were on average less (or more) violent in 2016. The solid black line regresses 2016 rates on 2015 rates. The dashed grey line is what to expect if there were no regression toward the mean. It has an intercept of zero and a slope of unity. The solid grey line is the overall mean in 2016. The most extreme town in 2015, past .025 on the x-axis, is much closer to the mean in 2016. The solid black regression line shows that this is true for all towns on average. The fact that 2015 and 2016 are consecutive years is immaterial; regression to the mean will be present between any two years.

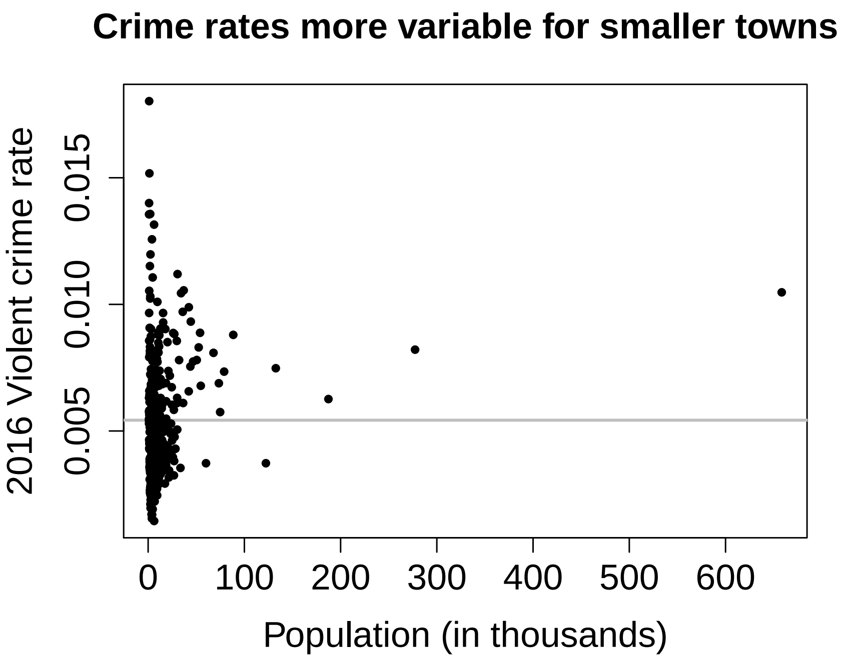


Figure 4.1: Rates of violent crime vs population in 2016 for all towns in Norway. The grey line shows the country-wide mean.

Figure 4.3 shows the distribution of the pooled violent crime rates for 2016. The solid black line is a beta distribution fit to these data.

4.2.1 Simulation study

We run a simulation study for validation. If we assume that the crime probability in town i is stationary we can pool the observed crime rates of all years and use their average, $\bar{\theta}_i$, as a reasonable "truth." This allows us to assess the performance of our estimator against known, realistic crime probabilities, which of course is impossible in the real data. The simulated crime report count in town i is $k_i \sim \text{Binomial}(\bar{\theta}_i, n_i)$, where n_i is the 2016 population of town i. Figure 4.4 shows a realization of this procedure. Although not a perfect replica of Figure 4.1—the real data do not have any rates below .0017—it looks fairly realistic.

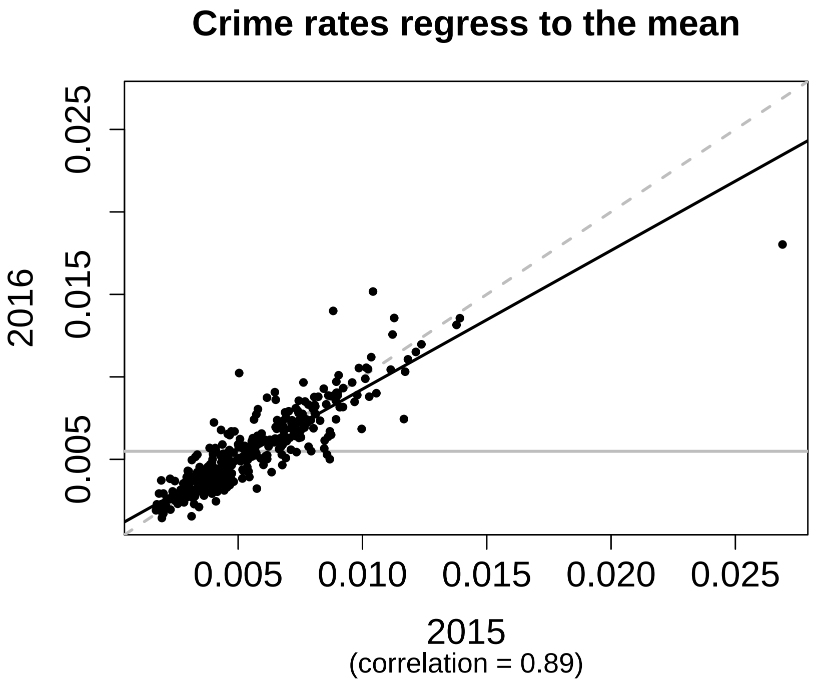


Figure 4.2: Regression to the mean from year to year. The plot compares 2016 and 2015; the black regression line shows that towns with high crime rates in 2015 tend to have lower crime rates in 2016, and vice versa for low crime rates. The grey dashed line shows what perfect correlation between 2015 and 2016 would look like.

4.3 Methods

4.3.1 Shrinkage estimates

We treat θ_i as the probability for a person to commit a crime in a given period. We model the total number of crime reports in the i-th town, k_i, as the number of successful Bernoulli trials among n_i, where n_i is the population of this town. As explained in the introduction, this suggests the following simple Bayesian model, also shown in Figure 4.5:

$$\begin{aligned}\theta_i|\alpha,\beta &\sim \text{Beta}(\alpha,\beta),\\ k_i|\theta_i &\sim \text{Binomial}(n_i,\theta_i).\end{aligned}$$

As mentioned the assumption of town exchangeability leads to this hierarchical model. This assumption might not be appropriate if we had reasons to think, for instance, that some regions are more prone to crime than others. In this case, region-specific priors might be better.

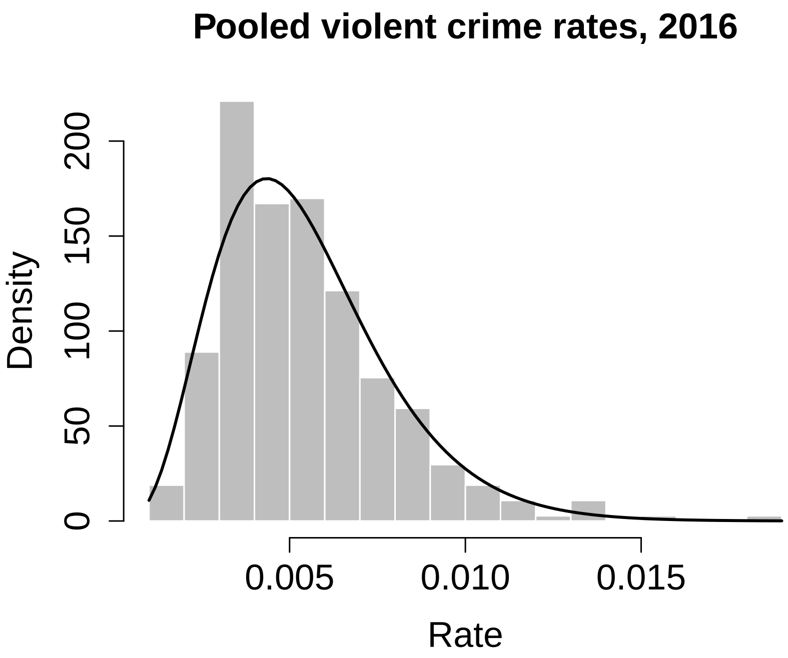


Figure 4.3: The distribution of violent crime rates in Norway, 2016. The black line describes the method-of-moments fit of a beta distribution to these data.

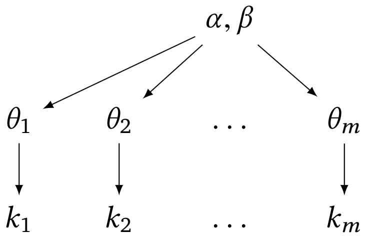


Figure 4.5: A graph describing our model. Crime counts, k_i, are (conditionally) i.i.d. binomials whose respective parameters, θ_i, are (conditionally) i.i.d. according to a common prior.

The posterior follows from the fact that the beta distribution is conjugate to itself with respect to the binomial likelihood. Generally, conjugacy means that the prior and posterior distributions belong to the same distributional family and usually entails that there is a simple closed-form way of computing the parameters of the posterior. Wasserman (2010, p. 178) shows a derivation of the posterior in the beta–binomial model:

$$\theta_i | k_i \sim \text{Beta}(\alpha + k_i, \beta + n_i - k_i).$$

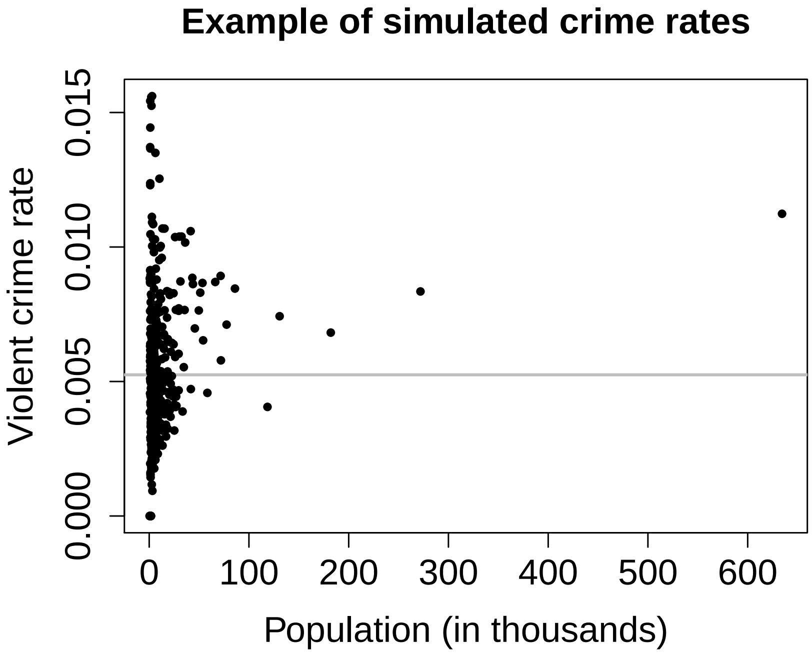


Figure 4.4: Funnel plot of a set of simulated crime rates

We will look into the relation between the parameters of the posterior to those of the prior in terms of successes and failures in the results section.

The shrinkage estimate for the crime probability in town i is the posterior mean

$$\hat{\theta}_i^s = \frac{\alpha + k_i}{\alpha + \beta + n_i}.$$

The maximum likelihood estimate for θ_i is the observed crime rate $\hat{\theta}_i = \frac{k_i}{n_i}$. In order to fix values of α and β, we pool the MLEs for all towns $\hat{\theta}_1, \dots, \hat{\theta}_m$ and fit a beta distribution to these data by the method of moments. We show the resulting fit in Figure 4.3. Because the expectation and variance of a Beta(α, β) are $\frac{\alpha}{\alpha+\beta}$ and $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$, respectively, the parameter estimates for the prior are

$$\beta = \frac{\alpha(1-\bar{\theta})}{\bar{\theta}}, \text{ and}$$
$$\alpha = \left(\frac{1-\bar{\theta}}{S^2} - \frac{1}{\bar{\theta}}\right)\bar{\theta}^2.$$

Here $\bar{\theta} = \frac{\sum_i \hat{\theta}_i}{m}$ and $S^2 = \frac{\sum_i (\hat{\theta}_i - \bar{\theta})^2}{m-1}$ are the sample mean and variance of the pooled MLEs.

Instead of estimating α and β from the data like this, which ignores any randomness in these parameters, we could have a prior distribution for the parameters themselves. This would yield a typical Bayesian hierarchical model. Note also that in forming the estimate for town i, we end up using its information twice: once in eliciting our prior and once in the likelihood. This is convenient because we need only to find one prior rather than one for each town where we exclude the ith town from the ith prior. This bit of trickery does not make much difference: we have several hundreds of towns and hence removing a single town does not affect the shape of the prior much.

The estimate $\hat{\theta}_i^s = \frac{\alpha+k_i}{\alpha+\beta+n_i}$ shrinks the observed, or MLE, crime rate toward the prior mean $\bar{\theta}$. We can rewrite so that $\hat{\theta}_i^s = \delta_i\bar{\theta} + (1-\delta_i)\hat{\theta}_i$, with $\delta_i = \frac{\alpha+\beta}{\alpha+\beta+n_i}$. Here δ_i directly reflects the prior's influence on $\hat{\theta}_i^s$, and we see that this influence grows as the town size, n_i, shrinks.

4.3.2 James-Stein estimates

For completeness we demonstrate empirically that the James–Stein estimator is superior to the MLE in terms of risk. If town i has a large enough population, we can consider the normal approximation to the binomial distribution and assume

$$\hat{\theta}_i = \frac{k_i}{n_i} \sim \mathcal{N}\left(\theta_i, \sigma_i^2\right),$$

where $\sigma_i^2 = \frac{\theta_i(1-\theta_i)}{n_i}$ is unknown. If we assume that towns are similar in terms of variance we can consider the pooled variance estimate

$$\sigma_P^2 = \frac{\sum_{i=1}^m (n_i - 1)\hat{\sigma}_i^2}{\sum_{i=1}^m (n_i - 1)},$$

where $\hat{\sigma}_i^2 = \frac{\hat{\theta}_i(1-\hat{\theta}_i)}{n_i} = \frac{k_i(n_i-k_i)}{n_i^3}$. The James-Stein estimator of crime probability for town i is then

$$\hat{\theta}_i^{JS} = \left(1 - \frac{(m-2)\hat{\sigma}_P^2}{\sum_{i=1}^m \hat{\theta}_i^2}\right)\hat{\theta}_i.$$

This is a shrinkage toward zero. It assumes that crime rates are probably not as high as they appear. This is different from our assumption that crime rates are probably not as far away from the average as they appear. It is simple to modify the above to shrink toward any origin. The Efron-Morris variant (Efron and Morris, 1973) shrinks toward the average:

$$\hat{\theta}_i^{JS} = \bar{\theta} + \left(1 - \frac{(m-2)\hat{\sigma}_P^2}{\sum_{i=1}^m (\hat{\theta}_i - \bar{\theta})^2}\right)(\hat{\theta}_i - \bar{\theta}).$$

We will use this variant so that the two methods shrink toward the same point.

4.3.3 Uncertainty intervals

We construct credible intervals from the posterior. A 95% credible interval contains .95 of the posterior density, and the simplest way to construct one is to place it between the .025 and .975 quantiles of the posterior. For the MLE we use the typical normal approximation (or Wald) confidence interval. There is to our knowledge no straight-forward way to construct confidence intervals for the JS estimator, so we will leave this as an exercise for the reader.

4.3.4 Global risk estimates

We use the total squared-error loss function,

$$L(\theta, \hat{\theta}^s) = \sum_{i=1}^{m} (\theta_i - \hat{\theta}_i^s)^2,$$

to measure the global discrepancy between the true rates $\theta = (\theta_i)_{i=1,\dots,m}$ and estimates $\hat{\theta}^s = (\hat{\theta}_i^s)_{i=1,\dots,m}$. We do the same for the maximum likelihood and James-Stein estimates $\hat{\theta} = (\hat{\theta}_i)_{i=1,\dots,m}$ and $\hat{\theta}^{JS} = (\hat{\theta}_i^{JS})_{i=1,\dots,m}$, respectively.

We will compare the expected loss, or risk, of the three estimators $R(\cdot) = E[L(\cdot)]$, confirming the well-known property that shrinkage estimators dominate the MLE. We obtain Monte Carlo estimates of risk by averaging $L(\cdot)$ across repeated simulations.

4.3.5 Coverage properties

For the credible interval $C^s = (a, b)$, we want to assess the coverage probability $\mathbb{P}(\theta \in C^s)$ and compare with $\mathbb{P}(\theta \in C^W)$ for the classical Wald confidence interval. We will not assess the James–Stein estimator in terms of coverage. Let $I(C_i)$, where $C_i = C_i^s$ or C_i^W, be the indicator function that is equal to unity if $\theta_i \in C_i$, and zero otherwise. We obtain MC estimates of coverage probability by averaging the mean internal coverage, $\frac{1}{m}\sum_{i=1}^{m} I(C_i^{\cdot})$, across repeated simulations. An uncertainty interval should be well-calibrated: if the size of the interval is 95% it should trap the true parameter .95 of the time.

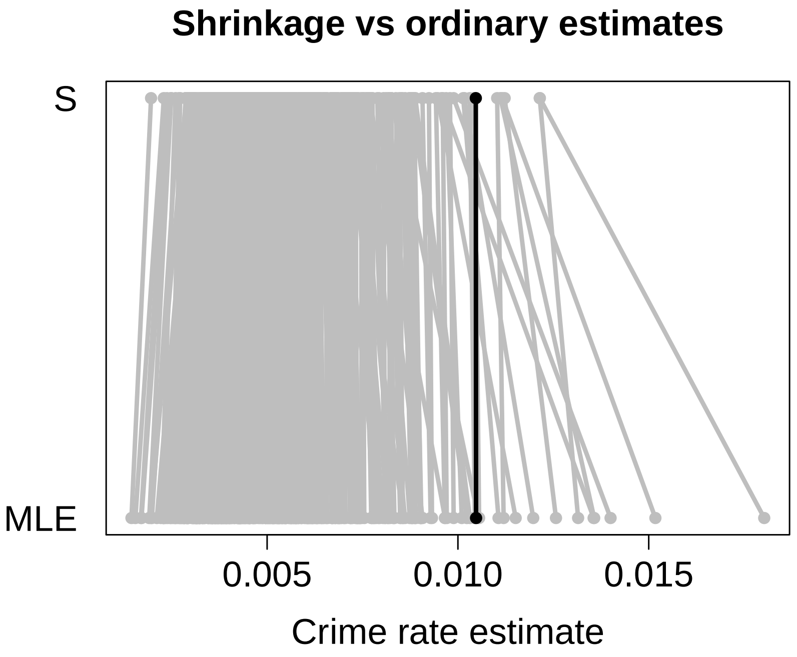


Figure 4.6: Comparing shrinkage and maximum likelihood estimates. Oslo, in black, is both close enough to the grand mean and large enough in size that the estimate does not change.

