Maximum entropy models in the analysis of genome-wide data in cancer research

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Châu Nguyễn

aus Namdinh, Vietnam

Köln, 2014

Berichterstatter: Prof. Johannes Berg

Prof. Joachim Krug

Prof. Leonid Mirny

Tag der letzten mündlichen Prüfung: 16.01.2015

Zusammenfassung

Diese Arbeit beschäftigt sich mit der Maximum-Entropie-Methode im Zusammenhang statistischer Modellierung. Anwendungsbeispiele sind dem aufkommenden Forschungsbereich der Krebsgenomik entnommen.

Wir beginnen mit einer kurzen Einführung in die Biologie von Krebserkrankungen in Kapitel 1. In Kapitel 2 diskutieren wir die Grundlagen der statistischen Modellierung, wobei wir eingehend die Maximum-Entropie-Methode besprechen. Insbesondere zeigen wir, dass viele statistische Modelle in einen einheitlichen Rahmen, basierend auf der Maximum-Entropie-Methode, gebracht werden können, der diese auf Probleme der statistischen Mechanik abbildet. In Kapitel 3 befassen wir uns mit einem bestimmten Maximum-Entropie-Modell, dem Ising-Modell, im Kontext des inversen Ising-Problems. Wir führen eine Bethe–Peierls-Näherung für das inverse Ising-Problem ein. Des Weiteren schlagen wir eine modifizierte Version der Molekularfeld-Näherung vor, welche auch für niedrige Temperaturen funktioniert. Die folgenden Kapitel wenden Maximum-Entropie-Modelle auf verschiedene Probleme aus dem Bereich der Krebsgenomik an. Eine direkte Anwendung des inversen Ising-Problems auf Daten zur Anzahl von Genkopien in Krebszellen wird in Kapitel 4 beschrieben. In Kapitel 5 erweitern wir die Konzepte der indirekten Korrelationen und der direkten Kopplungen des inversen Ising-Problems um den Einfluss der Anzahl von Genkopien auf die Expression von Genen in Krebszellen zu untersuchen. Wir zeigen, dass die Korrelationen in der Genexpression nicht unbedingt durch regulatorische Wechselwirkung zwischen Genen hervorgerufen werden müssen. Stattdessen können die Korrelationen in der Genexpression durch die Korrelationen in der Anzahl von Genkopien hervorgerufen werden, was auf der geometrischen Organisation des Genoms beruht. Wir zeigen, dass ein einfaches Maximum-Entropie Modell die Korrelationen in der Anzahl von Genkopien von den sogenannten “blanken Korrelationen” in den Genaktivitäten, welche nur den Effekt der regulatorischen Wechselwirkungen beschreiben, trennen kann. Kapitel 6 ist der Klassifizierung von Krebs gewidmet. Wir führen einen einfachen semi-überwachten Lernalgorithmus ein um eine Mischung aus paramagnetischen Modellen mit Ising-Spins dahingehend zu trainieren, Krebsmutationsprofile zu klassifizieren. Wir zeigen, dass dieser Lernalgorithmus, mit der Möglichkeit sowohl von den nicht zugeordneten Proben zu

lernen als auch eine falsche Zuordnung von Proben zu korrigieren, sowohl die überwachten als auch die unüberwachten Lernalgorithmen übertrifft. Die zwei Anhänge A und B fassen die jüngeren Studien über die Sensibilität und die Widerstandsfähigkeit von Krebszellen gegenüber Therapien zusammen.

Die Ergebnisse von Kapitel 3 wurden in H. C. Nguyen and J. Berg (2012a). "Bethe–Peierls approximation and the inverse Ising problem". *J. Stat. Mech.* P03004; and H. C. Nguyen and J. Berg (2012b). "Mean-field theory for the inverse Ising problem at low temperatures". *Phys. Rev. Lett.* 109, p. 50602 publiziert. Einige der Resultate aus Kapitel 6 wurden als Teil von The Clinical Lung Cancer Genome Project (CLCGP) and Network Genomic Medicine (NGM) (2013). "A genomics-based classification of human lung tumors". *Science Transl. Med.* 5.209, 209ra153 publiziert.

Abstract

This thesis studies the maximum entropy principle in statistical modelling. Applications are taken from the emerging field of cancer genomics.

We start with a short introduction to the biology of cancer in chapter 1. In chapter 2, we discuss general principles of statistical modelling. We discuss in detail the maximum entropy principle in statistical modelling. In particular, we show that many statistical models can be put in a unified framework based on the principle of maximum entropy, which maps them into problems of statistical mechanics. In chapter 3, we consider a particular maximum entropy model, the Ising model, in the context of the inverse Ising problem. We introduce a Bethe–Peierls approximation to the inverse Ising problem. We then also suggest a modification for the mean-field approximation to work at low temperatures. The following chapters apply maximum entropy models to different problems of cancer genomics. A direct application of the inverse Ising problem to gene copy-number data of cancer cells is described in chapter 4. In chapter 5, we extend the concepts of indirect correlations and direct couplings of the inverse Ising problem to investigate the influence of gene copy-numbers on gene expressions in cancer cells. We show that the correlations in gene expression need not be due to regulatory interactions between genes. Instead, correlations in gene expression of cancer cells can be induced by the correlations in their copy-numbers, which is due to the geometrical organisation of the genome. We show that a simple maximum entropy-model can disentangle copy-number-induced correlations and the so-called "bare-correlations" in gene expression, which capture the effect of regulatory interactions alone. Chapter 6 is devoted to cancer classification. We introduce a simple semi-supervised learning algorithm to train a mixture of paramagnetic models with Ising spins to classify cancer mutation profiles. We show that, with the capability of both learning from unlabelled samples and correcting mislabelled samples, this learning algorithm outperforms both the supervised and unsupervised learning algorithms. The two appendices A and B summarise recent studies on sensitivity and resistance of cancer cells to therapy.

The results of chapter 3 were published in H. C. Nguyen and J. Berg (2012a). "Bethe–Peierls approximation and the inverse Ising problem". *J. Stat. Mech.* P03004; and H. C.

Nguyen and J. Berg (2012b). "Mean-field theory for the inverse Ising problem at low temperatures". *Phys. Rev. Lett.* 109, p. 50602. Some results of chapter 6 were published as a part of The Clinical Lung Cancer Genome Project (CLCGP) and Network Genomic Medicine (NGM) (2013). "A genomics-based classification of human lung tumors". *Science Transl. Med.* 5.209, 209ra153.

Acknowledgements

In writing these lines, I am very grateful to my supervisor Johannes Berg. Over the years, I have learned from him many things, from how to pose a scientific problem to how to make scientific figures. I have also learned to be patient and to be humorous. Looking back, he has listened to my ideas however innocent they were. He has been very patient with my mistakes (sometimes repeatedly). His sense of humour has expanded in our group and has relaxed the stress of Ph.D. students.

Together with Johannes Berg, my mentors, Prof. Joachim Krug and Prof. Hartmut Monien also had encouraging advice. Comments and advice from my former supervisors, Prof. Markus Müller and Prof. V. Lien Nguyen, have been always very helpful. I also thank Prasanna, Nico, Nina, Nhung, Van-Anh for discussions and comments on different parts of this thesis. Nico also helped me with translating the abstract.

Our collaboration with Roman Thomas was very enjoyable. I have learned much about clinical biology through discussions with him and people in his group, Martin, Julie, Danila, Sandra, any many others. The contribution from the patients in the Clinical Lung Cancer Genome Project (CLCGP) and Network Genomic Medicine (NGM), who have contributed their data, played a vital role in my thesis. I consider my thesis as dedicated to them.

My friends in our group, Nico, Prasanna, Simon, Nina, Nandita, Filippos, Joachim, Bhavin have made my time in Cologne memorable. Our scientific and non-scientific chats during lunches and dinners, together with Michael Lässig and members of his group, Mara, Simone, Christian, Daniel, Torsten, Armita, Donate, Stephan and Stephane were very joyful. I am thankful to Frau Christa Stitz for her enthusiastic help, which saved me from a lot of bureaucratic problems.

The financial support from the DFG through the SFB-680 project, the emed-SMOOSE project and the Bonn–Cologne Graduate School are acknowledged.

Contents

Chapter 1

Introduction to the biology of cancer

More is different.
P. W. Anderson

In this chapter, we discuss some aspects of cancer biology to set the basis for the quantitative discussions in the following chapters. We keep the contents at the minimal level and more detailed discussions are given as separate additional remarks so that interested readers may have quick references. We close the chapter with a brief introduction to several data types emerging in cancer genomics research that are most relevant to our analysis throughout this thesis.

1.1 The human body homeostasis

In many aspects, a human body can be regarded as a spatial system of some trillions of cells of different types living in a common environment. These cells are organised in tissues, tissues make organs and organs make systems; all are working in harmony to maintain the body as a whole.

Remark: ***Cells in the human body***

Cells are the building blocks of the human body, and arguably of all *organisms*. A cell is surrounded by the *cell membrane*, which separates the intracellular space from the extracellular space. Inside the intracellular space is the *cytoplasm*, where billions of biochemical processes are taking place at any time.

Proteins are the main elements of all biochemical reactions
Proteins are sequences of amino acids folded in specific spatial conformations. There are millions of them in a cell. Proteins can interact with each other, changing their chemical details or their conformations, or forming protein complexes. Virtually all biological processes in a cell are performed by proteins and protein complexes: they break up nutrient molecules for energy;

they carry other substances to places where they are needed; they also catalyse the synthesis of fundamental materials for the cell, including the proteins themselves.

Genes are the templates for protein synthesis

Proteins are not very stable molecules; old proteins are degraded, and new proteins are constantly synthesised. A protein is synthesised according to a very stable template – the *genes*. Genes are specific sequences of nucleotides lying on a great polymer chain, the DNA molecule (deoxyribonucleic acid), which (in eukaryotes such as humans) are protected in the nucleus of the cell, separated from the cytoplasm by the nuclear membrane.

A DNA molecule is a double-stranded sequence of four *nucleotides*, or *bases*, namely adenine (A), thymine (T), guanine (G) and cytosine (C). Along a single strand, the nucleotides are linked via covalent bonds and the two strands are bound by hydrogen bonds complementarily, A-T, G-C. In normal conditions, together with some special proteins, the double stranded DNA molecules are twisted and folded multiple times in a very condensed structure, appearing as *chromosomes*, which were first observed by biologists of the eighteenth century. Chromosomes appear in pairs. There are 23 pairs of chromosomes in the nucleus of a human cell, known as the *human genome*.

Protein synthesis

The sequence of a gene on the DNA molecule is first copied to a shorter *messenger RNA* sequence via special proteins – the *RNA polymerases* – in a process known as *transcription*. Similar to DNA, a messenger RNA is also a sequence of the four nucleotides, but with thymine (T) replaced by uracil (U). Messenger RNA are single-stranded, much shorter and more dynamic than DNA. After transcription, the messenger RNA sequences are further processed and transported out of the nucleus to cellular organelles called ribosomes in the cytoplasm. At the ribosomes, they are *translated* into the amino acid sequences, where every three nucleotides (a *codon*) make an amino acid. The amino acid sequences then fold into a specific spatial conformations, giving rise to mature proteins. When a gene is actively making proteins, we say the gene is *expressed* or *active*, otherwise it is said to be *silenced* or *inactive*.

Gene expressions are controlled by *gene regulation mechanisms* at multiple levels. The expression of gene A (i.e., protein A) can seek and bind specifically to a short sequence before the starting point of gene B on the DNA, blocking (or promoting) the RNA polymerases from transcribing the gene. This is called *transcriptional regulation*. Alternatively, a protein can degrade the mRNA sequence of another gene or modify another protein in a way to deactivate or activate its function; the former mechanism is called *post-transcriptional regulation* and the latter mechanism is called *post-translational regulation*.

Cell stability and flexibility

Protein reactions, including those with DNA and with other substrates in the cell plasma form a complex network of biochemical reactions. The network has multiple alternative pathways, complicated feedback and feedforward pathways, which aim to give the cell both stability and flexibility. Stability means that the network is stable under small disturbances or fluctuations from inside the cell (*intrinsic fluctuations*) or from the environment outside (*extrinsic fluctuations*). Flexibility means the network can change to an other state with other active genes in a necessary response to large disturbances. Importantly, in a multicellular organism, cell flexibility also implies the ability to *response* to external signals from the other cells of the body. An external signal is usually a protein (*factor*) generated by some specific cell in the body that binds onto a protein on the surface of another cells (*receptor*). Upon the binding of the factor, the receptor changes its

characters and ignites a cascade of biochemical reactions into the cells (called *signalling pathway*), which changes the expression profile and therefore the biological state of the cell. We will see that such signalling pathways are of great importance for the integration of a multicellular organism.

Cells of the body are embedded in a slowly flowing *extracellular fluid*, which is connected throughout the body, forming the living environment of the cells. The concentrations of small dissolved molecules in the extracellular fluid are highly homogeneous and stable. The narrow ranges of allowed variations of the concentrations of different elements in the extracellular fluid in table 1.1 illustrate this stability.

	normal value	normal range	short-term nonlethal limit	unit
Oxygen	40	35–40	10–1000	mm Hg
Carbon dioxide	40	35–45	5–80	mm Hg
Sodium ion	142	138–146	115–175	mmol/l
Potassium ion	4.2	3.8–5.0	1.5–9.0	mmol/l
Calcium ion	1.2	1.0–1.4	0.5–2.0	mmol/l
Chloride ion	108	103–112	70–130	mmol/l
Bicarbonate ion	28	24–32	8–45	mmol/l
Glucose	85	75–95	20-1500	mg/dl
Body temperature	37.0	37.0	18.3–43.3	°C
Acid-base	7.4	7.3–7.5	6.9–8.0	pH

Table 1.1: The composition of the extracellular fluid as from Guyton and Hall 2006.

The stability of the extracellular fluid is maintained by the cells themselves through the activities of all organs and systems of the body. The circulating system keeps the extracellular fluid in flowing state; the respiratory system regulates oxygen and carbon dioxide concentrations; the kidneys control the water volume of the body and the ion concentrations; etc. Either directly or indirectly, virtually every cell contributes to maintaining constant living conditions for the whole population of cells, ensuring the stability of the tissues and therefore the whole body.

The cell populations in tissues normally remain in local equilibrium, but this equilibrium is very dynamic. Cells age and die; other cells continuously divide giving rise to new cells to replace the old ones. The cell renewal process needs to be very strictly regulated to ensure the tissue equilibrium. Cells can only grow and divide when they receive signals to grow and divide. They have to die when they receive signals to die. Even the ways they die is carefully controlled and organised, which are usually very soft deaths, avoiding any harm to the other cells (*apoptosis*). These control signals are often the responses of the body to the deviations of the cell population from equilibrium.

Remark: *Tissue architectures and cell renewal*

Tissues usually consist of cells of several types. Cells can be glued together by direct contacts via proteins hanging in their membrane. They can also secrete special proteins (collagens) to form a network of filaments called *extracellular matrix* which mechanically supports the tissues. Tissue architectures are classified into two classes: *epithelial* and *connective*.

Epithelial architecture and connective architecture

Epithelial architecture appears in organ surfaces. Skin, gut lumen and lung squamous are examples of epithelial architecture. They are usually thin and dense layers of cells. The cell walls are directly attached to each other via transmembrane proteins. In the upper layer, cells usually have enzyme production activities. At the base of epithelial tissues, cell walls have *integrins*, which go through the cell membrane and connect to the *cytoskeleton* (which are the "bones" that support the whole mechanical shape of the cells) at one end and attach to the extracellular matrix of the connective architecture lying below.

Below the epithelial layer is the connective architecture, which consists largely of extracellular matrix with cells scattering among. The extracellular matrix generates the tensile forces of the tissue. The space in the network is filled with proteins and polysaccharide, which provide compression resistance. The cells are attached to the collagen filaments of the extracellular matrix via integrins.

Cells in the connective architectures are of very different types. The endothelial cells and pericytes form blood vessels which bring nutrients and oxygen in, carbon dioxide and waste products out of the tissue, maintaining the local stability of the extracellular fluid. Nerve cells provide a way of sensing and feedback control on the tissue from the nervous system. Cells from the immune system protect the tissue from invasion by aliens such as bacteria, or digest old dying non-function cells to clean up the tissue environment.

Cell renewal

Cells age and die. Aged cells are usually prone to *apoptosis*, where a special series of reactions is activated resulting in cell destruction without releasing toxic proteins that are harmful to the tissue. Dying cells or their debris are also taken up by the immune system. Dying cells of the epithelial layers such as the skin or the gut lumen are usually directly detached and washed out.

New cells to replace the dying ones come about by cell divisions. Interestingly, not all cells in a multicellular organism can divide. Cells are organised into classes of *differentiated cells* and *stem cells*. Differentiated cells are functional, but they do not divide. The population of these differentiated cells is relied on the divisions of stem cells. A stem cell can perform a *symmetric division*, giving rise to two new stem cells, or an *asymmetric division*, giving rise to one stem cell and one differentiated cell. In reality, one type of stem cells can often differentiate into different types of differentiated cells. The differentiation process is also usually multi-step with cells of different levels of "stemness".

To maintain the homeostasis of the population, cell renewals are strictly regulated. Cells can only divide when they received signals that allow them to divide; new cell can only survive when they received signals that allow them to survive. There are also inhibitors that block the cell divisions or induce cell deaths. Cell renewal is strictly controlled by the balance of many positive and negative control signals.

The dynamical equilibrium of the body is usually known as the *human body homeostasis*.

Depending on the context, the homeostasis may refer to the stability of the extracellular fluid, or the equilibrium of the cell population. On the other hand, these are ultimately two sides of the same problem. In any case, a disturbance in the human body homeostasis is always either a cause or a consequence of some disease (Guyton and Hall 2006). In particular, the disturbance in cell population homeostasis is the ultimate nature of cancer, the topic of the following sections.

1.2 Tumours in the human body

Tumour as a disease is the disturbance of cell population homeostasis. Tumours are abnormal tissue bulks developing within normal tissues. Such tissue bulks contain uncontrolled growing cells.

Tumours are very diverse; they differ enormously in their levels of abnormality. Some tumours are slightly hyper-growing bulks of cells; other than that they are rather normal and can integrate well to the tissue. These are called *hyperplastic*. When the cells in a tumour appear abnormal in their shapes, their nuclei and their cytoplasmic activities, the tumour is called *dysplastic*. When such tumours archive macroscopic sizes, they are called *polyps* or *adenomas*. Adenomas however do not invade nearby tissues, respecting their boundary; they are thus called *benign* tumours, which can normally be removed by surgery. This is to distinguish from *malignant tumours*, which invade into nearby tissues, destroying the basal layers and can migrate to distal parts of the body via the blood streams or lymph streams to seed new tumours – a process called *metastasis*. Surgery and other treatments usually fail to treat malignant tumours because of their "roots" in the nearby tissues and because of their metastatic seeds. The explosive growth of metastasis tumours all over the body seriously disrupts the functions of different tissues of the body, causing pain and death of the organism (Weinberg 2007).

The word *cancer* refers to malignant tumours in clinical contexts. However, in cancer research, the terms "cancer" and "cancer cells" are usually also used for benign tumours and their cells. Following the convention of cancer research, we will also use "cancer" and "cancer cells" for a general tumour and tumour cells.

It is generally believed that tumours progress through at least some of the different levels of abnormality. Consistent evidence is available for some well-studied cases such as colorectal tumours; extension to all types of tumours is very difficult to verify due to the diversity and the complexity of the disease: any tumour is different from any other (Weinberg 2007).

A tumour does usually resemble, to some extent, the tissue and even the cell type from which it originates. Tumours and cancers can therefore be named according to their original tissues and their original cell types.

Tumours that develop in epithelial layers are called *carcinomas*, which make up 90% of clinical cases (Weinberg 2007). We can also distinguish two cell types in the epithelial layers: cells that do not have enzymatic secretory functions and those do have. Carcinomas of the former are called *squamous cell carcinomas*, while carcinomas of the latter are called *adenocarcinomas*.

There are also non-epithelial tumours. Tumours that form in the connective architectures such as the cartilage are called *sarcomas*. *Leukemias* are tumours in the blood-forming system (*hematopoietic*); *lymphomas* are tumours in the immune system. Tumours that form in the nervous systems are collectively called *neuroectodermal* tumours, which include, for example, *gliomas* and *glioblastomas*.

These classifications are by no mean clear-cut. In fact, they are sometimes not applicable, for example, when the original cell type of a tumour cannot be definitely determined. *Small cell lung cancer* is one example. Moreover, more often than not, many tumours such as those grow in lungs contain mixtures of squamous cell carcinomas and adenocarcinomas, rather than a simple homogenous type (Weinberg 2007).

1.2.1 Tumours are of genetic nature

If we zoom-in to look at the cells of a tumour, we find that they are physiologically different from normal cells. They are different in volumes, stiffnesses, shapes, surface structures, etc. In fact, it is always found that tumour cells either express abnormal proteins, or fail to express important proteins, or accumulate normal proteins at abnormal levels. As a result, different pathways of biochemical reactions are seriously disturbed, changing the behaviours of the cells in the tissue.

Today, it is clear that the alterations in the physiology of tumour cells are down to the alterations in their genetic contents. Indeed, it was found that in tumour cells, many genes that code for proteins that are crucial to the cell physiology are modified, or *mutated*. In fact, the genetic contents of tumour cells are usually *very* seriously damaged: not a few, but thousands of genes are modified – a phenomenon called *genetic instability*, which we will discuss in more detail in a following section.

Today, it is also clear that a tumour is generally initiated by a single cell. *The accumulation of the mutations in some key genes of a cell in the body during the living time of the organism alters the cell physiology in a way that allows it to escape from the cell homeostasis regulation of the body.* The cell divides and divides, initiating a clonal population of tumour cells; and there a primary tumour emerges and grows.

The life of a tumour is however not easy. Indeed, as we have learned, cells in a human body suffer from very strict regulations. *During the growth of a tumour, the genetic materials*

of the tumour cells need to change many times to adapt to and overcome such strict constraints. In other words, a tumour keeps evolving by means of mutation and selection in the somatic environment (*somatic evolution*).

1.2.2 Somatic mutations and the tumour founder cells

It is worth noting that essentially all normal cells of a human body share nearly identical genetic content. Indeed, starting from a single fertilised egg, cells continuously divide forming the whole organism. Cell divisions continue during the whole life of the organism: new cells are continuously born balancing old dying ones during cell renewals. A crucial point is that during such processes the genetic contents are highly conserved: the DNA molecules of new cells are meant to be the exact copies of the DNA molecules of the old ones.[1]

In reality, little changes in the genetic contents in the cells of a human body – or *somatic mutations* – nevertheless do occur. Somatic mutations can be due to the chemical instability of DNA molecules under environmental damages, or errors in DNA replications during cell divisions, which somehow survive the DNA repair mechanisms. In most cases, somatic mutations have no important effects on the cells, and the cells and their descendants will be continuously washed out from the body during cell renewals. In some other cases, somatic mutations are accumulated until some fatal mutations alter a cell in a way that allows it to proliferate; a tumour then starts.

It has been difficult to measure the somatic mutation rate, which is now estimated to be 10^{-6}–10^{-7} per gene per division (Frank and Nowak 2004). Suppose a human body consists of some 10^{13} cells, which all originated from a single cell, then there have been no less than 10^{13} cell divisions (when cell deaths are ignored). Then on the way to get to our mature body, we should have encountered 10^{6}–10^{7} mutations for each gene (!) This is already very large despite that the number of cell divisions was underestimated. Luckily a mutation in one gene is usually not sufficient to cause tumour; several mutations are required. The requirement for several simultaneous mutations is essential, which radically changes the picture. Indeed, if we assume that some three specific mutations are independently needed to initiate a primary proliferation, the above estimation would yield less than 10^{-5} cells that may have all the three mutations.[2]

The somatic mutation rate is significantly elevated through the exposure to *mutagens*.

[1] One may also wonder how different cell types in a human body arise despite their identical genetic content. This is due to a process called *cell differentiation*, during which the genomes of the cell are *epigenetically* modified, i.e., some chemical groups are attached to the nucleotide without changing the basic sequence. The result of these epigenetic modifications is that selective set of genes are silent in specific cell-types.

[2] Nevertheless, there is a speculation that most of us do indeed develop some covert tumours during our life-time (Greaves 2014).

Physical mutagens include ionising radiation and ultraviolet radiation. Chemical mutagens are very diverse, ranging from soot to some insecticides. Ionising radiations may directly cause double-stranded breaks, a very fatal type of damages to the DNA molecules. Ultraviolet radiation and chemical mutagens on the other hand chemically modify the bases of DNA molecules. The results of these modifications may be the conversion of the bases to other bases, or inducing interactions between the bases. The interactions between the bases can form obstacles which interfere with the movements of the replication machinery on the DNA molecule during replication, causing insertions, deletions or single- and double-stranded breaks in the synthesised sequences (Alberts et al. 2010, Weinberg 2007).

As the somatic mutation rate increases, tumours are more likely to occur – a fact that has been well-confirmed in *tumour epidemiology*. Thus, mutagens are also regarded as *carcinogens* – agents that cause cancer. Tobacco is well-known to cause tumours, in particular lung tumours. In the US, it was estimated that tobacco causes 90% of lung cancer deaths in male, and 75%-80% in female (Hecht 1999). There are some 55 mutagens in the tobacco smoke which are known to cause tumours when administrated on laboratory animals. The mutagens usually form covalent bonds with the bases of the DNA molecules, called DNA adducts, which normally block the replication machinery causing disruption of the synthesised DNA molecules (Hecht 1999).

1.2.3 Genetic instability and tumour hypermutations

It can be seen under a microscope that the chromosomes of a tumour cell usually appear in very different abnormal conformations. Here and there, parts of a chromosome or whole chromosomes are missing; other parts present in multiple copies; the chromosome fragments have strange shapes, sometimes stuck at the ends, see Hanahan and Weinberg 2011 and the references therein. Given the chaotic picture, it is obvious that thousands of genes may be deleted, other thousands may be amplified. When we zoom into their sequences, we also find excessive point mutations (i.e., alterations of single bases), short and long insertions, deletions all over the genome. Tumours are said to express *genomic instability*.[3]

There are two proposed mechanisms by which genomic instability may arise (Negrini et al. 2010). *The mutator hypothesis* is rather natural: at some point sooner or later in the development of the tumour, the tumour cells confer mutations which render its *DNA caretaker machinery* defective, of which the normal function is to keep the DNA integration

[3]Chromosomal instability is the major class of genomic instability. There are also other forms of genomic instability such as *micro-satellite instability*, which we will not discuss here, see Popat et al. 2005 for example. Moreover, it was also found that large parts of the DNA chains are usually also *epi-genetically* modified. We also skip the discussion of epigenetic modifications; interested readers are redirected to Dawson and Kouzarides 2012, Virani et al. 2012.

in check and to perform reparation when necessary. The obvious consequence of such defections is that an enormous number of DNA damages or DNA replication errors are left un-repaired. In reality, it was however found that in many cases, mutations in the DNA repair machinery are rare and late, despite the well established chromosomal instability of the tumour cells. This called for a second hypothesis: *oncogene-induced genomic instability*, where abnormal proteins made by mutated genes in tumour cells may directly interfere with the replication process, blocking of the replication machinery and causing replication errors (Negrini et al. 2010).

Regardless of the origin of genomic instability, it is generally believed that the high mutation rate it implies is necessary for the tumour to find its *evolutionary path* to escape from the strict constraints imposed by the human body environment.

1.2.4 Somatic selection and the evolution of a tumour

As we have already mentioned, there are multiple mechanisms that allow the body to regulate the cell homeostasis, which directly suppress or eradicate tumours. On its way to malignancy, a tumour keeps *evolving* to overcome the regulation of the body. Interestingly, in many aspects the evolution of a tumour resembles the familiar evolutionary process of a normal biological population (Okasha 2012): on one hand, somatic mutations continuously give rise to new *genotypes*; on the other hand, the cells carrying genes that allow them to overcome the body regulation divide faster than the others, driving the dynamics of the genetics of the whole population toward escaping the regulation. *In that sense, tumour development is an evolutionary process by means of mutation and selection in somatic context (Frank 2007, Greaves and Maley 2012).*

The evolution model of tumour development forms a very comprehensive picture of tumourigenesis, clarifying and systemising the hallmarks of cancer. One may recall the words from Theodosius Dobzhansky (Dobzhansky 1973),

"Nothing in biology makes sense except in the light of evolution."

Having conceptually identified the tumour development with an evolution process, we may attempt to extrapolate the sentence to the case of tumour biology,

"Nothing in tumour biology makes sense except in the light of tumour evolution."

