Chapter

Dimensionality and Structure in Cancer Genomics: A Statistical Learning Perspective

Jacob Bradley

Abstract

Computational analysis of genomic data has transformed research and clinical practice in oncology. Machine learning and AI advancements hold promise for answering theoretical and practical questions. While the modern researcher has access to a catalogue of tools from disciplines such as natural language processing and image recognition, before browsing for our favourite off-the-shelf technique it is worth asking a sequence of questions. What sort of data are we dealing with in cancer genomics? Do we have enough of it to be successful without designing into our models what we already know about its structure? If our methods do work, will we understand why? Are our tools robust enough to be applied in clinical practice? If so, are the technologies upon which they rely economically viable? While we will not answer all of these questions, we will provide language with which to discuss them. Understanding how much information we can expect to extract from data is a statistical question.

Keywords: dimensionality, sparsity, high-dimensional statistics, cancer genomics, biomarkers, learning theory

1. Introduction

This chapter should be equally approachable to those with a background in machine learning/statistics and those with a more biological background. Beginning with a contextualisation of cancer genomics as the starting point for drug and biomarker discovery, we will attempt to convince the reader that statistical theory serves as the backbone and language of modern developments in machine learning. In order to facilitate those with less experience in biology, we will provide a very brief introduction to the types of data encountered in sequencing-based studies and the opportunities and problems they present. After providing some terminology and useful concepts from high-dimensional statistics, we will discuss how these concepts arise naturally in the context of cancer genomics, with some illustrative examples of how different techniques may be employed in translational scientific research. We will conclude by providing sketches of some modern developments and a description of the transition from what can loosely be termed statistical learning to what nowadays is referred to as machine learning.

1.1 Cancer genomics in drug discovery

Since the success of the Human Genome Project [1], sequencing technologies have improved at an exponential rate, both in terms of cost per megabase sequenced and the number of individuals who have had some portion of their genome sequenced (although the cost remains higher in practice than often reported) [2]. This has introduced an invaluable new resource for biomedical research in general. For the study of cancer, a disease of the genome, the ability to rapidly and cheaply sequence normal and tumour-derived DNA has transformed basic research, birthing the field of cancer genomics. This is beginning to impact frontline clinical oncology [3]. Whole genome sequencing is not yet standard of care for the generic cancer patient, but access to in-depth genetic data is becoming more common. Initiatives such as the 10,000/100,000 Genomes Projects [4] and The Cancer Genome Atlas [5] have given researchers access to large clinical datasets with a variety of accompanying omics data.

Understanding the genomic landscape of cancer genomes is critical to the drug discovery pipeline [6], particularly in pre-clinical identification of targets and biomarkers. Knowledge of the location and associated products of oncogenes (genes in which mutation can cause a cell to become cancerous) can allow for intelligent selection of druggable sites and identification of tumour suppressor genes (genes that under normal circumstances prevent uncontrolled cell division) gives options for therapies which may replace patients' defective cell cycle control mechanisms. Alongside new drugs, it is becoming increasingly common for therapies to be offered alongisde genomic biomarkers, which may stratify patients who are more likely to benefit from the treatment [7, 8].

These new sources and types of data allow researchers a greatly expanded toolbox with which to investigate the causes and development of cancer, but also present a unique set of challenges. The number of covaraiates in omics datasets causes a variety of theoretical and practical problems for classical statistical analysis, a problem often referred to as the curse of dimensionality [9].

1.2 Statistical learning and machine learning

Informally, the field of high-dimensional statistics attempts to address theoretical and computational problems associated with datasets in which the number of covariates (in our case this may refer to chromosomal locations or genes) is comparable to or greater than the number of samples available. In these settings results such as the central limit theorem that rely on divergence of the sample size independent of the dimensionality are often not of much use [10]. This is often the case in cancer genomics.

Recent decades have seen much excitement around the application of machine learning methods to a wide variety of high-dimensional problems. Particular progress has been made in automated image recognition and natural language processing (NLP). This progress has come via the development of specialised techniques to exploit the **structure** inherent in each data type (e.g. convolutional neural networks for image recognition [11] and word embedding for NLP [12]), but also from a vastly increased pool of data on which to train models. These data resources have typically been collected online, where there exists an abundance of labelled and unlabelled images and pieces of text.

It is hoped that similar strides forward can be anticipated in biology, but it is important to acknowledge the current gap in data availability between cancer genomics and the other machine learning disciplines mentioned above. In the next section we will discuss typical types of biological data encountered in cancer

genomics (including sequencing-based omics technologics that may not strictly be genomics, such as gene expression profiling), their dimensionality and typical availability. While efforts to deploy machine learning architectures are certainly producing results in some cases [13, 14], an important takeaway is that in many cases, we are not yet in a situation where the data-heavy deep learning approaches that have revolutionised image recognition will be applicable to cancer genomics problems.

That is not to say that we cannot do anything! In fact, it is often instructive to try and make headway in situations where a 'data-heavy, structure-light' approach is unsuitable, and these sorts of investigations can have a profound impact on the design of more sophisticated models [15]. As a final point, readers approaching without a significant backlog of machine learning expertise will find that an understanding of statistical terminology will aid comprehension of the machine learning literature which has them as its basis.

2. Omics and biological data

2.1 DNA sequencing

Cancer genomics is underpinned by the ability to sequence DNA cheaply and quickly. DNA is organised into chromosomes, along each of which many genes are arranged, with further non-coding regions interspersed in-between. The fundamental units of DNA are nucleotide bases, of which there are four varieties (labelled C, G, T and A). These are organised in groups of length three called codons, which code for the production amino acids. Codons are arranged in sequences such that their amino acids when joined in chain form proteins—the products of genes.