4.4 Results

4.4.1 Official SSB data

We focus on violent crimes in the year 2016. Figure 4.6 shows the effect of shrinking the observed crime rates toward the prior mean. We see that the more extreme estimates shrink toward the center. The city with highest crime rate according to the maximum likelihood estimate is Havsik ($\hat{\theta} = 0.018$), a small town with slightly more than 1000 inhabitants ($n = 1054$). After shrinkage, Havsik still ranks first, but the shrinkage estimate is much lower ($\hat{\theta}^s = 0.012$). Similarly the town with the lowest crime rate is Selbu ($\hat{\theta} = 0.0017$), another small town ($n = 4132$). Selbu's shrinkage estimate is higher than the MLE by more than 40% ($\hat{\theta}^s = 0.0024$). Oslo, shown in black, is a big city ($n = 658390$) and the difference between the two estimates is null ($\hat{\theta} - \hat{\theta}^s = 7 \times 10^{-6}$).

Figure 4.7 is a quantile–quantile plot of the 2016 violent crime rates against the fitted prior. There is some very slight deviation around the tails, but overall it looks like a nice fit.

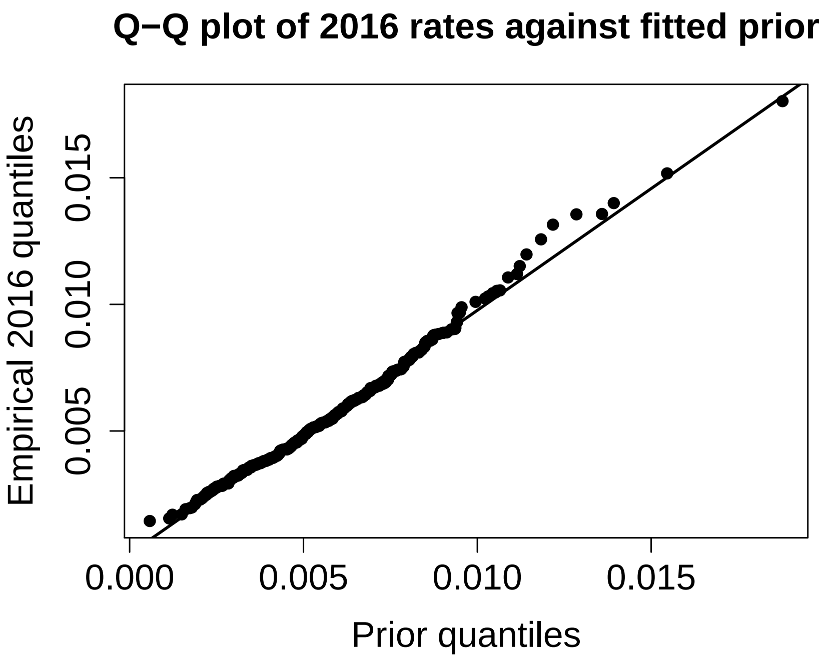


Figure 4.7: Quantile–quantile plot of 2016 crime rates against the fitted prior. The solid line describes a perfect fit.

By shrinking toward the ensemble we add some information—we use the term informally—to the observed rate. We can quantify this by looking at the form of the beta distribution, so far taken for granted in this treatment. Its density function is

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)},$$

where the beta function in the denominator is simply the normalizing constant

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\, \mathrm{d}t.$$

A natural interpretation is that this is a distribution over the probability of success, i.e. crime, in a sequence of Bernoulli trials with $\alpha - 1$ successes and $\beta - 1$ failures (cf. the binomial distribution). Hence we can interpret the posterior for town i as a distribution over the probability of success in a series of Bernoulli trials with $\alpha' = \alpha + k_i$ successes and $\beta' = \beta + n_i$ failures (ignoring the -1 for convenience). In our data we have that $\alpha \approx 5$ and $\beta \approx 917$; it is as though we add the information of 922 extra trials in the binomial sense. In other words we add a priori 922 inhabitants, including five criminals, to each town. Figure 4.8 shows α' and β' (gray and black) relative to the number of successes (k_i) and failures (n_i - k_i) for each town in the 2016 data. For the smaller towns, there is

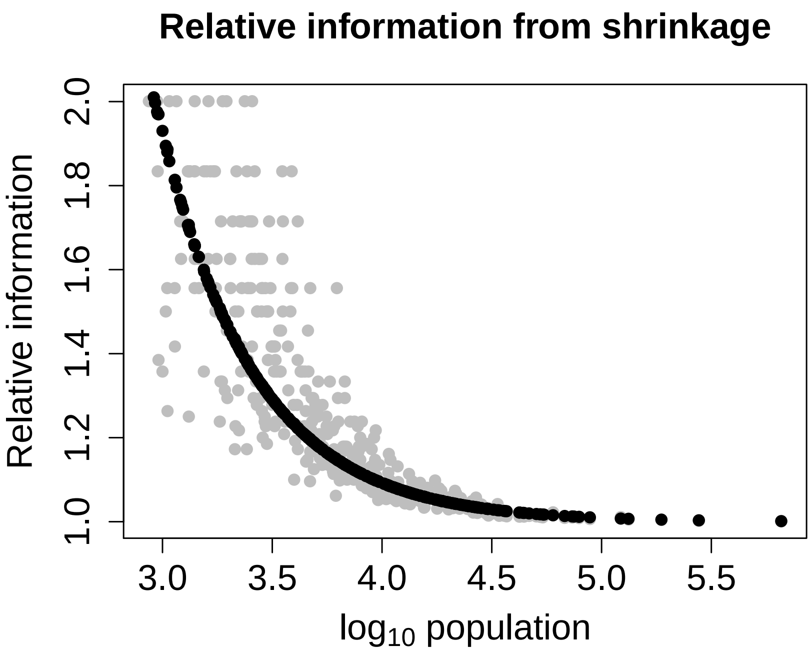


Figure 4.8: Relative information in the posterior mean compared to the MLE. The figure shows $(\alpha + k_i)/k_i$ in grey and $(\beta + n_i - k_i)/(n_i - k_i)$ in gray. These represent the added information in terms of number of successes and number of failures added to the MLE to form the shrinkage estimate. For the smallest towns, we practically double the information.

double the information in the shrinkage estimate, while for larger towns there is no practical increase. Naturally the value of this extra information depends on the degree to which the prior is relevant.

Figure 4.9 shows the ten most violent towns according to shrinkage estimate along with their 95% credible intervals. The official, or MLE, crime rate is shown as a red point. We see some change in ordering. For Hasvik—a small and presumably quiet village in northern Norway—the MLE is so implausible that it is outside the credible interval. For Oslo—the biggest city in Norway—the estimate doesn't change.

Figure 4.10 shows historical data for the three most violent and the three least violent towns in 2016, according to official crime rate. We show shrinkage estimates in red and official statistics in black. The vertical bars are 95% uncertainty intervals. The shrinkage estimate is usually more conservative, at least for the more violent towns, but the trends remain similar for both estimates. The credible intervals are shorter than the classical confidence intervals. We

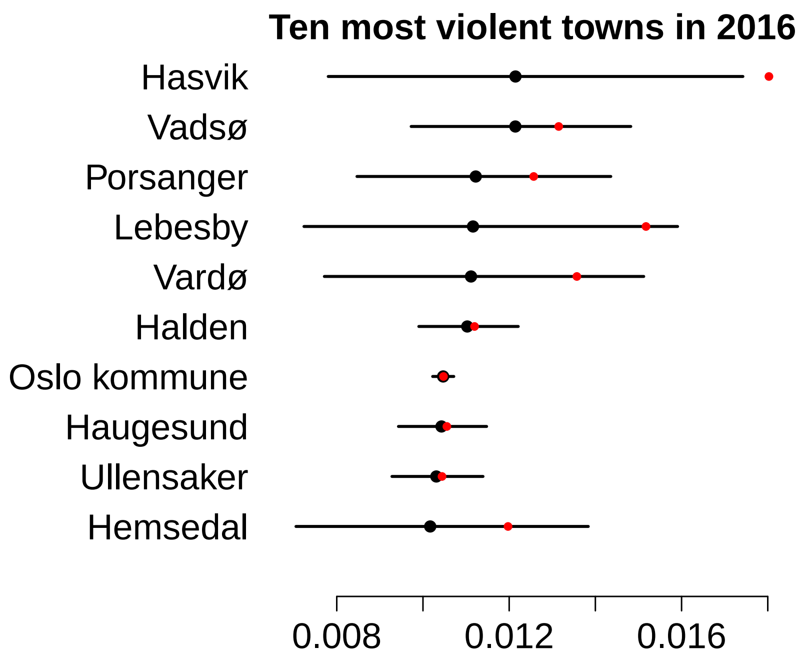


Figure 4.9: The ten towns with the highest crime rate, ordered by shrinkage estimate. The bars are 95% credible intervals. MLEs shown in red.

will see that in spite of this their coverage is better under simulation. It is interesting that the three most violent towns are all in Finnmark: Norway's largest and most sparsely populated county.

4.4.2 Simulated data

To obtain MC estimates of risks we run 100 000 simulations for each of our two experiments. Figure 4.11 shows kernel density estimates of the distributions of global loss. Our shrinkage estimates show lower global risk than maximum likelihood: $\hat{R}(\theta, \hat{\theta}^s) = 0.00054$ versus $\hat{R}(\theta, \hat{\theta}) = 0.00066$. The James–Stein estimates fall almost exactly between the two with $\hat{R}(\theta, \hat{\theta}^{JS}) = 0.00059$. We might have observed better results for JS had we used a variant of JS that allows unequal variances. Note that we fixed θ_i for this experiment, so we are only assessing the risk function in a single point.

Figure 4.12 presents estimated coverage probabilities in the same manner as Figure 4.11. The grey line shows the nominal coverage of .95. The coverage probability of the credible interval for the shrinkage estimator, $\hat{\mathbb{P}}(\theta \in C^s) = 0.917$, is closer to the nominal value than that of the the standard interval,

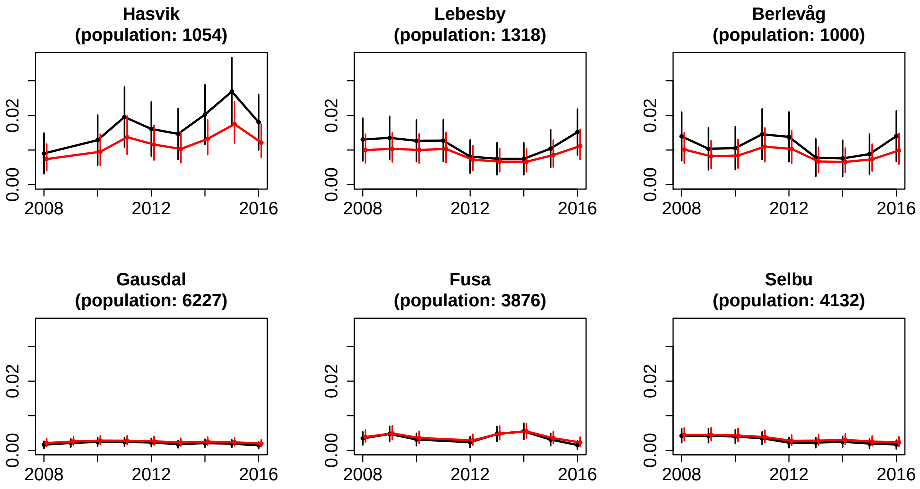


Figure 4.10: Historical data for the three most violent and the three least violent towns in 2016, ordered by official crime rate (MLE). The official statistics are drawn in black, and shrinkage estimates in red. The vertical bars indicate confidence and credible intervals, respectively.

$\hat{\mathbb{P}}(\theta \in C^W) = 0.898$. There is however still room for improvement.

4.5 Conclusion

This case study shows a simple method for simultaneous estimation of all town-specific crime rates in a country. The method is Bayesian in spirit, although we take some shortcuts with our prior. It is known that under squared-error loss the posterior mean is the optimal decision w.r.t. a given prior. In other words it minimizes Bayes risk, and is called the Bayes estimate. The theory gives us that Bayes estimates are admissible (Wald, 1947), and thus cannot be dominated. The risk estimates of our simulation agree with this. Our analysis provides an estimate of the crime probability with favorable frequency properties in terms of mean squared error and coverage.

Our simulations show that the Bayesian credible intervals from this treatment are narrower and have better coverage than the standard Wald confidence interval. Hence we get better information about the location of θ_i. Brown et al. (2001) show extensively that the Wald confidence interval for the binomial proportion behaves erratically for extreme values of p, for varying values of n, and for (un)lucky combinations of the two. Our result is interesting but quite narrow. Generalizing it requires more work.

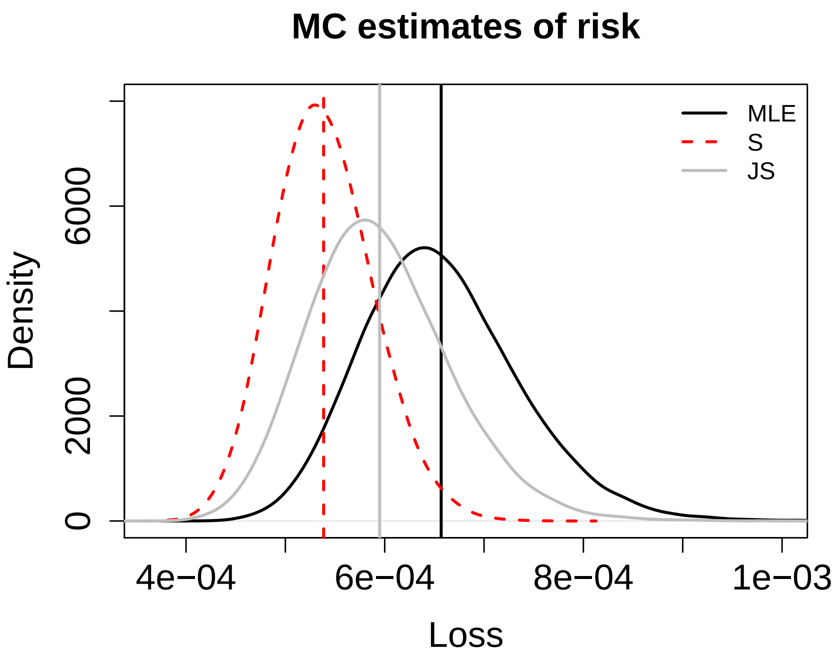


Figure 4.11: Distributions of $L(\theta, \hat{\theta})$ (solid black), $L(\theta, \hat{\theta}^s)$ (dashed red), and $L(\theta, \hat{\theta}^{JS})$ (solid grey). Vertical lines estimate the risk.

Smaller towns are over-represented among the most and least violent towns in the official Norwegian data. Mathematically this has to be the case. Applying shrinkage methods to these data we get more conservative estimates for these variable and often extreme quantities. At the same time it seems that variance is not the only factor that places some of these small towns among the most violent. As Figure 4.9 shows, the top and bottom three in 2016 show a certain stability year by year. Hasvik in Finnmark has never ranked especially low since 2008. Small towns in the north are often ranked high for violence. There could be many reasons for this and we leave further analysis to the criminologists.

These simple and useful estimation methods are best understood by practical examples. We encourage readers and students to actively follow this tutorial by playing with the available code and data. We used a single prior for all towns. It would be an interesting extension to use a mixture of beta distributions to account for any heterogeneity due to different latent rate levels. In this case, an EM algorithm could be used to assign each town to a class. Or, since Finnmark seems to be a special case, we might estimate per-county priors. It is also possible to include Bayesian multiple testing procedures to infer a list of cities likely to have true crimes rate above some given threshold. There is a temporal aspect to these data that we have not looked into. It would be possible

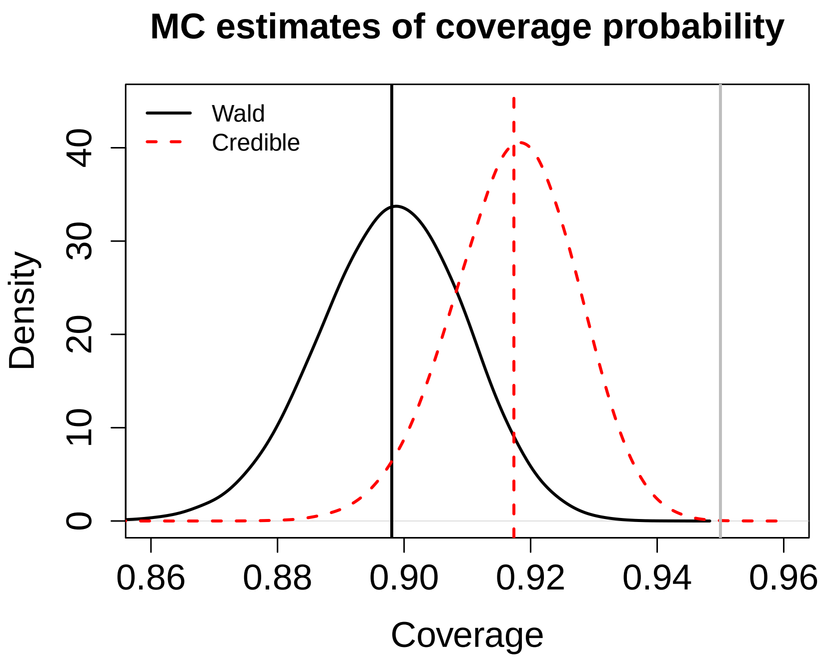


Figure 4.12: Distributions of the internal coverage $\frac{1}{m}\sum_{i=1}^{m} I(C_i^W)$ (solid black) and $\frac{1}{m}\sum_{i=1}^{m} I(C_i^s)$(dashed red). Vertical lines estimate coverage probability. The grey line shows the nominal coverage of .95.

to start out with a country-wide prior, but after this let the prior for one year be the posterior from the previous. Interested readers can find other ideas for further development in Robinson (2017). Gelman and Nolan (2017) also discuss a similar project to this one in their manual for statistics teachers.