We do not try to push such an extrapolation to extreme. What it implies here is that a hallmark of cancer should have some rationale under the light of tumour evolution.

The hallmarks of cancer have been summarised in the landmark papers by Hanahan and Weinberg (Hanahan and Weinberg 2011, Hanahan and Weinberg 2000). The *hallmarks*

of cancer in the context of tumour evolution consists of sustaining proliferative and survival signals; resistant to anti-growth and apoptosis signals; modifying stress responses; modifying cellular metabolism; invading the immune system; activating indefinite replications; inducing angiogenesis and *activating invation and metastasis*. Interested readers may find a more detailed discussion of these hallmarks in the following additional remark, or in Hanahan and Weinberg 2011, Hanahan and Weinberg 2000 and the references therein.

Remark: *The hallmarks of cancer*

Sustaining survival and growth signals

Starting from a normal cell, tumour cells will not be able to survive and grow if they do not receive special signal proteins that allow them to survive and grow. To promote their development, tumour cells implement several strategies to sustain the proliferative and survival signals. For example, they can activate the genes that code for these signals, which are often silent in normal differentiated cells. Alternatively, they can distort the signalling pathways that sense and relay the signals into the cell nucleus in a way that the pathway is consecutively activated.

Resistant to anti-growth and apoptosis signals

Anti-growth and apoptosis signals are examples of negative regulation of cell proliferation. In order to develop, tumours evolve to resist to these anti-growth and apoptosis signals. More interestingly, while being resistant to, tumour cells may actively produce these anti-growth and apoptosis signals as a mean of suppressing the normal cells in the neighbouring tissues.

Modifying stress responses

Tumour cells have abnormal genetic contents and express abnormal proteins. This is necessary for the tumour cells to invade the body regulation, but also seriously disturbs the normal living physiology of the cells. These disturbances may lead to cell cycle arrest or cell apoptosis as stress responses. As an example, protein p53 (coded by gene TP53) is activated as a result of DNA damages. Activated p53 proteins act as transcription factors which mediate the DNA repair machinery. When the damage is so serious that repair is not possible, p53 alternatively activates the apoptosis program. In order to develop, a tumour needs to invade these lethal stress responses. In particular, to avoid the destiny driven by p53, many tumours often evolve to mutate the gene TP53 so that the protein can no longer function properly. In fact, TP53 is non-functional in more than half of tumours of all kinds.

Modifying cellular metabolism

Cellular metabolism of glucose is a process which produces energy and necessary materials for the cell. Glucose after being transported into the cell is converted to pyruvate via a process called *glycolysis*. Pyruvates are then transported to the mitochondria where they are oxidised by oxygen. Both of these two steps produce energy in the form of highly energetic substances such as ATP (adenosine triphosphate). The latter process gives much more energy in comparison to the former; the price to pay is that it requires oxygen (*aerobic metabolism*) while the former does not (*anaerobic metabolism*). Anaerobic metabolism is important in *hypoxia* environment (i.e., lacking oxygen). However, it has been observed long ago that anaerobic metabolism is favoured in tumour cells even in the oxygen-sufficient environment. The phenomenon is known as *aerobic glycolysis*, which can happen quite early in tumour development. The exact rationale of tumour

aerobic glycolysis is yet to be clarified, but according to an old hypothesis, glycolysis is important to tumour cells perhaps because it gives rise to many other intermediate products which are important for synthesising diverse biomolecules necessary to build a new cell (Hanahan and Weinberg 2011).

Activating indefinite replications

Already in the 1960s it was known that a normal cell can divide only a finite number of times. A cell culture after a number of generations would come into a *senescence* state, where the cells remain alive but can no longer divide. Genetic techniques can kick them back into cell-cycles for some more generations. But soon after that, the culture comes to a second state called *crisis*, where most of the cells die with heavily damaged genomes. Few of them, however, survive the crisis and become *immortal* in the sense that they can divide indefinitely many times.

Today, it is known that the repeated sequences that cap the two ends of chromosomes – the *telomeres* – play a central role in this phenomenon. Normal DNA replication machineries cannot replicate those very ends of the chromosomes. As a result, the telomeres are shortened and shortened after each replication. When the caps can no longer be formed, the chromosome ends are unprotected and exposed to damage.

Stem cells are the source of cells for cell renewal; we expect that they must be able to divide indefinitely. How do they do that? In stem cells, there is a special protein, the *telomerase*, that takes the role of adding the repeated sequences to the DNA molecules after replications. Telomerase is deactivated at some point in cell differentiation.

Many tumours emerge from normal differentiated cells, where telomerase has been deactivated. It can be argued that without overcoming the cell senescence and the crisis, a tumour can never get to the sizes big enough to threaten the life of the organism. Although a normal differentiated cell can live up to some 70-80 generations before senescence, which already gives rise to some 2^{70} cells – larger than the number of cells of any body, we should not forget that cells in a tumour are dying and washing out very fast. A simple correction to the above calculation shows that with such a limit in the number of generations, tumours are indeed limited to very small sizes. Suppose the birth rate is $b = 1.4$ divisions per cell per day (Chmielecki et al. 2011), the time it takes for 70-80 generations is $T = 70/1.4 = 50$ days. If the death rate is $d = 1.3$ deaths per cell per day (Chmielecki et al. 2011), by the 50th day the tumour reaches $e^{(b-d)T} \approx 150$ cells – far too small for any harm.

Cancer genomics research found that telomerase is (re)activated in 80-90% of tumours (Harley 2008). It can be argued that the reactivation of telomerase is a frequently used strategy of cancer cells to become immortal. This raises the hope of targeting telomerase for cancer treatment (Harley 2008, Williams 2013).

Invading the immune system

Tumour is different from infectious diseases in that the tumour cells are from the organisms themselves, and not from the outside. While the infectious factors such as bacteria or viruses are usually easily recognised and destroyed by the powerful immune system of the body, we expect that tumour cells are not recognisable. One may be surprised when first learning that the immune system can also pinpoint and destroy tumour cells, sometimes very efficiently.

Evidences from laboratory experiments and clinical records have confirmed that the immune system have important prevention of tumours from developing in the body (Hanahan and Weinberg 2011, MacKie et al. 2003, Servan-Schreiber 2009). It is believed that the immune system

recognises tumour cells because of their abnormal features on the surface of tumour cells (abnormal proteins, missing immune suppressing complexes...) In considering the immune response as a promising cancer therapy, a large body of research has enumerated abnormal surface proteins (*surfaceome*) as targets for the immune system (Scott et al. 2012). Tumour-specific targets, defined as those antigens only expressed in tumour cells but not in normal cells, are yet to be found. However, there are a number of non-specific targets that invoke stronger immune responses in tumour cells than in normal cells.

Inducing angiogenesis

Cells require a connection to the blood streams for nutrients and waste transport; in fact, a cell in the human body cannot live at more than few micrometers away from a micro blood vessel. Because tumour cells grow and replicated faster, nutrients are even more essential for tumour cells than for normal cells. Indeed, it is now clear that solid tumours need to recruit blood vessels, without which they cannot grow to a macroscopic size and would be of no harm.

To recruit blood vessels (*angiogenesis*), tumours utilise the mechanisms by which blood vessels develop in newly forming organ or in wound healing processes (Jain and Carmeliet 2001, National Cancer Institute 2011). These involve mutual interactions between tumour cells and the cells that form the blood vessels. One of the well-known molecules which realises such interactions is the VEGF signalling protein (*vasculature epidermal growth factor*). The tumour cells activate and produce VEGF proteins; or they can also elicit other cells to do so. VEGF signalling proteins from the region of the tumour diffuse and bind to the VEGF receptors on the surface of the cells that form the blood vessels (*endothelial cells*). These endothelial cells are therefore excited and divide to form new blood vessels toward the tumour.

Tumour angiogenesis is conceptually promising for cancer therapy. By inhibiting, for example, VEGF, we expect that the development of tumour is called to a halt. Angiogenesis is a character of neoplastic tissues, therefore we also expects that angiogenesis inhibitors do no or little harm to normal stable tissues. In fact, it was hoped that cancer will be cured in few years after the discovery of angiogenesis molecules (Jain and Carmeliet 2001). Unfortunately, things were more complicated. Angiotherapy did not show high efficacy in clinical trials and/or tumours developed resistance to anti-angiogenesis agents. More importantly, after treatments, anti-angiogenesis resistant tumours were even more aggressive than before therapy, which led to the withdrawal of some FDA-approved angiogenesis inhibitors (Bergers and Hanahan 2008, Bottsford-Miller et al. 2012). Nevertheless, angiotherapy continues to be an active area of research and there is hope that such obstacles may be overcome at some point in the future, see also Bergers and Hanahan 2008, Bottsford-Miller et al. 2012, Cook and Figg 2010, Hanahan and Weinberg 2011.

Activating invasion and metastasis

Most of tumours eventually turn malignant, aggressively invading the nearby tissues and sending out tumour cells to colonise other parts of the body. Invasion and metastasis are highly complicated, of which many aspects are still elusive (Chambers et al. 2002).

Cells in a malignant tumour lose cell-cell adhesion molecules which held the cells together, and which also transduce antigrowth control signals from cell to cell. Alterations in integrins which help the cell attach to the extracellular matrix are also observed. Inside the cells, the cytoskeletons and other related proteins are altered. These changes are believed to be necessary for the cells to perform the very complicated metastasis process: damaging the basal layer, detach-

ing from the tumour, crawling through the extracellular matrix, leaking into the blood vessels, circulating throughout the body, leaking out of the blood vessels, and starting a new tumour at a new organ (Chambers et al. 2002, Wirtz et al. 2011).

Although metastasis is ubiquitous, the exact evolutionary forces that drive a tumour into the metastasis state are still a subject of debate (Chen et al. 2011). Note that if a cell of a non-malignant tumour at some point gains the ability to migrate (for metastasis), it is likely to move out of the tumour and leaves behind the static core of the tumour. In terms of tumour evolution, metastatic cells in a tumour have lower fitness than the other non-metastatic cells. Therefore, they cannot expand to dominate the tumour. Then how does a tumour become metastatic? The solution may lie in the fact that the transformation from non-metastasis to metastasis is more complicated than a one-step transformation sketched above: the complicated multi-step nature of the process and the spatial heterogeneity may be important factors that influence the evolution of a tumour toward metastasis (Quail and Joyce 2013).

1.2.5 Tumour heterogeneity, clonal interactions and cancer stem cells

The picture of the evolution of a tumour above is simple and clear. However there are complications behind the scene, which make tumour evolution a hard evolutionary problem. These complications are of multi-faceted: the exceedingly high mutation rate of tumour genomes, weak competence between clones, the environment heterogeneity, etc. One of the very important consequences is that there are often multiple clones with complicated interactions living together in the bulk of a tumour.

Tumour cell cooperation seems to be popular. Cooperation in sustaining growth signals and angiogenesis signals are prototype examples: few cells in a tumour produce those signals which bring the benefit not only to themselves but also to all the cells in the population. Cooperation in metabolism is another example: some tumours are divided into subpopulations with different metabolism schemes, where waste products of this subpopulation can serve as nutrients for the other subpopulation, see Hanahan and Weinberg 2011 and the references therein. There is also direct interference between tumour cell clones: it has been known that a primary tumour can secrete signals to suppress the growth of metastasised tumours; as a result, removing the former may boost the growth of the latter (Chambers et al. 2002).

It has been difficult to access the tumour heterogeneity experimentally. Molecular measurements performed on tumours are normally at the level of population-average. Any tumour specimen contains thousands of cells which are different from each other. By sequencing the specimen, for example, we are somehow accessing the population-averaged sequence; this calls for care in analysis and interpretation of the experiment data. Although there have been attempts in bioinformatics to infer the sequences of different clones in a tumour based on population-averaged measurements, this is still a difficult problem (Carter et al.

2012, Fischer et al. 2014, Oesper et al. 2013).

At the heart of cancer heterogeneity is the cancer stem cell hypothesis. The starting point of the hypothesis was the observation that only a minority of cells in a tumour can initiate new tumours when injected in laboratory animals (*xenograft*); the vast majority cannot (Beck and Blanpain 2013, Nguyen et al. 2012, Reya et al. 2001). The former was defined as *cancer stem cells*. Different surface markers for cancer stem cells were subsequently identified in some tumour types. The cancer stem cell hypothesis ventures that a tumour consisting of millions of cells is maintained only by a much smaller population of cancer stem cells, much in a similar way that normal tissues are maintained by normal stem cells. This hypothesis clearly has a great implication for cancer research. Since the cancer stem cell population is small, population-average measurements or even multiple-sampling measurements could not reveal them. This hinders understanding the genetics of cancer stem cells; and difficulties in designing therapies that address cancer stem cells follow. Much of current therapies are believed to address normal tumour cells instead of cancer stem cells. Cancer stem cells are less sensitive or hidden from therapy by the natural organisation of a tumour, which may also explain the invariable resistance of tumours to therapies. The analysis of cancer stem cells in different tumours and deeper understanding of their nature are waiting for future research, which may be critical to advances in cancer research and cancer therapy in the near future.

Remark: *Towards a more complete picture: tumour ecosystem*

The concept of tumour evolution shapes a very comprehensive picture of tumour development. Yet, there is an important conceptual limitation. In particular, the evolution model of tumours sets the tumour cells in the centre of the picture, which may lead to underestimate of the role of the responses of the body to the development of the tumour. Here, by the *body*, we mean not only the normal cells in the tumour or in its neighbourhood (which form the *tumour stroma*), but also the body as a whole. Recently, there seems to be a shift from the evolution model of tumours toward *the ecosystem model of tumours* (Quail and Joyce 2013): instead of thinking of a tumour as a population of tumour cells, we think of a tumour as a composite, co-evolving system of tumour cells and the body. This view is supported by abundant evidences of the mutual interactions between tumours and different elements of the body, of which the results are both promoting and preventing tumour development. It also changes the approach to cancer therapy: instead of addressing the tumour cells, one may target different components of the tumour ecosystem. In fact, the angiotherapy, which inhibits endothelial cells, already uses this approach. With the view of tumour ecosystem, therapeutic targets are now extended to include many other elements, e.g., the cells of the immune system or the fibroblasts.

Among the different cells that influence the development of a tumour, the cells from the immune system with their paradoxical roles are being of central interest (Visser et al. 2006). We learned that the immune system prevents a tumour from emergence and growth, but it also promotes tumour development in several ways. It has been known for a long time that a tumour

causes inflammation, followed by the infiltration of innate immune cells such as macrophages, neutrophils and mast cells into the neoplasia region. Macrophages and neutrophils are phagocytes, which are important to protect the body from the infection of alien agents such as bacteria or viruses. It was believed that the infiltration of the immune cells into a tumour reflects the attempt of the immune system in eradicating the tumour. Surprisingly, this belief turned out not quite true! There is evidence showing that the activities of the cells also promote tumour development in many ways. The recruited immune cells generate many important signalling proteins such as growth factors (e.g., TNFα) and vascular epidermal growth factor (VEGF), which are essential for the tumour growth. These cells also secrete MMP proteins (*matrix metalloproteinase*) that modify the extracellular matrix and cell adhesion proteins, supporting angiogenesis and cell migration in metastasis. Their highly reactive products are also thought to contribute to tumour genomic instability. Moreover, through multiple feedback interactions, the macrophages and the mast cells (which belong to the *innate arm* of the immune system) also inhibit the attempt of other components of the immune system in eradicating the tumour (which are often found to be of the *adaptive arm* of the immune system). Experiments have confirmed that suppressing the innate arm of the immune system does indeed slow down tumour progression, see Visser et al. 2006 and the references therein.

Further research is going on to reveal the exact activities of different immune components and their mutual interactions in response to a tumour (Gajewski et al. 2013). Once we understand and are able to harness the tumour immune response, the next generation immunotherapy will appear to be among the most promising approaches for tumour treatment and prevention (Couzin-Frankel 2013, Gajewski et al. 2013, Hanahan and Weinberg 2011, Scott et al. 2012, Visser et al. 2006).

Let us come back to the story of the blood vessels of tumours. We learned that angiogenesis inhibitors founded the basis for angiotherapy. Things go further than that. It was known long ago that tumour blood vessels are different from normal blood vessels. Tumour vessels are highly irregular. The fact that the slots on the walls of tumour blood vessels are much larger than that of normal blood vessels inspires the idea of nano-medicine. The idea is as follows: if the drug molecules are larger than normal vessel slots and smaller than the tumour vessel slots, they can get out of tumour blood vessels but not of normal blood vessels. Using nano-particles with appropriate sizes as drug transporters therefore may lead to the accumulation of drugs in the tumour but not in normal tissues, thereby increasing drug specificity (Grossman and McNeil 2012).

Pericytes are cells that form the outer cover to support the endothelium of normal blood vessels. Recent research reveals diverse functions of pericytes with close interactions with the endothelial cells (Bergers and Song 2005). In particular, pericytes have growth signalling activity, playing important roles in angiogenesis and the stabilisation of neovasculatures. Interestingly, in normal tissues, pericytes still function when the density is reduced by even 90%. In tumours, pericytes are normally sparse, and only loosely attached to the endothelia. Still, they are believed to be essential to the tumour blood vessels. Indeed, inhibition of pericytes leads to the destabilisation of tumour blood vessels (while normal blood vessels can cope with such slight reduction of pericytes), making them more susceptible to vessel destroying therapies. In experiments, pericyte inhibitors indeed improve the efficacy of endothelial-targeted agents and angiogenic therapy, see Bergers and Song 2005 and the references therein.

Fibroblasts are cells imbedded in and maintaining the extracellular matrix in normal tissues.

Fibroblasts are recruited to tumours in very early stages. Under the interactions with tumours and other cells in the tumour stroma, fibroblasts become active, proliferating and synthesising a large amount of extracellular matrix. The active fibroblasts secrete MMPs (*matrix metalloproteinases*) which degrade the extracellular matrix, thereby remodelling the extracellular matrix. These activities are thought to be important to tumour angiogenesis and metastasis activation. Fibroblasts also secrete growth factors, which directly affect tumour growth and cell motility. In addition, they are also sources of signalling molecules that regulate the activity of the immune system. Therapies that address fibroblasts are also under current investigation (Kalluri and Zeisberg 2006).

Many other cell types in the neighbourhood of the tumour also contribute to the tumour development, see Pienta et al. 2008, for example. Many other signalling interactions between different cell types in a tumour are still unknown. We are waiting for a more complete picture and great potential advances in cancer therapy fostered by the on-going researches in the field (Gatenby 2009, Gatenby and Gillies 2008, Merlo et al. 2006).

As a last note for this section, I would like to mention that given the complex nature of the interactions between the tumour and the body, high level responses of the body to tumours should be seriously considered. Likewise, the effects of alternative and complement therapies such as diets, traditional medicines, sports, meditations, etc. should be looked into scientifically. In the future, highly integrative therapies are perhaps to be expected.

1.3 Cancer genomics and statistical modelling

The above discussed picture of tumour development is the knowledge of centuries of cancer research in conjunction with biology and other scientific areas. Most significantly, one should mention the influence of molecular biotechnology in the last two, three decades. During the time, molecular biotechnology has been developing in a very fast pace in many different aspects, a very important consequence of which was the starting of the era of genomic research. A molecular biology measurement nowadays can consist of thousands of parallel specific simpler measurements. We are now able to measure the expression levels of a large part of a genome (in terms of messenger RNA concentrations). We are also able to measure the copy-numbers of most genes in parallel, or even to access to the sequences of the whole genome of an organism. These *genomic measurements* have provided a way for researchers of cancer research to look into the defects in the molecular machinery of cancer cells.

Cancer genomic research comes with big data. Here we will discuss several data types that are most relevant to our analysis. In particular, we summarise the ideas of a gene expression measurement (by microarrays) and a copy-number measurement. Although the exact protocol and detailed technology vary from experiment to experiment, these basic ideas remain relatively the same.

The gene expression measurement

In this thesis, gene expression measurement refers to the measurement of messenger RNA concentrations. Here we are interested in gene expression measurements by microarrays. In these measurements, messenger RNA are first extracted from a tumour sample and reverse-transcribed to the short complementary DNA sequences (*oligonucleotides*) by specific enzymes. The DNA sequences are then labelled by fluorescent agents. These labelled DNA short sequences are subsequently hybridised with their complementary sequences that have been prepared at specific places on an array (*microarray*). By scanning the fluorescent intensities on the array, the concentrations of the messenger RNA can be inferred. For the details of the measurement of gene expressions by microarrays, see Parmigiani and Garett 2003 for example.

The gene copy-number measurement

Copy-numbers refer to the numbers of copies of genes in the genome of a cancer cell, which are often different from two (diploid) because of genetic instability. Microarrays have been adapted to measure the copy-numbers of cancer cells. The procedure remains very similar to that of the gene expression measurement. Generally, DNA molecules are first cut into oligonucleotides by special enzymes and amplified by polymerase chain reactions (PCR). The oligonucleotides are then labelled by fluorescent agents and these labelled oligonucleotides are then hybridised with their complementary sequences prepared on an array. The array is then scanned to measure the fluorescent intensities. By this construction, these fluorescent intensities are directly related to the numbers of copies of the short sequences present in the sample. For the details of the measurement of copy-number by microarrays, see Huang et al. 2004, Pinkel et al. 1998.

DNA sequencing

Besides gene expressions and copy-numbers of cancer cells, we will also work with sequencing data. However, we will be concerned only with gene-specific sequencing instead of who genome sequencing. One of the classic technologies for sequencing is called Sanger sequencing. The DNA segment to be sequenced may first be amplified if necessary. One then lets the sequences replicate under the activity of polymerases in a medium with sufficient nucleotides. A small fraction of the nucleotides in the medium are modified so that they do not allow DNA elongation. The replicated sequences therefore are stopped at a random base if by chance a modified nucleotide is added at that position. Sequences of different lengths can be separated using the so-called gel-electrophoresis effect. If the

modified nucleotides are labelled with fluorescent agents of four different colours which correspond to four different nucleotides, then the ending bases of the random-stopped sequences can be identified. From these data, the sequence of the DNA segment can be inferred. For a more detailed description of the basis of Sanger sequencing, see Alberts et al. 2010.

Although we already mentioned in the last section, it is also worth emphasising again that often both the gene expression measurement and the copy-number measurement mentioned above only capture the population-averages of the messenger RNA concentrations and the DNA copy-numbers as they both require biological samples which contain thousands of cells. This calls for care in the interpretation of the results.[4]

Together with these molecular techniques, analytical methods to analyse the emerging data have been developing in a very fast pace. We will have short reviews of such analytical methods in the specific contexts of the following chapters before describing our corresponding approaches.

References

Alberts, B., D. Bray, K. Hopkin, A. D. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter (2010). *Essential cell biology*. Garland Science.

Beck, B. and C. Blanpain (2013). "Unravelling cancer stem cell potential". *Nat. Rev. Cancer* 13.10, pp. 727–38.

Bergers, G. and D. Hanahan (2008). "Models of resistance to anti-angiogenic therapy". *Nat. Rev. Cancer* 8, pp. 592–603.

Bergers, G. and S. Song (2005). "The role of pericytes in blood-vessel formation and maintenance". *Neuro. Oncol.* 7.4, pp. 452–64.

Bottsford-Miller, J. N., R. L. Coleman, and A. K. Sood (2012). "Resistance and escape from aniangiogenesis therapy: clinical implication and future stategies". *Journall Clin. Oncol.* 30.32, pp. 4026–4034.

Carter, S. L. et al. (2012). "Absolute quantification of somatic DNA alterations in human cancer". *Nat. Biotechnol.* 30.5, pp. 413–21.

Chambers, A. F., A. C. Groom, and I. C. MacDonald (2002). "Dissemination and growth of cancer cells in metastatic sites". *Nat. Rev. Cancer* 2, pp. 563–572.

Chen, J., K. Sprouffske, Q. Huang, and C. C. Maley (2011). "Solving the puzzle of metastasis: the evolution of cell migration in neoplasms". *PLoS One* 6.4, e17933.

Chmielecki, J. et al. (2011). "Optimization of dosing for EGFR-mutant non-small cell lung cancer with evolutionary cancer modeling". *Sci. Transl. Med.* 3.90, 90ra59.

Cook, K. M. and W. D. Figg (2010). "Angiogenesis inhibitors: current strategies and future prospects". *CA Cancer J. Clin.* 60, pp. 222–243.

[4] Although single cell technologies are being developed, they are still expensive for routine experiments.

Couzin-Frankel, J. (2013). "Cancer immunotherapy". *Science* 342.6165, pp. 1432–3.

Dawson, M. A. and T. Kouzarides (2012). "Cancer epigenetics: from mechanism to therapy". *Cell* 150.1, pp. 12–27.

Dobzhansky, T. (1973). "Nothing in biology makes of sense except in the light of evolution". *Am. Biol. Teach.* 35.3, pp. 125–129.

Fischer, A., I. Vázquez-García, C. J. R. Illingworth, and V. Mustonen (2014). "High-definition reconstruction of clonal composition in cancer". *Cell Rep.* 7.5, pp. 1740–52.

Frank, S. A. (2007). *Dynamics of cancer: incidence, inheritance, and evolution*. Princeton University Press.

Frank, S. A. and M. A. Nowak (2004). "Problems of somatic mutation and cancer". *BioEssays 26*, pp. 291–299.

Gajewski, T. F., H. Schreiber, and Y.-X. Fu (2013). "Innate and adaptive immune cells in the tumor microenvironment". *Nat. Immunol.* 14.10, pp. 1014–22.

Gatenby, R. A. (2009). "A change of strategy in the war on cancer". *Nature* 459, pp. 508–509.

Gatenby, R. A. and R. J. Gillies (2008). "A mircoenvironment model of carcinogenesis". *Nat. Rev. Cancer* 8, pp. 56–61.

Greaves, M. (2014). "Does everyone develop covert cancer?" *Nat. Rev. Cancer* 14, pp. 209–210.

Greaves, M. and C. C. Maley (2012). "Clonal evolution in cancer". *Nat. Rev. Cancer* 481, pp. 306–313.

Grossman, J. H. and S. E. McNeil (2012). "Nanotechnology in Cancer Medicine". *Phys. Today* 65.8, p. 38.

Guyton, A. C. and J. E. Hall (2006). *Textbook of medical physiology*. Elsevier Saunders.

Hanahan, D. and R. A. Weinberg (2011). "The hallmarks of cancer: the next generation". *Cell* 144.5, pp. 646–674.

Hanahan, D. and R. A. Weinberg (2000). "The hallmarks of cancer". *Cell* 100.1, pp. 57–70.

Harley, C. B. (2008). "Telomerase and cancer therapeutics". *Nat. Rev. Cancer* 8.3, pp. 167–79.

Hecht, S. S. (1999). "Tobacco smoke carcinogens and lung cancer". *J. Cancer Natl. Inst.* 91.14, pp. 1194–1210.

Huang, J. et al. (2004). "Whole genome DNA copy number changes identified by high density oligonucleotide arrays". *Hum. Genomics* 1.4, pp. 287–99.

Jain, R. K. and P. F. Carmeliet (2001). "Vessels of death or life". *Sci. Am.* 285.6, pp. 38–45.

Kalluri, R. and M. Zeisberg (2006). "Fibroblasts in cancer." *Nat. Rev. Cancer* 6.5, pp. 392–401.

MacKie, R. M., R. Reid, and B. Junor (2003). "Fatal melanoma transferred in a donated kidney 16 years after melanoma surgery". *N. Engl. J. Med.* 348.6, pp. 567–8.

Merlo, L. M. F., J. W. Pepper, B. J. Reid, and C. C. Maley (2006). "Cancer as an evolutionary and ecological process". *Nat. Rev. Cancer* 6, pp. 924–935.

National Cancer Institute (2011). *Angiogenesis inhibitors*. Fact Sheet.

Negrini, S., V. G. Gorgoulis, and T. D. Halazonetis (2010). "Genomic instability – an evolving hallmark of cancer". *Nat. Rev. Mol. Biol.* 11, pp. 220–228.

Nguyen, L. V., R. Vanner, P. Dirks, and C. J. Eaves (2012). "Cancer stem cells: an evolving concept". *Nat. Rev. Cancer* 12.2, pp. 133–43.

Oesper, L., A. Mahmoody, and B. J. Raphael (2013). "THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data". *Genome Biol.* 14.7, R80.

Okasha, S. (2012). "Population genetics". In: *The Stanford Encyclopedia of Philosophy*. Ed. by N. Zalta.

Parmigiani, G. and E. S. Garett (2003). *The Analysis of Gene Expression Data: Methods and Software*. Springer.