The aim of sequencing is to read, base by base, the information content of DNA. This was originally done by Sanger sequencing, a procedure to infer the base composition of a piece of DNA one base at a time. High-throughput sequencing automates this process via the following workflow:

1. DNA is isolated from a sample and amplified (replicated many times) to ensure good signal.

2. Purified DNA is broken into many pieces of manageable length.

3. These short strands are sequenced individually and simultaneously by an automated process similar to Sanger sequencing.

4. These short sequences are matched to a reference human genome to identify where the DNA in the original sample differed from that reference.

2.1.1 Tumour/normal variants

In cancer, some subset of cells accumulate mutations, via random misreplication of DNA during cell division or exposure to some external mutagen (e.g. cigarette smoke, UV light). Tumour cells therefore contain DNA with a different sequence to that of the patients' typical sequence. To understand this two samples are collected, one from the tumour and one from normal tissue, and both are sequenced. The sequences are compared and this produces a list of locations at which mutations have occured: these mutations can have a variety of types (replacements, insertions, etc.) and can have vastly differing functional implications.

In simplest setting, we could express a tumour's mutational profile as a vector, with each component corresponding to whether the tumour-derived and normal sequences match at that point. How long would this vector be? The human genome contains approximately 3×10^9 base locations. This is the dimensionality (which we will refer to later on as p) of naively presented genomic data. We often like to compare the dimensionality of a dataset with the number of samples (which we will later call n) to which we can expect to have access. In this case, unless we have access to tumour profiling for more than a third of all humans on the planet, we can never hope that these numbers will be comparable. We could make a small gain by listing all codons in the genome, labelling a component as one if the codon has been functionally altered by mutations and zero otherwise. Here though we would still have $p = 10^9$.

We could simplify our data further. Decades of biological research has focused on cataloguing the locations of genes across the genome. We might consider as covariates each of the (approximately 2×10^4) genes, and represent each sample as a vector where each component refers to (a) whether or not the gene contained a functional mutation; (b) how many such mutations were present; or (c) some other representation of the severity of collective mutations presents in the gene, drawing upon known biology. It is important to appreciate the trade-off we have made here: we have imposed an external notion of structure onto our data and in return have greatly reduced the dimensionality (by five orders of magnitude), but in exchange have lost resolution and thus potential information. This gain/sacrifice will be reflected when we choose to make even further structural assumptions in order to construct sensible models.

2.1.2 Heterogeneity and depth

Another important concern for those dealing with cancer genome data is that tumours are often highly heterogeneous. Different sub-populations of cells have different mutation profiles, which fit into an evolutionary hierarchy within the tumour's history. The importance of understanding the role of heterogeneity is beginning to be appreciated in a clinical context, and this has implications for the type of data that are used. In the context of the high-throughput sequencing pipeline, the relevant quantity is depth: identifying not just one but a variety of tumour sequences at a genomic locus along with the proportion in which they occur means thinking very hard about how best to express that data.

2.2 Gene expression

It is often not just the sequence of a gene which it relevant in a tumour, but the level of gene expression. The way that this is most often estimated is via the proxy of RNA transcript abundance: RNA is a similar molecule to DNA that is produced during the process of DNA being 'read', and acts as a messenger for sequences that should be converted to protein. Abundances of different RNA transcripts can be measured using procedures based on DNA sequencing. This will in general give data with the same dimensionality as gene-based mutation data, but is of a different type. Measured values are continuous to represent concentrations of gene products, rather than discrete 'mutated/not mutated' values. This has implications as to the sort of structural assumptions we can make about the data that we observe, and the models that will be best suited to capitalise on that structure.

3. Dimensionality and structure in statistical learning theory

Now familiar with the most relevant biological concepts, we turn to the mathematical theory of high-dimensional statistics, which has experienced a surge of interest in the last two decades. This is the language with which we will be attempting to interrogate issues of inference and prediction in cancer genomics. Informally, we may think of high-dimensional statistics to be concerned with the realm in which the dimensionality of our input data, p, is comparable to or greater than the number of training samples n we have available. In this regime the classical asymptotic theory of statistics, which generally relies on an assumption of fixed dimension and considers limiting behaviour as $n \to \infty$, may fail to apply. Classical results such as the law of large numbers and central limit theorem are not applicable.

3.1 What is high-dimensional statistics?

We often consider a very generic setup, in which we have paired data $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$. We model each of these pairs as being drawn from a joint probability distribution $P_{X \times Y}$, which gives the probability of observing any combination of observation x and label y. For now we make no assumptions about the nature of the y_i labels: they may be continuous values (regression), discrete values (classification) or more complicated objects such as is the case in survival analysis. We assume that $x_i \in \mathcal{X} \subset \mathbb{R}^p$ for each $1 \leq i \leq n$, so that our observed values are vectors of length p and each element is a real number (possibly restricted to some subset such as the positive reals—this is what $\mathcal{X}$ specifies). We refer to p as the dimension and n as the sample size of our data. We wish to fit some model $\mathcal{M}$ to the data. This could be in order to make some inference about the parameters of the distribution $P_{X \times Y}$, which will hopefuly shed light on the effect of each of the covariates contained in an observation x. Alternatively, we might be trying to predict future values of y from unlabelled observations as accurately as possible. These two aims are often distinguished by the umbrella terms statistical inference and statistical learning.