In this treatment we have moved from descriptive figures typical of official statistics to model-based inferential statistics, estimating a crime probability rather than reporting a crime count. This allows us to account for variance and perhaps avoid over-interpreting noise, and hence avoid small-schools-type mistakes. We believe that probabilistic thinking can enrich descriptive statistics and aid in their interpretation.

4.5.1 A note on -omics small data

As mentioned in the beginning of this chapter, shrinkage methods are central in modeling high-dimensional data. The crime rate estimation above can be seen as the estimation of a high-dimensional mean. There are some references in this document that apply shrinkage methods to -omics data, the most prominent of which is probably Smyth (2004). Two others are Johnson et al. (2007) and Hausser and Strimmer (2009). Penalized maximum likelihood with ridge or lasso penalties, as used in Chapter 2 are applications of the James–Stein-type idea of shrinking toward zero. There is a Bayesian equivalent to both these methods, almost all such methods for estimating many regression coefficients are based on treating these coefficients as exchangeable (Gelman et al., 2014). The experiments in the current chapter show that there are situations where there is a clear advantage to using informative priors rather than simply applying James–Stein shrinkage. It is difficult to say what such a prior would be in eg. the setting of Chapter 2, but it may be an idea worth developing.

5

Discussion

"Besides, he'd explained at length, there was no such thing as absolute control, not in a fully functioning universe. There was just a variable amount of lack of control."

–Terry Pratchett, *Darwin's Watch*

In this chapter I will examine and discuss some of the issues involved doing data analysis in our small-data setting. These are issues that figure prominently in any real-world data analysis, but that for various (often legitimate) reasons tend to get glossed over when writing an article. To begin with I will make more clear what I take to be small data and present a simple data generation scheme that I will use throughout to make my points. Finally I will summarize what I believe to be important small-data analysis considerations.

5.1 Small data definition

Figure 5.1 below shows a distribution over around 1200 sample sizes for human-derived transcriptomics experiments.[1] The middle 9/10 of this distribution lies between 6 and 106 observations. There are an about 19–20 thousand

[1] Data source: `https://www.ebi.ac.uk/gxa/experiments?organism=Homo+sapiens`

protein-coding genes in the human genome.

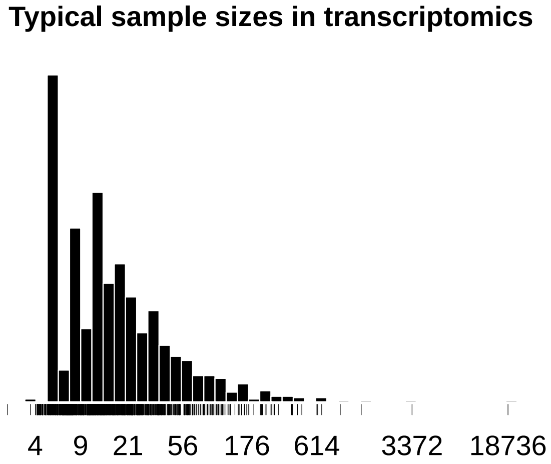


Figure 5.1: Sizes of human-derived transcriptomics data uploaded to the EMBL-EBI Expression Atlas between 2014 and 2018 (n=1178). Comprises both microarray and RNA-Seq experiments. The histogram is taken over the logarithm of sample sizes, but the numbers on x-axis shows the actual sizes. The rug is jittered on the log scale.

The NOWAC metastasis data in Chapter 2 comprised 88 observations and around 12 000 genes. According to Figure 5.1 this is quite a large sample. At the same time, it seems that comparing the blood cell gene expression levels of a case and a healthy control is a much more variable and low-signal setting than comapring expression levels in cancer tissue cells to healthy tissue cells. The NOWAC metasasis data is **typical small data:** thousands to tens-of-thousands of predictors, usually fewer than a hundred observations, and a low signal-to-noise ratio.

5.2 Small data problems in predictive models

In Section 2 we set out to build a model predicting metastasis agnostically with no pre-conceived notion of which genes would make a good predictor set. This is in the modern “data mining” spirit where the data shall give up her truths if modeled vigorously. It is a desirable approach in an exploratory setting such as NOWAC, since little is known about the signal of metastasis in blood cell genetic expression. However the small data setting is fraught with subtle pitfalls and there is reason to moderate one's expectations for the biological insights and generated hypotheses of a 100% agnostic approach. In this section I point out some of the problems involved.

5.2.1 Simulated data

I will consider a simple logit model of the form

$$\begin{aligned} \mathrm{x} &= [x_1, \ldots, x_p], x_i \sim N(0,1), \\ z &= \sum_{i=1}^{k} x_i, \\ \log \frac{p}{1-p} &= \beta_0 + \beta_1 z. \end{aligned} \tag{5.1}$$

The log odds of success, in the binomial sense, is a simple linear function of the variable z, the sum of the first $k \le p$ predictors. The basic identities for the normal distribution give us that $Z \sim N(0,k)$ and that $\log \frac{p}{1-p} \sim N(\beta_0, \beta_1^2 k)$. Control of β_0 and β_1 respectively yields control of the expected proportion of successes and the variation in odds. Large β_1 yields a large difference between success and failure, large β_0 yields a large proportion of successes. The last $p-k$ predictors are noise.

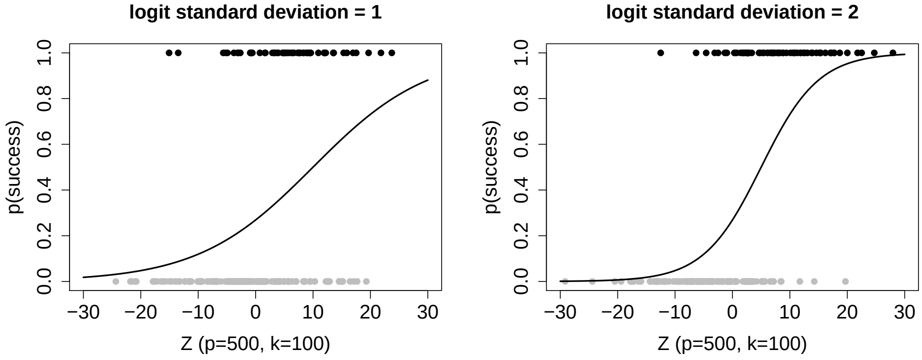


Figure 5.2: Examples of data simulated from model 5.1.

Figure 5.2 shows two example high-dimensional datasets simulated from the model in Equation 5.1. Code listing 5.1 implements this simulation scheme in R. The standard formulation for the slope $\beta_1 = 1/\sqrt{k}$ scales the log odds to unity variance. Similarly a $\beta_1 = s/\sqrt{k}$ formulation scales to an arbitrary standard deviation s.

Code Listing 5.1: R function for simulating data according to Equation 5.1

```
generate_data <- function (
    n,                              # number of observations
    ngenes = 100,                   # this is k above
    dim = 500,                      # while this is p
    intercept = -1,
    slope = 1/sqrt(ngenes)
) {
  x <- matrix(rnorm(n*dim, mean=0), nrow=n)
  z <- rowSums(x[, 1:ngenes])

  y_p <- 1/(1+ exp(-(intercept + slope*z)))
  y <- rbinom(n, 1, y_p)

  cbind(y, x)
}
```

5.2.2 Estimating prediction error

We evaluate prediction models by estimating the expected prediction error on new data. Perhaps the most reliable route to an unbiased estimate is to predict data to which we do not have access during modeling, the gold-standard procedure being to lock some portion of your data completely away during modeling (Hastie et al. (2009) tentatively sugest 25%) and to use this to estimate prediction error as a final step.

This is seldom feasible in a small-data setting. To make a very broad argument, suppose we take accuracy—the proportion of correct success/failure predictions—as evaluation metric. (To make this into an error measure take 1 - accuracy.) Suppose the true accuracy of our model is p, and that we wish to be certain of our estimated accuracy to within .01 with 95% confidence. This way we can eg. separate an accuracy of .90 from one of .91. For the normal approximation confidence interval for the binomial proportion this roughly corresponds to setting $2\sqrt{\frac{p(1-p)}{n}} < 1/100$, or $n > 40\,000p(1-p)$. Here n is the size of the holdout set. If we take this to be 25% of our data, we need $4n$ observations in total.

Figure 5.3 above shows the minimum sample size $4n$ as a function of p. Assuming a true accuracy of .95 we need a sample of 7600 observations. This is highly idealized for several reasons. As already mentioned in Chapter 4, it is well-known that the normal approximation confidence interval is poorly calibrated, especially so for large/small probabilities (Brown et al., 2001). Indeed, as p

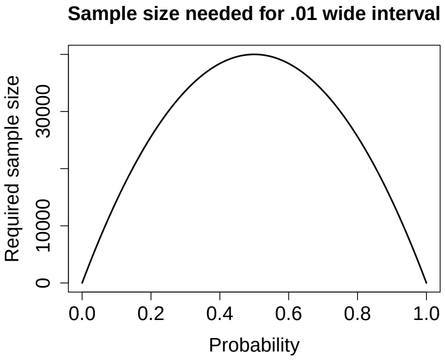


Figure 5.3: Required sample size for a .01 wide confidence interval estimating a binomial proportion.

gets small/large the skewness of the binomial distribution tends to ±∞. Also the the closer p gets to 0 or 1, the less useful an error of ± .1 in its estimate. And perhaps most important, estimating a binomial proportion is not the same as evaluating a prediction model: the proportion is one parameter to be estimated directly from some data; a prediction model introduces many more parameters each with some amount of estimation variance, all of which affects the accuracy estimate. Hence 7600 here should be seen as a very naive, absolute minimum sample size. Of the 1200 datasets in Figure 5.1, all but one are smaller than 7600 observations. The NOWAC data of Chapter 2, at 88 observations, is larger than 93.5% of the EMBL-EBI datasets.

There are various alternatives to using a held-out portion of data for evaluation. Here I will compare four estimators of out-of-sample error: (i) **split-sample** validation by the 25% rule suggested by Hastie et al. (2009); (ii) the Efron and Gong (1983) **optimism bootstrap**; (iii) **k-fold cross validation**; (iv) **repeated k-fold cross validation**. We defined the optimism bootstrap in detail in Chapter 2. Briefly it entails fitting the model on all data, calculating the apparent error (predictions on the training data), and correcting the optimistic bias in this score by the bootstrap. The k-fold cross validation is the very common practice of randomly partitioning the data in k equal-sized portions and averaging over the use of each as validation split. As the name suggests, repeated k-fold cross validation entails averaging over the k-fold cross validation score for some number of repetitions of this procedure. For reasons that we will get back to, accuracy is not always a great metric for evaluating predictive models. I will instead instead focus on Brier's score, defined in Chapter 2 as the mean squared error between predicted probabilities and known success/failure outcomes.

Figure 5.4 shows two simulation studies comparing the above methods of estimating prediction error. On the left is a low-dimensional situation with only two predictors, both associated with the outcome; on the right is a high-dimensional situation with 100 predictors associated with the outcome and 900 noise predictors. For each I repeatedly generate 150 observations from model 5.1 with zero slope and unit intercept. I fit a ℓ_1 penalty logistic regression (Friedman et al., 2010), choosing the shrinkage size λ by five-fold cross validation, and estimate Brier's score by the four different methods above. The densities show the distribution of estimated out-of-sample score. The "true" out of-sample Brier score of the fitted model, represented with a vertical grey line, is calculated in a separate sample of 100 000 observations.

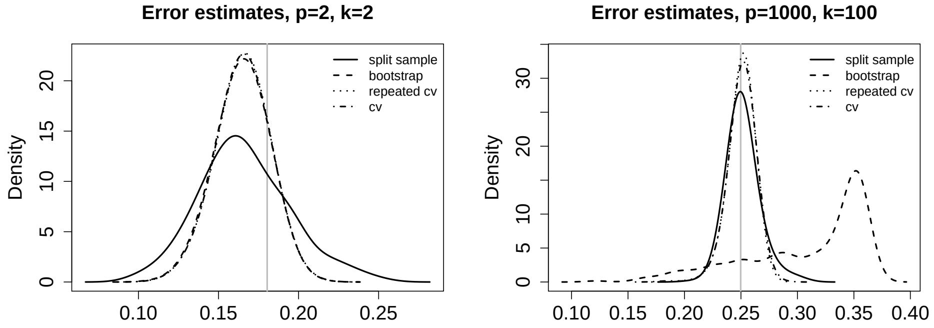


Figure 5.4: Densities over out-of-sample Brier score estimates for four different estimators. On the left is a situation with plenty observations per model parameter, on the right is a small-data situation. The vertical grey line shows the true quantity to be estimated.

In the low-dimensional situation we see that the alternatives to split-sample evaluation behave similarly to one another. In the high-dimensional situation the two cross-validation procedures perform similarly to one another, but the bootstrap procedure completely breaks down. The resampling procedures—ignoring for now the bootstrap problem, which I will get back to below—generally have tighter distributions, indicating lower variance over split-sample validation. The **relative efficiency** of two estimators T_1 and T_2 of the same quantity is the ratio of their variances $\frac{\text{Var}(T_1)}{\text{Var}(T_2)}$. This ratio roughly corresponds to the relative sample size needed for T_1 to be as efficient as T_2, and all else being equal we should prefer the less variable estimator (Mosteller and Tukey, 1977). Table 5.1 below shows the relative efficiency of split-sample validation against the other three methods.

Generally you need two to four times as many observations doing split-sample validation as with resampling methods. This is in line with Breiman (1992) who found that his variant of the bootstrap at a given sample size was about as good as split-sample with a validation set two times larger in a simulation setting

	Bootstrap	Cross validation	Repeated CV
Low-dimensional	3.5	3.6	3.6
High-dimensional	.057	2.0	2.6

Table 5.1: Relative efficiency of split-sample validation against bootstrapping, 10-fold cross validation, and repeated 10-fold cross validation.

with 40 predictors and 60, 160, or 600 observations. In the 60×40-sized data we are closer to the land of small data, and as Breiman states, “This is a land strange to asymptopia.” Kim (2009) finds repeated cross validation the best general-purpose validation method, stating that it reduces variance enough over single cross validation to be worth the extra computation. This is based on simulations in a setting where there is plenty of data per predictor. Molinaro et al. (2005) provides a study based on -omics type data and come to similar conclusions. Both of these articles focus on misclassification error instead of a continuous score.

Now I return to the bootstrap problem. I have introduced a methodological flaw to the experiments above to illustrate a source of internal variance and over-estimation of optimism; the problem I refer to in the preamble to Chapter 2. The nesting of a cross-validated tuning of the shrinkage size λ inside of a bootstrap resampling procedure leads to undershrinkage. Since the bootstrap samples with replacement from the original data, some observations will be repeated. These may end up in both training and test folds in the inner procedure and have a disproportionate influence on the model: essentially we overfit the repeated observations. To illustrate I have repeated the cross validation and bootstrap estimations above and recorded the λ chosen. To demonstrate that the problem indeed comes from observations repeated in both test and train folds I have also done the bootstrap with a modified nested cross-validation that removes test fold observations that also occur in the train fold.

Figure 5.5 illustrates: The cv-in-bootstrap method chooses a much lower λ than regular cross validation, undershrinking and overfitting. Removing duplicates from the test folds fixes this problem. The deduplicated version of bootstrap chooses a λ mostly in line with cross validation, although more variable. This is likely because there is less information available in the bootstrap sample; it contains about .6 of the original sample. The test folds are also much smaller due to deduplication; we have removed about .6 of the observations there.[2]

[2] These two arguments come from the following rough approximation: a given observation has a chance of $1 - 1/n$ to not be chosen as the ith observation in a bootstrap sample. From this there is a chance of $(1-1/n)^n \approx e^{-1} \approx .37$ to not be chosen at all. Similarly the chance for a given observation of not being in a given train fold out of k is $(1-1/n)^{n(k-1)/k}$, which by even rougher approximation $\approx e^{-1}$.

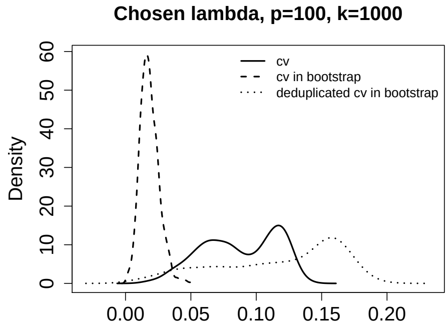


Figure 5.5: Choosing a λ by cross validation, by cross validation nested in a bootstrap sample, and by a deduplicated version of cross validation in bootstrap.

The effect I describe here is less pronounced in the low-dimensional setting, likely because there are fewer directions in which a particular observation can be strange. It is not clear to me how common this mistake is, but we committed it working on Chapter 2. The simulations in this chapter are by no means comprehensive and serve as toy problems to point out certain issues. Under the assumption that the particular issues for a given method affect the different models similarly, the λ-selection fix should make the penalized likelihood results in Chapter 2 comparable to the analyses in the appendix to that chapter. I briefly touch on the selection of λ again at the end of Section 5.3.

5.2.3 Measures of prediction error

Accuracy, the proportion of correct dichotomous 0-or-1 predictions, is a very common choice of *scoring rule* for assessing a prediction. There are variants of accuracy such as sensitivity and specificity, F_1 statistic, etc. I call these variants of accuracy because they are all based on counting “exactly correct” success or failure predictions.