Pienta, K. J., N. McGregor, R. Axelrod, and D. E. Axelrod (2008). "Ecological therapy for cancer: defining tumours using an ecosystem paradigm suggests new opportunities for novel cancer treatments". *Transl. Oncol.* 1.4, pp. 158–164.

Pinkel, D. et al. (1998). "High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays". *Nat. Genet.* 20.2, pp. 207–11.

Popat, S., R. Hubner, and R. S. Houlston (2005). "Systematic review of microsatellite instability and colorectal cancer prognosis". *J. Clin. Oncol.* 23.3, pp. 609–18.

Quail, D. F. and J. A. Joyce (2013). "Microenvironmental regulation of tumor progression and metastasis". *Nat. Med.* 19.11, pp. 1423–37.

Reya, T., S. J. Morrison, M. F. Clarke, and I. L. Weissman (2001). "Stem cells, cancer, and cancer stem cells". *Nature* 414.6859, pp. 105–11.

Scott, A. M., J. D. Wolchok, and L. J. Old (2012). "Antibody therapy of cancer". *Nat. Rev. Cancer* 12.4, pp. 278–87.

Servan-Schreiber, D. (2009). *Anticancer–a new way of life*. Viking Penguin.

Virani, S., S. Virani, J. A. Colacino, J. H. Kim, and L. S. Rozek (2012). "Cancer epigenetics: a brief review". *ILAR J.* 53.3-4, pp. 359–69.

Visser, K. E. de, A. Eichten, and L. M. Coussens (2006). "Paradoxical roles of the immune system during cancer development". *Nat. Rev. Cancer* 6.1, pp. 24–37.

Weinberg, R. A. (2007). *The biology of cancer*. Garland Science.

Williams, S. C. P. (2013). "No end in sight for telomerase-targeted cancer drugs". *Nat. Med.* 19.1, p. 6.

Wirtz, D., K. Konstantopoulos, and P. C. Searson (2011). "The physics of cancer: the role of physical interactions and mechanical forces in metastasis". *Nat. Rev. Cancer* 11.7, pp. 512–22.

Chapter 2

Probability, information and statistical modelling

More is the same...
L. P. Kadanoff

We start this chapter with a short review of some aspects of probability theory and information theory. We then discuss the general workflow of a statistical model, starting from making parametric assumptions to parameter estimation. Having done so, we concentrate on a specific strategy of building a statistical model, namely the maximum entropy reasoning. We will then discuss several linear models in a comprehensive view of the maximum entropy reasoning. Some interesting technical topics are given as additional remarks.

2.1 Probability and information

2.1.1 Probability

In statistical modelling, we often realise a discrete probabilistic space (Rosenthal 2010) by imagining a *system* C which takes *random states* (or *configurations*) in a discrete *sampling space* Ω according to a *probability distribution* $p : \Omega \to [0, 1]$. A subset of Ω is called an *event*. The probability of observing an event E is given by

$$\mathrm{P}(E) = \sum_{\mathcal{C} \in E} p(\mathcal{C}). \tag{2.1}$$

The family of all subsets of Ω is the *family of events* of Ω, denoted by $\mathscr{M}$. Note that P is a (set) function from $\mathscr{M}$ to $[0, 1]$. The triplet $(\Omega, \mathscr{M}, \mathrm{P})$ constitutes a probabilistic space (Rosenthal 2010).

When we know that a non-empty event E happens, C is limited to the elements of E. The probability for another event E' to happen therefore also changes. The probability of

E' conditioned on E is formally defined as

$$\mathrm{P}(E'|E) = \frac{\mathrm{P}(E' \cap E)}{\mathrm{P}(E)}. \tag{2.2}$$

A real function from Ω to $\mathbf{R}$, $Q : \Omega \to \mathbf{R}$, will be called an *observable*. The mean value of an observable Q is defined as

$$\langle Q \rangle = \sum_{\mathcal{C}} p(\mathcal{C}) Q(\mathcal{C}). \tag{2.3}$$

In general, a function from Ω to a measure space is called a *random variable* (Rosenthal 2010). In this extension, the system C itself can be considered as a random variable defined by the identity map on Ω. Note that we will distinguish the different notations for random variables (which are functions) and their values; for example, x, y, ϕ and s are values of random variables X, Y, Φ and σ.

Of continuous sampling spaces, we consider only finite dimensional Euclidean spaces. With little care when taking the continuum limit, these definitions are also applied for Euclidean sampling spaces. In particular, the probability distribution $p(\mathcal{C})$ for a discrete sampling space turns into a probability density function $p(x)$ in the continuum limit, and the probability of an event E is given by

$$\mathrm{P}(E) = \int_E \mathrm{d}\, x p(x), \tag{2.4}$$

where the event E is limited to Lebesgue integrable sets. The summation in the definition of mean values of observables is also turned into an integral accordingly

$$\langle Q \rangle = \int_\Omega \mathrm{d}\, x Q(x). \tag{2.5}$$

The definition of conditional probabilities (2.2) is still valid when $\mathrm{P}(E) \neq 0$. Little care is needed when $\mathrm{P}(E) = 0$: one will need to discretise the sampling space Ω and take the continuum limit of the ratio $\mathrm{P}(E' \cap E)/\mathrm{P}(E)$ properly. For the formal theory of probability, we refer to Feller 1950, Hu 1967, Rosenthal 2010. In the following, if not noted otherwise, we implicitly assume that Ω is discrete for the sake of convenience. The discussions can always be adapted to continuous spaces with minor modifications.

2.1.2 Information: entropy and relative entropy

Information theory all started with the theory of practical communication with Shannon's seminal paper (Shannon 1948). The implication of the theory gradually turns out much

more broad, penetrating into many different areas of science, in particular physics (Cover and Thomas 2006). Here, we concentrate on some aspects of information theory which will be useful for our discussions, in particular, the measures of information, the *entropy function* and the *Kullback–Leibler divergence* (or the *relative entropy function*).

Consider a finite element sampling space Ω, with all subsets of Ω as events. A probability measure P on such a finite space is completely determined by a distribution function $p : \Omega \to \mathbf{R}$, defined by $p(\mathcal{C}) = \mathrm{P}(\{\mathcal{C}\})$ with $\sum_{\mathcal{C}\in\Omega} p(\mathcal{C}) = 1$. We denote by $\mathscr{P}(\Omega)$ the set of all possible probability distribution on Ω.

The information entropy is a functional $S : \mathscr{P}(\Omega) \to \mathbf{R}$, defined by[1]

$$S[p] = \sum_{\mathcal{C}\in\Omega} -p(\mathcal{C}) \ln p(\mathcal{C}). \tag{2.6}$$

For an axiomatical (and more general) definition of entropy, see Khinchin 1957. In information theory, the information entropy measures the minimal average length of a codeword to code an event, see Cover and Thomas 2006. In general, the information entropy is usually interpreted as a measure of the unpredictability of a probability distribution. The latter interpretation merges the information entropy with the thermodynamic entropy in statistical mechanics, see Landauer 1961, Merhav 2010, Sagawa 2012, Zurek 1990 for detailed discussions.

Let p and q be two distributions on Ω, we define the *Kullback–Leibler divergence* from q to p as

$$\mathrm{KL}(p|q) = \sum_{\mathcal{C}\in\Omega} p(\mathcal{C}) \ln \frac{p(\mathcal{C})}{q(\mathcal{C})}. \tag{2.7}$$

The Kullback–Leibler divergence has a very important property: for any distributions q and p, $KL(p|q) \geq 0$, and for any distribution p, $KL(p|p) = 0$.

The Kullback–Leibler divergence is also known as the *relative entropy*, which measures how much a distribution q is different from a distribution p (Cover and Thomas 2006). In information theory, the Kullback–Leibler divergence can be interpreted as the cost (redundant code length) of using a wrong distribution for the source in coding (Cover and Thomas 2006). In statistical mechanics, the Kullback–Leibler divergence has a relation to works in thermodynamic processes (Bavaud 2009).

[1] In general, the entropy function S is defined on probability simplexes, and it is continuous at the boundary of the simplexes.

2.2 Statistical modelling

Consider a probabilistic space Ω with some specific event family $\mathscr{M}$, but the probability measure P is unknown. By *statistical modelling*, we mean to experimentally estimate the probability measure P on Ω, often through estimating the corresponding probability distribution function p.

We will see that to make the estimation for the probability distribution function p, one usually proceeds in two steps: making assumptions about the mathematical form of the distribution, which usually contains unknown parameters (*parametric assumptions*), then estimating the parameters by fitting the model to some standard sample data called *training data*.

2.2.1 Assumptions

A series of independent states of C, each called a *sample*, constitutes a *training dataset*. A dataset of M samples, $D = \{\mathcal{C}^1, \mathcal{C}^2, \ldots, \mathcal{C}^M\}$, reproduces the probability distribution p approximately in the sense that

$$p(\mathcal{C}) \approx \frac{1}{M} \sum_{\mu=1}^{M} \delta(\mathcal{C}, \mathcal{C}^\mu), \tag{2.8}$$

where δ is the Kronecker delta function. The approximation becomes exact in the limit of infinite number of samples. In practice, however, the sampling space Ω is very large while the sizes of datasets are much smaller. This estimation for p is therefore not very useful. A practical solution is to restrict the searching space to some subspace of the space of all distributions on Ω by *parametric assumptions*.

Parametric assumptions

A parametric assumption restricts the probability distribution to some specific functional form with some unknown parameters. For example, *assuming* that the heights X of people in a population are normally distributed restricts the form of the probability distribution of the heights to that of $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}$, which contains two parameters μ and σ. Often, the parameters of a parametric distribution have specific interpretations, bringing important insights into the problem. In this case, μ and σ are simply the mean and the standard deviation of the heights of the people.

Making such parametric assumptions is rather subjective. In some rare cases, the assumptions can be derived from some general theory of the phenomena investigated. Most of the time, there is however no clear basis for such parametric assumptions. Therefore

most of statistical models are *phenomenological*.

Parameter estimation

The parametric assumptions identify the model distribution up to some unknown parameters. To make quantitative predictions, one needs to estimate the values of the parameters in the model.

Maximum likelihood estimation (ML)

A classic criterion for estimating the parameters for a model, seemingly intuitive, is the *maximum-likelihood criterion*: the parameters are chosen so that the probability of observing the training dataset is maximised.

Consider again a random system C taking states in sampling space Ω. We are to estimate the probability distribution $p(\mathcal{C})$ on the basis of an experimental dataset, $D = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_M\}$, which are M states independently sampled from the sampling space. Suppose under the parametric assumptions, the distribution of the system is $p_\lambda(\mathcal{C})$, where λ collectively denotes the set of parameters. Then the maximum-likelihood criterion for parameter estimation is stated as

$$\lambda^{ML} = \underset{\lambda}{\operatorname{argmax}} \ln p_\lambda(D), \tag{2.9}$$

where $\ln p_\lambda(D)$ is the logarithm of the probability of observing the data D, referred to as the *log-likelihood function*, with the logarithm is used for the sake of convenience. Note that under the assumption that the samples are independent instants of the sampling space, we have

$$\ln p_\lambda(D) = \sum_{\mu=1}^{M} \ln p_\lambda(C^\mu). \tag{2.10}$$

Therefore, the log-likelihood function is scaled linearly with the size M of the training dataset.

Bayesian modelling

Bayesian inference takes a different view on the parameters in the parameteric distribution: as one does not know exactly where the best model (i.e. the best values for the parameters for the model distribution) is in the space of parametric models defined by the parameteric assumptions, one oughts to assign "a degree of belief" to every point in the space of the

parametric models which quantifies how likely the model is at that point.[2] Cox showed that with plausible axioms for the arithmetics of this degree of belief, it is a probability measure on the space of parametric models (Cox 1946). In fact, the key-point in Bayesian inference is to identify "degree of belief" with probability, and the arithmetics of degree of belief with the arithmetics of probability. This degree of belief is also called the *prior distribution* for the parameters due to the reason which will be clear shortly. Note that the prior distribution is an (further) assumption – a *quantitative assumption* (MacKay 2002).

The starting point of Bayesian inference is to assume a prior distribution $p(\lambda)$ for the model parameters. With this quantitative assumption, the probability of observing state $\mathcal{C}$ can readily be found by marginalising overall the space of parameters,

$$p(\mathcal{C}) = \int \mathrm{d}\,\lambda p(\lambda) p(\mathcal{C}|\lambda), \tag{2.11}$$

where we assumed that the space of parameters is continuous, which is often the case in practice. Note that changing the philosophy from maximum-likelihood estimation to Bayesian inference, we also change the notation for the model distribution from $p_\lambda(\mathcal{C})$ to $p(\mathcal{C}|\lambda)$, exploiting the notation of conditional probability.

The prediction (2.11) is entirely based on the quantitative assumption, namely the prior distribution $p(\lambda)$. This is called the *prior prediction*. It seems that, by making a quantitative assumption, we have based our prediction on the prior knowledge (quantitative assumption), disregarding the ability to adapt the model to a specific situation with specific training data as in the maximum likelihood approach. Not quite so! The whole strength of Bayesian inference is that the degree of belief can be updated when evidence such as a training dataset is available. Having identified the degree of belief with probability, the rule of updating the belief is simply expressed in Bayes' theorem,

$$p(\lambda|D) = \frac{p(D|\lambda)p(\lambda)}{p(D)}, \tag{2.12}$$

where the probability of observing data D (or the *evidence*) is

$$p(D) = \int \mathrm{d}\,\lambda p(D|\lambda) p(\lambda). \tag{2.13}$$

The quantity $p(\lambda|D)$ is also referred to as the *posterior distribution* of model parameters. Therefore, the probability of observing state $\mathcal{C}$ under the evidence D is also updated ac-

[2] "A density of belief" is applied for continuous parameters in a similar way to probability density in continuous spaces.

cordingly,

$$p(\mathcal{C}|D) = \int \mathrm{d}\,\lambda p(C|\lambda)p(\lambda|D). \tag{2.14}$$

The quantity $p(\mathcal{C}|D)$ is referred to as the *posterior prediction*.

We see that by identifying the degree of belief with probability, the Bayesian approach provides us with a very clear way of going from assumptions, incorporating evidence from training data to deriving predictions. These steps were performed entirely within the framework of probability theory, in particular Bayes' theorem. There are, however, two technical difficulties in Bayesian modelling. The first is that it is often rather difficult to make quantitative assumptions, namely to choose a prior distribution for the parameters. In most of the cases, this is done on the basis of subjective convenience. Secondly, Bayesian modelling is often computationally more difficult than the maximum likelihood approach. In practice, approximations are almost always needed in performing Bayesian inference. Interestingly, under certain approximations, Bayesian inference mathematically reduces to the maximum likelihood estimation.

Maximum a posteriori approximation (MAP)

Maximum a posteriori approximation (MAP) assumes that for large enough training datasets, the posterior distribution $p(\lambda|D)$ is sharply peaked at some value of parameters (which, for the sake of simplicity, is assumed to be unique). This makes it plausible to approximate the posterior distribution by a delta function,

$$p(\lambda|D) \approx \delta(\lambda - \lambda^{MAP}), \tag{2.15}$$

where

$$\lambda^{MAP} = \underset{\lambda}{\operatorname{argmax}}\, p(\lambda|D). \tag{2.16}$$

By inserting the delta function approximation (2.15) into (2.14), we obtain

$$p(\mathcal{C}|D) = p(\mathcal{C}|\lambda^{MAP}), \tag{2.17}$$

which is essentially the prediction within maximum-likelihood estimation with λ^{ML} replaced by λ^{MAP}.

Note that $p(\lambda|D)$ and $p(D,\lambda)$ are different by only a normalisation factor $p(D)$, which does not depend on λ,

$$p(D,\lambda) = p(\lambda|D)p(D). \tag{2.18}$$

That means one can also write

$$\lambda^{MAP} = \underset{\lambda}{\operatorname{argmax}}\, p(D, \lambda). \tag{2.19}$$

Then the joint probability $p(D, \lambda)$ can be decomposed into the likelihood $p(D|\lambda)$ and the prior $p(\lambda)$,

$$p(D, \lambda) = p(D|\lambda)p(\lambda), \tag{2.20}$$

therefore

$$\lambda^{MAP} = \underset{\lambda}{\operatorname{argmax}}\, p(D|\lambda)p(\lambda). \tag{2.21}$$

Since logarithm is a monotonic function, one can also write

$$\lambda^{MAP} = \underset{\lambda}{\operatorname{argmax}} \left\{ \ln p(D|\lambda) + \ln p(\lambda) \right\}. \tag{2.22}$$

In comparison to the maximum-likelihood estimation (2.9), MAP modifies the log likelihood function by the logarithm of a prior distribution.

One sees immediately that the two values λ^{ML} and λ^{MAP} are identical when $p(\lambda)$ is constant with respect to λ (*flat prior*).[3] When the prior is not flat, as $\ln p(D|\lambda)$ scales linearly with the size of the training dataset (see equation (2.10)) while the log-prior $\ln p(\lambda)$ does not, MAP estimation λ^{MAP} also asymptotically tends to λ^{ML} at the large dataset limit. Hence, when the training dataset is small, the difference between maximum-likelihood estimation and maximum a posteriori approximation is important. A limited training dataset causes a phenomenon called *overfitting* in maximum-likelihood estimation, where the likelihood function does not express a sharp peak and the likelihood maximiser is un-certain. In Bayesian modelling, inference with limited data is driven by prior knowledge. Mathematically, the posterior probability can still maintain a relatively sharp maximal peak because of the contribution of the prior. Therefore, by incorporating appropriate prior knowledge, the overfitting problem can be avoided, or at least reduced. MAP is therefore also referred to as *regularised maximum-likelihood estimation*.

Maximum a posteriori approximation is nothing but the saddle point approximation known in physics. A next step of the approximation can be expanding the posterior distribution to the second order around the maximum, which results in the so-called Laplace approximation (known in physics as Gaussian approximation); for the details, see Bishop 2006, MacKay 2002.

[3] These flat priors may be *improper* in the sense that they are not normalisable, since they span over infinite spaces.

2.3 Maximum entropy principle and linear models

In this section, we introduce a heuristic strategy for making parameteric assumptions – *the maximum entropy reasoning*. Using the maximum entropy reasoning approach, we setup a unified framework for many statistical models, including Gaussian model, linear regression, logistic classification and some other related models. During this section, interesting and important aside topics such as model regularisation, subset selection, will also be discussed as additional remarks.

Maximum entropy reasoning

Suppose we are to estimate the distribution p for a random system C with a configuration space Ω. In practice, the system under investigation is usually not yet clearly understood, and the parameteric assumptions that restrict the mathematical form of the distribution $p(\mathcal{C})$ are not easy to made. However, usually we have the idea of which *observables* (of which the values can be estimated from the data) are most relevant to the system. For example, one may not know the distribution of a random variable but its mean and variance are naturally considered important. Yet, we want to know about its *distribution*, and not just the mean values of such observables. Maximum entropy modelling helps bridging a set of mean values of observables and a parametric distribution.

Let us recall that an *observable* is simply a (measurable) function that maps Ω to $\mathbf{R}$,

$$\begin{aligned} Q: \quad & \Omega \to \mathbf{R} \\ & \mathcal{C} \mapsto Q(\mathcal{C}). \end{aligned} \tag{2.23}$$

The average of Q, with respect to the distribution of the system p is defined as

$$\langle Q \rangle = \sum_{\mathcal{C} \in \Omega} p(\mathcal{C}) Q(\mathcal{C}). \tag{2.24}$$

We will also sometimes refer to the average of an observable as the observable itself, which meaning we imply for the term 'observable' will be always clear from the contexts. In our case, the distribution $p(C)$ is unknown, but the average of an observable can be estimated from a *training dataset*,

$$\langle Q \rangle^D = \frac{1}{M} \sum_{\mu} Q(\mathcal{C}^\mu), \tag{2.25}$$

where the dataset contains M independently sampled states of the system, $D = \{\mathcal{C}_\mu\}$, with $\mu = 1, 2, \ldots, M$. Now, suppose we can point out K observables that are most relevant to the system, $\{Q_k\}$, with $k = 1, 2, \ldots, K$. These observables make up K corresponding

estimated averages,

$$\langle Q_k \rangle = \langle Q_k \rangle^D . \tag{2.26}$$

We will further assume that these K observables capture the key aspects of the system. We will refer to them as *sufficient statistics* of the system, and try to find a parametric distribution for the system that is consistent with the values of these observables.[4] There are still many distributions that are consistent with these constraints, which one should be chosen? The maximum entropy criterion says: one should choose from those the distribution that maximises the information entropy,

$$S[p(\mathcal{C})] = -\sum_{\mathcal{C}\in\Omega} p(\mathcal{C}) \ln p(\mathcal{C}). \tag{2.27}$$

Mathematically, we need to conditionally maximise the entropy (2.27) with constraints (2.26) and the normalisation condition,

$$\sum_{\mathcal{C}\in\Omega} p(\mathcal{C}) = 1. \tag{2.28}$$

This maximisation problem can be solved by introducing $K+1$ Lagrange multipliers for (2.26) and for (2.28). The solution then can be shown to take the form of the Boltzmann distribution,

$$p(\mathcal{C}) = \frac{1}{Z} \exp\{-H(\mathcal{C})\}, \tag{2.29}$$

with the Hamiltonian

$$H(\mathcal{C}) = -\sum_k \lambda_k Q_k(\mathcal{C}), \tag{2.30}$$

where the Lagrange multipliers $\{\lambda_k\}$ are found by solving the constraints (2.26), and the partition function Z is determined by the normalisation condition (2.28).

As an example, consider a system C consisting of N binary variables, $C = (\sigma_1, \sigma_2, \ldots, \sigma_N)$, of which the parametric distribution is unknown. Suppose we are most interested in the first moments $\langle \sigma_i \rangle^D$ (magnetisations) and the second moments $\langle \sigma_i \sigma_j \rangle^D$ (correlations), and higher moments are of less importance. Claiming that these observables are the sufficient statistics of the system, we can write down the maximum entropy distribution as

$$p(\mathcal{C}) = \frac{1}{Z} e^{-H(\mathcal{C})}, \tag{2.31}$$

[4] This is identical with the concept of sufficient statistics in statistics: we will see that the likelihood function for the model parameters will be the function of the average values of these observables, and only these observables.

with

$$H(\mathcal{C}) = -\sum_{(i,j)} J_{ij} s_i s_j - \sum_i h_i s_i, \tag{2.32}$$

where J_{ij} and h_i are the Lagrange multipliers determined by the constraints

$$\langle \sigma_i \rangle = \langle \sigma_i \rangle^D \tag{2.33}$$

$$\langle \sigma_i \sigma_j \rangle = \langle \sigma_i \sigma_j \rangle^D . \tag{2.34}$$

In that way, the pairwise interaction Ising Hamiltonian with couplings J_{ij} and fields h_i was derived from the choice of observables and the maximum entropy principle. The problem of inferring the couplings J_{ij} and fields h_i is from (2.33) and (2.33) is known as the inverse Ising problem, which we will discuss in detail in chapter 3.

We will refer to the Lagrange multipliers λ_k as *conjugate parameters* with respect to the corresponding observables. For example, in the inverse Ising problem, J_{ij} and h_i are conjugate parameters of $\langle \sigma_i \rangle$ and $\langle \sigma_i \sigma_j \rangle$, respectively. Often, the conjugate parameters have specific interpretations and the analysis of conjugate parameters reveals important information about the system. Take the inverse Ising problem for example. Although the connected correlation (or covariance) $\langle \sigma_i \sigma_j \rangle - \langle \sigma_i \rangle \langle \sigma_j \rangle$ somehow measures the dependency between two variables σ_i and σ_j, this kind of dependency can be indirect. Indeed, in the case of a pairwise interacting Hamiltonian, if both σ_i and σ_j interact strongly with a third spin σ_k, there can be a strong connected correlation between σ_i and σ_j despite there might be no direct coupling between them. This *direct coupling*, or *direct dependency*, is measured by the conjugate parameters J_{ij}; more precisely, if $J_{ij} = 0$, then σ_i and σ_j are independent given all the other variables.

Maximum entropy reasoning and its relation to statistical physics date back to 1957 with the seminal papers by Jaynes (Jaynes 1957a,b). Since then there has been a lot of development in applications of maximum entropy models in physics and engineering (Jaynes 1989, Pressé et al. 2013, Shore and Johnson 1980). Here, we discuss various linear models in their relation to the maximum entropy reasoning. We will find that the maximum entropy reasoning provides a comprehensive unified view for the many different models, which formally map to problems in statistical physics.

2.3.1 Gaussian model

Consider the case where the configuration C consists of N real valued random components, $C = \{X_i\}$, where $i = 1, 2, \ldots, N$. The simplest sufficient statistics in this case could be the means $\langle X_i \rangle$ and the correlations $\langle X_i X_j \rangle$ (including the "self-correlations" $\langle X_i X_i \rangle$).

The maximum entropy model defined by these observables is

$$H(x_1, x_2, \ldots, x_N) = -\sum_i b_i x_i + \frac{1}{2}\sum_{ij} A_{ij} x_i x_j, \tag{2.35}$$

where b_i and A_{ij} are conjugate parameters. The signs and the factor of $1/2$ are for the sake of convenience. Note that we have the symmetry $\langle X_i X_j \rangle = \langle X_j X_i \rangle$, which implies a symmetric constraint on $\{A_{ij}\}$, $A_{ij} = A_{ji}$. In similarity to the Ising model discussed in the previous section, A_{ij} measures the strength of direct interaction or direct dependency between variable X_i and X_j.

For the sake of convenience, we introduce a simple graphical presentation for this set of observables in figure 2.1. Each node represents a component variable. All the observables involved one variable (e.g. $\propto x_i$, $\propto x_i^2$) are implicit. An observable involved two variables (e.g. $\propto x_i x_j$) is represented by a link that connects the two variables. Our model in this case is represented by a fully connected graph.

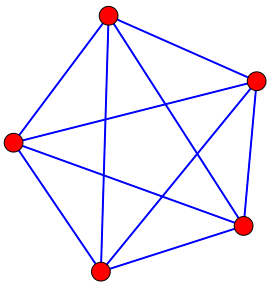


Figure 2.1: **A graphical presentation of Gaussian model.** Each node represents a variable. All the observables involved one variable (e.g. $\propto x_i$, $\propto x_i^2$) are implicit. An observable involved two variables (e.g. $\propto x_i x_j$) is represented by a link that connects the two variables. The Gaussian model is represented by a fully connected graph.

The corresponding Boltzmann distribution of the Hamiltonian (2.35) is nothing but the *multivariate Gaussian distribution*,

$$p(\{x_i\}) = \frac{1}{Z} \exp\left\{ +\sum_i b_i x_i - \frac{1}{2}\sum_{ij} A_{ij} x_i x_j \right\}, \tag{2.36}$$

with the partition function

$$Z = \frac{(2\pi)^{N/2}}{[\det A]^{1/2}} \exp\left\{\frac{1}{2}\sum_{ij}(A^{-1})_{ij}b_i b_j\right\}. \tag{2.37}$$

From the partition function, the averages $\langle X_i \rangle$ and $\langle X_i X_j \rangle$ (or the covariance $\langle X_i X_j \rangle - \langle X_i \rangle \langle X_j \rangle$) can then be easily obtained as

$$\langle X_i \rangle = \frac{1}{Z}\frac{\partial Z}{\partial b_i} \text{ and } \langle X_i X_j \rangle = \frac{1}{Z}\frac{\partial^2 Z}{\partial b_i \partial b_j}, \tag{2.38}$$

which yield

$$\langle X_i \rangle = (Ab)_i, \tag{2.39}$$

$$\langle X_i X_j \rangle - \langle X_i \rangle \langle X_j \rangle = (A^{-1})_{ij}. \tag{2.40}$$

The conjugate parameters b_i and A_{ij} are determined by solving (2.39) and (2.40) with the averages on lefthand sides estimated from data. The solutions are

$$b = \Sigma^{-1}\mu, \tag{2.41}$$

$$A = \Sigma^{-1}, \tag{2.42}$$

where we have denoted the *sample means* and the *sample covariances* by $\mu_i = \langle X_i \rangle^D$ and $\Sigma_{ij} = \langle X_i X_j \rangle^D - \langle X_i \rangle^D \langle X_j \rangle^D$.

Unfortunately, equation (2.42) often suffers from over-fitting: the sample covariance matrix Σ is usually singular. Indeed, it is easy to show that $\text{rank}(\Sigma) \leq \min\{M-1, N\}$, where M is the number of samples in the dataset, which is often smaller than N. There are several known ways to get around this difficulty, the following additional remark presents a simple one – the method of *pseudo-counts*.