In many statistical models we have a vector β of parameters with at least the same dimension as our data ($\beta \in \mathbb{R}^q$, $q \geq p$). In generalised linear models (GLMs) the likelihood of an observation y depends upon the data x_i solely via the inner product $x_i^T \beta$, so that each component of β corresponds to the relative importance of its associated covariate. Classically, we would attempt to to estimate the parameter β via our observation through a procedure such as likelihood maximisation. However, it is clear in this context that if p is comparable to or larger than n then we have very little chance of accurately inferring the parameter vector β. For example, we canot expect to simultaneously learn about the effect of 20 covariates if we only have 10 observations: we say here that the model is unidentifiable.

High-dimensional statistics attempts to gauge what we can do in regimes such as these. One is approach is to assume the data has some low-dimensional structure. This means that we can embed our data in a lower dimensional space such that the smaller representation of our data contains all or most of the necessary information about the joint distribution $P_{X \times Y}$. We will discuss some common structural assumptions. The simplest and most interpretable is sparsity.

Definition 3.1. (Sparsity): 'Relatively few covariates are important'.

Given a vector $\beta \in \mathbb{R}^p$ parameterising a model, we say β is k-sparse, for $k \leq p$, if at most k elements of β are non-zero, that is

$$|\beta|_0 := \sum_{j=1}^{p} \mathbf{1}\left\{\beta_j \neq 0\right\} \leq k.$$

We can say a model $\mathcal{M}$ parameterised by a vector β is k-sparse if the vector β is k-sparse.

Sparsity is a useful assumption to make for a variety of reasons. We are reducing the number of parameters that we must estimate—for a k-sparse model, we need only estimate k parameters. Before we do so we need to decide which k parameters are allowed to be non-zero, that is, to which k-dimensional subspace (out of $\binom{p}{k}$ choices) our parameter belongs. In practice this is not a huge issue—some powerful theory from the field of convex optimisation allows for efficient training of sparse models (see the LASSO estimator below). Finally, sparse models are interpretable. A small number of covariates selected for importance can be useful in hypothesis refinement.

3.1.1 Sparse data vs. sparse models

It is worth at this point drawing a distinction between two phenomena in statistics and data science both referred to as 'sparsity', both of which are exhibited in cancer genomics. The first is sparse *data*, in which almost all observed data points have the same value (typically zero). Mutation data displays this trait—the rate at which mutations occur in the genome varies widely across and within cancer types, but rarely exceeds 100 Mut/Mb, that is one mutation per 10^4 nucleotide base pairs [16]. This sparsity is exploited in the way that tumour/normal DNA data is stored, in file formats such as VCF (variant called format) and MAF (mutation annotated format). Many programming languages and data science packages have data structures optimised for sparse data, and it is also often possible to optimise learning and algorithms for sparse data. However, here we will focus on sparse *models*. These are models where it is assumed that only a small subspace of the covariate space is relevant, via assumptions such as the one described above.

This notion that there is some sparse representation of data but that it may not translate directly to a subset of our covariates motivates the more general principle of Sufficient Dimension Reduction (SDR). Sparsity restricts our attention to some small subspace of the covariate space $\mathbb{R}^p$. More generally, we may insist on some important smaller subspace, but one that does not depending on a specific representation of our data x. The definition of SDR is somewhat more technical, so those without mathematical background may find it easier to skip.

Definition 3.2. (Sufficient Dimension Reduction): 'Some small representation of our data contains all the important information'.

Given (X, Y) drawn from probability distribution $P_{X\times Y}$, we say there exists a sufficient dimension reduction of size d^* if there exists some function $S : \mathbb{R}^p \to \mathbb{R}^{d^*}$ with $d^* < p$ such that Y is conditionally independent of X given $S(X)$, that is,

$$Y \perp\!\!\!\perp X \mid S(X)$$

For an observation x, the image $S(X)$ is a d^*-dimensional representation of x. As a special case we have linear sufficient dimensional reduction if the function S is a linear projection $A^* : \mathbb{R}^p \to \mathbb{R}^{d^*}$.

Picking apart this definition, conditional independence means that Y only depends on X through some low-dimensional image. Note that, in contrast to sparsity, we have not made reference to a linear model parameter β. In fact, in the

context of a generalised linear model where Y depends on X only through some function of $\beta^T X$, we can simply take $S(X) = A^* X = \beta^T$ and see that Y admits a sufficient dimensionality condition with $d^* = 1$. SDR, therefore, is a helpful notion in settings in which we need to apply a non-linear model structure. Methods based on finding sufficient dimension reduction projections by searching through spaces of projections [17] in combination with non-linear base classifiers are beginning to show promise in a variety of domains including the analysis of high-dimensional medical data [18].

3.1.2 Techniques in high-dimensional statistics: Selection and regularisation

It is all very well imposing assumptions of low-dimensional structure onto our data. How can we now exploit this to produce models that reflect the structural assumptions we have made? One answer is regularisation. Regularisation refers to some penalisation process being applied to the parameters of our model. The intuition is that, given some model parameter β of size greater than or equal to the dimension p of our data, and thus of comparable magnitude to our number of samples, we have enough degrees of freedom when fitting the model that we can be guaranteed to produce almost perfect training set results without having done anything more than memorise our data. Therefore we must place restrictions on our parameter, and the trick is to do this as part of the model fitting process by combining a regularisation term to the loss function of our learning procedure (ideally in such a way as to preserve what is known as loss convexity, which allows efficient model fitting).

Regularisation is applied in practice across a whole range of model types, but is easiest to understand in the context of linear regression, so in the discussion that follows we will restrict ourselves to this setting.