Philosophically, a dichotomous prediction poses some problems in the realistic case where both outcomes are possible for a given predictor value. That a person smokes increases the probability that they contract lung cancer, but it is not given. How then can we in good conscience predict a 0-or-1 probability for lung cancer conditional on smoking? If we wish to admit some uncertainty in outcome, a probability model is more prudent.

Brier's score (see Chapter 2) is a rule that measures the calibration of a predicted probability as the average quadratic distance between the predicted probability and the known 0-or-1 outcome. Other rules exist for measuring probability calibration such as binomial deviance.

Concordance probability (see also Chapter 2) is the probability that a randomly chosen success will rank higher (in predicted probability of success) than a randomly chosen failure. In the machine learning literature this is known as **AUC**. AUC is formulated slightly differently, but the two measures are equivalent (Hanley and McNeil, 1982), and it is a very popular scoring rule. It is more akin to the accuracy-like rules in that it is still counting something; in this case concordant pairs of successes and failures.

Below I provide a small demonstration of the dangers of counting scores. I generate 100 observations from model 5.1 with k real predictors and k noise predictors. To these I fit two penalized regression models: one using only the k true predictors and one using all 2k predictors. Clearly we should prefer the model that only uses the true predictors. I investigate two different settings: one where there are few observations per predictor (k=30) and one where there are plenty observations per predictor (k=2). The model coefficients are $\beta_0 = 0$, and $\beta_1 = 1/\sqrt{k}$. The slope coefficient for each predictor is hence smaller in the k=30 situation. Over 5000 simulations I register whether the three scores preferred the true k-predictor model over the worse 2k-predictor model. At the end I have the probability for choosing the best model by each score for the two settings. Table 5.2 below summarizes.

	Accuracy	**Concordance/AUC**	**Brier score**
Low-dimensional	.57	.67	.73
High-dimensional	.73	.79	.93

Table 5.2: Probability of choosing the essentially correct no-noise model over the worse half-noise model with three different scoring rules.

We see that the Brier score discriminates best between the models, especially in the high-dimensional case. Accuracy is the worst, which should not be surprising, there is no reason to believe that accuracy is an adequate measure for the relative performance of probability models (Harrell and Lee, 1985). Accuracy is only affected by whether probabilities falls above or below a threshold. Whence this threshold is in fact an important issue which ideally should be separate from the modeling procedure. A very common convention is to use .5 as a threshold, presumably because it is the optimal decision *if the cost of a false positive is the same as that of a false negative.* I see no good reason for assuming this, and no good reason that the person who builds the prediction model is the one who should make the decision. AUC does better than accuracy as it admits that there may be other thresholds. Hand and Anagnostopoulos (2013) show that AUC is incoherent in that it depends on the empirical distribution of predicted probabilities. This implies that the optimal decision under AUC depends not on the probability of success or failure, but on how this probability was estimated.

Accuracy, AUC, and other counting-based relatives are improper rules that suffer from the fact that an arbitrarily small change in model can result in an arbitrarily large jump in score, which is not the case with Brier score. Hanczar et al. (2010) demonstrate that small-data estimates of these measures are only weakly correlated with the truth. It is impossible to tell which model will optimize one of the counting-based scores but it there is a fair chance that it is not the one that best fits the data.

5.3 Shrinkage and prior information

The crime rate shrinkage project in Chapter 4 illustrates the value of shrinkage and incorporating prior information in a small-data analysis. It is in many ways a classical application of partial pooling of information for a proportion where we borrowed about a thousand citizens worth of information from the ensemble to improve the high-variance estimates for very small towns. A very successful application of this idea in transcriptomics is the Smyth (2004) linear models and empirical Bayes methods for microarrays (LIMMA). The LIMMA approach is to do partial pooling of variance estimates in a manner similar to our crime rate shrinkage, leading to more stable (less variable) inference.

Using high-quality prior information (in a general, not necessarily Bayesian sense) can be very valuable in the small-data setting. In the simulation in Section 5.2.2 above there was .15 of an observation available per parameter in the full model. If we somehow had an idea that the first 100 predictors were more likely to hold information than the last 900, we could guide the model towards this perhaps expending fewer degrees of freedom on the last 900 parameters by penalizing these more. In whole-blood gene expression studies such information is scarce as the biological mechanisms are not well-understood. There is reason to doubt any single published signature set: we saw in Chapter 2 that such sets can be very unstable.

Penalized maximum likelihood regression models—with penalties such as ridge, lasso, and other relatives (Cessie and Houwelingen, 1992; Tibshirani, 1996; Zou and Hastie, 2005)—are also examples of shrinkage models. They are perhaps the best approach to deal with a small-n-large-p situation. They can be framed as Bayesian arguments with a certain prior on the coefficients (Goldstein (1976) for instance shows the correspondence of an exchangeable normal prior on β and ridge regression). The choice of a particular penalty corresponds to a prior assumption about the data, eg. whether the model should be sparse or dense, or whether we expect groups of predictors to work together. The strength of the penalty, λ, is often tuned to the data in a frequency argument. Greenland (2000) cautions against treating λ as a tuning parameter and advises that it should reflect the precision of background information. A very strong penalization then represents a very strong belief in our prior assumptions.

5.4 Standardized process

In Chapter 3 we described the Standard Operating Procedure (SOP) for finding and removing technical outliers in the NOWAC microarray material. Such quality assessment (QA) work receives relatively little attention although, as our simulations showed, the decisions made here impact the results downstream.

In a large research project such as NOWAC it is likely that different people end up doing QA at different times for different data sets. In such a setting it can be difficult to understand what has happened to the data and for which reasons. A research project usually comprises many temporary positions such as Ph. D. candidates and postdoctoral researchers. A constant churn means that new people will have to do a certain amount of detective work to gain knowledge of the material, and a certain amount of knowledge is lost with the departure of temporary researchers.

Dr. Hilary Parker recently presented her idea of *opinionated analysis development* as a way of codifying best practices in data analysis (Parker, 2017). Particularly the work focusses on the process of developing the technical artifact of data analysis. In academia this artifact is often an article for publication; in QA this should ideally be a report that details which observations have been excluded and the reasons for excluding them. Dr. Parker lists three main features of an analysis: it should be (i) reproducible and auditable; (ii) accurate; and (iii) collaborative.

We took the NOWAC SOP from a script based on a single analysis and moved it toward a standard tool—the `nowaclean` package—and a standardized process—the accompanying article and vignette. To encourage reproducibility and auditability we use literate programming (Knuth, 1984) with `knitr` and R markdown (Xie, 2014). Informal peer review has evolved as a way to sanity check both the decisions made in the analysis and the analysis itself. This is especially important considering the essential value of each observation in a small data setting. Data is hard to come by, and no observation should be excluded unless there is convincing reason to do so. This in turn makes the report artifact central as a tool for auditing analyses and disseminating knowledge.

Making the original script into a package shifts the intellectual burden from the error-prone process of copy-paste programming to the development of a convincing and comprehensible analysis artifact. A centralized, open-source package enables the analysis process of QA to be kept accurate and collaborative, because it enables modular and testable code and simplifies the process of making contributions to the process.

These are lessons well-known in software engineering that are only beginning to make their way into data analysis and academia. Such concerns are likely to grow more important as research projects grow larger and more interdisciplinary.

5.5 Conclusion: small data in practice

The small data setting in biomedical research is one with tens to hundreds of observations and thousands of measurements for each observation (in extreme cases hundreds of thousands). The economical, practical, and ethical restrictions on access to human participants makes the value of each observation central. The high-dimensional nature of these data makes high variance and weak signal a fact of any statistical analysis.

The choice of validation method and scoring rule are very important because an uninformed choice will introduce unnecessary variance. If we are not careful we may end up wasting our valuable data. Resampling validation is two to four times as efficient as split sample validation. Single k-fold cross validation (CV) performed quite well in my simulations, but repeated cross validation reduces variance further; Molinaro et al. (2005) found as few as ten repetitions helpful. There is a lot of theory around the bootstrap, which may recommend it over repeated CV. We must then pay particular attention to the danger of nested procedures. The cost of resampling is computation time, which can be considerable if the model fitting procedure itself is compute-intensive.

In a setting where a given predictor value can be associated with different outcomes it is prudent to model a probability rather than learn a classification rule. Optimization of accuracy and accuracy-like scoring rules is inadequate for assessing a probability model, and this may be exacerbated in high-dimensional, low signal-to-noise settings. These rules add unnecessary noise and even in simple settings it is wasteful to optimize for accuracy or AUC. Keep in mind that this pertains not only to mathematical optimization, but also of post-hoc model comparison, which amounts to the same thing. AUC is useful in its alternative interpretation as concordance probability but should not be trusted on its own. When the proper response is a probability—and in the biomedical small data setting it arguably is—the model should be optimized for calibration of these probabilities.

When possible it is desirable to incorporate prior information, biological or otherwise, to guide the modeling procedure and reduce variance in model estimation. Informative priors can be very valuable in this respect, but useful information is scarce in human-derived blood sample transcriptomics data. Published signature gene sets may be of some value, but variable selection procedures can be quite unstable, as we saw in Chapter 2 and its appendix. It is not obvious how to translate them to priors or otherwise build them into a model. The typical parallel structure of high-dimensional biomedical data such as -omics lends itself very well to hierarchical model type shrinkage. Penalized-likelihood-type shrinkage is another technique for managing variance and may often be the best bet when prediction is the goal and prior information is scarce.

/6

Future work

"Ponder Stibbons had once got one hundred percent in a prescience exam by getting there the previous day."

–Terry Pratchett, *Unseen Academicals*

The results in **the metastasis prediction work** are necessarily exploratory and uncertain. We have by now tried a fair few different methods and evaluation schemes and need fresh data to validate our findings. I believe this should be done by careful a-priori consideration of the 108 selected genes to remove obvious noise candidates followed by validation in new data perhaps based on hierarchical modeling where we take prior information from the fold change estimates in our data (see Figure 2.6). This would also allow us to integrate the likely notion that stratification still is important, which should be investigated. We have already started investigating the biological interpretation of these genes.

The NOWAC SOP article has a couple of avenues for improvement. We plan to improve the R package. We have confined ourselves to reporting the SOP as it is and has been applied to NOWAC data, there is some question whether it is general enough for others to use. An investigation might be needed. The microarray as a platform is on its way out, and the general move seems to be toward using RNA-Seq. We are uncertain to what extent our approach is applicable in an RNA-Seq setting. Some adaptation is probably needed, which suggests another avenue for investigation.

When people present work from the NOWAC blood transcriptomics material it is a common question **whether observed differential expression is not simply a stress effect due to the presumably considerable psychological stress of being told that you may have breast cancer.** A small study was conducted to investigate this question where two groups of NOWAC participants were invited to contribute blood samples for gene expression analysis: one group who had experienced health-related psychological stress after either having an abnormality on their screening mammogram, or after discovering a breast change, and one control group of NOWAC participants attending a routine gynecological examination at their local gynecologist. In this work, where I provide statistical analyses as joint first author with Associate Professor Karina Standahl Olsen, we assess differences in gene expression by LIMMA moderated t-tests (Smyth, 2004) for single genes and globaltest (Goeman et al., 2004) for stress-related gene sets curated from the literature. We find no significant stress effect, though there might be a question about statistical power. Dr. Olsen presented this work as an abstract at The 7th Conference on Epidemiology and Registry-Based Health Research - NordicEpi 2015 (Olsen et al., 2015).

SAM (Tusher et al., 2001) did very well in the Holsbø et al. (2018) simulations of data where followup time influences expression levels. This procedure does not explicitly model time, which makes the result slightly surprising. Around the same time as I performed that simulation study Kang and Song (2017) published an extension of SAM that uses weighted observations for estimating the "fudge factor," s_0, in Equation 2.5. Their results suggest that such a weighting yields a more robust inference in noisy situations. The same type of thinking might be applied to a setting where effect sizes depend on followup time. In the prospective part of NOWAC blood samples are collected at enrollment. If the followup time is long it is unlikely there is much trace of a systemic response to the disease, which might be diagnosed years later. If the sample is fairly recent it is more reasonable to look for such a response. I present a semi-weighted SAM procedure that accounts for followup time prospective biological samples. Rather than taking $\bar{x}_{\text{case}} - \bar{x}_{\text{control}}$, where $\bar{x}$ is the group mean, as effect size I suggest using a weighted group mean, $\bar{x}(w) = \frac{1}{\sum_j w(x_j)} \sum_i w(x_i)x_i$, for the cases. If we estimate the weights as $w(x_i; \sigma) = \theta(\frac{t_i}{\sigma})$, with θ the standard normal density function, and t_i the followup time of the ith sample, $\bar{x}(w)$ is the Nadaraya–Watson estimate of x_i as a function of t_i at the time $t_0 = 0$. This has the nice interpretation that it is the expected gene expression at the time of diagnosis. Similarly weighting the variance estimate s_i in Equation 2.5 and leaving s_0 alone yields a **followup-weighted SAM procedure for prospective case–control studies**. I have some promising early simulation results for this.

A

`nowaclean` vigntette

On the following recto page I have included the `nowaclean` vignette that belongs to the article in Chapter 3 as it looked at the time of writing.

Outlier detection with *nowaclean*

Einar Holsb*

May 25, 2017

Abstract

This vignette shows the use of the *nowaclean* *R* package, which implements the standard operating procedure for detecting and removing technical outliers in the NOWAC microarray material.

1 nowaclean

1.1 Installation and loading

We'll be using the development version of *nowaclean*, which is hosted on GitHub.[1] To install from GitHub you need to install *devtools*.

```
install.packages("devtools")
```

Once you have installed *devtools*, you can use it to install *nowaclean* from its GitHub repository.

```
devtools::install_github("3inar/nowaclean", build_vignettes=T)
```

Once it is installed, you can use *nowaclean* like you would any other *R* package.

```
library(nowaclean)
```

To view *nowaclean* on github (for instance for bug reports, etc.), visit https://github.com/3inar/nowaclean.

2 Loading and Preprocessing

First to load the dataset; we have suppressed the huge text dump that happens when you load the *lumi* package:

```
library(lumi)  # Required to access LumiBatch objects
datapath <- "~/Downloads/sop_data.rda"
load(datapath)
```

This is a typical data set from the Norwegian Women and Cancer study. These are anonymized data that are freely available from the UiT Dataverse https://opendata.uit.no/. The reference is *Einar Holsb, 2017, "Supporting data for "A Standard Operating Procedure for Outlier Removal in Large-Sample Epidemiological Transcriptomics Datasets"", doi:10.18710/FGVLKS, UiT Open Research Data Dataverse, DRAFT VERSION.*

*einar@cs.uit.no

[1] https://github.com

2.1 Remove blood type probes

In some situations we remove 38 probes related to genes in the human leukocyte antigen (HLA) system. These are usually expressed strongly and have high variance, which affects multivariate analyses. Specifically we have seen that they might dominate the variance-covariance pattern in the PCA transformation of the data, and as such other patterns might be obscured. The `blood_probes` function returns the nuIDs of these probes.

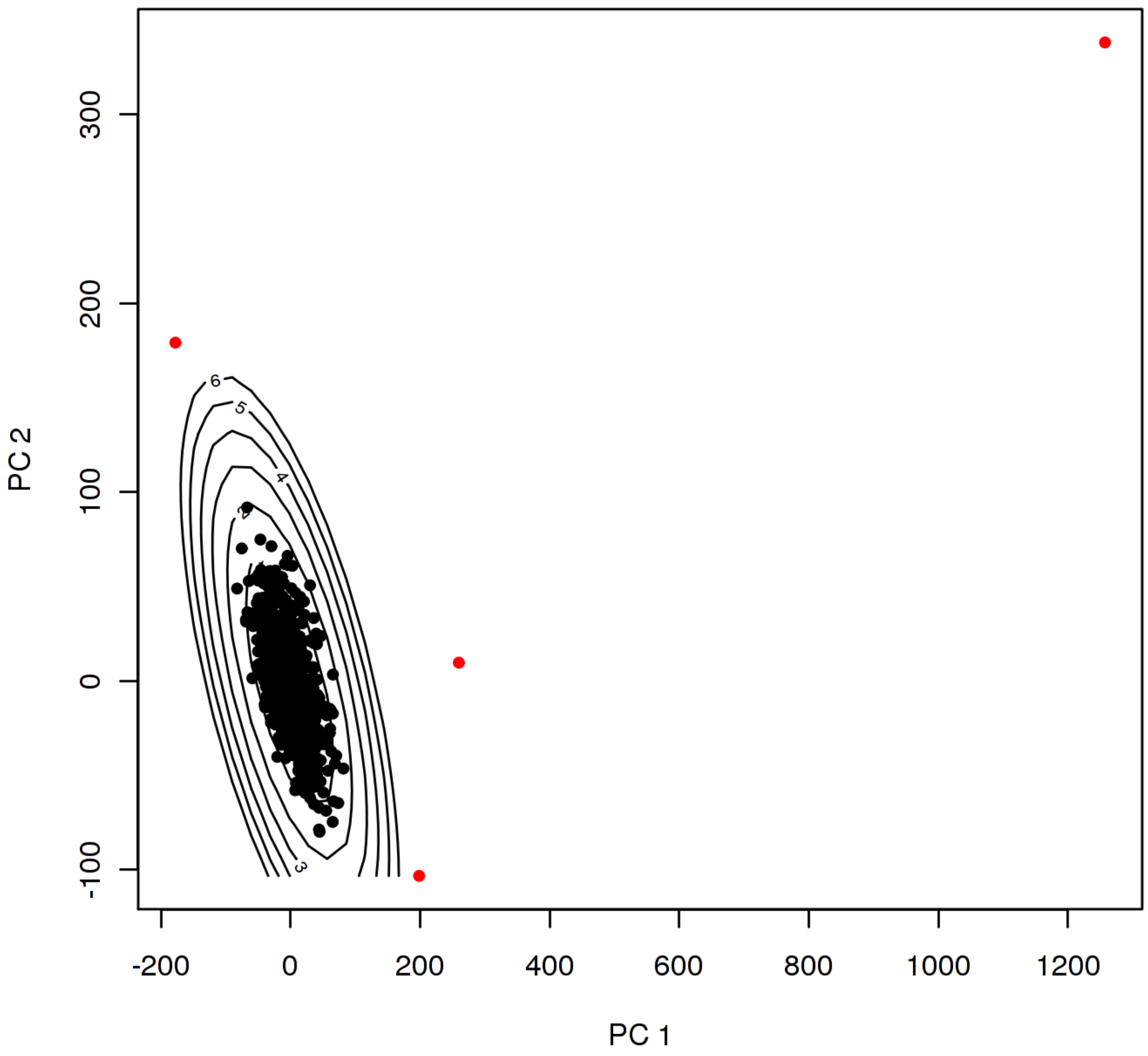
```
gene_expression <- gene_expression[!rownames(gene_expression) %in% blood_probes(), ]
```

3 Outlier detection

We find outliers by exploratory plotting and statistical measurements described more closely in the package documentation. We will be working on log_2-transformed data to ameliorate the higher variance we usually see for higher intensities and to make the expression levels more symmetrical.