Remark: ***Model regularisation: pseudo-counts***

Imagine that we appended priorly to the real training dataset D (which we used to estimate the means μ and the covariances Σ) another artificial dataset D_0. Now let us estimate the average observables from the combined data set $D \cup D^0$,

$$\langle X_i \rangle^{D \cup D_0} = (1-\gamma)\langle X_i \rangle^D + \gamma \langle X_i \rangle^{D_0}, \tag{2.43}$$

$$\left\langle X_i X_j \right\rangle^{D \cup D_0} = (1-\gamma)\left\langle X_i X_j \right\rangle^D + \gamma \left\langle X_i X_j \right\rangle^{D_0}, \tag{2.44}$$

where $\gamma = M_0/(M_0 + M)$ with M_0 and M denoting the numbers of samples in the datasets D_0

and D, respectively. Note that when γ tends to 1, the estimations are driven by dataset D_0, and when γ tends to 0, the estimations are driven by dataset D.

Suppose we have some idea of how the parameters $\{b_i\}$ and $\{A_{ij}\}$ should be, say around $\{b_i^0\}$ and $\{A_{ij}^0\}$. We then assume that dataset D_0 was designed so that it gives the exact estimates for the means $\langle X_i \rangle^{D_0}$ and the covariances $\left\langle X_i X_j \right\rangle^{D_0}$ corresponding to this set of parameters. Therefore, at $\gamma = 1$, with the estimated averages (2.43) and (2.44) driven by D_0, equations (2.39) and (2.40) have the solutions $b_i = b_i^0$ and $A_{ij} = A_{ij}^0$ (therefore they are *not singular*). Now imagine we reduce γ down to some intermediate values between 1 and 0, keeping the equations non-singular. As the data averages are compromises of the two datasets D and D_0, as seen from equations (2.43) and (2.44), the solution now is expected to be a compromise between our *prior guess* b^0 and A^0 and the *suggested solution* by the real dataset D. Thus, the singularity is remedied, and we still hopefully retain the properties of the solution suggested by D. In practice, we usually have to guess b_i^0 and A^0, and choose γ small enough so that the particular values of our guess are (hopefully) not so important.

2.3.2 Linear regression

In a regression problem, we are interested in the response of a variable Y, which is referred to as *response*, with respect to the other N variables $\{X_i\}$, which are referred to as *signals*. In this case, we are interested in systems with sampling space $\mathbf{R}^{N+1}$, $C = (Y; \{X_i\})$, where $i = 1, 2, \ldots, N$.

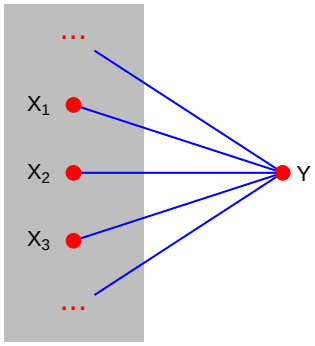


Figure 2.2: **Graphical presentation of a linear regression model.** Each link represents a two-variable observable (coupling) between a component of the signals and the response. The signals are imbedded in a box, indicating the model for the signals is implicit.

We now define our maximum entropy model by specifying a set of sufficient statistics. These are the mean of the response $\langle Y \rangle$, the second moment of the response $\langle Y^2 \rangle$ (which determines the variance $\langle Y^2 \rangle - \langle Y \rangle^2$), and the covariances of the response with all the signals $\langle Y X_i \rangle$, where $i = 1, 2, \ldots, N$. With those observables, the maximum entropy

Hamiltonian is partly specified,

$$H(y;\{x_i\}) = H^X(\{x_i\}) - \sum_i b_i x_i y - ay + \frac{1}{2}cy^2, \tag{2.45}$$

where we introduced the conjugate parameters a, $\{b_i\}$ and c. The signs and the factor $1/2$ are for the sake of convenience. The term $H^X(\{x_i\})$ is meant to model the signals $\{X_i\}$ (jointly) in a way we do not have to specify now; the point is that it does not depend on a, $\{b_i\}$, and c.

The corresponding Boltzmann distribution is

$$p(y;\{x_i\}) = \frac{1}{Z}\exp\left\{-H^X(\{x_i\}) + \sum_i b_i x_i y + ay - \frac{1}{2}cy^2\right\}, \tag{2.46}$$

where Z is the partition function.

If we define $\alpha = a/c$, $\beta_i = b_i/c$ and $\sigma^2 = 1/c$, it is straightforward to show that

$$p(y|\{x_i\}) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y - \alpha - \sum_i \beta_i x_i)^2}{2\sigma^2}\right\}, \tag{2.47}$$

which is the statistical model defining the linear regression.

Temporally keeping the old variables a, b and c_i, we come back to (2.45) to calculate the partition function,

$$Z = \int_{\mathbf{R}^N}\prod_i \mathrm{d}\, x_i \left(\frac{2\pi}{c}\right)^{1/2}\exp\left\{-H^X(\{x_i\}) + \frac{(a + \sum_i b_i x_i)^2}{2c}\right\}. \tag{2.48}$$

Note that the partition function $Z(a, b, \{c_i\})$ is the generating function of observables $\langle Y\rangle$, $\left\langle Y^2\right\rangle$ and $\langle YX_i\rangle$,

$$\langle Y\rangle = \frac{1}{Z}\frac{\partial Z}{\partial a},\ \left\langle\frac{Y^2}{2}\right\rangle = -\frac{1}{Z}\frac{\partial Z}{\partial c} \text{ and } \langle YX_i\rangle = \frac{1}{Z}\frac{\partial Z}{\partial b_i}. \tag{2.49}$$

Calculating these derivatives of Z using (2.48), after some algebraic manipulations, we arrive at

$$\langle YX_i\rangle - \langle Y\rangle\langle X_i\rangle = \sum_j \beta_j(\left\langle X_i X_j\right\rangle - \langle X_i\rangle\left\langle X_j\right\rangle), \tag{2.50}$$

and

$$\alpha = \langle Y \rangle - \sum_i \beta_i \langle X_i \rangle , \tag{2.51}$$

$$\sigma^2 = \left\langle \left(Y - \sum_i \beta_i X_i \right)^2 \right\rangle , \tag{2.52}$$

where we used the newly defined variables $\alpha = a/c$, $\beta_i = b_i/c$, $\sigma^2 = 1/c$.

The averages $\langle Y \rangle$, $\langle Y X_i \rangle$ in (2.50) and (2.51) are estimated from data, as demanded by the maximum entropy reasoning. But solving for α, $\{\beta_i\}$ also requires observables such as $\langle X_i \rangle$ and $\langle X_i X_j \rangle$. Taking the values estimated from data for these observables, from the maximum entropy reasoning point of view, is equivalent to model the signal X_i with some distribution that contains *at least* the first and second moments of the signals, namely $\langle X_i \rangle$ and $\langle X_i X_j \rangle$, as sufficient statistics observables. However, taking the averages $\langle X_i \rangle$ and $\langle X_i X_j \rangle$ from the data like that does not mean that the model for the signals is Gaussian. A proper model for the signals can include higher moments, such as $\langle X_i X_j X_k \rangle$, independently; the point is that those deviations from a Gaussian model are *not relevant* to the linear response Y.[5] On the other hand, if for some reason, we know the exact model of the signals, especially when it is simple, we can directly use it to calculate $\langle X_i \rangle$ and $\langle X_i X_j \rangle$; this results in the so-called *mixture models*, which will also be discussed later in this chapter.

Regression with a categorical variable

So far we assumed that the signal variables $\{X_i\}$ are continuous, so that observables of the forms $\langle X_i \rangle$, $\langle X_i X_j \rangle$ make sense. Now suppose we want to study the dependency of the heights of people in a population on their nutrient conditions and on their origins (European, American, African...) The origins of the people are examples of *categorical variables*. Such a categorical variable can take some K different values, but no addition and multiplication can be defined among them.

We will split the categorical variable from the other variables to treat it differently, so that the joint configuration is written as $C = (Y; \Phi, \{X_i\})$, where Φ stands for the categorical component of K levels taking values in $\{1, 2, \ldots, K\}$ while Y and X_i are real-valued random variables.

If Φ has only two levels, observables such as $\langle \Phi \rangle$, $\langle X_i \Phi \rangle$ and $\langle Y \Phi \rangle$ still make sense. Therefore, the linear regression model discussed above still applies. Here we are con-

[5] If we insist a Gaussian model for the signal, we found ourselves back at the Gaussian model for the joint response-signal configuration $(Y, \{X_i\})$.

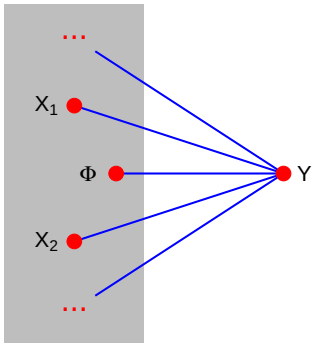


Figure 2.3: **Graphical presentation of linear regression with categorical variables.** This is similar to graphical presentation of linear regression in general. The signal Φ is separated to indicate its categorical nature.

cerned with the cases where Φ has more than 2 levels. In these cases, no multiplication or summation can be defined among the categorical variables and between the categorical variables and the other continuous variables. In replacing for non-sensical observables such as $\langle Y\Phi\rangle$, we simply take observables $\langle\delta(\Phi,t)Y\rangle$ with $t = 1,2,\ldots,K$. Overall, the set of observables now consists of $\langle YX_i\rangle$, $\langle Y\delta(\Phi,t)\rangle$ and $\langle Y^2\rangle$ with $t = 1,2,\ldots,K$ and $i = 1,2,\ldots,N$. Note that we do not need to include the observable $\langle Y\rangle$ anymore because it would be redundant since $\langle Y\rangle = \sum_t \langle\delta(\Phi,t)Y\rangle$. The Hamiltonian for the model can now be written down directly,

$$H(y;\phi,\{x_i\}) = H^X(\phi,\{x_i\}) - \sum_i b_i x_i y - \sum_t a_t\delta(\phi,t)y + \frac{1}{2}cy^2. \tag{2.53}$$

Note that we again left the part related to the signals alone H^X unspecified. Going through similar calculations as above, starting with the partition function

$$Z = \sum_\phi \int_{\mathbf{R}^N} \prod_i \mathrm{d}\,x_i \left(\frac{2\pi}{c}\right)^{1/2} \exp\left\{-H^X(\phi,\{x_i\}) + \frac{(\sum_t a_t\delta(\phi,t) + \sum_i b_i x_i)^2}{2c}\right\}, \tag{2.54}$$

with

$$\langle YX_i\rangle = \frac{1}{Z}\frac{\partial Z}{\partial b_i},\ \langle Y\delta(\Phi,t)\rangle = \frac{1}{Z}\frac{\partial Z}{\partial a_t} \text{ and } \left\langle\frac{Y^2}{2}\right\rangle = -\frac{1}{Z}\frac{\partial Z}{\partial c}, \tag{2.55}$$

we can show that, in similarity to (2.50),

$$\langle YX_i\rangle - \sum_t \frac{\langle\delta(\Phi,t)Y\rangle\langle\delta(\Phi,t)X_i\rangle}{\langle\delta(\Phi,t)\rangle} = \sum_j \beta_j \left(\langle X_iX_j\rangle - \sum_t \frac{\langle\delta(\Phi,t)X_i\rangle\langle\delta(\Phi,t)X_j\rangle}{\langle\delta(\Phi,t)\rangle}\right), \tag{2.56}$$

and in similarity to (2.51) and (2.52),

$$\alpha_t = \frac{\langle Y\delta(\Phi,t)\rangle}{\langle\delta(\Phi,t)\rangle} - \sum_i \beta_i \frac{\langle X_i\delta(\Phi,t)\rangle}{\langle\delta(\Phi,t)\rangle}, \tag{2.57}$$

$$\sigma^2 = \left\langle \left(Y - \sum_i \beta_i X_i - \sum_t a_t \delta(\Phi,t) \right)^2 \right\rangle, \tag{2.58}$$

where we used $\alpha_t = a_t/c$, $\beta_i = b_i/c$ and $\sigma^2 = c$.

We see that with the presence of the categorical variable Φ, the equations for the linear regression parameters maintain their forms with a modified definition of covariances. Moreover, we see that the effect of a categorical variable of K levels in a linear regression should be thought of as the effect of introducing K different constant intercepts in the response axis (instead of a single constant as in the linear regression with continuous variables).

Note that the standard way to treat categorical variable in linear regression is to do the so-called *dummy coding* procedure. Dummy coding works by replacing one categorical variable by $K-1$ binary variables; and then we are back at the linear regression with categorical variables. The interpretation of coefficients in linear regression with dummy coding needs to be taken with care, however. Within the maximum entropy framework, we have introduced a more direct way of treating categorical variables with simpler interpretation.

REMARK: *Model regularisations: ridge and lasso*

In linear regression, the key equation to be solved for $\{\beta_i\}$ is

$$r = \Sigma\beta, \tag{2.59}$$

with

$$r_i = \langle YX_i\rangle - \langle Y\rangle\langle X_i\rangle, \tag{2.60}$$

$$\Sigma_{ij} = \left\langle X_iX_j\right\rangle - \langle X_i\rangle\left\langle X_j\right\rangle, \tag{2.61}$$

if the signals have no categorical component, or

$$r_i = \langle YX_i\rangle - \sum_t \frac{\langle\delta(\Phi,t)Y\rangle\langle\delta(\Phi,t)X_i\rangle}{\langle\delta(\Phi,t)\rangle}, \tag{2.62}$$

$$\Sigma_{ij} = \left\langle X_iX_j\right\rangle - \sum_t \frac{\langle\delta(\Phi,t)X_i\rangle\left\langle\delta(\Phi,t)X_j\right\rangle}{\langle\delta(\Phi,t)\rangle}, \tag{2.63}$$

if there is a categorical component in the signals.

Noting that the generalised covariance matrix Σ is a positive matrix, then the solution to the

linear equation (2.59) can be written as an optimisation problem,

$$\beta^* = \underset{\beta}{\operatorname{argmin}}\, S(\beta), \tag{2.64}$$

with the objective function

$$S(\beta) = \frac{1}{2}\beta^t \Sigma \beta - r^t \beta. \tag{2.65}$$

Equation (2.59) can be found back by setting the derivatives of S with respect to β_i to zero. Provided the covariance matrix Σ is invertible, equation (2.59) has the solution

$$\beta^* = \Sigma^{-1} r, \tag{2.66}$$

which is also the minimiser of (2.65).

As we already discussed in the context of Gaussian models, due to the insufficiency of training data, the covariance matrix Σ is usually degenerate or close to degenerate. The consequence of this degeneracy is that equation (2.59) no longer has one unique solution. In the space of β (which is $\mathbf{R}^N$), the set of solutions expands over a linear manifold (a linear subspace plus a constant vector) in the degenerate directions of Σ. Along these degenerate directions, which span to infinity, the objective function $S(\beta)$ is flat. To avoid this flat stationary set, a common procedure is to deform the objective function by a *regularisation*. Note that although regularisations such as ridge or lasso can be thought as special choices of prior distributions for the model parameters, such choices are poorly motivated. For that reason, we will call them 'model regularisations', instead of 'priors for parameters'.

In *ridge regularisation*, the regularisation term is proportional to the Euclidean norm (l_2-norm) of β. The objective function to be minimised then changes to

$$\tilde{S}(\beta) = S(\beta) + \frac{\gamma}{2}\sum_i \beta_i^2, \tag{2.67}$$

with some positive γ. As this new objective function is still quadratic in β_i, the minimisation is straightforward. By inspecting the derivatives of $\tilde{S}(\beta)$, we find that the stationary equation (2.59) changes accordingly,

$$r = (\Sigma + \gamma I)\beta. \tag{2.68}$$

As $\Sigma + \gamma I$ is non-degenerate for any arbitrary positive γ, the solution is uniquely determined,

$$\beta^* = (\Sigma + \gamma I)^{-1} r. \tag{2.69}$$

Such a parabolic regularisation directly deforms the flat solution set, shrinking the solution towards zero (*shrinkage*).

In *lasso regularisation*, the regularisation term is proportional to the l_1-norm of β. Now the objective function is

$$\tilde{S}(\beta) = S(\beta) + \gamma \sum_i |\beta_i|, \tag{2.70}$$

with some positive γ. This problem is still convex, which means standard optimisation techniques such as gradient descent can be applied directly. However it is no longer quadratic, so the computation can be rather tedious. But there came a surprising news: under some rather general conditions, there exists a simple algorithm to recover the whole solution of the problem for all γ

with the same computational expense as that of ridge regression (Efron et al. 2004, Hastie et al. 2009).

Similar to ridge regression, lasso regression is also a shrinkage procedure. But it has another very important property: *subset-selection* (Hastie et al. 2009, Tibshirani 1996). For some large enough value of γ, it may happen in regression with lasso that one or several components of β may vanish *exactly*. In effect, we retain *only* components of the signals that are relevant to the response. This is very important in practice for understanding the phenomena under investigation (Hastie et al. 2009, Tibshirani 1996).

2.3.3 Logistic regression

In linear regression, the configuration $C = (Y; \{X_i\})$ has a continuous response variable Y. What if the response variable is binary or categorical? For simplicity, we assume no categorical component in the signals $\{X_i\}$.

Binary logistic regression

In this case $y \in \{-1, +1\}$. Now $\langle Y^2 \rangle$ is no longer informative because it is just identically 1. The set of observables is reduced to the mean of the response $\langle Y \rangle$, and the correlations with the signals $\langle YX_i \rangle$, where $i = 1, 2, \ldots, N$. The Hamiltonian of the model is then

$$H(y; \{x_i\}) = H^X(\{x_i\}) - \sum_i b_i x_i y - ay, \tag{2.71}$$

in similarity to (2.45).

It is also easy to derive the probability of the response given the signals,

$$P(Y = +1 | \{X_i = x_i\}) = \frac{1}{1 + \exp\{+\alpha + \sum_i \beta_i x_i\}}, \tag{2.72}$$

$$P(Y = -1 | \{X_i = x_i\}) = \frac{1}{1 + \exp\{-\alpha - \sum_i \beta_i x_i\}}, \tag{2.73}$$

where $\alpha = 2a$, $\beta_i = 2c_i$, which identify the model with the logistic regression model (Bishop 2006).

Coming back to the Hamiltonian (2.71), keeping in mind that $y = \{-1, +1\}$, the corresponding partition function can be shown to be

$$Z = \int_{\mathbf{R}^N} \prod_i \mathrm{d}\, x_i e^{-H^X(\{x_i\})} 2 \operatorname{ch}(a + \sum_j b_j x_j). \tag{2.74}$$

Note that

$$\langle YX_i \rangle = \frac{\partial Z}{\partial c_i} \text{ and } \langle Y \rangle = \frac{\partial Z}{\partial a}, \tag{2.75}$$

the self-consistent equations can be written easily,

$$\langle Y \rangle = \left\langle \text{th}(a + \sum_j b_j X_j) \right\rangle, \tag{2.76}$$

$$\langle Y X_i \rangle = \left\langle \text{th}(a + \sum_j b_j X_j) X_i \right\rangle. \tag{2.77}$$

Unfortunately, with the averages taken from data, this system of equations for a and $\{b_i\}$ is non-linear and difficult to solve. In practice, it easier turning the problem to an optimisation problem and then apply various optimisation techniques. We will not go into the details of these technical points, which are covered in Hastie et al. 2009, for example.

Remark: *The pseudo-likelihood method for the inverse Ising problem*

In the inverse Ising problem, we want to estimate the couplings J_{ij} and the fields h_i in a pairwise interacting Ising Hamiltonian of N spins,

$$H(\{s_i\}) = -\sum_{(ij)} J_{ij} s_i s_j - \sum_i h_i s_i, \tag{2.78}$$

on the basis of a dataset of some M sampled configurations of the spins, $D = \{C^\mu\}$ with $\mu = 1, 2, \ldots, M$, where $C^\mu = (s_1^\mu, s_2^\mu, \ldots, s_N^\mu)$. The maximum likelihood reasoning requires that J_{ij} and h_i satisfy a system of equations,

$$\langle \sigma_i \rangle = \langle \sigma_i \rangle^D, \tag{2.79}$$

$$\left\langle \sigma_i \sigma_j \right\rangle = \left\langle \sigma_i \sigma_j \right\rangle^D. \tag{2.80}$$

which identify the model averages $\langle \sigma_i \rangle$ and $\left\langle \sigma_i \sigma_j \right\rangle$ with their corresponding sample averages.

In the next chapter, we will discuss various methods to solve this problem. Here, in the context of regression models, we explain the idea of the pseudo-likelihood approach (Besag 1975, Ravikumar et al. 2010), which has become attractive because of its robustness against sampling noise (Aurell and Ekeberg 2012).

The idea is rather simple. Imagine we pick a spin σ_i, and think of it as a regression function of the other spins $\{\sigma_j\}$ ($j \neq i$). The regression reveals the coupling strength J_{ij} of the spin σ_i to the others σ_j for $j \neq i$. The response, spin i, is binary, so the problem is a binary logistic regression. The inverse Ising problem of N spins is reduced to N regression problems and can benefit from various developed regression techniques, in particular subset-selection techniques such as lasso.

If we apply a subset selection regularisation such as lasso to the problem, those spins j that do not directly interact with spin i are expected to be excluded, $J_{ij} = 0$. For every spin, we can point out the set of spins that interact with i, together with the strength of interaction J_{ij}. Note that the estimated couplings J_{ij} are in general asymmetric, $J_{ij} \neq J_{ji}$. Combining the two values is a little bit arbitrary, one can take their average, for example. In practice, this simple idea works surprisingly well in reconstructing the network of interactions of an Ising model (Aurell and Ekeberg 2012, Ravikumar et al. 2010).

As we discussed, solving equations (2.86) and (2.87), or equivalently maximising the corresponding likelihood function, is hard. Adding lasso regularisation to the minimisation makes the problem even harder. The problem is however still convex and there are recent new techniques to solve (Aurell and Ekeberg 2012, Friedman et al. 2010, Ravikumar et al. 2010).

Categorical logistic regression

This picture of logistic regression can be easily extended to the case where Y is a categorical response of K levels. We will use the symbol Φ for the response to emphasise its categorical nature. In this case, the natural set of observables can be $\langle \delta(\Phi, t) \rangle$ and $\langle \delta(\Phi, t) X_i \rangle$, where $t = 1, 2, \ldots, K$ and $i = 1, 2, \ldots, N$, leading to the Hamiltonian

$$H(\phi; \{x_i\}) = H^X(\{x_i\}) - \sum_t \sum_i b_i^t x_i \delta(\phi, t) - \sum_t a_t \delta(\phi, t). \tag{2.81}$$

Note that since the set of observables $\langle \delta(\Phi, t) \rangle$ are not independent, namely we have a constraint $\sum_t \langle \delta(\Phi, t) \rangle = 1$, we will need to set a constraint on the set of conjugate variables $\{a_t\}$ in order for them to be completely determined.

The corresponding conditional probability of the response given the signals in this case can be found to be

$$P(\Phi = \phi | \{X_i = x_i\}) = \frac{\exp\{\sum_t \sum_i b_i^t x_i \delta(\phi, t) + \sum_t a_t \delta(\phi, t)\}}{\sum_{\phi'} \exp\{\sum_t \sum_i b_i^t x_i \delta(\phi', t) + \sum_t a_t \delta(\phi', t)\}}. \tag{2.82}$$

The formula looks more complicated that it actually is: the summation over delta Kronecker can be performed directly,

$$P(\Phi = \phi | \{X_i = x_i\}) = \frac{\pi_\phi \exp\{\sum_i b_i^\phi x_i\}}{\sum_{\phi'} \pi_{\phi'} \exp\{\sum_i b_i^{\phi'} x_i\}}, \tag{2.83}$$

where we denoted $\pi_\phi \propto e^{a_\phi}$ (the exact proportional coefficient is irrelevant).

The self-consistent equations for the multivariate logistic regression can also be found easily. Note that the partition function is

$$Z = \int_{\mathbf{R}^N} \prod_i \mathrm{d}\, x_i e^{-H^X(\{x_i\})} \sum_\phi e^{\sum_i b_i^\phi x_i + a_\phi}, \tag{2.84}$$

and

$$\langle \delta(\Phi, t) \rangle = \frac{1}{Z} \frac{\partial Z}{\partial a_t} \text{ and } \langle \delta(\Phi, t) X_i \rangle = \frac{1}{Z} \frac{\partial Z}{\partial b_i^t}, \tag{2.85}$$

we can show that

$$\langle \delta(\Phi, t) \rangle = \left\langle \frac{\pi_t e^{\sum_i b_i^t X_i}}{\sum_{t'} \pi_{t'} e^{\sum_i b_i^{t'} X_i}} \right\rangle, \tag{2.86}$$

$$\langle \delta(\Phi, t) X_i \rangle = \left\langle X_i \frac{\pi_t e^{\sum_i b_i^t X_i}}{\sum_{t'} \pi_{t'} e^{\sum_i b_i^{t'} X_i}} \right\rangle, \tag{2.87}$$

where $\pi_t \propto e^{a_t}$. This is again a system of non-linear equations, techniques to solve these equations can be found in textbooks about logistic regressions, e.g. Hastie et al. 2009.

Note that in all regression models we did not explicitly specify the distribution of the signals H^X, but matching of certain observables with the data is required. As discussed, this may help avoiding a wrong model for the signals. On the other hand, it may result in complicated non-linear equations for the parameters, such as (2.86) and (2.87). Explicitly introducing a model for signals can help simplifying these equations. This is the advantage of using *mixture models* over *logistic regressions*.

2.3.4 Mixture model classification

Gaussian mixture model

In the model for categorical logistic classification (2.81), let us assume that H^X is Gaussian. The Hamiltonian of the model is modified to[6]

$$H(\phi; \{x_i\}) = \frac{1}{2} \sum_{ij} A_{ij} x_i x_j - \sum_t \sum_i b_i^t x_i \delta(\phi, t) - \sum_t a_t \delta(\phi, t). \tag{2.88}$$

Note that we do not need to include the mean of X in the set of observables, since they can be directly derived from $\langle \delta(\Phi, t) X_i \rangle$,

$$\langle X \rangle = \sum_t \langle \delta(\Phi, t) X_i \rangle . \tag{2.89}$$

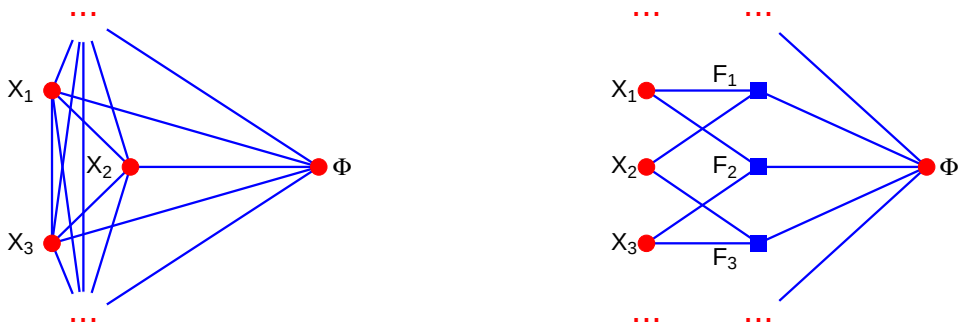
In this case, the conditional probability density of $\{X_i\}$ given $\Phi = t$ can be found to be

$$P(\{x_i\} | \Phi = t) = \frac{[\det A]^{1/2}}{[2\pi]^{N/2}} \exp \left\{ + \sum_i b_i^t x_i - \frac{1}{2} \sum_{ij} A_{ij} x_i x_j \right\}. \tag{2.90}$$

This equation can be interpreted as if the configurations $\{X_i = x_i\}$ are generated from K different Gaussian models, each with different mean (or conjugate parameters b_i^t) but

[6]Here we implicitly suppose that the signals X_i are continuous, so that a Gaussian distribution can be assumed. When the signals X_i are binary, we will have a mixture of Ising models, which will be discussed in the next section. Note that logistic regression however applies to both: whether X_i are continuous or binary, the logistic regression still takes the same form; the only thing that matters is that Y is binary/categorical.

shared covariance (or conjugate parameters A^{-1}); thus the name *Gaussian mixture model*. Strictly, a Gaussian mixture model allows the covariance matrices of the Gaussian component distribution to depend on t as well; although this can be done by introducing observables such as $\langle \delta(Y,t) X_i X_j \rangle$, we avoid that complication here.