In linear regression we have a model $\mathcal{M}_\beta$, parameterised by β, given by

$$\mathcal{M}_\beta : Y_i = X_i^T \beta + \varepsilon, \tag{1}$$

for some noise ε. We are saying that Y can be approximated by a linear combination of the components of X, with the relative weightings of each component given by the components of β. The loss of our model (a measure of how inaccurately it is predicting across all our data) is given by

$$\mathcal{L}(\mathcal{M}_\beta) = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - X_i^T \beta\right)^2. \tag{2}$$

In general we choose β to minimise this loss for an optimal model, but suppose we wish to find an optimal k-sparse model, that is one for which β is k-sparse. Rather than minimising over all possible choices of β, we are minimising the loss over all values of β that are also k-sparse:

$$\min_{\beta \in \mathbb{R}^p, \quad |\beta|_0 \leq k} \left\{ \mathcal{L}(\mathcal{M}_\beta) \right\}. \tag{3}$$

Here we face a computational difficulty: we have to separately check each subset of covariates of size k and minimise on that set of possible parameters, then compare them all to find the best. What we do to circumvent this is include a penalisation term for β, which encourages sparsity alongside the loss function in our optimisation. An obvious choice would be the L0 'norm', $|\beta|_0$, which counts non-zero coefficients. In practice this is not computationally feasible (to be technical, the problem is non-convex and so NP-hard), so instead we use the the L1 norm $|\beta|_1$

given by $\sum_{j=1}^{p}|\beta_j|$. While this does not explicitly encode sparsity, it turns out that in practice it does produce sparse solutions. This process of replacing a non-convex problem with an easier one is in general called convex relaxation.

Technique 3.1. (Regularisation for Sparsity: L1/LASSO): Given the setup above, L1 regularised estimation (known in the case of linear regression as the LASSO estimator [19]) selects β solving the following optimisation

$$\min_{\beta \in \mathbb{R}^p} \left\{ \sum_{i=1}^{n} \mathcal{L}\left(\mathcal{M}_\beta\right) + \lambda|\beta|_1 \right\}$$

where λ is a positive number chosen to specify how strongly we want to encourage sparsity: different values of λ will produce different k s in the ouput. A particularly attractive feature of the LASSO selector is that it acts simultaneously as a variable selection and model fitting procedure.

To take stock, we have begun with an assumption that some small subset of our covariates are important in predicting the response Y. This assumption might have come from necessity due to data availability, from knowledge of the biological system we are modelling, or from both. We will discuss these possibilities in more depth in the next chapter. We have taken a simple model, and altered it to express this structure, and have done so in a way that is computationally feasible.

The specific form of the regularisation we employ can have very subtle effects on the traits it encourages in models, which should motivate us to be very careful when translating the biological knowledge we want to express into our learning systems. For example, adding an identical regularisation term but replacing the L1 norm with the L2 norm ($|\beta|_2 = \sqrt{\sum \beta_i^2}$) does not produce sparse models, but rather models that do not contain large coefficients. The corresponding structural assumption for this is slightly more technical (we can assert a multivariate Gaussian prior on the parameter space for β). This can be applied in a wide variety of high-dimensional situations, often alongside other forms of regularisation, as a combatant to over-fitting (typically via cross-validation).

Technique 3.2. (Regularisation for Dimension: L2/Ridge Regression): L2 regularised estimation (known as ridge regression in the linear setting [19]) selects β solving the following optimisation

$$\min_{\beta \in \mathbb{R}^p} \left\{ \sum_{i=1}^{n} \mathcal{L}\left(\mathcal{M}_\beta\right) + \lambda|\beta|_2 \right\}$$

where again λ is a positive value that can be selected by cross-validation to reduce overfitting.

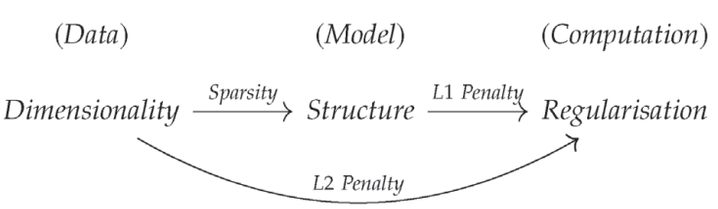

Figure 1.
An example of a high-dimensional workflow, where high dimensionality is addressed via the imposition of model structure, in this case sparsity. This is translated into a computationally tractable extension of standard regression model fitting via an L1 penalty. Dimension-induced overfitting is simultaneously managed via L2 regularisation. If sparsity is a reasonable structural assumption, that is few covariates have genuine impact, L2 regularisation should have a relatively small impact.

Figure 1 describes the workflow of modelling high-dimensional data. The data dimensionality, as discussed in the previous chapter, is the underlying problem, which we address with structural assumptions informed from a mixture of external knowledge and practicality, which are then transformed into a feasible computational problem. Intuition around the biological and also statistical context are applied at each step.

For those unsatisfied with the abstract nature of the discussion above, we now attempt to provide more concrete examples.

4. Cancer genomics questions in the language of high-dimensional statistics

4.1 Biomarker/driver gene identification

We have discussed some of the terminology associated with high-dimensional statistics. We can now express some cancer genomics questions in the same language. We have data with a very high dimensionality p: bases, codons or genes ($p \approx 3 \times 10^9$, 1×10^9 and 2×10^4 respectively) and we would like to predict some outcome, be it a survival value, biomarker signature or other phenotype. Due to the resources and time required to perform whole genome or exome sequencing we often face restrictions in the number of samples at our disposal. The popular Cancer Genome Atlas resource [5], for example, contains sequencing data for around 20,000 tumour/normal matched samples. Even if all of these samples were relevant to our study, and we were trying to predict some phenotype Y using gene-level data, we would be working in the $p \approx n$ regime. If we were using codon or nucleotide level information, we would be well into the $p >> n$ regime. In the following we will assume we are working with some gene-level covariates, and investigate what sort of structural assumptions we may wish to make in order to fit tractable and robust models.