First we examine PCA-transformed data. The contour lines show distance to the center of the data in number of standard deviations.

```
expression <- log2(t(exprs(gene_expression))) # transpose for samples by probes
prc_all <- prcout(expression)
plot(prc_all)
```

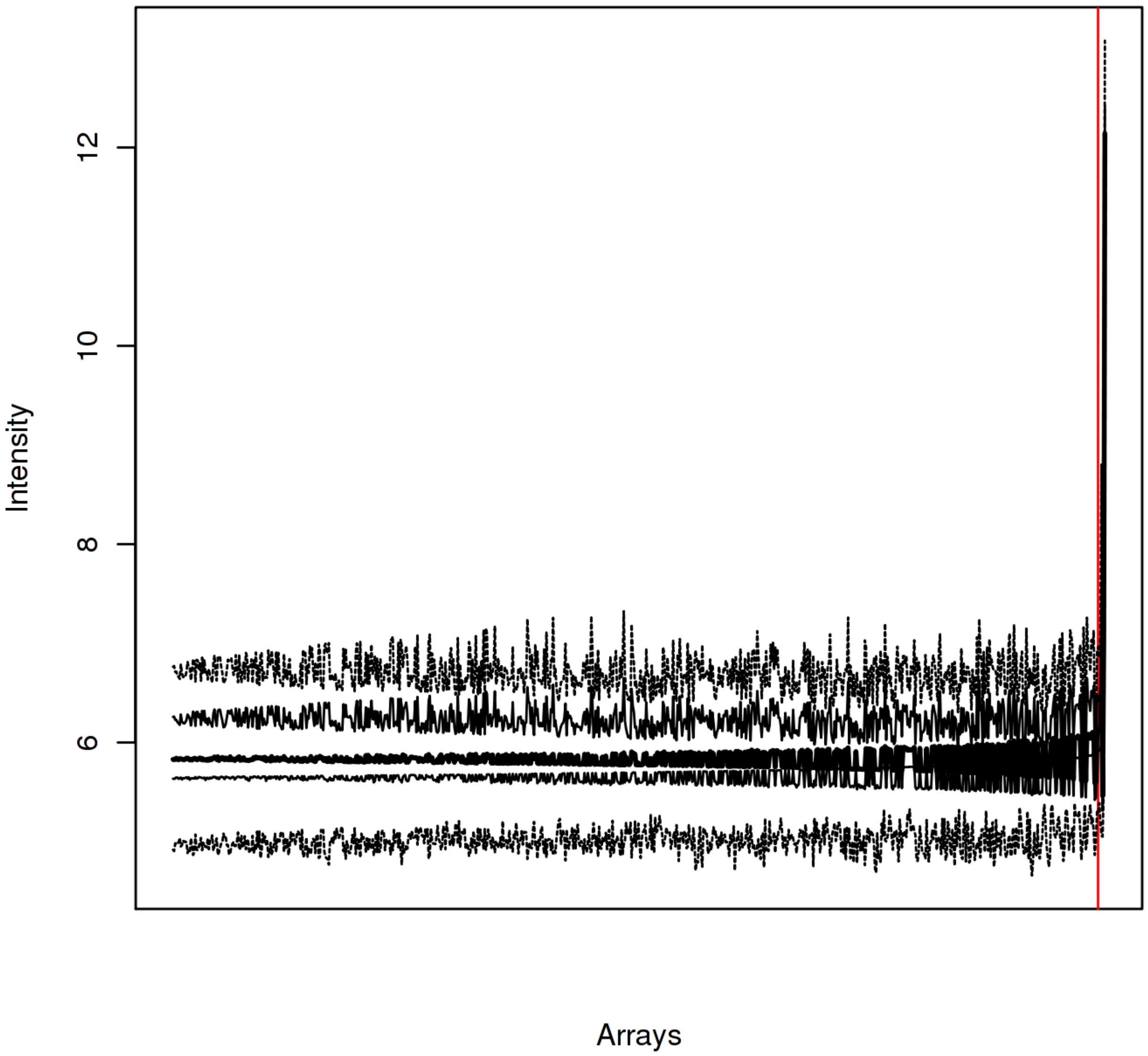


The points marked in red are three or more standard deviations away from the main bulk of the data: they look quite astonishing. Let's keep these red points as possible outliers.

```
pca_outliers <- predict(prc_all, sdev=3)
pca_outliers

## [1] "137" "396" "500" "705"
```

Next we investigate some boxplots.

```
boxo <- boxout(expression)
plot(boxo)
```

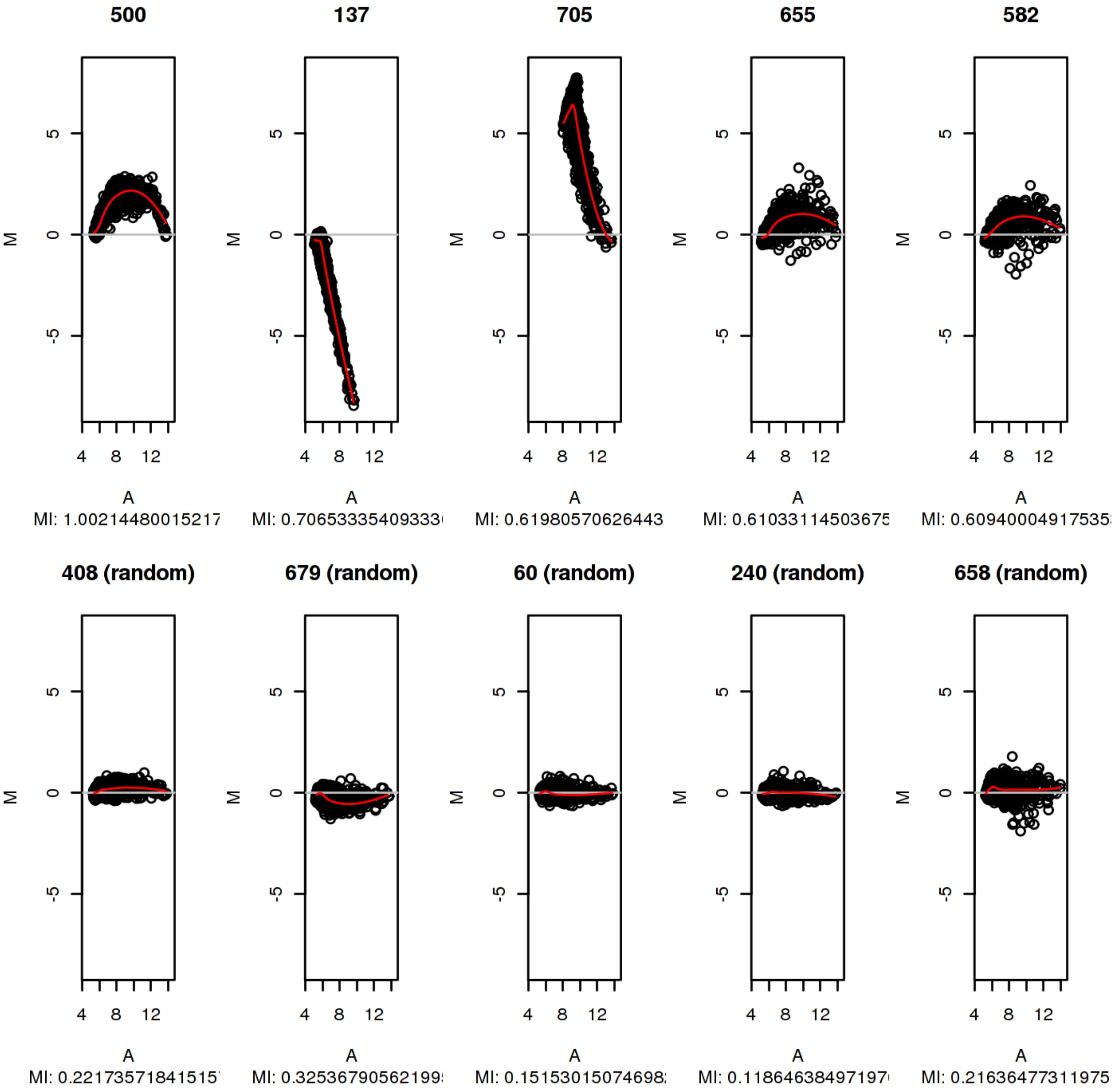


Points on the lines in this plot represent the box and whiskers of the regular `boxplot` function for your arrays. The lines represent the first and third quartiles, the median (ie the standard box), and the most extreme points that fall within 1.5 times the interquartile range (ie the whiskers/fences). As default the arrays are sorted by size of ks statistic (distance to pooled empirical distribution function).The red line demarks the cutoff for outlier or not.

```
boxplot_outliers <- predict(boxo, sdev=3)
boxplot_outliers

## [1] "11"  "137" "396" "500" "705" "827"
```

The final detection method we use is the MA-plot. Let's plot the worst candidates and compare to some random samples. Badness is here defined in terms of mutual information between M and A statistics.

```
maout <- mapout(expression)
plot(maout, nout=5, lineup=T) # lineup = T  compares against some
```



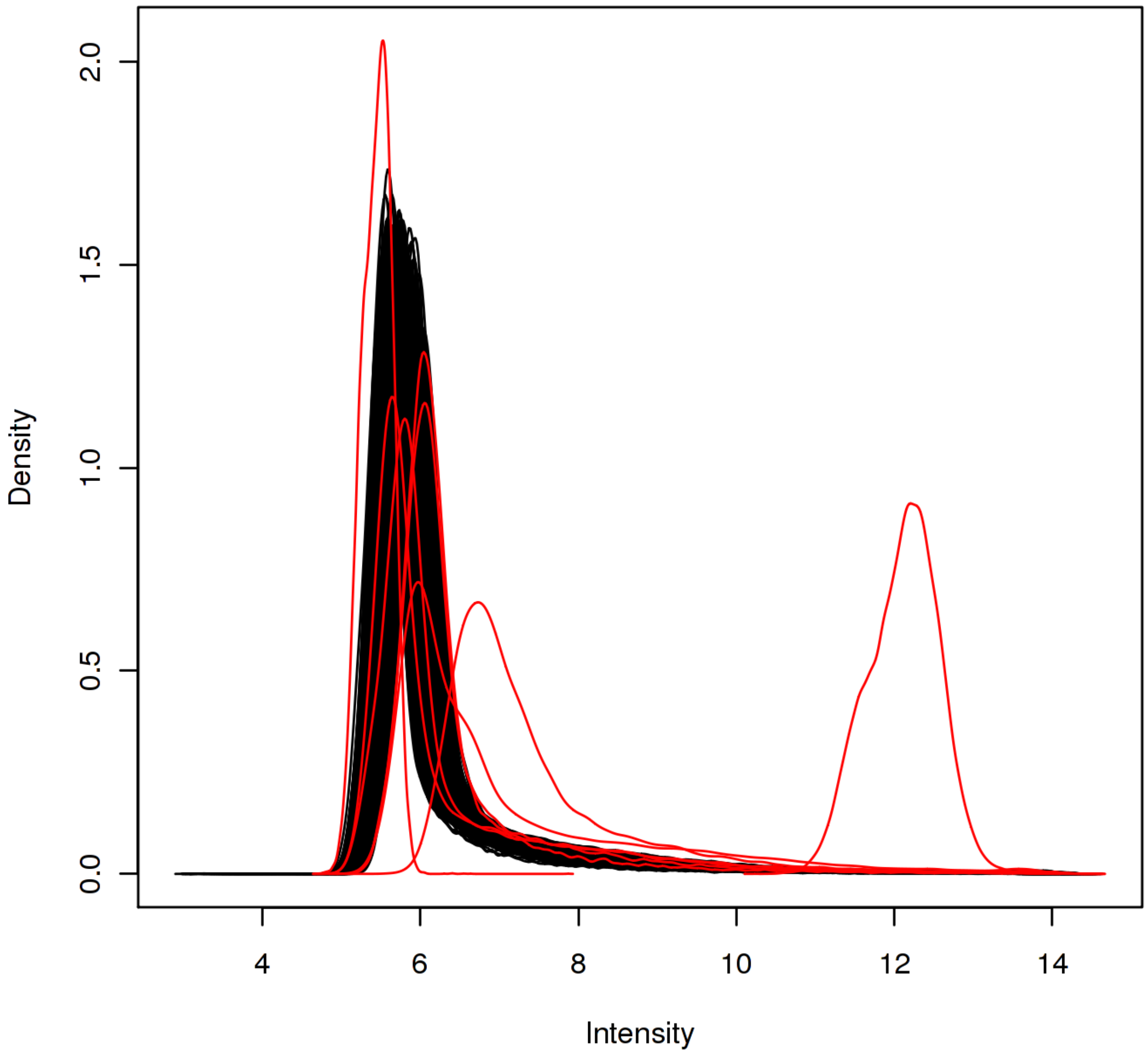
```
                        # randomly chosen observations

mapoutliers <- predict(maout, sdev=3)
mapoutliers

## [1] "137" "500" "582" "655" "705"
```

Let's now combine all outlier vectors.

```
outliers <- unique(c(mapoutliers, boxplot_outliers, pca_outliers))
outliers

## [1] "137" "500" "582" "655" "705" "11"  "396" "827"
```

These are the densities of expression values for all samples, proposed outliers in red:

```
densities <- dens(expression)
plot(densities, highlight=outliers)
```



As we can see, all of the clearly strange densities in this plot are marked as outliers; some of the candidates look fine however.

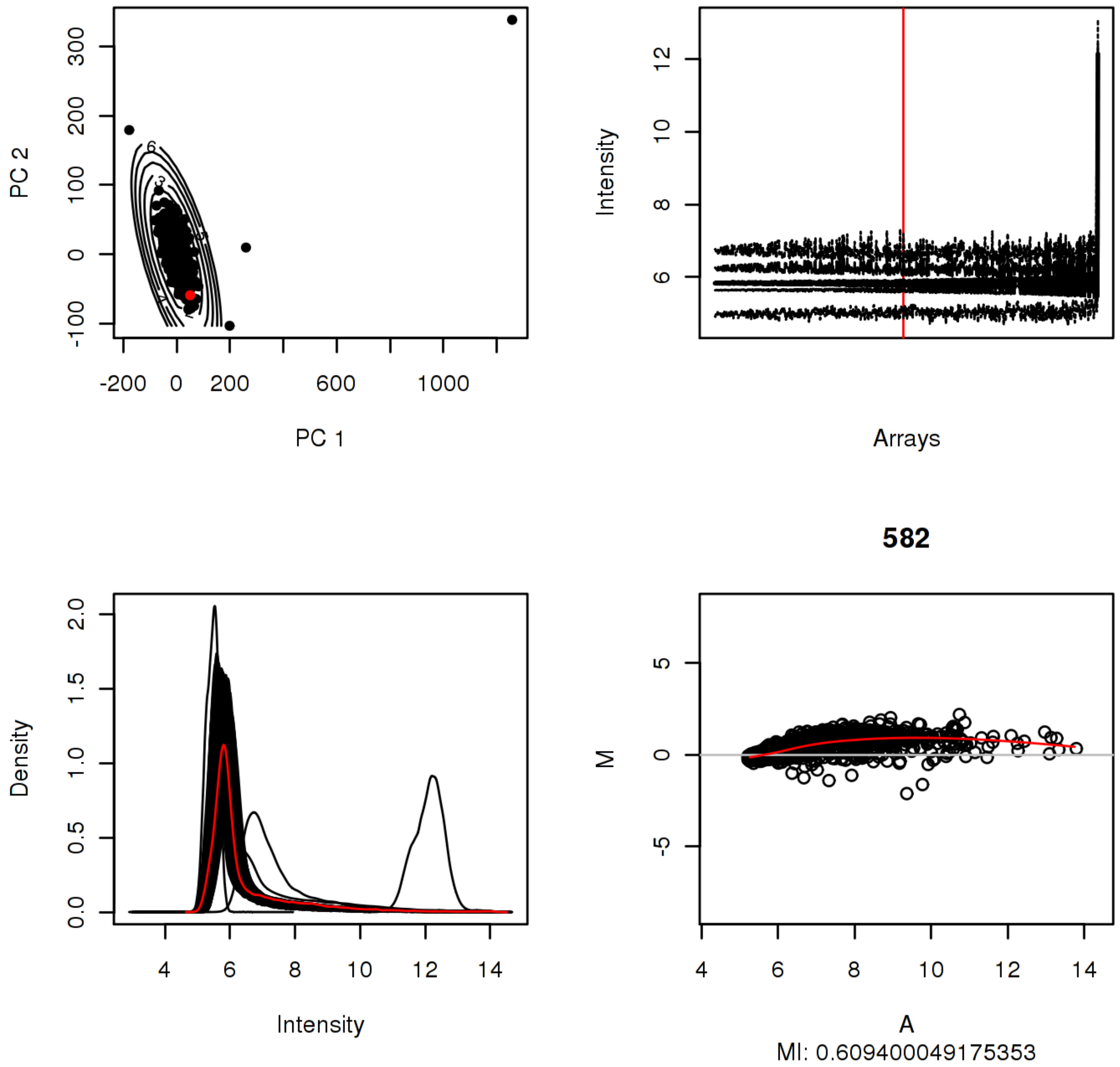
4 Manual outlier removal

So now we have a list of 8 candidate outliers that we suspect are technical outliers. This section will examine each of them and we'll make a decision to either keep or remove them as need be. Note that I would usually use the actual sample names instead of indexing the outlier vector with numbers. This is to be absolutely certain that I'm looking at what I think I'm looking at. I suggest others do the same. However, these data are anonyimzed, there are no sample names, and strings of row numbers will have to do.