(a) (b)

Figure 2.4: **Graphical presentation of Gaussian mixture models.** (a) Gaussian mixture model with shared covariance. The signals are now explicitly modelled by a Gaussian distribution. Overall, the graph is a fully connected graph. (b) Gaussian mixture model with different covariance matrices for different Gaussian components. The three-variable observables are introduced (squared nodes).

Now, the partition function (2.84) can be evaluated,

$$Z = \sum_t e^{a_t} \frac{[2\pi]^{N/2}}{[\det A]^{1/2}} \exp\left\{ \frac{1}{2} \sum_{ij} (A^{-1})_{ij} b_i^t b_j^t \right\}. \tag{2.91}$$

Then equations (2.86) and (2.87) can also be written down explicitly in terms of the newly introduced parameters A_{ij},

$$\langle \delta(\Phi,t) \rangle = \frac{e^{a_t} e^{\frac{1}{2}\sum_{ij}(A^{-1})_{ij} b_i^t b_j^t}}{\sum_{t'} e^{a_{t'}} e^{\frac{1}{2}\sum_{ij}(A^{-1})_{ij} b_i^{t'} b_j^{t'}}}, \tag{2.92}$$

$$\langle \delta(\Phi,t) X_i \rangle = \frac{e^{a_t} e^{\frac{1}{2}\sum_{ij}(A^{-1})_{ij} b_i^t b_j^t}}{\sum_{t'} e^{a_{t'}} e^{\frac{1}{2}\sum_{ij}(A^{-1})_{ij} b_i^{t'} b_j^{t'}}} \sum_j (A^{-1})_{ij} b_j^t. \tag{2.93}$$

These equations look rather complicated, but if we set

$$\pi_t = \frac{e^{a_t} e^{\frac{1}{2}\sum_{ij}(A^{-1})_{ij} b_i^t b_j^t}}{\sum_{t'} e^{a_{t'}} e^{\frac{1}{2}\sum_{ij}(A^{-1})_{ij} b_i^{t'} b_j^{t'}}}, \tag{2.94}$$

which is nothing but the probability of $\Phi = t$, marginalised over $\{X_i\}$, that is $\pi_t = \langle \delta(\Phi, t) \rangle$, equation (2.93) becomes fairly simple,

$$\frac{\langle \delta(\Phi, t) X_i \rangle}{\langle \delta(\Phi, t) \rangle} = (Ab^t)_i. \tag{2.95}$$

This resembles equation (2.39), which relates the means with its conjugate partners in a Gaussian model.

Finally, we need to find the equations for A_{ij}. Because

$$\langle \delta(\Phi, t) X_i \rangle = \frac{1}{Z} \frac{\partial Z}{\partial b_i^t}, \text{ then } \langle \delta(\Phi, t) X_i X_j \rangle = \frac{1}{Z} \frac{\partial^2 Z}{\partial b_i^t \partial b_j^t},$$

so

$$\langle X_i X_j \rangle = \sum_t \frac{1}{Z} \frac{\partial^2 Z}{\partial b_i^t \partial b_j^t}, \tag{2.96}$$

where we have used $\langle X_i X_j \rangle = \sum_t \langle \delta(\Phi, t) X_i X_j \rangle$. From that we find

$$(A^{-1})_{ij} = \langle X_i X_j \rangle - \sum_t \frac{\langle \delta(\Phi, t) X_i \rangle \langle \delta(\Phi, t) X_j \rangle}{\langle \delta(\Phi, t) \rangle}. \tag{2.97}$$

This equation, together with (2.95), completes the necessary equations to determine the model parameters b_i^t and A_{ij}. The Gaussian mixture model has a significantly larger number of parameters in comparison to the logistic regression, namely $N(N+1)/2$ elements of A_{ij}, which may cause over-fitting.[7] In particular, equation (2.97) is usually degenerate and cannot be inverted to find A_{ij}; regularisation of (2.97) is often needed.

Paramagnetic Ising mixture model

The other example where the equations for parameters in the logistic regression (2.86) and (2.87) can be simplified by introducing an explicit model for signals is the paramagnetic Ising mixture model. In this case, we are dealing with binary signals, $\sigma_i \in \{-1, +1\}$. Note that we denote binary signals by σ_i instead of X_i, which denote continuous signals. The simplest Hamiltonian for the signals (so that the model can be solved analytically) is the paramagnetic Ising model. Note that including observables $\langle \sigma_i \rangle$ into the model for the categorical logistic regression (2.81) is redundant because $\sum_t \langle \delta(\Phi, t) \sigma_i \rangle = \langle \sigma_i \rangle$. That

[7] This comparison is actually not quite fair. A mixture model has no implicit observation or implicit parameter, whereas in logistic regressions, over-fitting can come from the implicit model for the signals.

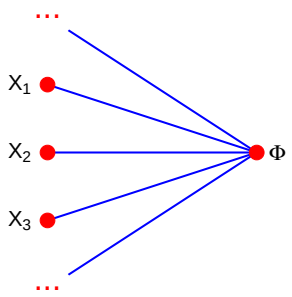


Figure 2.5: **Graphical presentation of paramagnetic Ising mixture model.** The signals now are explicitly modelled by a non-interacting Ising model (paramagnetic model).

means, the total Hamiltonian is simply

$$H(\phi, \{s_i\}) = -\sum_t \sum_i b_i^t s_i \delta(\phi, t) - \sum_t a_t \delta(\phi, t). \tag{2.98}$$

This Hamiltonian describes a system of N Ising spins σ_i coupled to a Potts spin Φ. For this system, the partition function (2.84) (with the integral replaced by the sum over discrete values $\{+1, -1\}$) can also be found explicitly,

$$Z = \sum_\phi e^{a_\phi} \prod_i 2\,\mathrm{ch}(b_i^\phi). \tag{2.99}$$

This partition function results in simple equations for the model parameters

$$\langle \delta(\Phi, t) \rangle = \pi_t, \tag{2.100}$$

$$\frac{\langle \delta(\Phi, t) X_i \rangle}{\langle \delta(\Phi, t) \rangle} = \mathrm{th}(b_i^t), \tag{2.101}$$

where

$$\pi_t = \frac{e^{a_t} \prod_i 2\,\mathrm{ch}(b_i^t)}{\sum_{t'} e^{a_{t'}} \prod_i 2\,\mathrm{ch}(b_i^{t'})}. \tag{2.102}$$

Equations (2.100) and (2.101) with the left-hand sides estimated from data, can be solved directly for π_t (or equivalently a_t) and b_i^t. We will make use of this model to classify lung cancer samples into different types based on molecular biological data in chapter 6.

2.4 Beyond parametric and probabilistic models

This chapter was devoted to probabilistic models, and in particular parametric models. It would be however incomplete without mentioning non-parametric models, or more generally, non-parametric analytical methods, as well.

Non-parametric methods refer to those that do not assume an explicit form of the distribution for the system under consideration, instead, they rely on general features of probability measures. Examples of non-parametric analytical methods are Spearman correlation, sign test, or rank test (Higgins 2004).

There is also a great effort in developing Bayesian non-parametric models. The non-parametric Bayesian approach is a natural extension of Bayesian modelling discussed in this chapter: instead of assigning a belief to every point in the space of parametric distributions defined by the parametric assumptions, one assigns a belief to every point in the space of all possible distributions. Since the space of all possible distribution is often of high cardinality, Bayesian non-parametric models require advanced techniques of distributions in such high cardinal spaces (e.g., Gaussian processes or Dirichlet processes) (Orbanz and Teh 2010).

Besides parametric and non-parametric models, there are also non-probabilistic or semi-probabilistic models. Linear classification, K-means clustering, hierarchical clusterings and neural networks are examples (Bishop 2006). They do not explicitly rely on probabilistic statements. However, usually one can find some relations in one way or another to the probabilistic basis behind these models, which are often too complicated to exploit directly.

References

Aurell, E. and M. Ekeberg (2012). "Inverse Ising inference using all the data". *Phys. Rev. Lett.* 106, p. 90601.

Bavaud, F. (2009). "Information theory, relative entropy and statistics". In: *Formal Theories of Information*. Ed. by G. Sommaruga, pp. 54–78.

Besag, J. (1975). "Statistical analysis of non-lattice data". *J. R. Stat. Soc. Ser. D* 24.3, pages.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Cover, T. M. and J. A. Thomas (2006). *Elements of Information Theory*. Wiley & Son.

Cox, R. T. (1946). "Probability, frequency and reasonable expectation". *Am. J. Phys.* 14.1, p. 1.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). "Least angle regression". *Ann. Stat.* 32.2, pp. 407–499.

Feller, W. (1950). *An Introduction to Probability Theory with Applications, Volume 1*.

Friedman, J., T. Hastie, and R. Tibshirani (2010). "Regularization paths for generalized linear models via coordinate descent". *J. Stat. Softw.* 33.1, pp. 1–22.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning*. Springer.

Higgins, J. J. (2004). *Introduction to modern non-parametric statistics*. Cengage Learning.

Hu, S.-T. (1967). *Elements of Real Analysis*. Holden-Day.

Jaynes, E. T. (1957a). "Infomation theory and statistical mechanics I". *Phys. Rev.* 106.4, p. 620.

Jaynes, E. T. (1957b). "Information theory and statistical mechanics II". *Phys. Rev.* 108, p. 171.

Jaynes, E. T. (1989). *E.T. Jaynes: Papers on Probability, Statistics, and Statistical Physics*. Springer Science & Business Media, p. 434.

Khinchin, A. I. (1957). *Mathematical Foundations of Information Theory*. Dover Publications.

Landauer, R. (1961). "Irreversibility and heat generation in the computing process". *IBM J. Res. Dev.* 5.3, pp. 183–191.

MacKay, D. J. C. (2002). *Information theory, Inference and Learning Algorithms*. Cambridge University Press.

Merhav, N. (2010). "Physics of the Shannon limits". *IEEE Trans. Inf. Theory* 56.9, pp. 4274–4285.

Orbanz, P. and Y. W. Teh (2010). "Bayesian nonparametric models". In: *Encycl. Mach. Learn.* Springer.

Pressé, S., K. Ghosh, J. Lee, and K. A. Dill (2013). "Principles of maximum entropy and maximum caliber in statistical physics". *Rev. Mod. Phys.* 85.3, pp. 1115–1141.

Ravikumar, P., M. J. Wainwright, and J. D. Lafferty (2010). "High-dimensional Ising model selection using l1-regularized logistic regression". *Ann. Stat.* 38.3, p. 1287.

Rosenthal, J. S. (2010). *A First Look at Rigorous Probability Theory*. World Scientific.

Sagawa, T. (2012). "Thermodynamics of information processing in small systems". *Prog. Theor. Phys.* 127.1, pp. 1–56.

Shannon, C. E. (1948). "A mathematical theory of communication". *Bell Syst. Techical J.* 27, pp. 379–423, 623–656.

Shore, J. and R. Johnson (1980). "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy". *IEEE Trans. Inf. Theory* 26.1, pp. 26–37.

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso". *J. R. Stat. Soc. Ser. B* 58.1, pp. 267–288.

Zurek, W. H., ed. (1990). *Complexity, Entropy and the Physics of Information*. Westview Press.

Chapter 3

The inverse Ising problem

We start this chapter with a review of the definition of the inverse Ising problem discussed in the context of maximum entropy models in chapter 2. After a quick review of some different methods for the problem, we explain how the Bethe–Peierls approximation can solve the inverse Ising problem, exactly in trees and approximately in general graphs. We will show that the solution given by the Bethe–Peierls approximation outperforms many known mean-field-like methods, including (naïve) mean-field approximation, Thouless–Anderson–Palmer (TAP) approximation, Sessak–Monasson expansion. We then study the inverse Ising problem at low temperatures, where the mean-field-like methods generally fail. We suggest a modification for the mean-field solution to work at low temperatures.

3.1 The inverse Ising problem

Consider an arbitrary system of N binary spins $\sigma_i \in \{-1, +1\}$, with $i = 1, 2, \ldots, N$. We denote a training dataset of M independently sampled configurations of the system by $D = \{s_i^\mu\}$, where $i = 1, 2, \ldots, N$ and $\mu = 1, 2, \ldots, M$. Very often we are most interested in the first moments $\langle \sigma_i \rangle$ and the second moments $\langle \sigma_i \sigma_i \rangle$ of these variables. If we consider these first two moments as sufficient statistics of the system, then the maximum entropy model of the system assumes a Hamiltonian with pairwise interactions,

$$H[\mathcal{C}] = -\sum_i h_i s_i - \sum_{(i,j)} J_{ij} s_i s_j, \tag{3.1}$$

with couplings J_{ij} and fields h_i, and $\mathcal{C} = (s_1, s_2, \ldots, s_N)$. The couplings J_{ij} and fields h_i are subject to the self-consistent equations,

$$\langle \sigma_i \rangle = \langle \sigma_i \rangle^D, \tag{3.2}$$

$$\langle \sigma_i \sigma_j \rangle = \langle \sigma_i \sigma_j \rangle^D, \tag{3.3}$$

where the left-hand sides are the averages of σ_i and $\sigma_i\sigma_j$ over the model distribution and the right-hand sides are their averages over the training dataset D (see also chapter 2). The latter equation can also be written in terms of the connected correlations,

$$\langle \sigma_i\sigma_j \rangle - \langle \sigma_i \rangle \langle \sigma_j \rangle = \langle \sigma_i\sigma_j \rangle^D - \langle \sigma_i \rangle^D \langle \sigma_j \rangle^D . \tag{3.4}$$

In the following, we will often denote the common values of both sides of these self-consistent equations by the same symbols, $m_i = \langle \sigma_i \rangle = \langle \sigma_i \rangle^D$ and $C_{ij} = \langle \sigma_i\sigma_j \rangle - \langle \sigma_i \rangle \langle \sigma_j \rangle = \langle \sigma_i\sigma_j \rangle^D - \langle \sigma_i \rangle^D \langle \sigma_j \rangle^D$.

The calculation of $\langle \sigma_i \rangle$ and $\langle \sigma_i\sigma_j \rangle$ as functions of h_i and J_{ij} is referred to as the *forward problem*, while the *inverse problem* is to solve equations (3.2) and (3.3) for J_{ij} and h_i. The calculation of the model averages $\langle \sigma_i \rangle$ and $\langle \sigma_i\sigma_j \rangle$ as functions of J_{ij} and h_i is in general a highly non-trivial task. In practice, Markov Chain Monte Carlo simulations are involved, where the configuration space is sampled stochastically via a Markov chain that gives rise to the Boltzmann measure as its equilibrium distribution (MacKay 2002, Mézard and Montanari 2009). This can be used to solve equations (3.2) and (3.3) by gradually updating step by step the parameters h_i and J_{ij} as

$$\triangle h_i = -\eta[\langle \sigma_i \rangle - \langle \sigma_i \rangle^D], \tag{3.5}$$

$$\triangle J_{ij} = -\eta[\langle \sigma_i\sigma_j \rangle - \langle \sigma_i\sigma_j \rangle^D], \tag{3.6}$$

where η is a small positive parameter called the *learning rate*. The minus signs in the equations account for the correct direction of changing h_i and J_{ij} so that $\langle \sigma_i \rangle$ and $\langle \sigma_i\sigma_j \rangle$ change towards $\langle \sigma_i \rangle^D$ and $\langle \sigma_i\sigma_j \rangle^D$, which can be derived formally by studying the likelihood function of h_i and J_{ij} explicitly (MacKay 2002). These update rules are known as the *Boltzmann machine update rules*. Such a solution to the inverse problem via Markov Chain Monte Carlo simulation is often rather slow, especially for large systems and at low temperatures.

There exist alternatively a number of approximate methods. The simplest approximation of all considers every pair of spins independently from the rest, resulting in the reconstructed couplings as

$$J_{ij} = \frac{1}{4} \ln \left[\frac{\big((1+m_i)(1+m_j)+C_{ij}\big)\,\big((1-m_i)(1-m_j)+C_{ij}\big)}{\big((1+m_i)(1-m_j)-C_{ij}\big)\,\big((1-m_i)(1+m_j)-C_{ij}\big)} \right], \tag{3.7}$$

and an additional heuristic argument yields the reconstructed fields as

$$h_i = \frac{1}{2} \ln \left[\frac{1+m_i}{1-m_i} \right] + \frac{1}{2} \sum_{j \neq i} \ln \left[\frac{1-m_j-C_{ij}/(1+m_i)}{1-m_i+C_{ij}/(1-m_i)} \right]. \tag{3.8}$$

This is known as *independent-pair approximation* (Roudi et al. 2009).

In the *mean-field approximation*, explicit formulas for m_i and C_{ij} as functions of h_i and J_{ij} can also be found and we are able to solve equations (3.2) and (3.3) explicitly under this approximation (Kappen and Rodíguez 1998). The solution takes an elegant form,

$$J_{ij} = -(C^{-1})_{ij}, \ (i \neq j), \tag{3.9}$$
$$h_i = \operatorname{arcth}(m_i) - \sum_{j \neq i} J_{ij} m_j, \tag{3.10}$$

where C^{-1} is the inverse of the connected correlation matrix C. We will derive this equation in section 3.2

In *Thouless-Anderson-Palmer (TAP) approximation*, one can also derive similar equations (Thouless et al. 1977),

$$J_{ij} + 2(J_{ij})^2 m_i m_j = -(C^{-1})_{ij}, \ (i \neq j), \tag{3.11}$$
$$h_i = \operatorname{arcth}(m_i) - \sum_j J_{ij} m_j + m_i \sum_j J_{ij}^2 (1 - m_j^2). \tag{3.12}$$

Recently, Sessak and Monasson introduced an expansion of the couplings J_{ij} with respect to the correlation C_{ij} (Sessak and Monasson 2009),

$$J_{ij} = J_{ij}^{loop} + J_{ij}^{IP} - \frac{C_{ij}}{(1 - m_i^2)(1 - m_j^2) - (C_{ij})^2}, \ (i \neq j), \tag{3.13}$$

where

$$J_{ij}^{loop} = (L_i L_j)^{-1/2} [M(I + M)^{-1}]_{ij}, \tag{3.14}$$

with $L_i = 1 - m_i^2$, $M_{ij} = C_{ij}(L_i L_j)^{-1/2}$ for $i \neq j$ and $M_{ii} = 0$ and the quantities J_{ij}^{IP} are the couplings calculated in the independence-pair approximation (3.7),

$$J_{ij}^{IP} = \frac{1}{4} \ln \left[\frac{\left((1 + m_i)(1 + m_j) + C_{ij}\right) \left((1 - m_i)(1 - m_j) + C_{ij}\right)}{\left((1 + m_i)(1 - m_j) - C_{ij}\right) \left((1 - m_i)(1 + m_j) - C_{ij}\right)} \right]. \tag{3.15}$$

The fields then can also be calculated,

$$h_i = \operatorname{arcth}(m_i) - \sum_j J_{ij} m_j + \sum_{j \neq i} K_{ij}^2 m_i L_j, \tag{3.16}$$

with $K_{ij} = C_{ij}/(L_i L_j)$, where we stopped at the second order in the correlation C_{ij} for the sake of simplicity (Sessak and Monasson 2009). Roudi *et al.* then pointed out that the term J_{ij}^{loop} is actually equivalent to the mean-field solution and gave a simple explanation for the

expansion (Roudi et al. 2009).

The inverse Ising problem is also of central interest among the machine learning community. There are many important contributions of the machine learning researchers to the inverse Ising problem. With the rising interests among the machine learning community in the so-called *message-passing algorithms*, in particular *belief-propagation*, Mézard and Mora introduced susceptibility propagation for the inverse Ising problem, where the update rules are simple hybridisations of the Boltzmann machine update rules (3.5) and (3.6) and the iterative procedure of the *belief-propagation* (Mézard and Mora 2009).

In the context of the forward problem, Yedidia has pointed out that the equations of belief-propagation actually describe the stationary points of the so-called Bethe free energy appearing in Bethe–Peierls approximation (Yedidia et al. 2001, 2003). It is also known long time ago that the connected correlation function within Bethe–Peierls approximation can be calculated via the so-called *linear response relation* (Welling and Teh 2004). Curiously, it took some time to realise that the correlations calculated from Bethe–Peierls approximation directly give a simple solution to the inverse Ising problem. We and others recently studied the Bethe–Peierls approximation in the context of the inverse Ising problem and showed that the resulted solution is exact in trees and outperforms many known methods in loopy graphs (Nguyen and Berg 2012a, Ricci-Tersenghi 2012). The next section is devoted to Bethe–Peierls approximation in the context of the inverse Ising problem.

Although these methods (which are referred to as mean-field-like methods) perform well at high temperatures (weak couplings), at low temperatures (strong couplings) they often fail to produce correct couplings. Other effort has been made in solving the problem at low temperatures. Several methods have been proposed, including adaptive cluster expansion (Cocco et al. 2009) and pseudo-likelihood maximisation (Ravikumar et al. 2010). Recently, we also introduced a hybridisation of clustering algorithms and the mean-field approximation to solve the inverse Ising problem at low temperatures (Nguyen and Berg 2012b), which will be the topic of the last section.

3.2 Bethe–Peierls approximation and mean-field-like methods

For the pedagogical reason, we start with a discussion of mean-field approximation. We will derive the mean-field approximation from the framework of variational approximate inference. We then discuss how the forward and the inverse Ising problem can be solved within the mean-field approximation.

3.2.1 Mean-field approximation

Consider a system of N Ising spins with the Hamiltonian (3.1), which amounts to the Boltzmann distribution

$$p(\mathcal{C}) = \frac{1}{Z} e^{-H(\mathcal{C})}, \tag{3.17}$$

where $\mathcal{C}$ is a configuration of the system, $\mathcal{C} = (s_1, s_2, s_3 \ldots, s_N)$, and Z is the partition function. The idea of the variational approximate inference is to find a surrogate distribution q for the Boltzmann distribution of the system (p), which should be simple enough for inference. The simplest distributions of all, the *mean-field family of approximate distributions* (Opper and Saad 2001), factorise over all spins,

$$q(\mathcal{C}) = \prod_i \frac{1 + m_i s_i}{2}, \tag{3.18}$$

where m_i are now the magnetisations of spins i calculated under the approximate distribution, which are generally only approximates for the exact magnetisations of the system. To determine the values of m_i so that q best approximates p, the variational principle states that the parameters of q should be chosen to minimise the Kullback–Leibler divergence from p to q, see chapter 2. In physics, instead of the Kullback–Leibler divergence, we often use a slightly different quantity, the *functional free energy* (Opper and Saad 2001). The two quantities are different only by an additive constant,

$$\mathrm{KL}(q|p) = F_p[q] - F_0, \tag{3.19}$$

where $F_p[q]$ is the functional free energy, which is a functional of q, and F_0 is the exact free energy of the system. Since F_0 does not depend of q, minimising $\mathrm{KL}(q|p)$ is equivalent to minimising $F_p[q]$ (both with respect to q). The functional free energy can be written in a similar way of thermodynamics equations,

$$F_p[q] = U_p[q] - S_p[q], \tag{3.20}$$

where $U_p[q] = \sum_{\mathcal{C}} q(\mathcal{C}) H(\mathcal{C})$ and $S_p[q] = -\sum_{\mathcal{C}} q(\mathcal{C}) \ln q(\mathcal{C})$. Ignoring the index p of F for simplicity, with the mean-field distribution (3.18), the functional free energy can be found easily,

$$\begin{aligned} F[\{m_i\}] &= -\sum_i h_i m_i - \sum_{ij} J_{ij} m_i m_j + \\ &\quad + \sum_i \left[\frac{1 + m_i}{2} \ln \frac{1 + m_i}{2} + \frac{1 - m_i}{2} \ln \frac{1 - m_i}{2} \right]. \end{aligned} \tag{3.21}$$

Taking the derivatives of $F[\{m_i\}]$ with respect to m_i and set them to zero, we obtain the *mean-field self-consistent equation*

$$\operatorname{arcth}(m_i) = \sum_{j \neq i} J_{ij} m_j + h_i. \tag{3.22}$$

In the forward problem, this equation is used to solve for the magnetisations m_i, which are the mean-field approximate magnetisations of the spins in the system.

In the inverse problem, however, we need to find h_i and J_{ij}. To do so, we also need to calculate the correlations $\langle \sigma_i \sigma_j \rangle$. In practice, it is more convenient to calculate the *connected correlations* instead,

$$C_{ij} = \langle \sigma_i \sigma_j \rangle - \langle \sigma_i \rangle \langle \sigma_j \rangle . \tag{3.23}$$

Since the mean-field distribution (3.18) factorises over the spins, the approximate connected correlation between any pair of spins would be zero under direct calculation. Is there any other way to have estimate values for the correlations? The trick is to use the so-called *linear response relation*, which relates the connected correlation function with the *susceptibility*,

$$C_{ij} = \frac{\partial m_i}{\partial h_j}. \tag{3.24}$$

Note that this linear response relation holds exactly under any general distribution. Taking the derivatives of (3.22) with respect to h_k, we have

$$\frac{C_{ik}}{1 - m_i^2} = \sum_{j \neq i} J_{ij} C_{jk} + \delta_{ik}. \tag{3.25}$$

This equation has a simple solution for J_{ij},

$$-J_{ij} + \frac{\delta_{ij}}{1 - m_i^2} = (C^{-1})_{ij}. \tag{3.26}$$

With the solutions J_{ij}, equation (3.22) can be used to solve for h_i, completing the mean-field solution to the inverse Ising problem.

The computation of the mean-field solution, which consists mainly the inversion of the connected correlation matrix, is among the fastest methods for the inverse Ising problem. In many cases, the mean-field approximation gives satisfactory estimates for J_{ij} and h_i. However, we will see that mean-field approximation often overestimates the couplings, especially at low temperatures (see also chapter 4), for which there is a better approximation: the Bethe–Peierls approximation.

3.2.2 Bethe–Peierls approximation

Since Bethe–Peierls approximation is exact when the underlying interacting topology is a *tree* (i.e., in (3.1) the summation over (i,j) is restricted to a subset of spin pairs which form a tree) (Bethe 1935), we first restrict our discussion to this case. We also first assume that although the coupling strengths are not known, this tree topology of interactions is known. Later, we will relax this assumption.

We construct the Bethe family of distributions as follows. To every spin i, we assign a distribution $b_i(s_i)$, and to every pairwise interaction (i,j), we assign a pairwise distribution $b_{ij}(s_i,s_j)$. These distributions are called *beliefs*, which are consistently normalised,

$$\sum_{s_j} b_{ij}(s_i,s_j) = b(s_i), \tag{3.27}$$

$$\sum_{s_i} b_i(s_i) = 1. \tag{3.28}$$

Noting that the Ising spins take values in $\{+1,-1\}$, these distributions can be parameterised as follows,

$$b_i(s_i) = \frac{1+m_i s_i}{2} \tag{3.29}$$

$$b_{ij}(s_i,s_j) = \frac{(1+m_i s_i)(1+m_j s_j) - \bar{C}_{ij} s_i s_j}{4}. \tag{3.30}$$

with

$$-1 \le m_i \le 1 \tag{3.31}$$

$$-1 + |m_i + m_i| - m_i m_j \le \bar{C}_{ij} \le +1 - \left|m_i - m_j\right| - m_i m_j \tag{3.32}$$

A distribution in the Bethe family, or a Bethe ansatz, is defined as

$$q(\{s_i\}) = \prod_i b_i(s_i) \prod_{(i,j)} \frac{b_{ij}(s_i,s_j)}{b_i(s_i) b_j(s_j)}, \tag{3.33}$$

where the first product is over all spins, and the second product is over all interactions. It can be shown that $q(\{s_i\})$ is properly normalised with one-spin marginals $b_i(s_i)$ and two-spin marginals $b_{ij}(s_i,s_j)$, provided the topology of interaction is a tree (Mézard and Montanari 2009). Using (3.33), we can write down the functional free energy easily,

$$F[\{m_i,\bar{C}_{ij}\}] = U[\{m_i,\bar{C}_{ij}\}] - S[\{m_i,\bar{C}_{ij}\}], \tag{3.34}$$

where the energy is

$$U[\{m_i, \bar{C}_{ij}\}] = -\sum_{(ij)} J_{ij}(\bar{C}_{ij} + m_i m_j) - \sum_i h_i m_i, \tag{3.35}$$

while the entropy is

$$S[\{m_i, \bar{C}_{ij}\}] = S_1[\{m_i, \bar{C}_{ij}\}] + S_2[\{m_i, \bar{C}_{ij}\}], \tag{3.36}$$

with

$$S_1[\{m_i, \bar{C}_{ij}\}] = \sum_i (1 - z_i) \sum_{s_i} \frac{1 + m_i s_i}{2} \ln \frac{1 + m_i s_i}{2}, \tag{3.37}$$

$$S_2[\{m_i, \bar{C}_{ij}\}] = \sum_{(i,j)} \sum_{s_i, s_j} \frac{(1 + m_i s_i)(1 + m_j s_j) + \bar{C}_{ij} s_i s_j}{4} \ln \frac{(1 + m_i s_i)(1 + m_j s_j) + \bar{C}_{ij} s_i s_j}{4}, \tag{3.38}$$

where z_i is the number of spins that connect to spin i (its *degree*). By setting the derivatives of $F[\{m_i, \bar{C}_{ij}\}]$ with respect to m_i and $\bar{C}_{ij}$ to zero, we have

$$h_i + \sum_{j \in \partial i} J_{ij} m_j = (1 - z_i) \operatorname{arcth}(m_i) + \sum_{j \in \partial i} \sum_{s_i, s_j} \frac{s_i + m_j s_i s_j}{4} \ln \frac{(1 + m_i s_i)(1 + m_j s_j) + \bar{C}_{ij} s_i s_j}{4}, \tag{3.39}$$

$$J_{ij} = \sum_{s_i, s_j} \frac{s_i s_j}{4} \ln \frac{(1 + m_i s_i)(1 + m_j s_j) + \bar{C}_{ij} s_i s_j}{4}, \tag{3.40}$$

where ∂i stands for the boundary set containing all spins that interact with spin i. In the forward problem, these equations (3.39) and (3.40) are to be solved for m_i and $\bar{C}_{ij}$, which define the best approximate distribution q according to the variational principle. In fact, equation (3.40) can be solve explicitly for $\bar{C}_{ij}$,

$$\bar{C}_{ij} = \frac{(1 - m_i^2 - m_j^2) t_{ij} + 2 m_i m_j}{1 + \sqrt{D_{ij}}} - m_i m_j, \tag{3.41}$$

where $D_{ij} = 1 - 2 m_i m_j t_{ij} - (1 - m_i^2 - m_j^2) t_{ij}^2$ and $t_{ij} = \operatorname{th}(2 J_{ij})$. By substituting (3.41) into (3.39), we have a self-consistent equation to be solved for m_i.