4.2 Sparsity by assumption: driver genes

Driver genes in the simplest sense are genes that, when mutated, will elevate risk of the development, progression or adaptation of a tumour [20]. They may be grouped roughly into oncogenes and tumour suppressors: oncogenes admit mutations giving some selective advantage to a cancer cell, while tumour suppressors in their standard form protect against aberrant cell growth or apoptosis evasion. Identifying driver genes (or driver sites within genes) among the extensive backdrop mutation in tumours is notoriously difficult. Selection pressures produce subtle and often non-obvious patterns of mutation density between neutral and non-neutral genes as well as distinct signatures for oncogenes and tumour suppressors [21]. Neglecting these difficulties for now, suppose we wish to infer some phenotype Y (again for simplicity we assume that this is continuous and single-valued). We do not have nearly enough data to fully explore the dependence of Y on all genes simultaneously—we have to assume that there are *relatively few relevant features/driver genes*. This is exactly a sparsity assumption—a regularisation method such as LASSO might be helpful. The advantages of this are twofold. We have identified a set of genes of interest, which might form the basis for some targeted prognostic panel, while simultaneously inferring a predictive structure on top of this list of genes. The added interpretability of our model given by assuming a structural restraint is useful when verifying our results in the lab. We have produced a manageable set of interesting genes that can be investigated on a more detailed individual basis.

4.3 Sparsity by necessity: gene panels for genome-wide biomarkers

Another justification for selecting some small set of genes/genomic loci to include in an investigative panel is that the cost and time to perform sequencing depends (approximately linearly) on the size of the subsection of the genome to be sequenced, and the depth at which it is sequenced. This means that in many practical or clinical environments, cost is a major factor. While the cost of whole genome sequencing has decreased at an impressive rate, it is far from being standard of care for cancer patients. It is therefore important that gene-panel style biomarkers are as small as possible, while maintaining enough accuracy that clinicians feel confident in acting upon predictions. This is a particular issue for genome-wide biomarkers, which have gained popularity in recent years, for example in cancer immunotherapy. Examples include tumour mutation burden [22] and indel burden [23], which report density of somatic mutation across the entire cancer genome. In this case all regions of the genome are relevant to greater or lesser extent (**Figure 2**)—the optimal panel for prediction would be the entire genome (or exome, depending on the specific biomarker). However, certain genes may be particularly relevant, for example by taking an active role in DNA repair mechanisms. When estimating such biomarkers, we therefore want to offset the positive predictive contributions of individual genes/loci against the added cost burden given by inclusion in the panel. Analyses of the impact of panel size on predictive power in theoretical and practical settings are becoming more common [24].

Suppose we have some set G of genes, where g refers to an individual gene with coding sequence of length n_g. Now let $P \subset G$ refer to a gene panel comprising a set of genes, and $\mathcal{M}_P$ be a model trained on some data with covariates included according to the gene panel P. Then we might wish to solve the optimisation problem

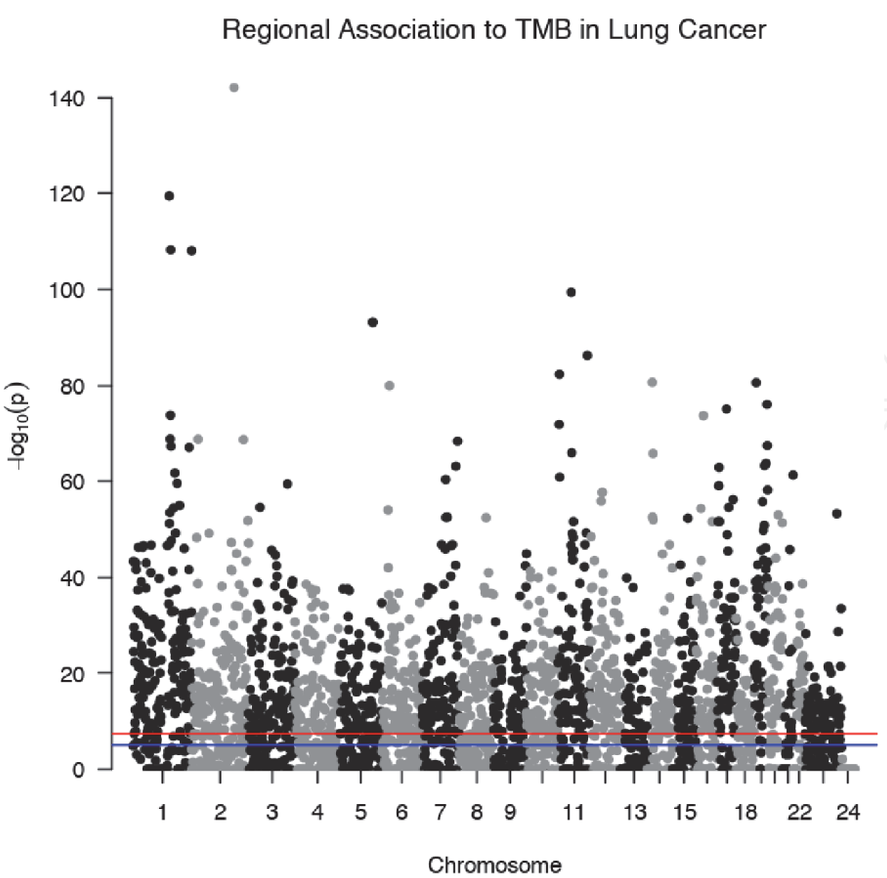

Figure 2.
For an additive, genome-wide biomarker such as TMB (tumour mutational burden), all genomic loci are significantly correlated with TMB (unlike in typical GWAS studies). How do we choose a subset that is not prohibitively large but can reliably estimate the marker via some predictive model?