4.1 582

This one looks fine. Maybe the MA plot is the reason it got flagged. I won't remove this.

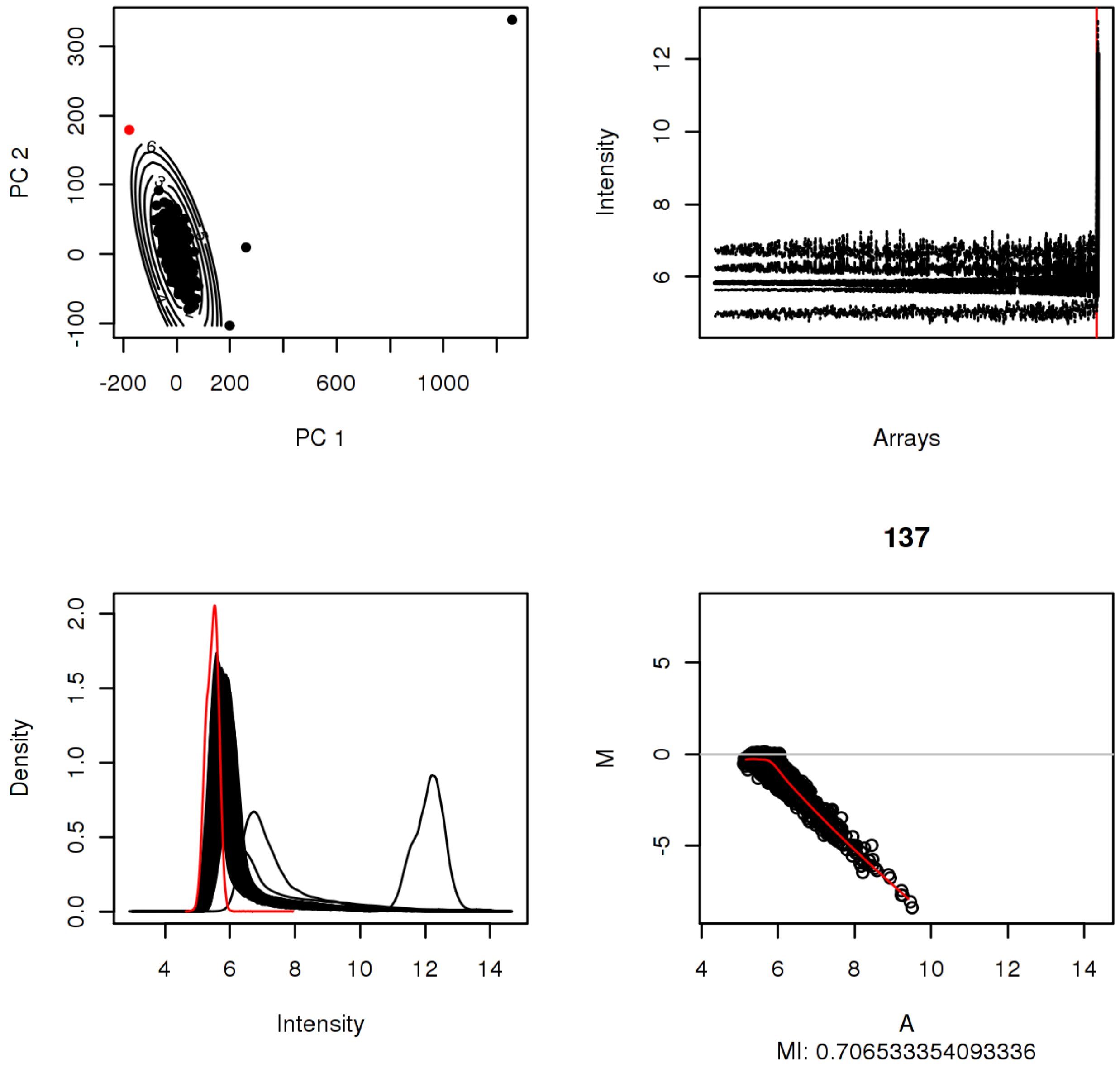
```r
highlight("582", pca=prc_all, box=boxo, dens=densities, ma=maout)
```



4.2 137

This one is clearly very strange in all the plots, I will remove this.

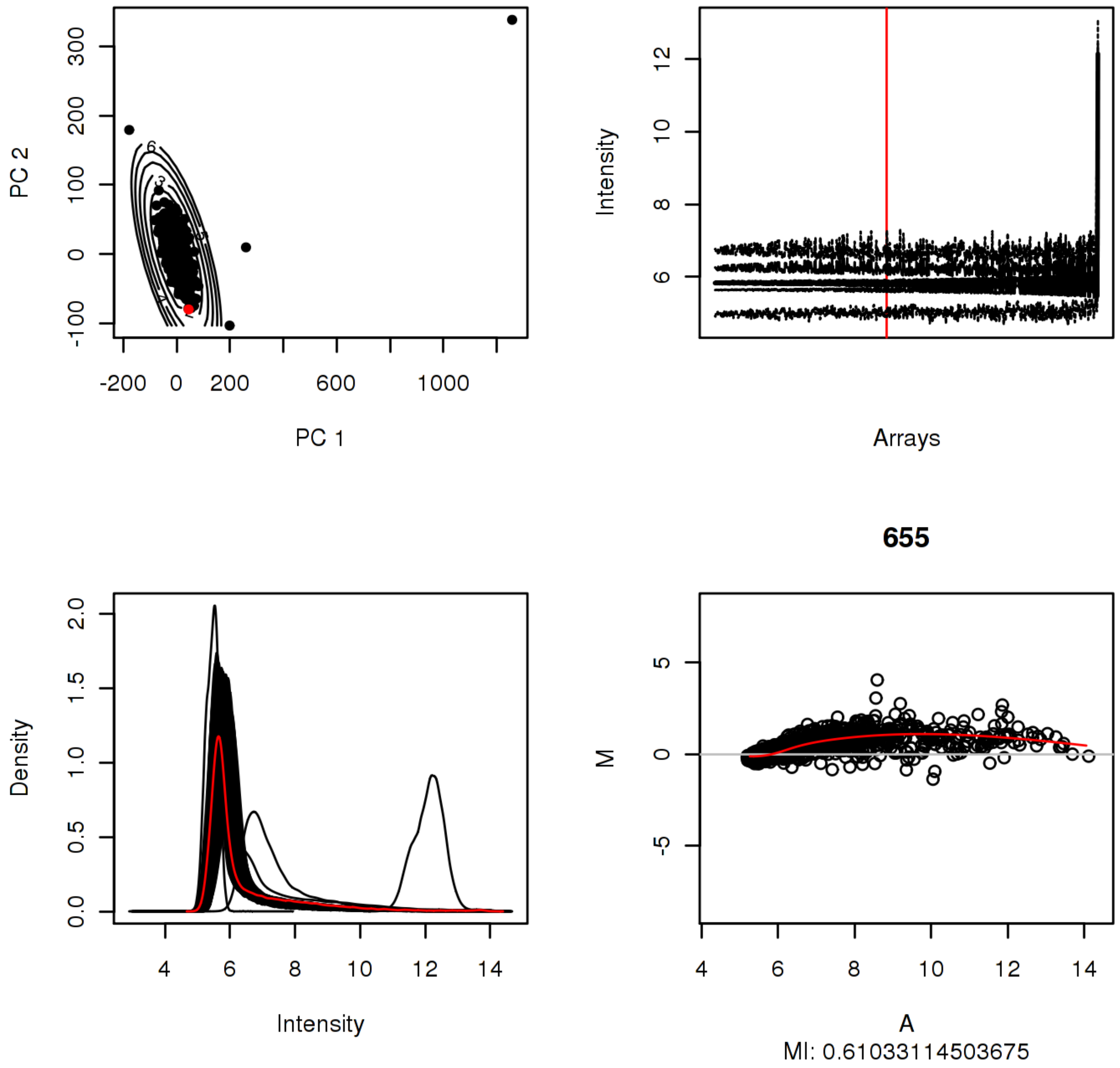
```
highlight("137", pca=prc_all, box=boxo, dens=densities, ma=maout)
```



```
for_removal <- "137"
```

4.3 655

```
highlight("655", pca=prc_all, box=boxo, dens=densities, ma=maout)
```

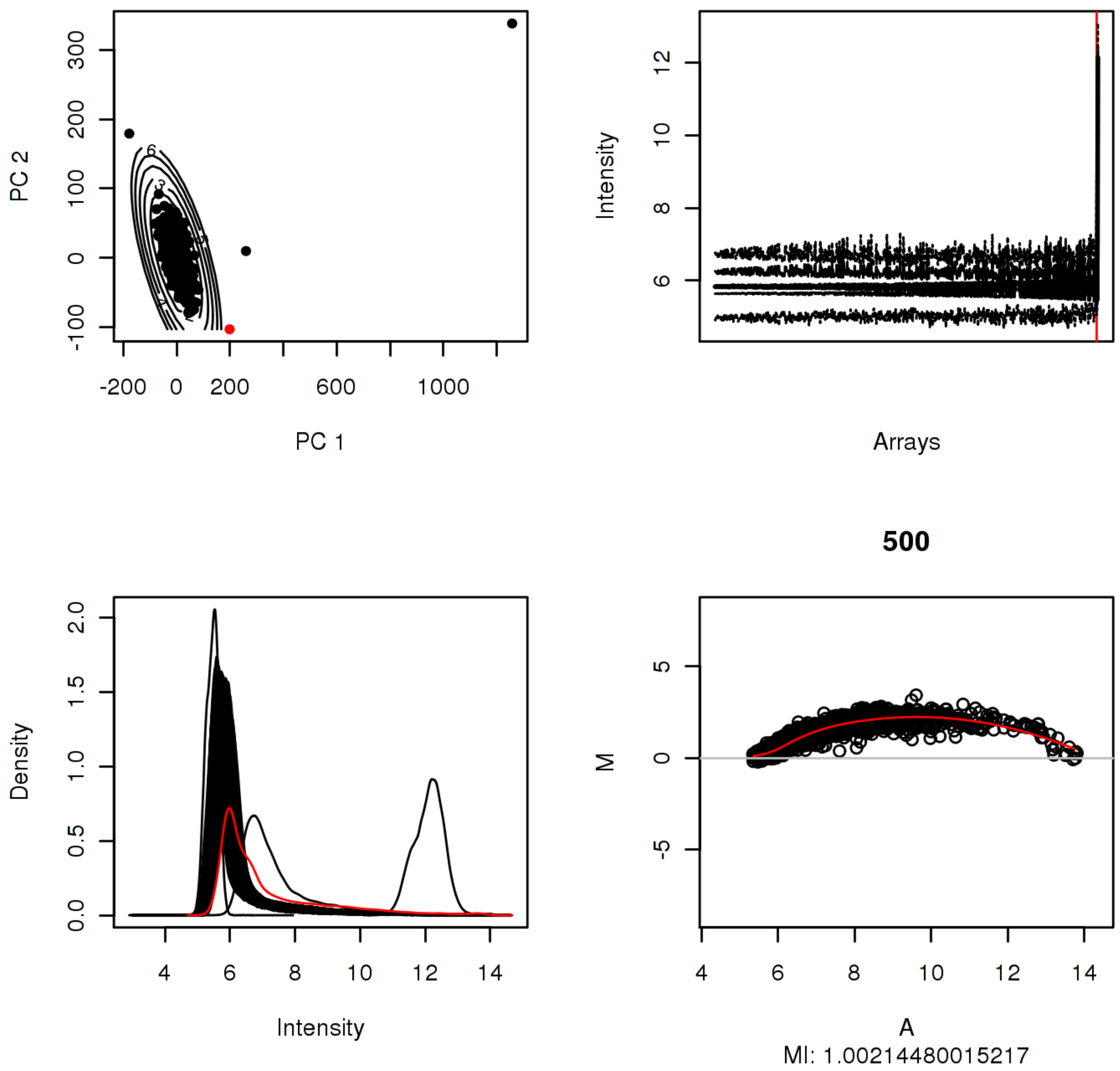


This one looks fine as well. Again it's probably the slightly high MI statistic.

4.4 500

This one once again looks strange in all the plots and I will take it out.

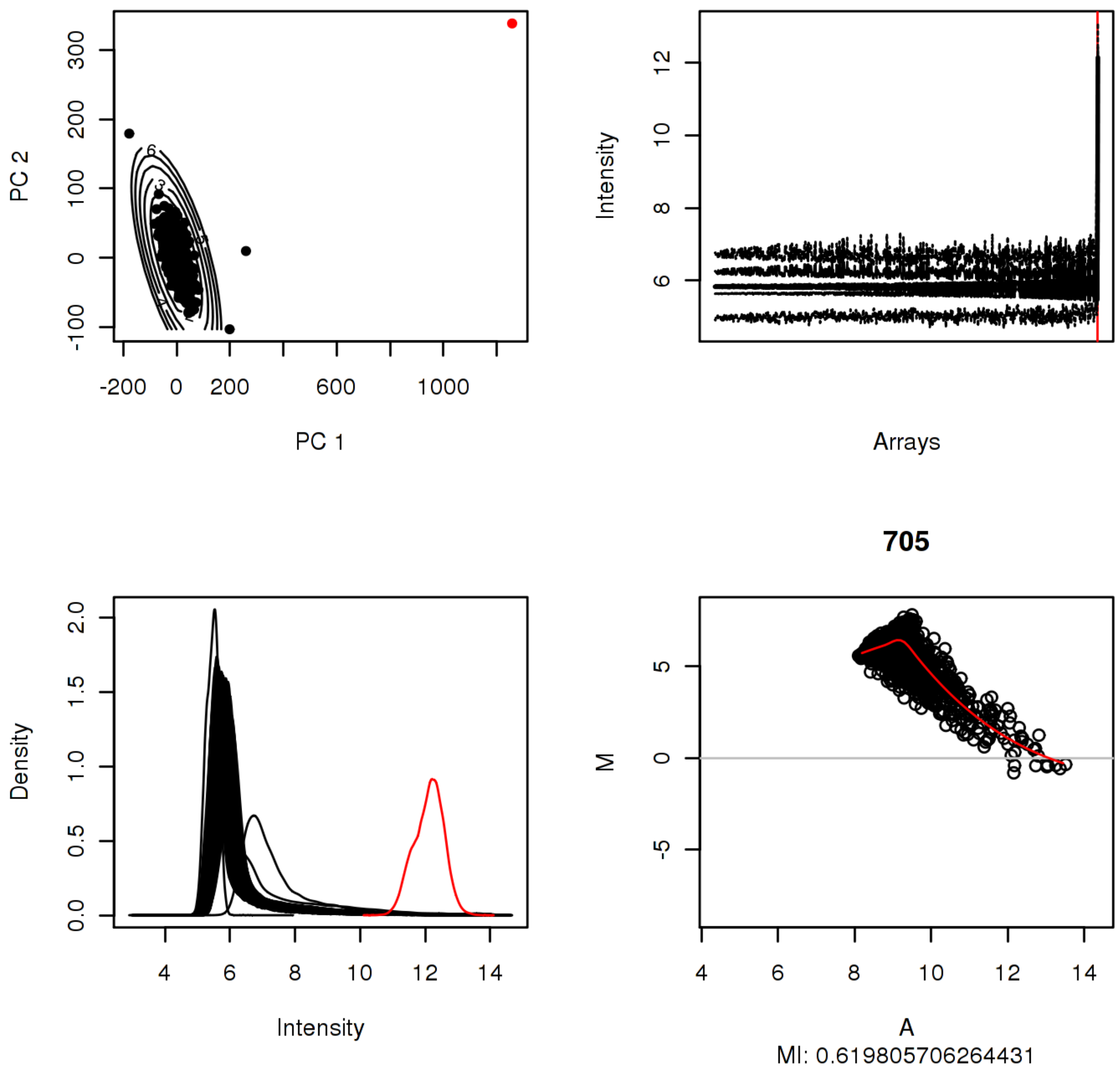
```
highlight("500", pca=prc_all, box=boxo, dens=densities, ma=maout)
```



```
for_removal <- c(for_removal, "500")
```

4.5 705

```
highlight("705", pca=prc_all, box=boxo, dens=densities, ma=maout)
```