In this case, the topology of interactions is a tree, the best approximate distribution is the exact one. It follows that m_i are the exact magnetisations and $\bar{C}_{ij}$ are the exact connected correlations of the system. However, $\bar{C}_{ij}$ are defined only for interacting pairs of spins,

therefore, in this way, only the connected correlations between interacting spins can be found. In the inverse problem, equations (3.39) and (3.40) can be used to calculate h_i and J_{ij} directly given that the values m_i and $\bar{C}_{ij}$ are identified with the magnetisations and the connected correlations estimated from data. In fact, this procedure turns out identical to the independent-pair approximation. If the topology of interaction is known (which is a tree, by our assumption), the procedure gives the exact values for the couplings and the fields.

Unfortunately, in most of the cases, even when the assumption that the topology of interaction is a tree is met, the concrete tree is not known. In fact, in many applications, the foremost interest is to infer the topology of interaction, while the values of coupling strengths are of less importance. Yet, we will show that given the topology of interaction is a tree, it is possible to infer the exact topology with the exact coupling strengths using another formula for the connected correlations, which are now defined for all pairs of spins. This is based on the linear response relation, which we already learned in the last section,

$$C_{ij} = \frac{\partial m_i}{\partial h_j}. \tag{3.42}$$

Keeping the linear response relation in mind, and with $\bar{C}_{ij}$ substituted from (3.41), taking the derivative of (3.39) with respect to h_j, we have

$$\begin{aligned} &\frac{1-z_i}{1-m_i^2}C_{ik} + \sum_{j\in\partial i}\sum_{s_i s_j}\frac{s_i}{4}\frac{1}{1+m_i s_i + m_j s_j + (\bar{C}_{ij}+m_i m_j)s_i s_j}\times \\ &\times\left\{\left[s_i - \frac{s_i s_j}{\sqrt{D_{ij}}}(m_i t_{ij} - m_j)\right]C_{ik} + \left[s_j - \frac{s_i s_j}{\sqrt{D_{ij}}}(m_j t_{ij} - m_i)\right]C_{jk}\right\} = \delta_{ik}, \end{aligned} \tag{3.43}$$

which can be solved to give

$$(C^{-1})_{ij} = \frac{\bar{C}_{ij}}{(\bar{C}_{ij})^2 - (1-m_i^2)(1-m_j^2)},\ (i\neq j), \tag{3.44}$$

$$(C^{-1})_{ii} = \frac{1-z_i}{1-m_i^2} - \sum_{j\in\partial i}\frac{1-m_j^2}{(\bar{C}_{ij})^2 - (1-m_i^2)(1-m_j^2)}, \tag{3.45}$$

where $\bar{C}_{ij}$ were extended to all pairs of spins so that $\bar{C}_{ij} = 0$ if i and j do not interact directly.

If $m_i = 0$ (due to the absence of the local fields h_i and spontaneous magnetisations),

we obtain particularly simple equations,

$$\begin{aligned}(C^{-1})_{ij} &= -\frac{1}{2}\,\mathrm{sh}(2J_{ij}), (i \neq j),\\ (C^{-1})_{ii} &= 1 + \sum_{j\in\partial i} \mathrm{sh}^2(J_{ij}).\end{aligned} \tag{3.46}$$

In the forward problem, where h_i and J_{ij} are known, after solving (3.39) and (3.40), equations (3.44) and (3.45) can be used to solve for C_{ij}. In the inverse problem, where m_i and C_{ij} are estimated from data, equation (3.44) is solved for $\bar{C}_{ij}$, and then (3.40) is solved for J_{ij}. Interestingly, even if we do not know the topology of interactions, given it is a tree, this procedure gives the exact topology of interaction and coupling strength: if i and j do not interact directly with each other, $(C^{-1})_{ij} = 0$, therefore $\bar{C}_{ij} = 0$ and $J_{ij} = 0$. Having J_{ij}, equation (3.39) can be used to calculate for h_i.

In the case of loopy graphs, the Bethe ansatz (3.33) is no longer properly normalised and cannot be considered as a probability distribution. However, one can still write down and minimise the Bethe functional free energy. The values m_i and $\bar{C}_{ij}$ found by solving (3.39) and (3.40) are the approximate magnetisations and the approximate connected correlations for the system. Moreover, solving (3.44) and (3.45) for C_{ij}, we will have another approximation for the connected correlations (which is a better one, see Opper and Winther 2004). Likewise, in the inverse problem, equation (3.44) can again be used to calculate for the coupling J_{ij} and (3.39) can be used to calculate for h_i with the magnetisations m_i and the connected correlations C_{ij} estimated from data, as if the topology of interactions were a tree.

Relation to belief propagation and susceptibility propagation

While both Bethe–Peierls approximation and belief propagation are widely used in physics and machine learning, their ultimate relation was pointed out by Yedidia only recently (Yedidia et al. 2001, 2003). Specifically, he showed that belief propagation emerges as an algorithm to minimise the Bethe functional free energy, where messages are simply the Lagrange multipliers to impose the normalisation condition for the beliefs.

This close relation opens another way to calculate the correlation function within belief-propagation without propagating the susceptibility. Upon the convergence of belief propagation, the magnetisations are known. In fact, the correlations between interacting spins can also be derived. These magnetisations and correlations can be identified with m_i and $\bar{C}_{ij}$ in the Bethe–Peierls approximation. Then, equations (3.44) and (3.44) can be used to find the connected correlations C_{ij}. For details of this relationship, we refer to (Opper and Saad 2001, Yedidia et al. 2001, 2003).

Benchmark the mean-field-like methods

To benchmark different methods for the inverse Ising problem, we apply them to simulated data. We set up a system of N spins with couplings J_{ij}^0 and fields h_i^0. To study the effect of the strength of the couplings and the field, we will also explicitly introduce the inverse temperature β, which effectively varies the strengths of the couplings and fields to βJ_{ij}^0 and βh_i^0, respectively. We then calculate the correlations C_{ij} and the magnetisations m_i either by direct enumeration for small systems (*infinite number of samples*) or by Markov Chain Monte Carlo simulation for large systems (*finite number of samples*). These correlations and magnetisations are then used to reconstruct the couplings βJ_{ij} and the fields βh_i by mean-field approximation (and TAP modification), Bethe–Peierls approximation and Sessak–Monasson expansion, which are all considered as mean-field-like methods. Susceptibility propagation is also performed to compare with Bethe–Peierls approximation. Concentrating on the reconstruction of the couplings, we access the quality of the reconstructed couplings by the quantity d, defined as

$$d(\{\beta J_{ij}\}, \{\beta J_{ij}^0\}) = \frac{\sum_{(i,j)} (\beta J_{ij} - \beta J_{ij}^0)^2}{\sum_{(i,j)} (\beta J_{ij}^0)^2}, \tag{3.47}$$

(note that in fact the β in the nominator and in the dominator cancel.)

Simulated data: infinite number of samples

We start with small systems ($N = 20$), where infinite sampling (configuration enumerating) to derive the exact magnetisations m_i and the exact correlations C_{ij} is possible. We consider two topologies of interaction for J_{ij}^0: a Cayley tree of degree $z = 3$ and a Sherrington–Kirkpatrick (SK)-model.

A Cayley tree of degree z is constructed by induction as follows. Starting from one node which constitutes the zeroth layer of the tree, we add the first layer that consists of z spins that are connected to the single spin of the zeroth layer; the second layer consists of $z(z-1)$ spins, where every $z-1$ spins connect to one spin in the second layer, and so on (see figure 3.1). Note that, every node has degree $z = 3$ except for the leaves, which are at the outermost layer. For every link, a coupling is drawn from a uniform distribution on the interval $[-1, +1]$ while the field at every spin is set to zero for the sake of simplicity.

Figure 3.2a shows the distances between the reconstructed couplings by mean-field approximation (and TAP modification), Sessak–Monasson expansion, susceptibility propagation and Bethe–Peierls approximation and the known original couplings, $d(\{\beta J_{ij}\}, \{\beta J_{ij}^0\})$, as functions of inverse temperature β. Note that in this case, since the magnetisations are

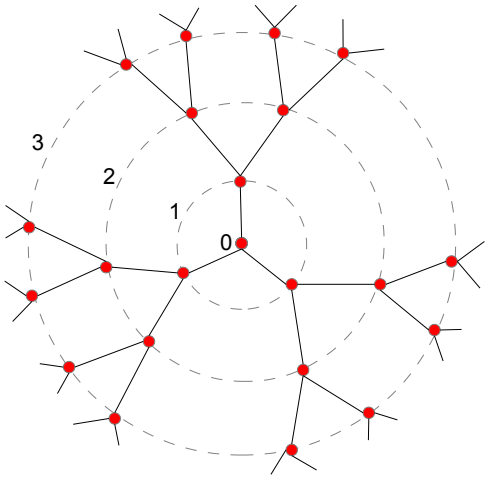


Figure 3.1: **Construction of a Cayley tree with $z = 3$.** Starting from one node which constitutes the zeroth layer of the tree, we add the first layer that consists of z spins that are connected to the single spin of the first layer. The second layer consists of $z(z-1)$ spins, where every $z-1$ spins connect to one spin in the second layer, and so on.

all zero, the TAP reconstruction is identical to that of mean-field. We see that while mean-field reconstruction and Sessak–Monasson expansion give approximate solutions, Bethe–Peierls approximation can reconstruct the couplings of the system exactly to the numerical accuracy. Susceptibility propagation indeed also gives the exact reconstruction, but only for small coupling strength (small β); for strong couplings, susceptibility propagation converges to a wrong stationary point with wrong couplings.

A Sherrington–Kirkpatrick (SK) model is a fully connected network of spins where each coupling is drawn from a Gaussian distribution with zero mean and unit standard deviation. We again consider the system at zero field. In this case, the topology is highly loopy, therefore Bethe–Peierls approximation is no longer exact. Still, Bethe–Peierls approximation yields a very good approximation in comparison to mean-field approximation and Sessak–Monasson expansion in the strong coupling regime, see figure 3.2b. Note that Sessak–Monasson expansion (at zero field) is also very competitive, which gives the best approximation when the couplings are small enough. At small couplings, susceptibility propagation again follows Bethe–Peierls approximation closely in quality, but fails to converge to the Bethe solution completely perhaps because of the numerical precision in computation. Susceptibility propagation again fails at strong couplings.

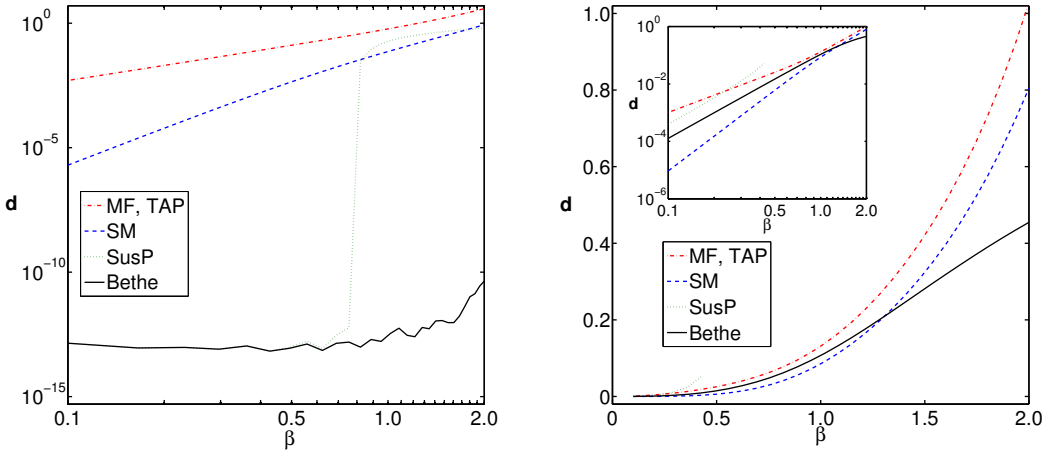


Figure 3.2: **Benchmark of the mean-field-like solutions of the inverse Ising problem of $N = 20$ spins (with infinite number of samples).** The deviations d of reconstructed couplings by mean-field (MF) and TAP approximation (TAP), Bethe–Peierls approximation (Bethe), Sessak–Monasson expansion (SM) and susceptibility propagation (SusP) from the true couplings are plotted as functions of inverse temperatures β. (a) In a Cayley tree: while mean-field approximation and Sessak–Monasson expansion are approximate, Bethe–Peierls reconstruction is exact. Susceptibility propagation is also exact at weak couplings (small β) but fails to converge to the correct couplings at strong couplings (large β). (b) In a SK model: while Sessak–Monasson expansion works the best for weak couplings, Bethe–Peierls reconstruction also gives good reconstructed couplings, and becomes the best reconstruction at strong couplings.

Simulated data: finite number of samples

We then turn to a more realistic situation where the number of spins are large and enumerating all the configurations of the system is not possible. The magnetisations and correlations in this case are estimated from M sampled configurations of the system obtained by Markov Chain Monte Carlo simulation.

We set up $N = 100$ spins on a random graph of fixed degree z, which can be simulated using an R-package provided by Viger and Latanpy (Viger and Latapy 2005). At large N, the graph resembles a Cayley tree: it does contain loops, but the loops are long (of the order of $\ln N$). The couplings J_{ij} are drawn from a uniform distribution on the interval $[-1, +1]$ and the fields h_i are drawn from a uniform distribution on the interval $[-0.3, +0.3]$. Figure 3.3a shows the quality of reconstruction of mean-field, TAP and Sessak–Monasson expansion as functions of inverse temperature β with $M = 5000$ samples. The figure clearly shows that in general, TAP approximation improves mean-field approximation, and Sessak–Monasson expansion significantly improves TAP. Bethe–Peierls approximation is still the best option especially at strong couplings. At weak couplings, the differences between the method are smeared out due to sampling noise. The effect of

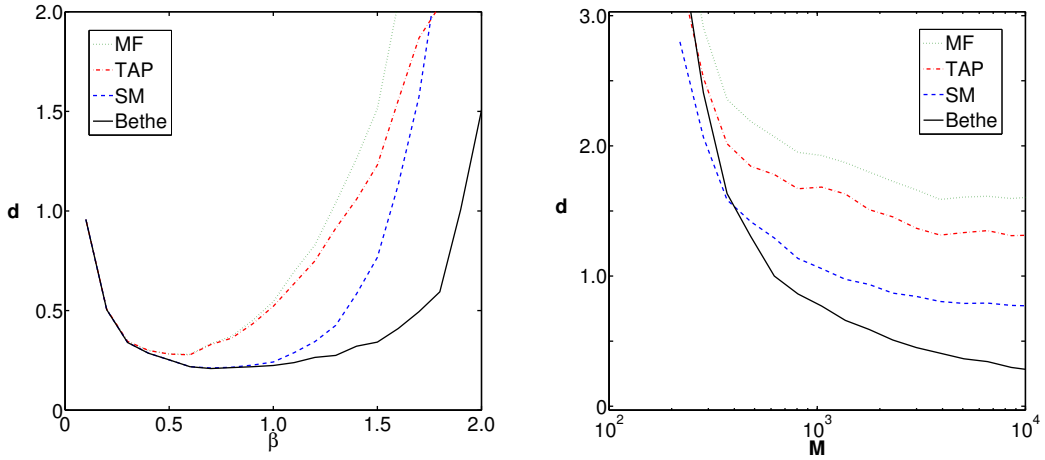


Figure 3.3: **Benchmark of the mean-field solutions of the inverse Ising problem for $N = 100$ spins on a random graph with fixed degree $z = 3$ (with finite number of samples).** (a) The deviations d of reconstructed couplings by mean-field (MF), TAP reconstruction (TAP), Bethe–Peierls reconstruction (Bethe), Sessak–Monasson expansion (SM) and susceptibility propagation (SusP) from the true couplings as functions of inverse temperatures β with $M = 5000$ simulated samples. While the differences between the reconstructed couplings by different methods are insignificant at weak couplings (small β), Bethe–Peierls reconstruction gives the best reconstructed couplings at relatively strong couplings (large β). (b) The deviations d from the true couplings of the reconstructed couplings by the same methods as functions of number of simulated samples M with $\beta = 1$. When the number of samples is small, the difference between the reconstructed couplings by different methods are insignificant. When the number of samples is relatively large, Bethe–Peierls reconstruction gives the best estimations for the couplings.

sampling noise is also illustrated in figure 3.3b. We see that the mean-field-like methods for the inverse Ising problem requires rather large number of samples for high quality reconstruction. The reconstruction quality quickly decreases as the number of samples in the dataset decreases. In addition, the differences between these methods are not significant when the numbers of samples are small.

3.3 The inverse Ising problem at low temperatures with mean-field approximations

While mean-field-like reconstructions perform well at high temperatures (weak couplings), at low temperatures the quality of the reconstructed couplings quickly gets worse. The breakdown of mean-field reconstructions can have different roots: the emergence of multiple thermodynamic states at a phase transition, an increasing correlation length at lower temperatures, or the freezing of the spins into a reduced set of configurations at low temperatures requiring more samples to measure the correlations between spins. To investigate this issue, we first consider a very simple case where mean-field theory is exact: the Curie–Weiss model. The zero-field Hamiltonian of N binary spins s_i is $H_J(\{s_i\}) =$

$-J/N\sum_{i<j}s_is_j$ with $J=1$. This corresponds to equal couplings $J^0_{ij}=J/N$ between all pairs of spins, a fact that is of course not known when reconstructing the couplings. M samples of spin configurations are taken from the equilibrium measure $\exp\{-\beta H_J(\{s_i\})\}/Z$, where β is the inverse temperature and Z is the partition function. One then can calculate the observed magnetisations $\bar{m}_i=\frac{1}{M}\sum_\mu s_i^\mu$ and connected correlations $\bar{C}_{ik}=\frac{1}{M}\sum_\mu s_i^\mu s_k^\mu-\bar{m}_i\bar{m}_k$, with $\mu=1,\dots,M$ denoting the sampled configurations. Note that in this section, we will need to carefully distinguish the Boltzmann distribution averages (m_i for the magnetisations and C_{ij} for the connected correlations) and the data average ($\bar{m}_i$ for the magnetisations and $\bar{C}_{ij}$ for the connected correlations).

The mean-field prediction for the magnetisations of the Curie-Weiss model is also given by the solution of the self-consistent equation (3.22),

$$m_i=\mathrm{th}(\sum_{j\neq i}J_{ij}m_j+h_i)\;, \tag{3.48}$$

where the couplings are rescaled with temperature $J_{ij}=\beta J^0_{ij}$. As we also discussed, the connected correlations follow from (3.48) by considering the linear response

$$\begin{aligned} C_{ik} &= \frac{\partial m_i}{\partial h_k}=(1-m_i^2)\left(\sum_{j\neq i}J_{ij}\frac{\partial m_j}{\partial h_k}+\delta_{ik}\right)\\ &= (1-m_i^2)\left(\sum_{j\neq i}J_{ij}C_{jk}+\delta_{ik}\right)\;. \end{aligned} \tag{3.49}$$

Inserting the observed magnetisations and correlations into (3.49) gives (Kappen and Rodíguez 1998)

$$\sum_{j\neq i}J_{ij}\bar{C}_{jk}=-\delta_{ik}+\bar{C}_{ik}/(1-\bar{m}_i^2)\;, \tag{3.50}$$

which can be solved directly for the couplings $J_{ij}=-(\bar{C}^{-1})_{ij}$ $(i\neq j)$ and the fields $h_i=\mathrm{arcth}\,\bar{m}_i-\sum_{j\neq i}J_{ij}\bar{m}_j$ using (3.48).

Figure 3.4a shows how well this reconstruction performs at different inverse temperatures β and different number of samples M. For $\beta<\beta_c=1$, the reconstruction error goes to zero with the number of samples as $M^{-1/2}$: since for the Curie–Weiss model the self-consistent equation (3.48) is exact, the reconstruction is limited only by fluctuations of the measured correlations resulting from the finite number of samples and by the finite system size.

Yet for $\beta>\beta_c$, the difference between the underlying couplings and the reconstructed couplings does not vanish with increasing number of samples. While the self-consistent equation (3.48) is still correct, the identification of its solutions with the observed magneti-

sations $\bar{m}_i$ is mistaken. For the ferromagnetic phase at $\beta > \beta_c$, there are two solutions of the self-consistent equation, denoted $m_i^{\pm} = \pm m$. The observed magnetisations are averages over these two thermodynamic states and they have nothing to do with either of the two solutions of (3.48). The same holds for the connected correlations C_{ij}^{+} and C_{ij}^{-} in the two states, and the observed correlations $\bar{C}_{ij}$.

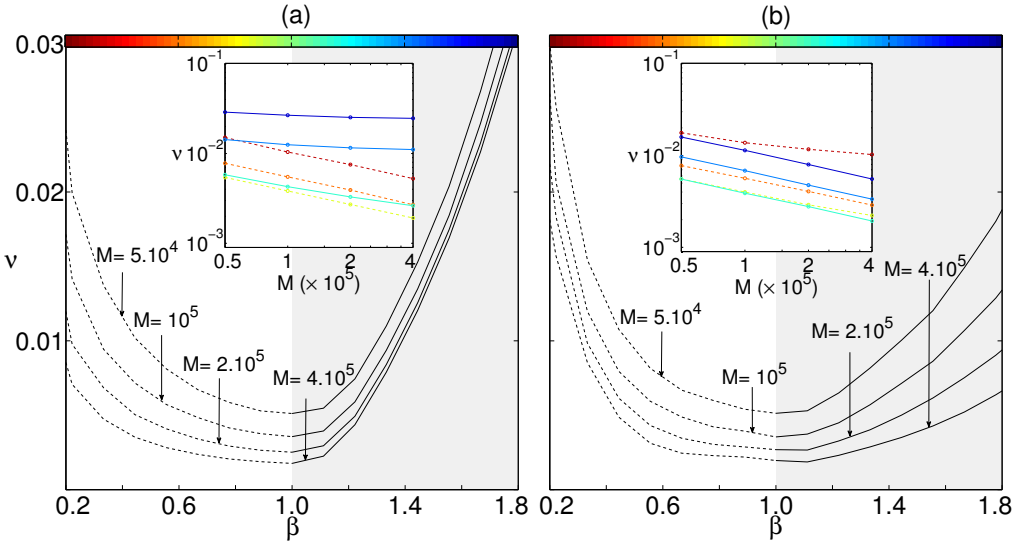


Figure 3.4: **Reconstructing couplings of the Curie–Weiss ferromagnet.** The root-mean-squared deviation between the reconstructed couplings and underlying couplings, $\nu = \sqrt{\frac{2}{N(N-1)} \sum_{i<j} (J_{ij}/\beta - J/N)^2}$, is plotted against the inverse temperature β for different numbers of configurations. The system size is $N = 100$. The insets show this deviation on a logarithmic scale versus the number of samples M at different inverse temperatures indicated by the colours of the curves ($\beta = 0.3, 0.58, 0.86, 1.14, 1.42, 1.7$). **a)** Reconstruction based on a single thermodynamic state breaks down in the low-temperature phase $\beta > 1$ and the deviation between reconstructed and underlying couplings does not vanish with an increasing number of samples M (see the blue curves in the inset). **b)** Reconstruction based on two thermodynamic states is asymptotically exact. The deviation between reconstructed and underlying couplings vanishes like $M^{-1/2}$ even at low temperatures.

A simple cure suggests itself: Since each sample stems from one of the two thermodynamic states, we divide the M configurations into those configurations with positive total magnetisation $\sum_i s_i^{\mu}$, and those with negative total magnetisation. Then the magnetisations in the two thermodynamic states can be calculated separately, giving $\bar{m}_i^{+} = \frac{1}{M_+} \sum_{\mu \in +} s_i^{\mu}$ and similarly for $\bar{m}_i^{-}$ and the connected correlations. Identifying these magnetisations with the solutions of the self-consistent equation (3.48), we obtain in the place of (3.50) *two*

sets of equations

$$\sum_{j\neq i} J_{ij}\bar{C}^{+}_{jk} = -\delta_{ik} + \bar{C}^{+}_{ik}/(1-(\bar{m}^{+}_{i})^2)\,, \tag{3.51}$$

$$\sum_{j\neq i} J_{ij}\bar{C}^{-}_{jk} = -\delta_{ik} + \bar{C}^{-}_{ik}/(1-(\bar{m}^{-}_{i})^2)\,. \tag{3.52}$$

Reconstructing the couplings using a single state only, by solving say (3.51), the observed positive magnetisation can be accounted for equally well by positive external fields (even though the samples were generated by a model with zero field), or alternatively, by ferromagnetic couplings between the spins. One finds that solving (3.51) leads to an underestimate of the couplings, and *positive* external fields calculated by (3.48) follow. Correspondingly, basing the reconstruction only on data from the down state by solving (3.52) also leads to an underestimate of the couplings, and large *negative* fields. This effect has already been noted in the context of the inverse Hopfield problem (Braunstein et al. 2011). We thus demand that the reconstructed fields obtained from either state are equal to one another

$$\sum_{j\neq i} J_{ij}(\bar{m}^{+}_{j} - \bar{m}^{-}_{j}) = \operatorname{arcth}\bar{m}^{+}_{i} - \operatorname{arcth}\bar{m}^{-}_{i} \tag{3.53}$$

and claim that jointly solving equations (3.51), (3.52), and (3.53) gives the correct mean-field reconstruction at low temperatures.