$$\min_{P \subset G} \{\mathcal{L}(\mathcal{M}_P)\} \text{ such that } |P| \leq L, \tag{4}$$

where $\mathcal{L}(\mathcal{M})$ is the loss of the model $\mathcal{M}$, $|P| = \sum_{g \in P} n_g$ is the total length of the gene panel P and L is some prescribed maximum panel length. Note the similarity with the LASSO setup described in Section 3.1. In the case of a linear model we can similarly reformulate the problem in terms of the parameter β, and solve the analogous problem.

Technique 4.1. (Weighted L1 Regularisation/LASSO): Here we select β satisfying the optimisation problem

$$\min_{\beta \in \mathbb{R}^{|G|}} \left\{ \mathcal{L}(\mathcal{M}_\beta) + \lambda \sum_{g \in G} n_g |\beta_g| \right\}$$

where we have again swapped the panel length bound L for the regularisation parameter λ. Since all the n_g values are positive, this is still a convex optimisation problem and thus can be solved efficiency as in the standard case. Choice of λ is less likely to be chosen via cross-fitting, as smaller values of λ will always improve predictive power. Instead λ will be chosen to control the size of the resulting gene panel.

4.3.1 Distinguishing causative mutations

It should again be noted that these are illustrations of how high-dimensional model construction is done. In reality many more subtleties may have to be taken into account. In the above a key caveat requiring understanding is the role of selective pressure in cancer-relevant genes [25], and how this affects the mutation rate in different sections of the genome [26]. One way this can be investigated is by looking at the relative predictive power of synonymous and non-synonymous mutations for genome-wide mutation burden [27]. The gold standard for identifying causative relationships between genotype and phenotype, however, remains with functional validation studies.

4.4 Survival prediction

No review of statistical learning in cancer genomics would be complete without a mention of survival prediction. Survival prediction is useful in a variety of situations, far beyond direct prognostic application. Hazard regression models based on genomic data have been useful in identifying therapeutic resistance [28] or general prognosis [29, 30] factors, which are of great interest to those developing drugs or attempting to understand which patients can expect to benefit from them. Regularisation-based techniques are perfectly adaptable to proportional-hazards style models [31], to which end there has much literature beyond what we have scope to discuss in this chapter.

5. Modern techniques in high-dimensional statistics and dimensionality reduction

We conclude with some examples from recent literature of techniques related to dimensionality reduction in modelling genomic data. The examples have been chosen to demonstrate the structure/regularisation workflow discussed in this chapter, and are small a set of examples rather than (anywhere near) an exhaustive list.

5.1 Regularised graphical models

In the regression examples discussed previously, the parameters of interest have represented the weighted effect of observed covariates on a label. In supervised and unsupervised cases, we are also often interesting in looking at how closely related different covariates are, through estimating the correlation matrix of the observation variable X. If we have an observation of dimensionality p, then the covariance matrix will be of size p^2, so problems of estimation from small n are even more confounded!

Two forms of regularisation are popular, often used in tandem. The first is a sparsity penalty applied to all matrix entries [32]. What does this correspond to structurally? It means that that most pairs of covariates are independent (or at least uncorrelated). This is a very relevant notion in network analysis, where variables are thought to affect each other in a way that can be described by some graphical structure. Sparsity of matrix elements then corresponds to sparsity of the graph describing the network. It is also not uncommon to sparsely penalise precision, defined by the components of the inverse covariance matrix [33].

Alternately (or in addition), we may wish to limit the number of distinct *patterns* of correlation, so that all covariates display a correlation profile that is made up of a combination of a relatively small set of base signatures. This structure may be fitted for by imposing rank-based regularisation [34, 35]. For those wanting a greater appreciation of the theory, the way this is imposed is another good example of convex penalty relaxation (as was achieved by switching from the L_0 to L_1 norm in sparsity regularisation), where here the nuclear norm is used as the convex relaxation of matrix rank.

5.2 Localised sparsity assumptions

We have made an extensive discussion of sparse models in this chapter. We might wonder if there are any generalisations to the assumption that relatively few of our covariates are important throughout all of our samples. One such generalisation would be that for some subsets of our samples sparsity assumptions hold, but that the important covariates may differ from subset to subset within our data. In a localised sparsity setting, we are often given some knowledge of the organisational structure of data, either in a discrete way through a prior partition of the samples or network structure, or in a continuous way through a measure of distance between samples (which may come directly from the input data). We can then fit linear models that are regularised towards sparsity, but where variable selection is allowed to vary between samples, and allowed to vary more between samples that are more distant. This has been applied to the prediction of drug toxicity based on differential gene expression data [36].

5.3 Variational autoencoders

For our final example we consider a notion of dimensionality reduction that is more general and that has been studied extensively in the machine learning literature. This nicely elucidates the grey border between statistical and machine learning, and the difficulties and opportunities available to biological research by embracing the latter.