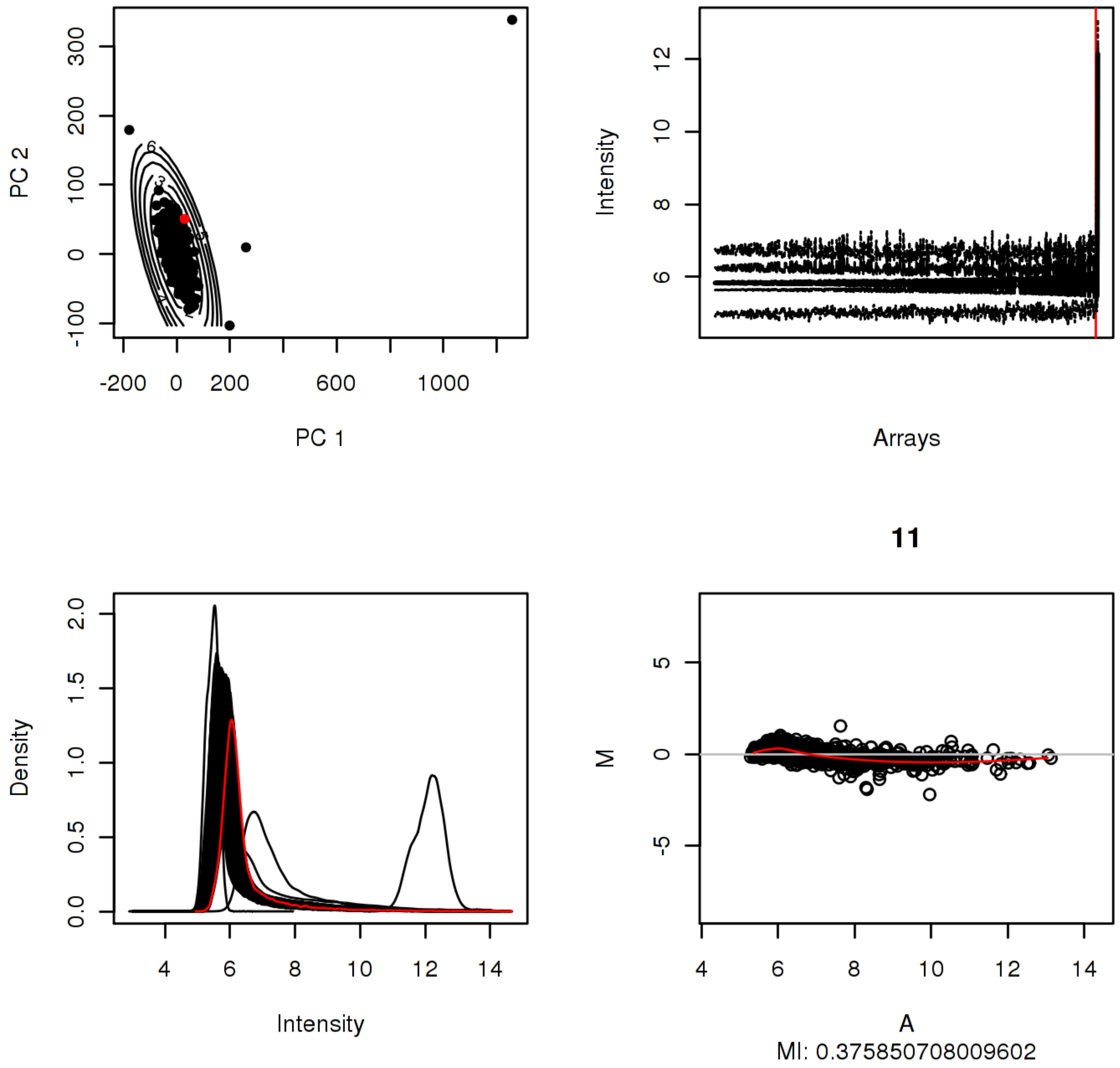


```
for_removal <- c(for_removal, "705")
```

Our most extreme point yet! Not only are the intensities pushed all the way to the right, there seems also to be some slight bimodality and other strangeness that the healthy samples don't exhibit.

4.6 11

```
highlight("11", pca=prc_all, box=boxo, dens=densities, ma=maout)
```



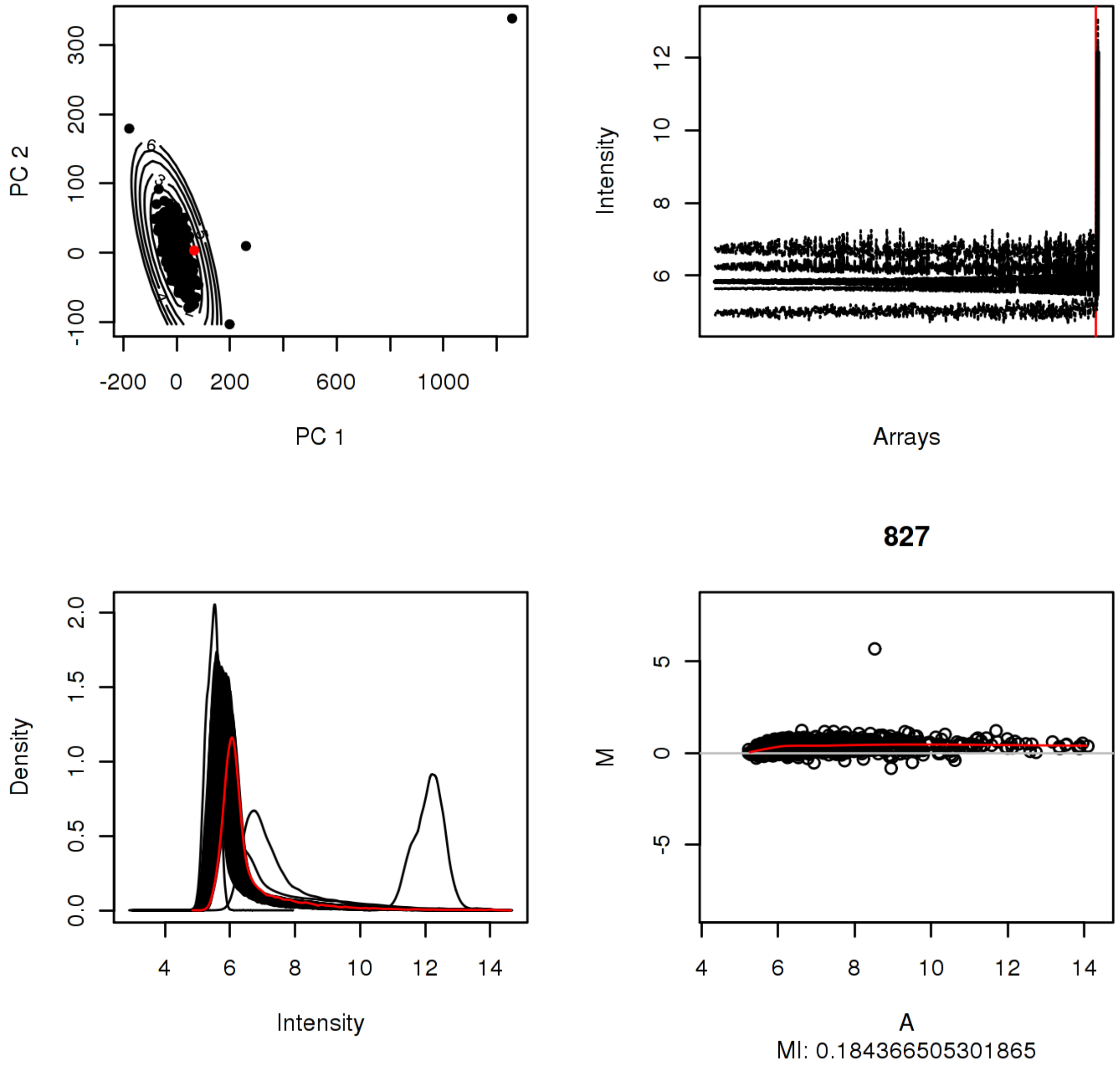
This one is more interesting, it's out there but not clearly broken. It looks as though the boxplots flagged it as outlier. Let's look at the lab info:

```
lab_info["11", ]

##    Ng/ul_RNA 260/280_RNA 260/230_RNA RIN Ng/ul_cRNA 260/280_cRNA 260/230_cRNA
## 11     55,85        2,06        1,83 8,1       1630          2,2          2,2

lab_thresholds

## [1] "Bad: RIN value < 7"            "Bad: 260/280 RNA ratio < 2"
## [3] "Bad: 260/230 RNA ratio < 1.7" "Good: 50 < Ng/ul RNA < 500"
```

It's not outside the predefined thresholds. Let's keep it.

4.7 827

```
highlight("827", pca=prc_all, box=boxo, dens=densities, ma=maout)
```

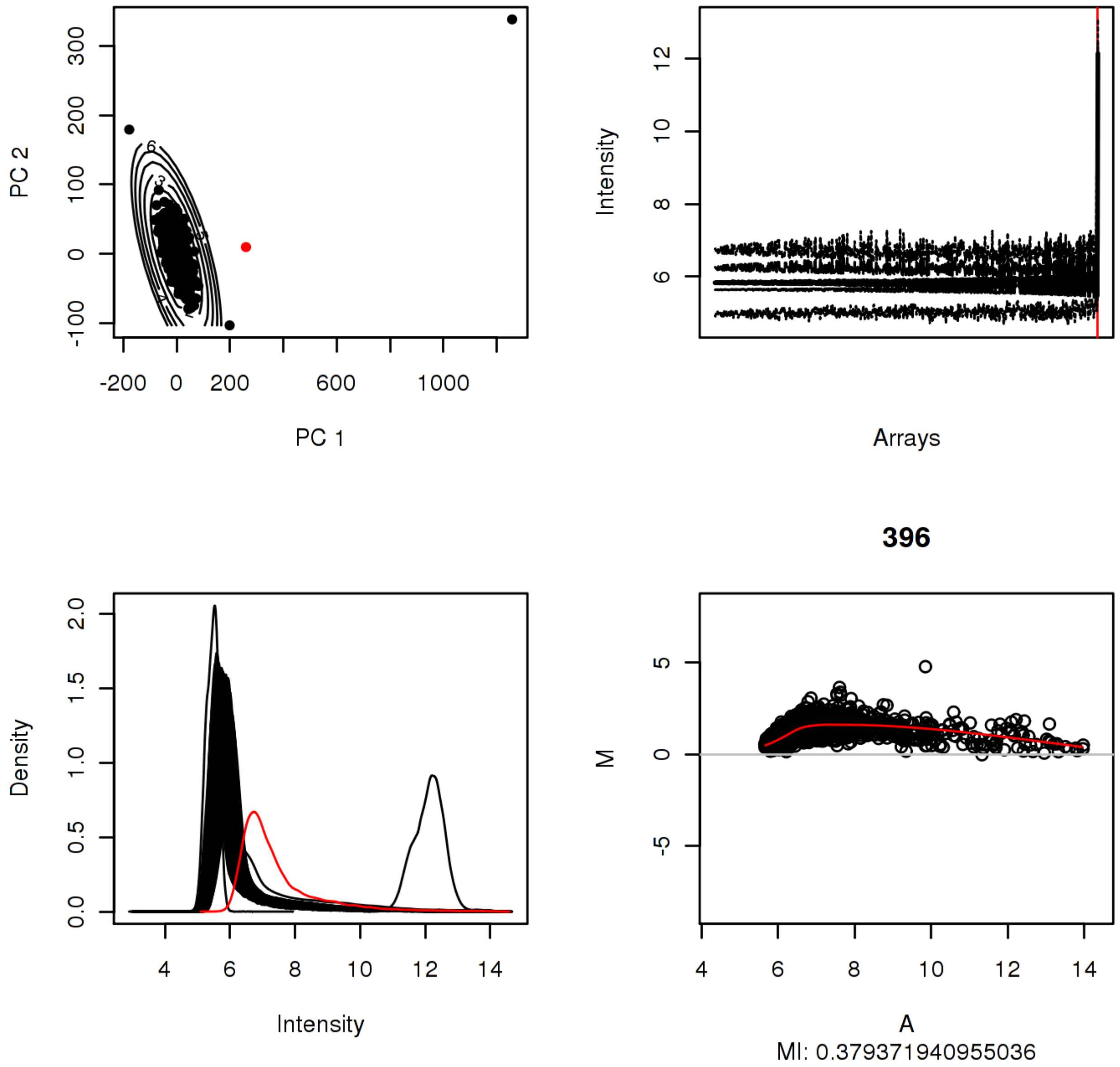


```
lab_info["827", ]

##     Ng/ul_RNA 260/280_RNA 260/230_RNA RIN Ng/ul_cRNA 260/280_cRNA 260/230_cRNA
## 827     94,44        2,08        1,72 8,2     2188,5        2,135        2,035

lab_thresholds

## [1] "Bad: RIN value < 7"           "Bad: 260/280 RNA ratio < 2"
## [3] "Bad: 260/230 RNA ratio < 1.7" "Good: 50 < Ng/ul RNA < 500"
```

This one is also slightly strange but not exactly outside the thresholds, so I'll keep it.

4.8 396

```
highlight("396", pca=prc_all, box=boxo, dens=densities, ma=maout)
```



```
lab_info["396", ]
## 	Ng/ul_RNA 260/280_RNA 260/230_RNA RIN Ng/ul_cRNA 260/280_cRNA 260/230_cRNA
## 396     125,5        2,16        1,38 9,0       1625          2,2          2,2
for_removal <- c(for_removal, "396")
```

This one is strange in three out of four plots and has a too-low 260/230 ratio.

4.9 Summary

```
outliers_manual <- for_removal
```

Normally we would do this whole process once more to make sure that the data look well behaved without these outliers. We won't include that in this document however.

5 Automated outlier removal

While we don't recommend doing automated removal it is interesting to see how many samples would be discarded with such an approach. We have already fit the different detection models, so let's just take everything flagged by predict() this time:

```
th <- 2
outliers_automatic <- c(predict(boxo, sdev=th), predict(prc_all, sdev=th),
                        predict(maout, sdev=th))
outliers_automatic <- unique(outliers_automatic)
outliers_automatic

##  [1] "11"  "97"  "105" "137" "175" "211" "232" "345" "386" "396" "421" "497" "500" "506"
## [15] "510" "513" "548" "580" "620" "649" "658" "700" "705" "769" "780" "810" "815" "827"
## [29] "2"   "33"  "37"  "51"  "110" "168" "181" "206" "221" "257" "270" "306" "317" "370"
## [43] "438" "482" "504" "509" "539" "561" "564" "582" "655" "699" "725" "728" "754" "757"
## [57] "777" "823" "831"

length(outliers_automatic)

## [1] 59
```

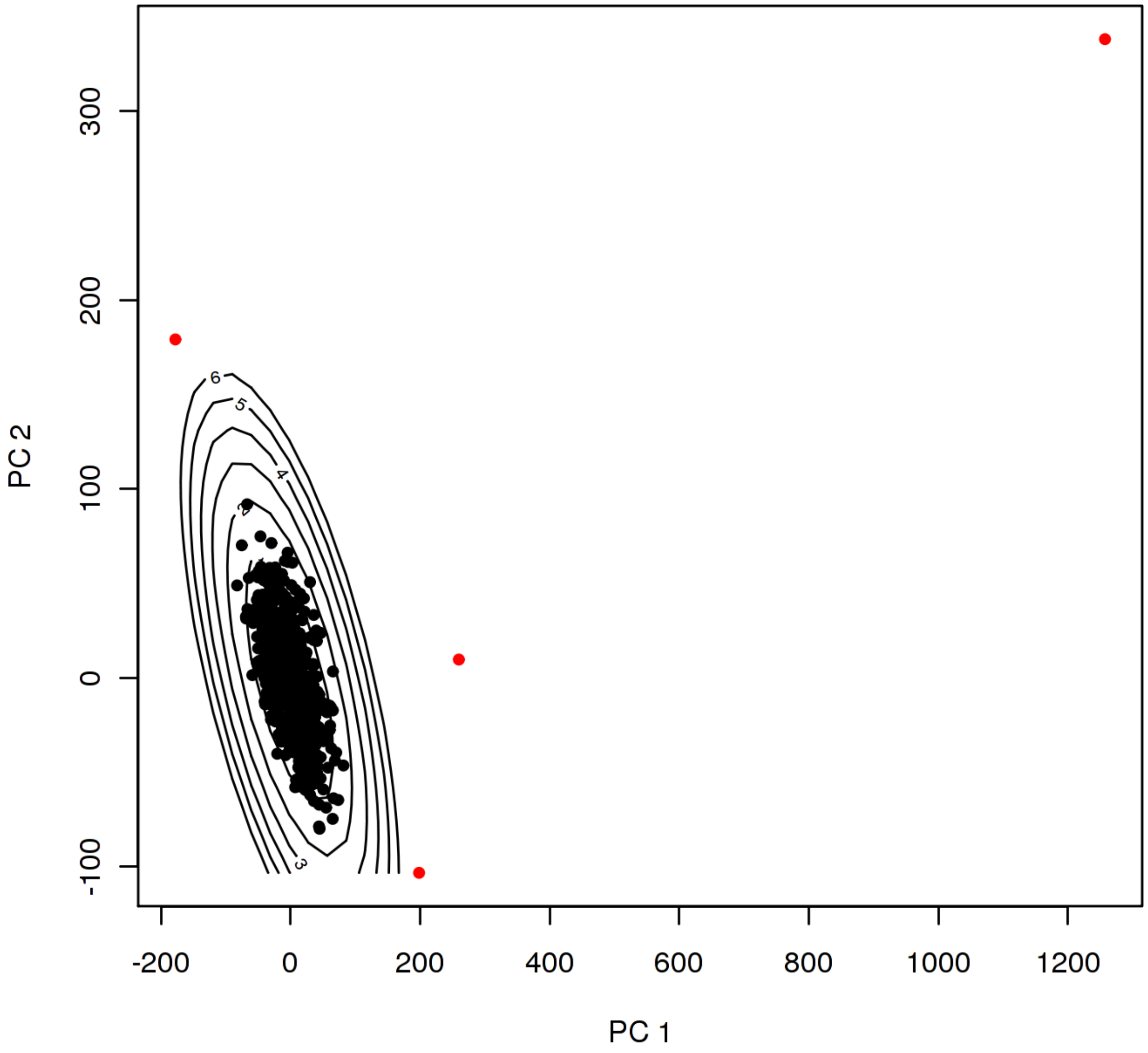
At the admittedly strict threshold of two standard deviations, this flags an astonishing 59 samples for removal. At the standard threshold of three standard deviations, only eight samples (the ones we looked at) would be flagged, but this is data dependent.