Already equations (3.51) and (3.52) are two linear equations per coupling variable, so in general there is no solution to these equations. However, we expect that the underlying couplings used to generate the M configurations actually solve these equations, at least up to fluctuations due to the finite number of configurations sampled and the finite size effect. For an overdetermined linear equation of the form $\mathbf{A}\cdot\mathbf{x} = \mathbf{b}$ with vectors of different lengths $\mathbf{x}$ and $\mathbf{b}$ and a non-square matrix $\mathbf{A}$, the Moore–Penrose pseudoinverse $\mathbf{A}^{+}$ (Moore 1920, Penrose 1955) gives a least-square solution $\mathbf{x} = \mathbf{A}^{+}\cdot\mathbf{b}$ such that the Euclidean norm $||\mathbf{A}\cdot\mathbf{x}-\mathbf{b}||_2$ is minimised. In this sense, the Moore–Penrose pseudoinverse allows to solve (3.51), (3.52), and (3.53) as well as possible. The linear equations (3.51), (3.52), and (3.53) can be written as a single matrix equation $\mathbf{J}\cdot\mathbf{A} = \mathbf{B}$, where $\mathbf{A}$ is the $N\times(2N+1)$ matrix $\left(\bar{\mathbf{C}}^{+},\bar{\mathbf{C}}^{-},\bar{\mathbf{m}}^{+}-\bar{\mathbf{m}}^{-}\right)$ and $\mathbf{B}$ is the $N\times(2N+1)$ matrix $(\bar{\mathbf{b}}^{+},\bar{\mathbf{b}}^{-},\tilde{\mathbf{m}}^{+}-\tilde{\mathbf{m}}^{-})$, with $\bar{b}^{+}_{ij} = -\delta_{ij} + \bar{C}^{+}_{ij}/(1-(\bar{m}^{+}_{i})^2)$ and analogously for $\bar{b}^{-}_{ij}$, and $\tilde{m}^{+}_{i} = \operatorname{arcth}\bar{m}^{+}_{i}$ and analogously for $\tilde{m}^{-}_{i}$. The Moore–Penrose inverse is calculated using singular value decomposition (Press et al. 2002) and right-multiplied with $\mathbf{B}$ to obtain the the optimal solution $\mathbf{J}$. In general, this matrix will not be symmetric, and we use $(J_{ij}+J_{ji})/2$, $i\neq j$ for the reconstructed couplings. The external fields can be computed for each state from $h^{+}_{i} = \operatorname{arcth}\bar{m}^{+}_{i} - \sum_{j\neq i} J_{ij}\bar{m}^{+}_{j}$, and analogously for h^{-}_{i}. Their averages

over the two states is used for the reconstructed fields.

Figure 3.4b shows how the reconstruction error now vanishes with $M^{-1/2}$ also in the ferromagnetic phase, albeit with a prefactor which grows as the temperature decreases. So while the mean-field reconstruction from many samples is still successful at low temperatures, more configurations are needed to obtain a certain reconstruction error: At very low temperatures, most spins will be in the same state (either up or down); the connected correlations are small as a result and require many samples for their accurate determination.

In practice, couplings between spins will not all be equal to each other like in the Curie–Weiss model. Ferromagnetic as well as antiferromagnetic couplings may be present in magnetic alloys, neurons have excitatory and inhibitory interactions, regulatory interactions between genes can either enhance or suppress the expression of a target gene. The Curie–Weiss ferromagnet is not a good model for all those cases where the couplings are of different signs and magnitudes. In fact, in models where all spins interact with each other via couplings that can be positive or negative (Sherrington and Kirkpatrick 1975), the low-temperature regime may be characterised not by two, but by many thermodynamic states (Mézard, Parisi, et al. 1987, Thouless et al. 1977). These so-called glassy states cannot be identified simply by the total magnetisation of each sample, as is the case for the ferromagnet. Nevertheless, configurations μ, μ' from the same thermodynamic state are typically close to each other, having a large overlap $(1/N)\sum_i s_i^\mu s_i^{\mu'}$. Glassy thermodynamic states thus appear as clusters in the space of configurations (Hed et al. 2001, Krazakala et al. 2007).

We use the k-means clustering algorithm (Bishop 2006) to find these clusters in the sampled spin configurations. Starting with a set of randomly chosen and normalised cluster centres, each configuration is assigned the cluster centre it has the largest overlap with. Then the cluster centres are moved to lie in the direction of the centre of mass of all configurations assigned to that cluster, and the procedure is repeated until convergence. We also tried out different algorithms from the family of hierarchical clustering methods, but found no significant difference in the reconstruction performance. Then magnetisations and connected correlations are computed for each cluster separately. Equations (3.51), (3.52), and (3.53) can be written for k thermodynamic states. The mean-field equation for each state and the condition that the external fields are equal in all states can be written again in the form of a matrix equation $\mathbf{J} \cdot \mathbf{A} = \mathbf{B}$. $\mathbf{A}$ is the $N \times (kN + k)$ matrix $\left(\bar{\mathbf{C}}^{\mathbf{1}}, \ldots, \bar{\mathbf{C}}^{\mathbf{k}}, \bar{\mathbf{m}}^{\mathbf{1}} - \langle \bar{\mathbf{m}} \rangle, \ldots, \bar{\mathbf{m}}^{\mathbf{k}} - \langle \bar{\mathbf{m}} \rangle\right)$ where $\langle \cdot \rangle$ denotes the average over clusters, $\langle \bar{\mathbf{m}} \rangle = (1/k)\sum_{a=1}^{k} \bar{\mathbf{m}}^{\mathbf{a}}$, and analogously for $\mathbf{B}$. The pseudoinverse of A can be computed in $\mathcal{O}(kN^3)$ steps (Press et al. 2002), so up to a factor of k coming from the number of clusters,

this is just as fast as the high-temperature mean-field reconstruction based on Gaussian elimination to invert the correlation matrix.

We test this approach using couplings drawn independently from a Gaussian distribution of zero mean and variance $1/N$ (the Sherrington–Kirkpatrick model, see Sherrington and Kirkpatrick 1975). Figure 3.5a shows the reconstruction at low temperatures getting better with the number of clusters k and configuration samples M.

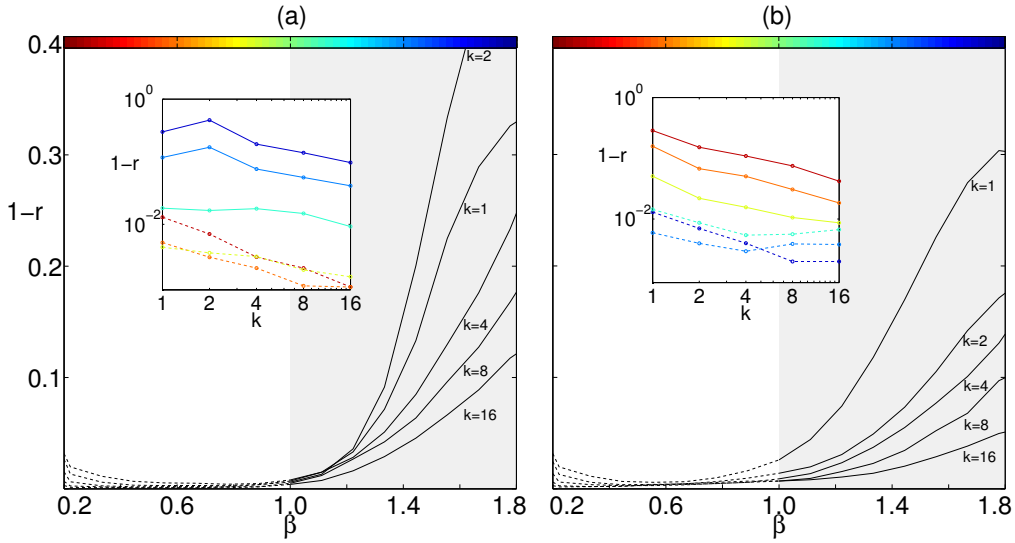


Figure 3.5: **Reconstructing couplings of the Sherrington–Kirkpatrick model.** The Pearson correlation coefficient r quantifies the correlation between reconstructed couplings and underlying couplings, $r = \frac{\frac{1}{N(N-1)}\sum_{i\neq j}(J_{ij}-\langle J\rangle)(J^0_{ij}-\langle J^0\rangle)}{\sqrt{\mathrm{var}(J)\,\mathrm{var}(J^0)}}$, where $\langle J\rangle$, $\mathrm{var}(J)$ are the mean and variance of the reconstructed couplings across bonds, and similarly for the underlying couplings. $r = 1$, or $1 - r = 0$, corresponds to perfect reconstruction. The main plots show $1 - r$ against the inverse temperature β for different numbers of clusters k. The insets show how $1 - r$ depends on the number of clusters k at different inverse temperatures indicated by the colours of the curves ($\beta = 0.3, 0.58, 0.86, 1.14, 1.42, 1.7$). The numbers of samples M are scaled with the numbers of clusters $M = k \times 5 \times 10^4$ to ensure a constant average number of states per cluster. The system size is $N = 100$. **a)** Reconstruction based on mean-field approximation. **b)** Reconstruction based on the TAP approximation with gradient descent.

A further improvement is possible. For disordered systems, the self-consistent equation (3.48) is not exact. An additional term is required, the so-called Onsager reaction term describing the effect a spin has on itself via the response of its neighbouring spins. The Thouless–Anderson–Palmer (TAP) equation (Thouless et al. 1977)

$$m_i = \mathrm{th}\Big(\sum_{j\neq i} J_{ij} m_j - m_i \sum_{j\neq i} J_{ij}^2 (1 - m_j^2) + h_i\Big) \tag{3.54}$$

turns out to be exact for models where all spins interact with each other. For each state a

we now obtain instead of (3.51)

$$\sum_{j\neq i} J_{ij}\bar{C}^a_{jk} = -\delta_{ik} + \bar{C}^a_{ik}/(1-(\bar{m}^a_i)^2) + \bar{C}^a_{ik}\sum_{j\neq i} J^2_{ij}(1-(\bar{m}^a_j)^2) - 2\bar{m}^a_i \sum_{j\neq i} J^2_{ij}\bar{m}^a_j\bar{C}^a_{jk}\,. \tag{3.55}$$

These equations are no longer linear in the couplings J_{ij} and cannot be solved by the pseudoinverse. A simple gradient descent method still allows to solve these equations in $\mathcal{O}(kN^3)$ steps per iteration. We define a quadratic cost function S for the couplings $\mathbf{J}$ by squaring the difference between lhs and rhs of equation (3.55) and summing over all spin pairs i,k and states a. Differences in the external fields $h^a_i = \operatorname{arcth}\bar{m}^a_i - \sum_{j\neq i} J_{ij}\bar{m}^a_j + \bar{m}^a_i\sum_{j\neq i} J^2_{ij}(1-(\bar{m}^a_j)^2)$ across thermodynamic states are penalized by an additional term $\sum_{i,a}\left(h^a_i - \langle h_i\rangle\right)^2$. The iterative prescription with rate η, $J_{ij} \leftarrow J_{ij} - \eta\partial S/\partial J_{ij}$, converges to a point near the solution of the TAP equation with small differences in the external fields across states (the deviations resulting from the finite number of samples and finite system size). Figure 3.5b shows how the reconstruction error asymptotically now tends to zero with growing k and M.

Mean-field theories exists beyond the Curie–Weiss, or the Sherrington–Kirkpatrick model discussed here (Opper and Saad 2001). We have shown that the use of mean-field methods to solve the inverse Ising problem at low temperatures hinges on our ability to reconstruct the thermodynamic states from the sampled data. With this proviso, the entire range of mean-field methods can be now be used, for instance, for the Bethe approximation or Kikuchi approximation (Kikuchi 1951).

3.4 Conclusions and outlook

In this section, we discussed the methods of the inverse Ising problem. We introduced Bethe–Peierls approximation to the mean-field-like methods for the inverse Ising problem. At low temperatures, using clustering techniques, we showed that the problem can also be solved by mean-field methods.

The inverse Ising problem and its variants have been applied to many different problems, in particular in quantitative biology, ranging from neural network (Schneidman et al. 2006), protein-protein interaction (Weigt et al. 2009) to gene regulatory networks (Bailly-Bechet et al. 2010). In chapter 4 of this thesis, we present an application of the inverse Ising problem to analyse a dataset of copy-numbers of cancer patients. Chapter 5 can also be considered as an extension of the inverse Ising problem to analyse the correlations in gene expression of cancer cells.

References

Bailly-Bechet, M., A. Braunstein, A. Pagnani, M. Weigt, and R. Zecchina (2010). "Inference of sparse combinatorial-control networks from gene-expression data: a message passing approach". *BMC Bioinformatics* 11.1, p. 355.

Bethe, H. A. (1935). "Statistical theory of superlattices". *Proc. R. Soc.* A 150, pp. 552–575.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Braunstein, A., A. Ramezanpour, R. Zecchina, and P. Zhang (2011). "Inference and learning in sparse systems with multiple states". *Phys. Rev. E* 83, p. 56114.

Cocco, S., S. Leibler, and R. Monasson (2009). "Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods". *Proc. Natl. Acad. Sci.* 106.33, pp. 14058–14062.

Hed, G., A. K. Hartmann, D. Stauffer, and E. Domany (2001). "Spin domains generate hierarchical ground state structure in $J = \pm 1$ spin glass". *Phys. Rev. Lett.* 86, pp. 3148–3151.

Kappen, H. J. and F. B. Rodíguez (1998). "Efficient learning in Boltzmann machines using linear response theory". *Neural Comput.* 10, pp. 1137–1156.

Kikuchi, R. (1951). "A theory of cooperative phenomena". *Phys. Rev.* 81, pp. 988–1003.

Krazakala, F., A. Montanari, F. Ricci-Tersenghi, G. Semerjian, and L. Zdeborová (2007). "Gibbs states and the set of solutions of random constraint satisfaction problems". *Proc. Natl. Acad. Sci.* 104.25, pp. 10318–10323.

MacKay, D. J. C. (2002). *Information theory, Inference and Learning Algorithms*. Cambridge University Press.

Mézard, M. and A. Montanari (2009). *Information, Physics, and Computation*. Oxford University Press.

Mézard, M. and T. Mora (2009). "Constraint satisfaction problems and neural networks: A statistical physics perspective". *J. Physiol. Paris* 103.1-2, pp. 107–113.

Mézard, M., G. Parisi, and M. Virasoro (1987). *Spin Glass Theory and Beyond*. World Scientific Publishing.

Moore, E. H. (1920). "On the reciprocal of the general algebraic matrices". *Bull. Amer. Math. Soc.* 26, pp. 394–395.

Nguyen, H. C. and J. Berg (2012a). "Bethe–Peierls approximation and the inverse Ising problem". *J. Stat. Mech.* P03004.

Nguyen, H. C. and J. Berg (2012b). "Mean-field theory for the inverse Ising problem at low temperatures". *Phys. Rev. Lett.* 109, p. 50602.

Opper, M. and D. Saad, eds. (2001). *Advanced Mean-field Methods: Theory and Practice*. The MIT Press.

Opper, M. and O. Winther (2004). "Variational linear response". In: *Advances in Neural Information Processing Systems 16*. Ed. by S. Thrun, L. K. Saul, and B. Schölkopf. Cambridge, MA: MIT Press, pp. 1157–1164.

Penrose, R. (1955). "A generalized inverse for matrices". *Math. Proc. Camb. Phil. Soc.* 51.03, pp. 406–413.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (2002). *Numerical Recipes in C++*. Cambridge University Press.

Ravikumar, P., M. J. Wainwright, and J. D. Lafferty (2010). "High-dimensional Ising model selection using l1-regularized logistic regression". *Ann. Stat.* 38.3, p. 1287.

Ricci-Tersenghi, F. (2012). "On mean-field approximations for estimating correlations and solving the inverse Ising problem". *J. Stat. Mech.* P08015.

Roudi, Y., E. Aurell, and J. A. Hertz (2009). "Statistical physics of pairwise probability models". *Front. Comput. Neurosci.* 3, p. 22.

Schneidman, E., M. J. Berry, R. Segev, and W. Bialek (2006). "Weak pairwise correlations imply strongly correlated network states in a neural population". *Nature* 440.7087, pp. 1007–12.

Sessak, V. and R. Monasson (2009). "Small-correlation expansions for the inverse Ising problem". *J. Phys. A* 42, p. 55001.

Sherrington, D. and S. Kirkpatrick (1975). "Solvable model of a spin-glass". *Phys. Rev. Lett.* 35.26, pp. 1792–1796.

Thouless, D. J., P. W. Anderson, and R. G. Palmer (1977). "Solution of a 'solvable model of a spin glass'". *Phil. Mag.* 35, p. 593.

Viger, F. and M. Latapy (2005). "Efficient and simple generation of random simple connected graphs with prescribed degree sequence". In: *Annual International Conference on Computing and Combinatorics*. Ed. by L. Wang. Springer, pp. 440–449.

Weigt, M., R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa (2009). "Identification of direct residue contacts in protein-protein interaction by message passing". *Proc. Natl. Acad. Sci.* 106.1, pp. 67–72.

Welling, M. and Y. W. Teh (2004). "Linear response algorithms for approximate inference in graphical models". *Neural Comput.* 16, pp. 197–221.

Yedidia, J. S., W. T. Freeman, and Y. Weiss (2001). *Characterization of belief propagation and its generalizations*. Tech. rep. 15. MERL.

Yedidia, J. S., W. T. Freeman, and Y. Weiss (2003). "Understanding belief propagation and its generalizations". In: *Exploring Artificial Intelligent in the New Millennium*. Ed. by G. Lakemeyer and B. Nebel. Morgan Kaufmann Publishers. Chap. 8, pp. 239–270.

Chapter 4

Gene copy-number correlations in cancer cells

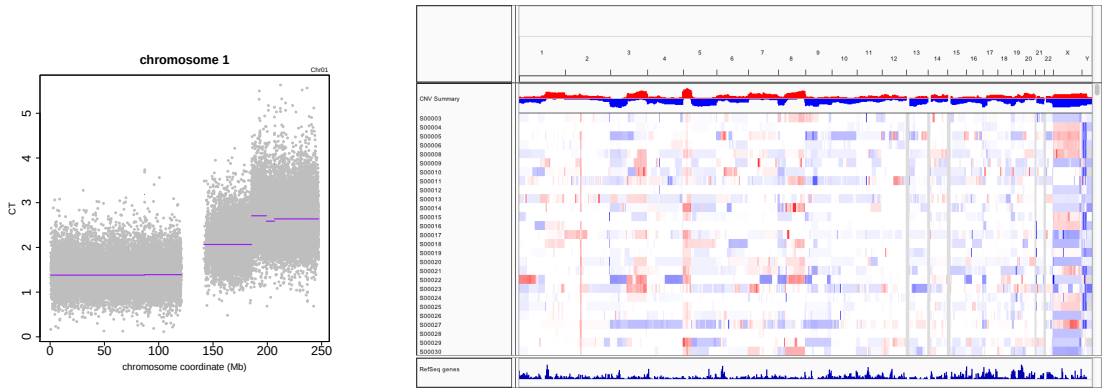
In this chapter, we study the correlations in gene copy-number among different loci of the genome of cancer cells. The matrix of correlations between copy-numbers shows a dense blocked form. We speculate that such a dense blocked correlation matrix can be explained by simpler, sparse direct interactions between the loci. Mapping the problem into the inverse Ising problem, we show that this is indeed the case. We also compare the inferred coupling matrix with the contact frequency map of chromosomes (Hi-C data) to show that strongest coupled loci indeed have higher mean contact frequency than highest correlated loci. Analysis of weaker couplings between distal parts of a chromosome is however not yet possible in this study.

4.1 The gene copy-number variation in cancer cells

Advances in biotechnology during the last decades have made it possible to measure the copy-numbers over a genome with relatively high resolution (Huang et al. 2004, LaFramboise 2009, Pinkel et al. 1998). To archive this, generally, DNA molecules are first cut into short sequences of nucleotides (*oligonucleotides*) by special enzymes. The oligonucleotides are amplified and then labelled with fluorescent agents. The labelled oligonucleotides hybridise with their complementary sequences (probes) prepared on an array. The array is then scanned to measure the fluorescent intensities for every probe. By construction, these fluorescent intensities are directly related to the copy-numbers of the oligonucleotides present in the sample (Huang et al. 2004, Pinkel et al. 1998).

The fluorescent intensities at the probes are very noisy measures of the DNA copy-numbers. Luckily, loci that are close to each other on a chromosome tend to have the same copy-numbers, forming long *segments* of chromosome with the same copy-number separated by so-called *breakpoints*. This provides a way to smoothen the fluorescent intensities with a piecewise constant function, significantly reducing noise in the data. This procedure is called *segmentation*, and the resulted data are called *segmented data*. In fig-

ure 4.1, panel (a) shows an example of raw copy-number data and segmented smoothening, while panel (b) shows a few samples in a segmented dataset in the Integrated Genomic Viewer (IGV) (Robinson et al. 2011, Thorvaldsdóttir et al. 2013). Note that the copy-numbers are often non-integers because a tumour sample usually contains different tumour clones as well as contamination by normal cells, making the copy-number measurement a population-averaged measurement (see chapter 1). Luckily, due to the discrete nature of copy-numbers, there exist impurity correction methods to eliminate the copy-numbers of contaminated normal cells and infer the copy-numbers of the main clone in the population and sometimes also the subclone structure of the tumour population (Carter et al. 2012, Fischer et al. 2014, Oesper et al. 2013). In this chapter, we will work with an impurity-corrected copy-number dataset provided by our collaborators at Department of Translational Genomics.



(a) (b)

Figure 4.1: **The copy-number data.** (a) The signal intensities of copy-number data (grey) and segmentation smoothening (example in Circular Binary Segmentation package developed by Olshen et al. 2004). (b) Few sample tracks of segmented copy-number data viewed in IGV with red colour for amplification and blue colour for deletion.

We observe from figure 4.1 that there are indeed many copy-number mutations in the copy-number profile of a tumour sample. Many of these copy-number mutations are random, but few *amplifications* or *deletions* recurrently appear across all samples or samples of a particular type of cancers. This observation has two implications. First, these recurrent amplifications and deletions suggest that the affected regions are important to tumour development, inviting further investigation to shed light on their biological effects. Second, since different types of cancer may have different recurrent patterns of copy-number mutations (and hence different biological properties), the copy-number data of a cancer sample can be used to classify the tumour into tumour class to inform treatments. Of course, in a way, these two implications are simply two sides of the same problem.

A number of statistical methods have been suggested to find recurrent patterns of copy-

number mutations in cancers or in a particular type of cancers. One important example is GISTIC developed by Mermel et al. 2011 and its variants. The ideas behind such statistical methods are often rather simple: suppose we sum up over all the copy-number profiles of a subtype of cancer, recurrent amplifications or deletions would stand out as peaks (see figure 4.1b, top track). Statistical tests can be used to assign a p-value to each peak to indicate how less likely such a peak emerges purely by chance. Such approaches have successfully revealed that many of the recurrent amplifications indeed contain oncogenes that play key roles in cancer development, and many of the recurrent deletions contain genes that are tumour suppressors. For example, EGFR and KRAS are recurrently amplified in lung cancers, prostate cancers and some other cancer types, while TP53 is recurrently deleted in cancers of many types, especially those in late stages of the disease.

From another perspective, we may ask another question: *are copy-number mutations correlated?* In physical terminology, while GISTIC-like methods are dealing with the "one-point function", we want to ask the question about the "two-point function". We speculate that, since copy-number mutations in cancer can arise because of tandem replication of different segments of chromosomes, the correlations between copy-numbers of different parts of a chromosome could be related to the spatial organisation of the genome.

Complementarily, recent advances, especially the landmark works by Lieberman-Aiden, have made it possible to capture the spatial organisation of DNA (Lieberman-Aiden et al. 2009). By locking the chromosome segments that are closest to each other by special proteins before cutting the genome by enzymes and analysing the resulted short sequences, the interacting segments of chromosomes are recorded. Such a measurement results in a *contact frequency matrix*, of which each element is the frequency of physical contact of two loci in the genome. The contact map thus shows rough picture of the geometrical organisation of the genome. An example of such a contact frequency matrix is shown in the terrain colour scale[1] in figure 4.2a. The study of the spatial organisation of genomes reveals an important picture, see figure 4.2b. The interacting partners of a human chromosome are often continuous segments, here referred to as *sectors*.[2] For the purpose of this simple study, we determine the sectors in a very simple way, see figure 4.2 and its caption. Slight difference in determining the sectors do not affect the results. For a more careful definition of sectors, see Lieberman-Aiden et al. 2009.

[1] Terrain colour scale: green and yellow hues indicate low frequencies, brown and white hues indicate high frequencies.

[2] According to Lieberman-Aiden et al. 2009, sectors are of two types A or B. Sectors of the same types tend to interact more frequently with each other, and sectors of different types interact less frequently with each other. Moreover, one type of sectors contains more genes and are more accessible (more open) than the other (more close). Further statistical analyses also reveal that such a compartment structure of the human genome is also correlated with the DNA replication timing and also the presence of the so-called fragile sites in the genome (De and Michor 2011).

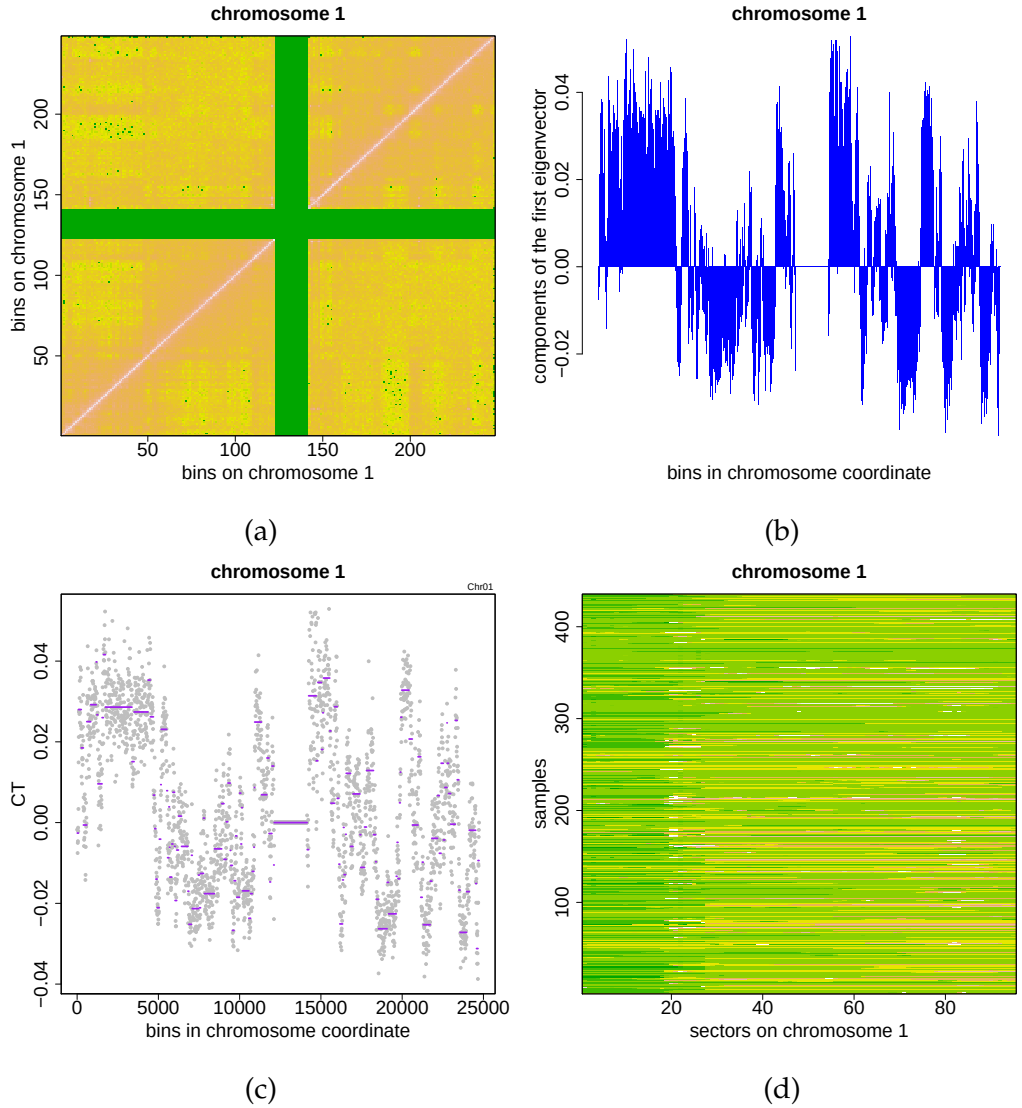


Figure 4.2: **Chromosome contact frequency map and its sector structure.** (a) The contact frequency matrix of chromosome 1 is shown as a terrain map (data from Lieberman-Aiden et al. 2009). Chromosome 1 is binned with the resolution of 1000 base-pairs, and the contact frequency of every pair of bins is presented. (b) The first eigenvector of the correlation matrix of the contact frequency, following Lieberman-Aiden et al. 2009. The components of the eigenvector have alternating signs, indicating the two types of sectors (compartments) on the chromosome; for more details, see Lieberman-Aiden et al. 2009. (c) We determine the sector of chromosomes by simply segmentation smoothening the eigenvector into segments (with Circular Binary Segmentation CBS by Olshen et al. 2004). (d) Copy-number matrix of sectors for all samples extracted from the segmented data.

We speculate that the sector structure may play a role in copy-number mutations in cancer cells. Therefore, we extract the copy-numbers of those sectors and study the correlations in copy-number across these sectors over cancer genomes, see figure 4.2c.