Variational autoencoders (VAEs) are a class of neural networks with a variety of architectures and sizes, but whose premise centres around producing an encoding/decoding framework between high-dimensional data and a lower dimensional

representation [37]. VAEs have an 'hourglass' shape: input data is fed into the network, and information is propagated through layers of progressively smaller size until a bottleneck is reached. The central layer will have some small number of latent nodes. Subsequent layers increase in size, reaching an output of dimension matching the input. VAEs are trained to reproduce the inputs with which they are trained as accurately as possible. We can then view the central latent nodes as an encoding of our input data [38]. This might (a) contain some insightful information and (b) be useful as lower dimensional input data for training other models.

In the context of cancer genomics [39], VAEs pose two challenges, illustrative of those that machine learning procedures in general must overcome to be useful in a basic research or clinical setting. Firstly, they are highly parameterised compared to the types of model discussed so far. We have discussed at length the balance between data availability and model size, and the significant extra effort necessary to extract information when information is scarce. One of the advantages of deep learning procedures is their versatility and lack of dependence on prior knowledge and assumptions of structure. The cost is that they are very data intensive, prohibitively so in some cases. Secondly, while a VAE's latent nodes may be informative within a network, there is no necessary guarantee that they will be interpretable by a human, nor that biologically relevant features will have been neatly allocated to a single node. Strategies to 'untangle' VAEs are necessary to make biologically relevant predictions [40].

6. Conclusions

The dimensionality of data in genomics is a sticking point that at its full potency is more debilitating than in any other research discipline [9]. Even at the current pace of increase of the availability of sequencing data, it will be a long time away (if ever) that the most powerful and general machine learning techniques will be at our disposal without recourse to the vast wealth of biological knowledge we as a species have accumulated. To properly use that knowledge, we need researchers who are able to speak the language of both camps. It is not sufficient that researchers in cancer genomics provide data and questions to researchers in machine learning, nor that machine learning researchers communicate back the output of their methods. Instead, methods need to be crafted bespokely by those who understand what features of cancer data are relevant, how those features manifest themselves and how to exploit them in a mathematically consistent way.

This entire workflow is quite easy to follow when the sort of structure we are insisting upon in our models is very simple. Even when a structural assumption can be motivated in a single sentence (see Definition 3.1), and a model is simple (such as in linear regression), a good design of learning procedure might not be immediately obvious. It can likely, however, be given a fairly ground-up description within a single book chapter. When the structural assumptions we really want to incorporate might well extend as far as our current appreciation of the mutational processes affecting tumours across heterogenuous cell populations, chromosomes, genes and codons, and the models we want to fit are similarly at the cutting edge of computational research, then the position of an interdisciplinary researcher may well require far more legwork to maintain.

As motivation for the above legwork, it should go without saying that cancer genomics in the machine learning age has potential to do a great deal of good in the long term. Yet uncovering a deeper understanding of how cancer works is not the only worthwhile goal. Designing procedures that can work *now* to be more effective, sometimes crossing a threshold between non-pracitcality and practicality (in

some part of the world), can have a more immediate benefit. In the clinic, the time scale and cost of data collection are not abstract mathematical problems, so designing a test that works with less data can be just as enabling as uncovering a new paradigm of cancer progression.

Acknowledgements

Many thanks to Timothy Cannings and Belle Taylor for their support and advice, to John Cassidy for suggestions of improvements, to Steven Bradley for proofreading and providing a non-technical reader's viewpoint, and to Morton for his invaluable contributions.

Conflict of interest

The author declares no conflict of interests.

Author details

Jacob Bradley[1,2]

1 School of Mathematics, University of Edinburgh, UK

2 Cambridge Cancer Genomics, Cambridge, UK

*Address all correspondence to: j.r.j.bradley@ed.ac.uk.com

References

[1] Lander E, Chen C, Linton L, Birren B, Nusbaum C, Zody M, et al. Initial sequencing and analysis of the human genome. Nature. 2001;**409**:860-921

[2] Sboner A, Xinmeng M, Greenbaum D, Auerbach R, Gerstein M. The real cost of sequencing: Higher than you think! Genome Biology. 2011;**12**:125

[3] Prokop J, May T, Strong K, Bilinovich S, Bupp C, Rajasekaran S, et al. Genome sequencing in the clinic: The past, present, and future of genomic medicine. Physiological Genomics. 2018;**50**:563-579

[4] Telenti A, Pierce L, Biggs W, di Iulio J, Wong E, Fabani M, et al. Deep sequencing of 10,000 human genomes. Proceedings of the National Academy of Sciences. 2016;**113**:11901-11906

[5] Weinstein JN, Collisson EA, Mills GB. The cancer genome atlas pan-cancer analysis project. Nature Genetics. 2013;**45**(10):1113-1120

[6] Raja R, Lee Y, Streicher K, Conway J, Wu S, Sridhar S, et al. Integrating genomics into drug discovery and development: Challenges and aspirations. Pharmaceutical Medicine. 2017;**31**: 217-233

[7] Weber B, Hager H, Sorensen B, Mcculloch T, Mellemgaard A, Khalil A, et al. EGFR mutation frequency and effectiveness of erlotinib: A prospective observational study in Danish patients with non-small cell lung cancer. Lung Cancer. 2013;**83**:224-230

[8] Awad K, Dalby M, Cree I, Challoner B, Ghosh S, Thurston D. The precision medicine approach to cancer therapy: Part 1 solid tumours. The Pharmaceutical Journal. 2019;**303**

[9] Barbour D. Precision medicine and the cursed dimensions. npj Digital Medicine. 2019;**2**. Article no. 4

[10] Martin W. High-Dimensional Statistics: A Non-Asymptotic Viewpoint. Cambridge, UK: Cambridge University Press; 2019