6 Summary

```
outliers_manual

## [1] "137" "500" "705" "396"
```

In the end we have four outliers we consider tecnical in nature. It's the ones you immediately feel strange about in the PCA plot:

```
plot(prc_all, highlight=for_removal)
```



```
# for experiments
save(outliers_manual, outliers_automatic, file="../dataset/outliers.rda")
```

7 Session info

- R version 3.1.2 (2014-10-31), x86_64-unknown-linux-gnu
- Locale: LC_CTYPE=en_US.UTF-8, LC_NUMERIC=C, LC_TIME=en_US.UTF-8, LC_COLLATE=en_US.UTF-8, LC_MONETARY=en_US.UTF-8, LC_MESSAGES=en_US.UTF-8, LC_PAPER=en_US.UTF-8, LC_NAME=C, LC_ADDRESS=C, LC_TELEPHONE=C, LC_MEASUREMENT=en_US.UTF-8, LC_IDENTIFICATION=C
- Base packages: base, datasets, graphics, grDevices, methods, parallel, stats, utils

- Other packages: Biobase 2.26.0, BiocGenerics 0.12.1, knitr 1.16, lumi 2.18.0, nowaclean 0.2.8
- Loaded via a namespace (and not attached): affy 1.44.0, affyio 1.34.0, annotate 1.44.0, AnnotationDbi 1.28.2, backports 1.1.0, base64 2.0, base64enc 0.1-3, BatchJobs 1.6, BBmisc 1.9, beanplot 1.2, BiocInstaller 1.16.5, BiocParallel 1.0.3, BiocStyle 1.4.1, biomaRt 2.22.0, Biostrings 2.34.1, bitops 1.0-6, brew 1.0-6, bumphunter 1.6.0, Cairo 1.5-9, checkmate 1.7.4, codetools 0.2-15, colorspace 1.2-6, DBI 0.6-1, digest 0.6.12, doRNG 1.6.6, entropy 1.2.1, evaluate 0.10, fail 1.3, foreach 1.4.3, genefilter 1.48.1, GenomeInfoDb 1.2.5, GenomicAlignments 1.2.2, GenomicFeatures 1.18.7, GenomicRanges 1.18.4, grid 3.1.2, highr 0.6, illuminaio 0.8.0, IRanges 2.0.1, iterators 1.0.8, KernSmooth 2.23-15, lattice 0.20-33, limma 3.22.7, locfit 1.5-9.1, magrittr 1.5, MASS 7.3-45, Matrix 1.2-6, matrixStats 0.50.2, mclust 5.2, memoise 1.1.0, methylumi 2.12.0, mgcv 1.8-12, minfi 1.12.0, multtest 2.22.0, nleqslv 3.0.1, nlme 3.1-128, nor1mix 1.2-2, openssl 0.9.4, pkgmaker 0.22, plyr 1.8.4, preprocessCore 1.28.0, quadprog 1.5-5, RColorBrewer 1.1-2, Rcpp 0.12.9, RCurl 1.95-4.8, registry 0.3, reshape 0.8.6, rngtools 1.2.4, Rsamtools 1.18.3, RSQLite 1.1-2, rtracklayer 1.26.3, S4Vectors 0.4.0, sendmailR 1.2-1, siggenes 1.40.0, splines 3.1.2, stats4 3.1.2, stringi 1.1.2, stringr 1.2.0, survival 2.39-4, tools 3.1.2, XML 3.98-1.4, xtable 1.8-2, XVector 0.6.0, zlibbioc 1.12.0

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory.*, pages 267–281. Akademiai Kiado.

Aristizábal-Pachón, A. F., de Carvalho, T. I., Carrara, H. H. A., de Andrade, J. M., and Takahashi, C. S. (2015). Detection of human mammaglobin a mrna in peripheral blood of breast cancer patients before treatment and association with metastasis. *Journal of the Egyptian National Cancer Institute*, 27(4):217–222.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300.

Best, M. G., Sol, N., Kooi, I., Tannous, J., Westerman, B. A., Rustenburg, F., Schellen, P., Verschueren, H., Post, E., Koster, J., Ylstra, B., Ameziane, N., Dorsman, J., Smit, E. F., Verheul, H. M., Noske, D. P., Reijneveld, J. C., Nilsson, R. J. A., Tannous, B. A., Wesseling, P., and Wurdinger, T. (2017). RNA-seq of tumor-educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics. *Cancer Cell*, 28(5):666–676.

Bolstad, B. M., Irizarry, R. A., Åstrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193.

Bøvelstad, H. M., Holsbø, E., Bongo, L. A., and Lund, E. (2017). A standard operating procedure for outlier removal in large-sample epidemiological transcriptomics datasets. *bioRxiv 144519*.

Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *Journal of the American Statistical Association*, 87(419):738–754.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability.

Monthey Weather Review, 78(1):1–3.

Brown, L. D., Cai, T. T., and DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, 16(2):101–133.

Cessie, S. L. and Houwelingen, J. C. V. (1992). Ridge estimators in logistic regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41(1):191–201.

Chi, K. R. (2016). The tumour trail left in blood. *Nature*, 532:269 – 271.

Cohen, J. D., Li, L., Wang, Y., Thoburn, C., Afsari, B., Danilova, L., Douville, C., Javed, A. A., Wong, F., Mattox, A., Hruban, R. H., Wolfgang, C. L., Goggins, M. G., Dal Molin, M., Wang, T.-L., Roden, R., Klein, A. P., Ptak, J., Dobbyn, L., Schaefer, J., Silliman, N., Popoli, M., Vogelstein, J. T., Browne, J. D., Schoen, R. E., Brand, R. E., Tie, J., Gibbs, P., Wong, H.-L., Mansfield, A. S., Jen, J., Hanash, S. M., Falconi, M., Allen, P. J., Zhou, S., Bettegowda, C., Diaz, L., Tomasetti, C., Kinzler, K. W., Vogelstein, B., Lennon, A. M., and Papadopoulos, N. (2018). Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science*, page eaar3247.

Cohen Freue, G. V., Hollander, Z., Shen, E., Zamar, R. H., Balshaw, R., Scherer, A., McManus, B., Keown, P., McMaster, W. R., and Ng, R. T. (2007). Mdqc: a new quality assessment method for microarrays based on quality control reports. *Bioinformatics*, 23(23):3162–3169.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., et al. (2016). A survey of best practices for rna-seq data analysis. *Genome biology*, 17(1):13.

Crick, F. (1958). On protein synthesis. In *Symposia of the Society for Experimental Biology*, volume 12, page 8.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561.

Du, P., Kibbe, W. A., and Lin, S. M. (2008). lumi: a pipeline for processing illumina microarray. *Bioinformatics*, 24(13):1547–1548.

Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica sinica*, 12:111–139.

Dumeaux, V., Børresen-Dale, A.-L., Frantzen, J.-O., Kumle, M., Kristensen, V. N., and Lund, E. (2008). Gene expression analyses in breast cancer epidemiology:

the norwegian women and cancer postgenome cohort study. *Breast Cancer Research*, 10(1):R13.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American statistical association*, 56(293):52–64.

Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press.

Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37(1):36–48.

Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors—an empirical bayes approach. *Journal of the American Statistical Association*, 68(341):117–130.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Fan, H. and Hegde, P. S. (2005). The transcriptome in blood: challenges and solutions for robust expression profiling. *Current molecular medicine*, 5(1):3–10.

Filipits, M., Rudas, M., Jakesz, R., Dubsky, P., Fitzal, F., Singer, C. F., Dietze, O., Greil, R., Jelen, A., Sevelda, P., et al. (2011). A new molecular predictor of distant recurrence in er-positive, her2-negative breast cancer adds independent information to conventional clinical risk factors. *Clinical Cancer Research*, pages clincanres–0926.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Frigessi, A., Bühlmann, P., Glad, I., Langaas, M., Richardson, S., and Vannucci, M. E., editors (2016). *Statistical Analysis for High-Dimensional Data*. Springer International Publishing.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian data analysis*. Chapman and Hall/CRC, third edition.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383.

Gelman, A. and Nolan, D. (2017). *Teaching statistics: A bag of tricks*. Oxford University Press.

Giuliano, M., Giordano, A., Jackson, S., De Giorgi, U., Mego, M., Cohen, E. N., Gao, H., Anfossi, S., Handy, B. C., Ueno, N. T., Alvarez, R. H., De Placido, S., Valero, V., Hortobagyi, G. N., Reuben, J. M., and Cristofanilli, M. (2014). Circulating tumor cells as early predictors of metastatic spread in breast cancer patients with limited metastatic dissemination. *Breast Cancer Research*, 16(5):440.

Goeman, J. J., Van De Geer, S. A., De Kort, F., and Van Houwelingen, H. C. (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99.

Goldstein, M. (1976). Bayesian analysis of regression problems. *Biometrika*, 63(1):51–58.

Greenland, S. (2000). When should epidemiologic regressions use random coefficients? *Biometrics*, 56(3):915–921.

Hanczar, B., Hua, J., Sima, C., Weinstein, J., Bittner, M., and Dougherty, E. R. (2010). Small-sample precision of ROC-related estimates. *Bioinformatics*, 26(6):822–830.

Hand, D. and Anagnostopoulos, C. (2013). When is the area under the receiver operating characteristic curve an appropriate measure of classifier performance? *Pattern Recognition Letters*, 34(5):492 – 495.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36.

Harrell, F. (2013). Regression modeling strategies. *as implemented in R package 'rms' version*, 3(3).

Harrell, F. E. and Lee, K. L. (1985). A comparison of the discrimination of discriminant analysis and logistic regression under multivariate normality. *Biostatistics: Statistics in Biomedical, Public Health and Environmental Sciences', North-Holland, New York, United States*, pages 333–343.

Hastie, T. and Tibshirani, R. (1990). Exploring the nature of covariate effects in the proportional hazards model. *Biometrics*, pages 1005–1016.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning, Second Edition*. Springer Series in Statistics. Springer New York, New York, NY.

Haury, A.-C., Gestraud, P., and Vert, J.-P. (2011). The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *Plos One*, 6(12):e28210.

Hausser, J. and Strimmer, K. (2009). Entropy inference and the james-stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research*, 10(Jul):1469–1484.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Holsbø, E. and Perduca, V. ((to appear) 2018). Shrinkage estimation of rate statistics. *Case Studies in Business, Industry and Government Statistics (CS-BIGS)*.

Holsbø, E., Perduca, V., Bongo, L. A., Lund, E., and Birmelé, E. (2018). Stratified time-course gene preselection shows a pre-diagnostic transcriptomic signal for metastasis in blood cells: a proof of concept from the nowac study. *bioRxiv 141325*.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379.

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127.

Kang, S. and Song, J. (2017). Robust gene selection methods using weighting schemes for microarray data analysis. *BMC bioinformatics*, 18(1):389.

Kauffmann, A., Gentleman, R., and Huber, W. (2008). arrayQualityMetrics—a Bioconductor package for quality assessment of microarray data. *Bioinformatics*, 25(3):415–416.

Kauffmann, A. and Huber, W. (2010). Microarray data quality control improves the detection of differentially expressed genes. *Genomics*, 95(3):138–142.

Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational statistics & data analysis*, 53(11):3735–3745.

Knuth, D. E. (1984). Literate programming. *The Computer Journal*, 27(2):97–111.

Kristensen, V. N., Lingjærde, O. C., Russnes, H. G., Vollan, H. K. M., Frigessi, A., and Børresen-Dale, A.-L. (2014). Principles and methods of integrative genomic analyses in cancer. *Nature Reviews Cancer*, 14(5):299–313.

Kwa, M., Makris, A., and Esteva, F. J. (2017). Clinical utility of gene-expression signatures in early stage breast cancer. *Nature Reviews Clinical Oncology*, 14(10):595.

Li, J. and Tibshirani, R. (2013). Finding consistent patterns: A nonparametric approach for identifying differential expression in rna-seq data. *Statistical Methods in Medical Research*, 22(5):519–536. PMID: 22127579.

Liew, C.-C., Ma, J., Tang, H.-C., Zheng, R., and Dempsey, A. A. (2006). The peripheral blood transcriptome dynamically reflects system wide biology: a potential diagnostic tool. *Journal of Laboratory and Clinical Medicine*, 147(3):126–132.

Lim, B. and Hortobagyi, G. N. (2016). Current challenges of metastatic breast cancer. *Cancer and Metastasis Reviews*, pages 1–20.

Liquet, B., Lafaye de Micheaux, P., Hejblum, B. P., and Thiébaut, R. (2015). Group and sparse group partial least square approaches applied in genomics context. *Bioinformatics*, 32(1):35–42.

Lund, E., Dumeaux, V., Braaten, T., Hjartåker, A., Engeset, D., Skeie, G., and Kumle, M. (2008). Cohort profile: the norwegian women and cancer study—nowac—kvinner og kreft. *International journal of epidemiology*, 37(1):36–41.

Lund, E., Holden, L., Bøvelstad, H., Plancade, S., Mode, N., Günther, C.-C., Nuel, G., Thalabard, J.-C., and Holden, M. (2016). A new statistical method for curve group analysis of longitudinal gene expression data illustrated for breast cancer in the nowac postgenome cohort as a proof of principle. *BMC Medical Research Methodology*, 16(1):28.

Ma, X.-J., Salunga, R., Dahiya, S., Wang, W., Carney, E., Durbecq, V., Harris, A., Goss, P., Sotiriou, C., Erlander, M., et al. (2008). A five-gene molecular grade index and hoxb13: Il17br are complementary prognostic factors in early stage breast cancer. *Clinical cancer research*, 14(9):2601–2608.

Marczyk, M., Krol, L., and Polanska, J. (2014). Automatic detection of outlying microarrays using multi-array quality metrics. In *IWBBIO*, pages 738–746.

Molinaro, A. M., Simon, R., and Pfeiffer, R. M. (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, 21(15):3301–3307.

Morris, C. N. (1983). Parametric empirical bayes inference: theory and applications. *Journal of the American Statistical Association*, 78(381):47–55.

Mosteller, F. and Tukey, J. W. (1977). Data analysis and regression: a second course in statistics. *Addison-Wesley Series in Behavioral Science: Quantitative Methods*.

Olsen, K. S., Holsbø, E., Rognmo, K., Krum-Hansen, S., and Lund, E. (2015). Stress related to a suspicious mammogram - potential transcriptomic effects. *The 7th Conference on Epidemiology and Registry-Based Health Research - NordicEpi 2015*.

Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F. L., Walker, M. G., Watson, D., Park, T., et al. (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine*, 351(27):2817–2826.

Parker, H. (2017). Opinionated analysis development. *PeerJ Preprints*, 5:e3210v1.

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology*, 27(8):1160.

Plancade, S., Rozenholc, Y., and Lund, E. (2012). Generalization of the normal-exponential model: exploration of a more accurate parametrisation for the signal distribution on illumina beadarrays. *BMC bioinformatics*, 13(1):329.

Robinson, D. (2017). *Introduction to Empirical Bayes: Examples from Baseball Statistics*. Gumroad.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

Sandanger, T. M., Nøst, T. H., Guida, F., Rylander, C., Campanella, G., Muller, D. C., van Dongen, J., Boomsma, D. I., Johansson, M., Vineis, P., et al. (2018).

Dna methylation and associated gene expression in blood prior to lung cancer diagnosis in the norwegian women and cancer cohort. *Scientific reports*, 8(1):16714.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467.

Shieh, A. D. and Hung, Y. S. (2009). Detecting outlier samples in microarray data. *Statistical applications in genetics and molecular biology*, 8(1):1–24.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1):1–25.

Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., et al. (2006). Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute*, 98(4):262–272.

Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004). *Bayesian approaches to clinical trials and health-care evaluation*, volume 13. John Wiley & Sons.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 197–206, Berkeley, Calif. University of California Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121.

Tversky, A. and Kahneman, D. (1971). Belief in the law of small numbers. *Psychological bulletin*, 76(2):105.

van Schooneveld, E., Wouters, M. C., Van der Auwera, I., Peeters, D. J., Wildiers, H., Van Dam, P. A., Vergote, I., Vermeulen, P. B., Dirix, L. Y., and Van Laere, S. J. (2012). Expression profiling of cancerous and normal breast tissues identifies micrornas that are differentially expressed in serum from patients with (metastatic) breast cancer and healthy volunteers. *Breast Cancer Research*, 14(1):R34.

Van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., Van Der Kooy, K., Marton, M. J., Witteveen, A. T., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *nature*, 415(6871):530.

Verweij, P. J. and Van Houwelingen, H. C. (1994). Penalized likelihood in cox regression. *Statistics in medicine*, 13(23-24):2427–2436.

Von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., Vandenbroucke, J. P., Initiative, S., et al. (2007). Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *PLoS medicine*, 4(10):e296.

Wainer, H. (2007). The most dangerous equation. *American Scientist*, 95(3):249.

Wald, A. (1947). An essentially complete class of admissible decision functions. *The Annals of Mathematical Statistics*, 18(4):549–555.

Wang, Z., Gerstein, M., and Snyder, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57.

Wasserman, L. (2010). *All of Statistics: A Concise Course in Statistical Inference*. Springer Publishing Company, Incorporated.

Welch, B. L. (1947). The generalization of student's problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.*, 9(1):60–62.

Xie, Y. (2014). knitr: A comprehensive tool for reproducible research in R. In Stodden, V., Leisch, F., and Peng, R. D., editors, *Implementing Reproducible Computational Research*. Chapman and Hall/CRC. ISBN 978-1466561595.

Xie, Y., Wang, X., and Story, M. (2009). Statistical methods of background correction for illumina beadarray data. *Bioinformatics*, 25(6):751–757.

Yang, S., Guo, X., Yang, Y.-C., Papcunik, D., Heckman, C., Hooke, J., Shriver, C. D., Liebman, M. N., and Hu, H. (2006). Detecting outlier microarray arrays by correlation and percentage of outliers spots. *Cancer informatics*, 2:117693510600200017.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the “degrees of freedom” of the lasso. *Ann. Statist.*, 35(5):2173–2192.