[11] Liu Q, Zhang N, Yang W, Wang S, Cui Z, Chen X, et al. A review of image recognition with deep convolutional neural network. In: Intelligent Computing Theories and Application. Proceedings of the 13th Intenrational Conference of Intelligent Computing. 2017. pp. 69-80

[12] Gutirrez L, Norambuena BK. A systematic literature review on word embeddings. In: Proceedings of the 7th International Conference on Software Process Improvement (CIMPS 2018). 2019. pp. 132-141

[13] Kussad Y, Kirkham D, Cassidy J, Patel N, Clifford H. Flatsomatic: A method for compression of somatic mutation profiles in cancer. 2019. Available from: https://arxiv.org/abs/1911.13259

[14] Kussad Y, Kirkham D, Cassidy J, Patel N, Clifford H. Learning embeddings from cancer mutation sets for classification tasks. 2019. Available from: https://arxiv.org/abs/1911.09008

[15] Bhlmann P, Kalisch M, Meier L. High-dimensional statistics with a view toward applications in biology. Annual Review of Statistics and Its Application. 2014;**1**:255-278

[16] Chalmers ZR, Connelly CF, Fabrizio D, Gay L, Ali SM, Ennis R, et al. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. Genome Medicine. 2017;**9**(1):34

[17] Omidiran D, Wainwright M. High-dimensional variable selection with sparse random projections: Measurement sparsity and statistical

efficiency. Journal of Machine Learning Research. 2010;**11**:2361-2386

[18] Cannings TI, Samworth RJ. Random-projection ensemble classification. Journal of the Royal Statistical Society, Series B: Statistical Methodology. 2017;**79**(4):959-1035

[19] Tibshirani R. Regression shrinkage selection via the lasso. Journal of the Royal Statistical Society, Series B. 2011; **73**:273-282

[20] Hanahan D, Weinberg R. The hallmarks of cancer. Cell. 2000;**100**: 57-70

[21] Brown A-L, Li M, Goncearenco A, Panchenko AR. Finding driver mutations in cancer: Elucidating the role of background mutational processes. PLoS Computational Biology. 2019; **15**(4):1-25

[22] Samstein R, Lee C-H, Shoushtari A, Hellmann M, Shen R, Janjigian Y, et al. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. Nature Genetics. 2019;**51**:02

[23] Tajlic S, Litchfield K, Xu H. Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: A pan-cancer analysis. The Lancet Oncology. July 2017;**18**:1009-1021

[24] Budczies J, Allguer M, Litchfield K, Rempel E, Christopoulos P, Kazdal D, et al. Optimizing panel-based tumor mutational burden (TMB) measurement. Annals of Oncology. 2019;**30**(9):1496-1506

[25] Bull K, Rimmer A, Siggs O, Miosge L, Roots C, Enders A, et al. Unlocking the bottleneck in forward genetics using whole-genome sequencing and identity by descent to isolate causative mutations. PLoS Genetics. 2013;**9**:e1003219

[26] Iengar P. Identifying pathways affected by cancer mutations. Genomics. 2017;**110**:12

[27] Chu D, Wei L. Nonsynonymous, synonymous and nonsense mutations in human cancer-related genes undergo stronger purifying selections than expectation. BMC Cancer. 2019;**19**:12

[28] Seagle B-L, Eng K, Yeh J, Dandapani M, Schultz E, Samuelson R, et al. Discovery of candidate tumor biomarkers for treatment with intraperitoneal chemotherapy for ovarian cancer. Scientific Reports. 2016; **6**:21591

[29] Zhang Y, Li H, Zhang W, Che Y, Bai W, Huang G. Lassobased coxph model identifies an 11lncrna signature for prognosis prediction in gastric cancer. Molecular Medicine Reports. 2018;**18**:10

[30] Guinney J, Wang T, Laajala TD. Prediction of overall survival for patients with metastatic castration-resistant prostate cancer: Development of a prognostic model through a crowdsourced challenge with open clinical trial data. The Lancet Oncology. 2017;**18**(1):132-142

[31] Benner A, Zucknick M, Hielscher T, Ittrich C, Mansmann U. High-dimensional cox models: The choice of penalty as part of the model building process. Biometrical Journal. 2010; **52**(1):50-69

[32] Witten DM, Tibshirani R. Covariance-regularized regression and classification for high dimensional problems. Journal of the Royal Statistical Society, Series B: Statistical Methodology. 2009;**71**(3):615-636

[33] Lin X, Huang X, Wang G, Tao W. Positive-definite sparse precision matrix estimation. Advances in Pure Mathematics. 2017;**07**:21-30

[34] Hu Z, Nie F, Tian L, Li X. A comprehensive survey for low rank regularization. Computing Research Repository. 2018. Available from: https://arxiv.org/abs/180804521

[35] Ye G, Tang M, Cai J-F, Nie Q, Xie X. Low-rank regularization for learning gene expression programs. PLoS One. 2013;**8**(12):1-9

[36] Yamada M, Takeuchi K, Iwata T, Shawe-Taylor J, Kaski S. Localized lasso for high-dimensional regression. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. Fort Laud-erdale, FL; 2017

[37] Diederik PK, Max W. Auto-encoding variational bayes. In: Proceedings of International Conference on Learning Representations. Scottsdale; 2013

[38] Zheng H, Yao J, Zhang Y, Tsang I, Wang J. Understanding vaes in fisher-shannon plane. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. 2019. pp. 5917-5924

[39] Way G, Greene C. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. Pacific Symposium on Biocomputing. 2018;**23**:80-91

[40] Kompa B, Coker B. Learning a latent space of highly multidimensional cancer data. Pacific Symposium on Biocomputing. 2020;**25**:379-390