Figure 4.3 shows the correlation matrix between the copy-numbers of the sectors on

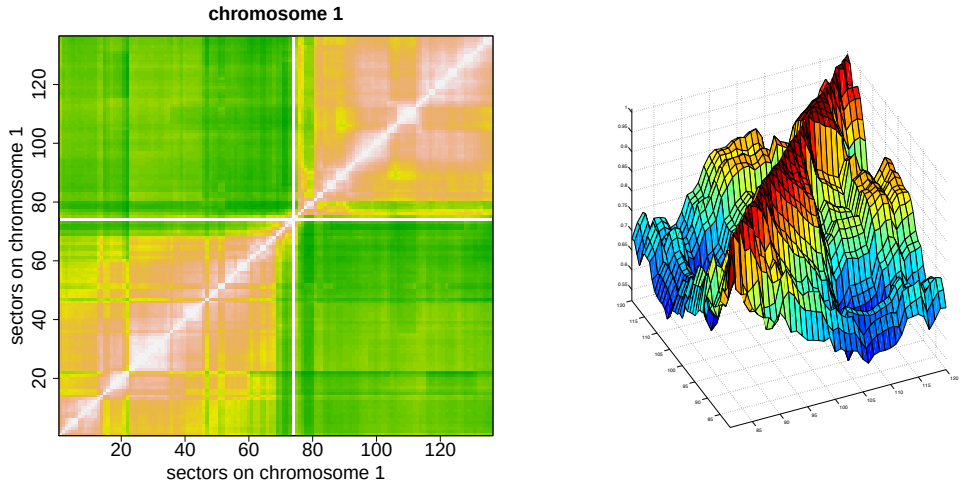


Figure 4.3: **The copy-number correlation matrix.** (a) The terrain map of the matrix of connected correlations in copy-number between different sectors on chromosome 1. (b) A small block on the diagonal of the correlation matrix.

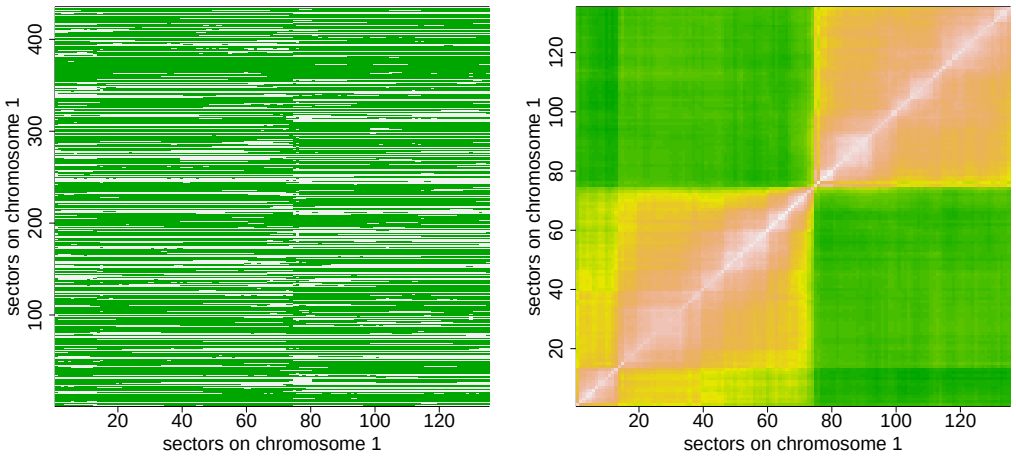
chromosome 1. Again, statistical methods with careful manipulations of the data may pinpoint which pairs of loci are more correlated than expected by chance. Such an approach has been discussed, for example, by Klijn et al. 2010.

However, these statistical models have an intrinsic difficulty: correlation matrices, by their nature, are often dense, that is, many sectors are highly correlated with each other. Indeed, having a closer look at the correlation matrix shown in figure 4.3, we find that it shows dense correlated blocks. But do these correlated blocks demonstrate dense direct interactions among the loci? (The exact meaning of the term 'direct interaction' will be explained in the following section.) We will show that such dense blocked form of the correlation matrix can be deceptive. In fact, the correlations can be explained by a very sparse interaction matrix with most interactions concentrating among the nearest neighbours.

4.2 Correlation and interaction: the inverse Ising problem

To understand the nature of the question we posed, let us go back to the Ising model for an analogy. Consider a linear chain of three Ising spins S_1, S_2 and S_3. Suppose S_1 and S_2 are coupled by a strong coupling, and so do S_2 and S_3. There is no coupling between S_1 and S_3. However, it is clear that there can be strong correlation between S_1 and S_3, which is induced by the shared interaction partner S_2. Therefore, the strong block of correlation between S_1, S_2 and S_3 in this case can be explained simply by their linear geometrical organisation.

This example of the Ising model elucidates the central point: it is important to distinguish the concept of correlation from that of interaction. In the context of Ising models, the problem of inferring the couplings (and fields) from the correlations (and magnetisations) is known as the inverse Ising problem, which we discussed in chapter 3. The inverse Ising problem and its variants have been successfully applied to disentangle the (indirect) correlations and (direct) interactions in the contexts of neural networks (Schneidman et al. 2006), protein-protein interaction networks (Weigt et al. 2009) and gene-regulatory networks (Bailly-Bechet et al. 2010). Here, we apply the inverse Ising problem to analyse the interactions that underlie the correlations in the dataset of copy-numbers of lung cancers.



(a) (b)

Figure 4.4: **Copy-number data in binary format.** (a) The copy-number of sectors on chromosome 1 for all samples in binary format. (b) The correlation matrix recalculated with binary data.

In order to map the problem into the inverse Ising problem, technically we have to map copy-number values to binary format. Fortunately, with the impurity-corrected dataset provided by our collaborators, this binarisation of copy-number values is rather straightforward: for any sector, we assign value $+1$ (*positive* or *affected*) if the copy-number is larger than 2.5 or smaller than 1.5, and assign value -1 (*negative* or *unaffected*) if otherwise. This results in a binary copy-number dataset and the correlation matrix is recalculated accordingly; these are shown in figure 4.4. (Here, we are consider the amplifications and deletions of the sectors jointly, similar approaches can be used to study the amplifications or the deletions of the sectors separately.)

4.2.1 The inverse Ising problem in mean-field approximation

Methods for the inverse Ising problem have been discussed in chapter 3. In the following, we will use with the simple mean-field reconstruction (with a minor modification described below) throughout, which gives the couplings between the spins as

$$J_{ij} = C^{-1}_{ij}, \text{ for } i \neq j, \tag{4.1}$$

where C is the connected correlation (covariance) matrix among the spins. We found that in our case, the differences between different mean-field-like methods are not significant as the number of samples are small (see chapter 3, figure 3.3). On the other hand, mean-field reconstruction has an important advantage compared to the others: it is easy to incorporate a model regularisation. In particular, it easy to show that a ridge regularisation (Hastie et al. 2009) for the mean-field reconstruction is simply given by

$$J_{ij} = -(C + \gamma I)^{-1}_{ij}, \text{ for } i \neq j, \tag{4.2}$$

where I is the identity matrix and γ is the regularisation coefficient (see also chapter 2).

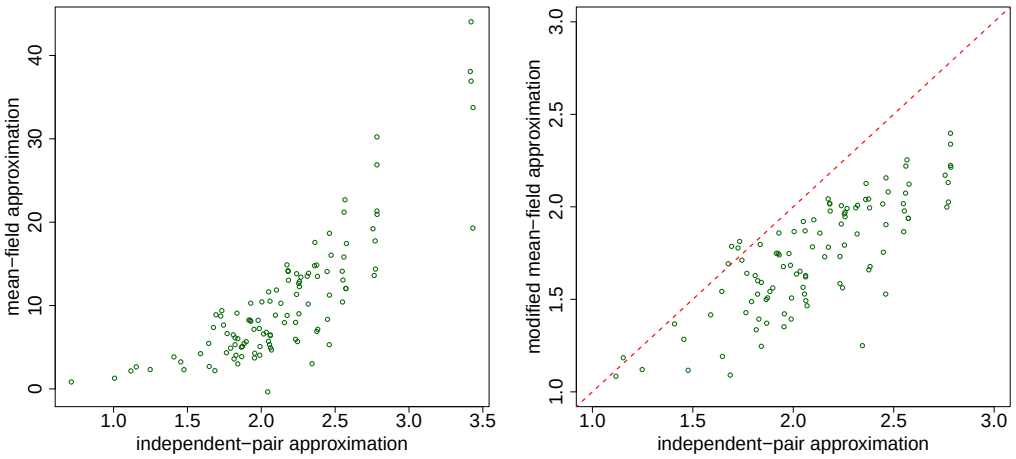
4.2.2 Analysing the copy-number data of lung cancers

With the binarised copy-number data (see figure 4.4), it is rather straightforward to infer the couplings using the mean-field approximation. We perform the inference for each chromosome separately. The results here are from the data of chromosome 1; results for other chromosomes are similar.

Modifying the mean-field solution

We infer the couplings J_{ij} using (4.2) with a small value of the regularisation coefficient, $\gamma = 0.01$. As expected, we find that the coupling matrix is much sparser than the correlation matrix with strong couplings concentrating on the subdiagonal, namely between nearest neighbouring spins. While these nearest neighbour couplings are expected to be strong, these mean-field reconstructed couplings turn out to be enormously large.[3] The fact that the couplings of the nearest neighbours are large means that considering a pair of neighbouring spins independently from the other spins can be a good approximation. Therefore, the independent-pair approximation (see chapter 3) is expected to be good for these couplings. Figure 4.5a shows the scatter plot between independent-pair reconstructed couplings and mean-field reconstructed couplings of nearest neighbours. This

[3] A coupling of 50 gives a difference in energy of 100, and therefore a difference in probability of e^{100}!



(a) (b)

Figure 4.5: **Modification of the mean-field solution.** The scatter plots between nearest neighbour couplings reconstructed by the independent-pair approximation versus the couplings reconstructed by the mean-field approximation ($\gamma = 0.01$) without arcsh modification (a) and with arcsh modification (b). While without arcsh modification, mean-field reconstructed couplings can be enormously large in comparison to independent-pair approximation, the arcsh transformation makes the mean-field reconstructed couplings very similar to that of the independent-pair approximation for nearest neighbours.

figure shows a clear difference between them.

How do we resolve this difference? Surprisingly, the answer lies in the Bethe–Peierls approximation. Let us recall from chapter 3 that Bethe–Peierls approximation suggests that the coupling matrix when the magnetisations are zero should be inferred as $J_{ij} = -1/2 \operatorname{arcsh}(2C^{-1}_{ij})$, for $i \neq j$, instead of $J_{ij} = -C^{-1}_{ij}$ (for $i \neq j$) as in mean-field approximation. We speculate that for strong couplings, this arcsh-modification would be generally needed. The reconstruction formula (4.2) then changes to

$$J_{ij} = -\frac{1}{2} \operatorname{arcsh}\left[2(C + \gamma I)^{-1}_{ij}\right], \text{ for } i \neq j. \tag{4.3}$$

Indeed, after this modification, the couplings between nearest neighbours reconstructed by independent-pair approximation and by mean-field are in better agreement, as shown in figure 4.5b. In the figure, one also observes that the modified mean-field reconstruction slightly underestimates the couplings. This is however expected because of the ridge regularisation, which shrinks the estimation towards zero (shrinkage – see chapter 2).

Analysing the coupling matrix

Figure 4.6b shows the terrain map of the reconstructed coupling matrix. We see that the coupling matrix is much sparser than the correlation matrix shown in figure 4.6a. The strongest couplings are concentrated on the near neighbours. The scatter plot between the correlations and the couplings in figure 4.6c shows that many pairs of sectors that are strongly correlated in copy-number actually have small direct couplings.

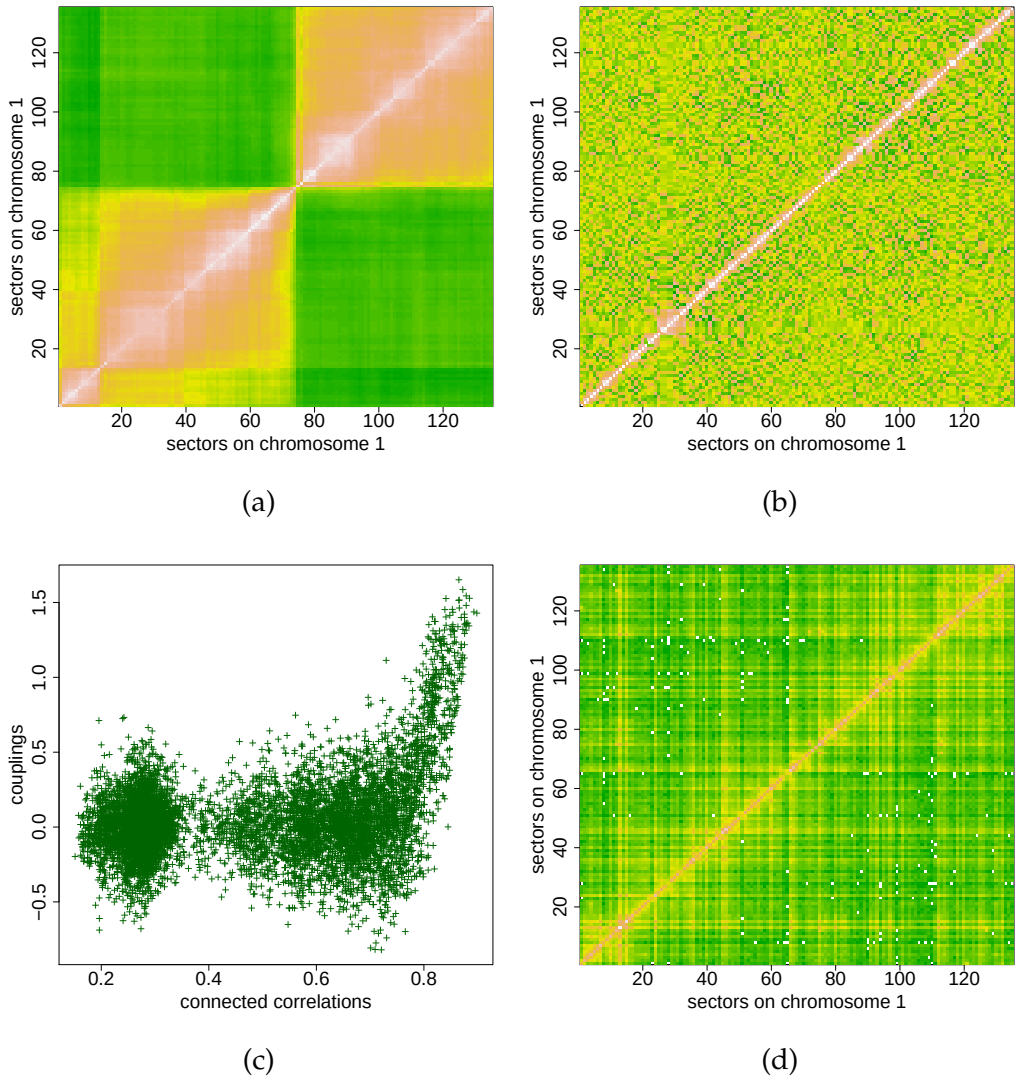


Figure 4.6: **Connected correlations, couplings and the contact frequencies.** (a) The connected correlation matrix is dense with large correlated blocks. (b) The coupling matrix is sparse with strong couplings concentrating mainly among close neighbours. (c) A scatter plot of the couplings versus the connected correlations shows that many pairs of sectors that have strong connected correlations are only weakly coupled. (d) For a comparison, we also plot the terrain map of the contact frequency matrix (in log scale, pairs that have contact frequency 0 are left white).

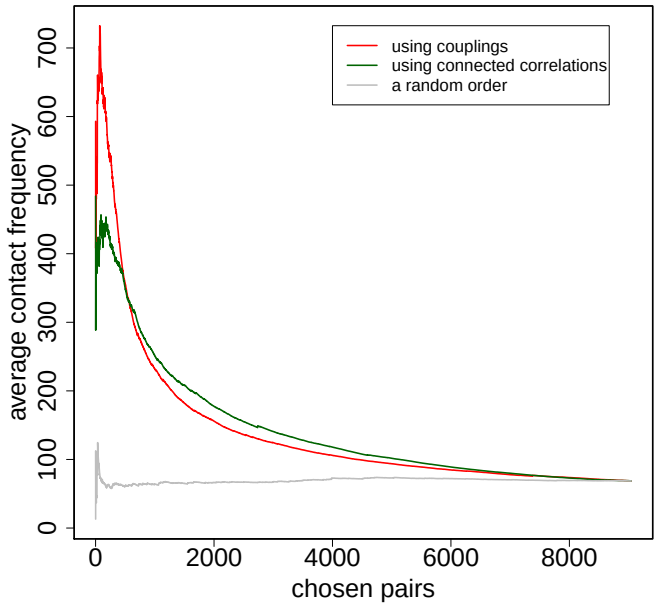


Figure 4.7: **Agreement of the coupling matrix with the contact frequency matrix.** The mean contact frequencies of the pairs of sectors chosen in decreasing coupling strengths (red) or decreasing connected correlations (green) and a random order (a permutation of all sector pairs) (grey).

Comparing the coupling matrix and the contact frequency matrix

As a first attempt to analyse the coupling matrix, we investigate the agreement between the coupling matrix and the contact frequency matrix of the chromosome measured independently using the Hi-C technique (Lieberman-Aiden et al. 2009) as described above[4], see figure 4.2. To have a comparison, we will also consider the agreement between the connected correlation matrix and the contact frequency matrix. We do so in the following way. First, all pairs of sectors are ordered according to their absolute coupling strengths (or correspondingly, their connected correlations). We use p_i^α to denote the ith pair, with α denoting either the ordering using couplings ($\alpha = 1$) or using connected correlations ($\alpha = 2$). The contact frequency of the pair is denoted by $x(p_i^\alpha)$. Suppose we choose n pairs of sectors with strongest couplings (or strongest correlations), we then ask: what is their mean contact frequency? This mean contact frequency of n chosen pairs can be calculated as

$$y^\alpha(n) = \frac{1}{n}\sum_{k=1}^{n} x(p_k^\alpha). \tag{4.4}$$

[4]This aligns with our speculation that pairs of DNA segments are jointly mutated because they are frequently in close proximity. However, in this study we are able to confirm this hypothesis yet.

The mean contact frequencies y^{α} are shown in figure 4.7a. For a comparison, the mean contact frequencies for n fist sector pairs of a random mutation of all the sector pairs are also presented. The figure shows that the mean contact frequency of n pairs of sectors that have strongest couplings is higher than that of n pairs of sectors that are most correlated when n is small ($n < 100$). When n is large, however, the mean contact frequency of n most correlated pairs is higher. Therefore, while the strong couplings inferred by the inverse Ising model are reliable, weak couplings are noisy. Further investigation of figure 4.6 shows that strong couplings are often among the close near neighbours, indicating the one-dimensional organisation of the chromosome.

4.3 Conclusions and outlook

In this chapter, we studied the correlation in copy-number of sectors of chromosomes (which are defined by their physical contact map). We emphasised the conceptual difference between the (indirect) correlations and the (direct) interactions. To accomplish this discrimination, we binarised the copy-number data and applied methods of the inverse Ising problem to infer the couplings that underlie the correlations observed. We showed that a very sparse coupling matrix, which is compatible with the one-dimensional organisation of chromosomes, can explain the dense, blocked-form copy-number correlation matrix. We compared the inferred coupling matrix and the connected correlation matrix with the contact frequency matrix of chromosome 1, where strong inferred couplings shows better agreement.

While the coupling matrix is mainly one-dimensional, we expect that occasional long range couplings between distal parts of a chromosome may also be important. Inferring these sparse long range couplings is still difficult as they are generally weak. Extending the analysis for the couplings between distal parts of a chromosome can be a fruitful project in the future.

References

Bailly-Bechet, M., A. Braunstein, A. Pagnani, M. Weigt, and R. Zecchina (2010). "Inference of sparse combinatorial-control networks from gene-expression data: a message passing approach". *BMC Bioinformatics* 11.1, p. 355.

Carter, S. L. et al. (2012). "Absolute quantification of somatic DNA alterations in human cancer". *Nat. Biotechnol.* 30.5, pp. 413–21.

De, S. and F. Michor (2011). "DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes". *Nat. Biotechnol.* 29.12, pp. 1103–8.

Fischer, A., I. Vázquez-García, C. J. R. Illingworth, and V. Mustonen (2014). "High-definition reconstruction of clonal composition in cancer". *Cell Rep.* 7.5, pp. 1740–52.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning*. Springer.

Huang, J. et al. (2004). "Whole genome DNA copy number changes identified by high density oligonucleotide arrays". *Hum. Genomics* 1.4, pp. 287–99.

Klijn, C., J. Bot, D. J. Adams, M. Reinders, L. Wessels, and J. Jonkers (2010). "Identification of networks of co-occurring, tumor-related DNA copy number changes using a genome-wide scoring approach". *PLoS Comput. Biol.* 6.1. Ed. by W. S. Noble, e1000631.

LaFramboise, T. (2009). "Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances". *Nucleic Acids Res.* 37.13, pp. 4181–93.

Lieberman-Aiden, E. et al. (2009). "Comprehensive mapping of long-range interactions reveals folding principles of the human genome". *Science* 326.5950, pp. 289–93.

Mermel, C. H., S. E. Schumacher, B. Hill, M. L. Meyerson, R. Beroukhim, and G. Getz (2011). "GISTIC 2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers". *Genome Biol.* 21, R41.

Oesper, L., A. Mahmoody, and B. J. Raphael (2013). "THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data". *Genome Biol.* 14.7, R80.

Olshen, A. B., E. S. Venkatraman, R. Lucito, and M. Wigler (2004). "Circular binary segmentation for the analysis of array-based DNA copy number data". *Biostatistics* 5.4, pp. 557–72.

Pinkel, D. et al. (1998). "High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays". *Nat. Genet.* 20.2, pp. 207–11.

Robinson, J. T., H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov (2011). "Integrative genomics viewer". *Nat. Biotechnol.* 29.1, pp. 24–6.

Schneidman, E., M. J. Berry, R. Segev, and W. Bialek (2006). "Weak pairwise correlations imply strongly correlated network states in a neural population". *Nature* 440.7087, pp. 1007–12.

Thorvaldsdóttir, H., J. T. Robinson, and J. P. Mesirov (2013). "Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration". *Brief. Bioinform.* 14.2, pp. 178–92.

Weigt, M., R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa (2009). "Identification of direct residue contacts in protein-protein interaction by message passing". *Proc. Natl. Acad. Sci.* 106.1, pp. 67–72.

Chapter 5

Gene expression correlations in cancer cells

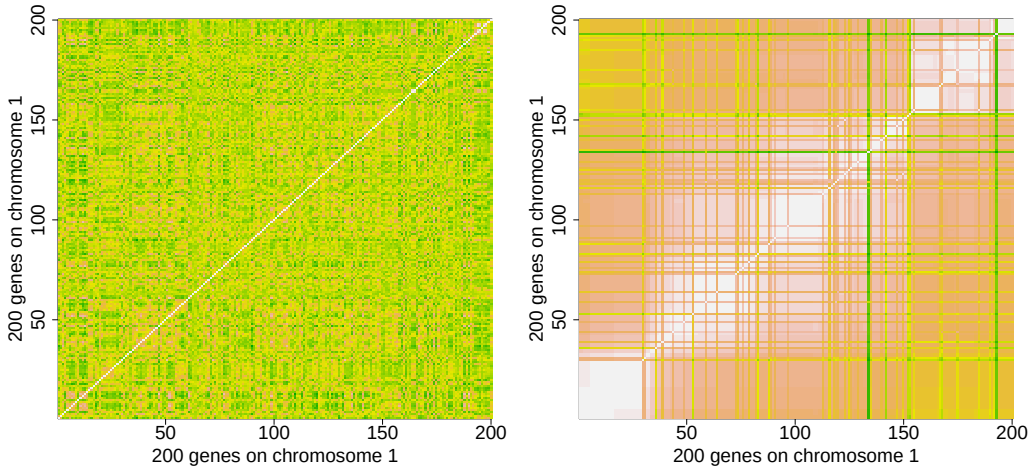
In this chapter, we extend the concepts of indirect correlation and direct interaction to study the influence of copy-number alterations on gene expression correlations. Specifically, we will show that correlations in gene expression level in cancer cells (called observed correlations) do not necessarily imply regulatory interactions: they can be simply induced by the correlations in gene copy-number. We also show that with a simple maximum entropy model it is possible to eliminate this induced correlation. This model brings the concept of bare correlation in gene expression, which captures the regulatory interactions alone. A Gene Ontology similarity analysis shows that gene pairs that are strongly correlated in terms of bare correlation are indeed more likely to be biologically related than those that are correlated in terms of observed correlation.

5.1 Gene expressions of cancer cells and the influence of copy-number alterations

Although direct high-throughput measurements of protein concentrations have been limited, high though-put measurements of messenger RNA (mRNA) concentrations as surrogates to gene expressions are rather popular (Parmigiani and Garett 2003). Due to the early development of microarrays for mRNA gene expression measurement (Lashkari et al. 1997, Schena et al. 1995), methods to analyse genome-wide gene expression levels of cancer cells have attracted interests over the last decades (Parmigiani and Garett 2003). These early analyses concentrated mostly on either molecular classification of tumours or cancer-related gene prediction (Golub 1999). Gene expressions were then used to analyse the regulatory networks of cancer cells (Rhodes et al. 2007). Such analyses were often based on defining clusters of co-expressed genes, or utilising different ways of reconstructing the gene regulatory networks based on gene expression correlations (Bonnet et al. 2010, D'haeseleer et al. 2000, Segal, Friedman, et al. 2004, Segal, Shapira, et al. 2003). For a thorough introduction to gene expression analysis, see Parmigiani and Garett 2003.

Later came the technology to measure gene copy-numbers on a genome-wide scale

with high resolution (J. Huang et al. 2004, LaFramboise 2009, Pinkel et al. 1998). While gene expressions can vary from cell to cell, gene copy-numbers are believed to remain rather stable in a clonal population. Therefore, cancer analysis based on copy-numbers are believed to be more reliable. Still, analyses at the genome-wide scale (such as GISTIC) are of high dimension and also subject to the curse of dimensionality and other problems, see also chapter 4.



(a) (b)

Figure 5.1: **Impacts of copy-number alterations on the correlations in gene expression.** (a) The correlation matrix of expressions of 200 genes on chromosome 1. A closer look reveals correlated blocks along the chromosome, i.e. genes that are close to each other tend to be correlated. (b) The correlation matrix of copy-number of the same 200 genes. We will show that the block structure of the correlation matrix of expressions partly originates from the block structure of the correlation matrix of copy-numbers.

A natural idea is to combine the copy-number data with the gene expression data to maximise the information for analysis. Many approaches to incorporating the two different data types have been proposed; although several formalisms emerged (N. Huang et al. 2012, Kirk et al. 2012, Masica and Karchin 2011, Menezes et al. 2009, J. Wang et al. 2012), there seems to be no convincing unique framework.

In this chapter, we highlight a very simple effect of copy-number alterations on gene expression in cancer cells. The effect is illustrated in figure 5.1a, which shows the terrain map of the correlation matrix of some 200 genes on chromosome 1, ordered in the chromosomal order. Although it is not easy to detect with the naked eyes, a careful look at the figure reveals that it has block structure: genes in the chromosome order tend to form correlated blocks, which seem to act as single units. Where do such blocks come from? It could be that genes that are close to each other on a chromosome tend to be regulatorily related (e.g., they may share the same transcription factors or the same chromatin

structures). However, we will show that there is another source that induces the correlation between genes that are close to each other on a chromosome, which is (incidentally) important for cancer cells.

Figure 5.1b shows the terrain map of the correlation matrix of copy-numbers of the same 200 genes. The correlation matrix of copy-numbers also has very clear blocks, reflecting the geometrical organisation of the chromosome (see also chapter 4). Putting the two correlation matrices side-by-side, we speculate that the blocks in the gene expression correlation matrix may be in fact some "images" of the blocks in the copy-number correlation matrix. This turns out to be true and the phenomenon can be explained in a rather simple way. If gene A and gene B are close to each other in the genome, then they should be strongly correlated in gene copy-numbers as shown in the copy-number correlation matrix. On the other hand, the expression of gene A is correlated with its copy-number and the expression of gene B is also correlated with its copy-number. Therefore, even without any biological relation (a shared transcription factor or shared chromatin structure), the expression of gene A and the expression gene B may still be correlated. The phenomenon becomes even more obvious in a comparison with an Ising model of four spins, as shown in figure 5.2.

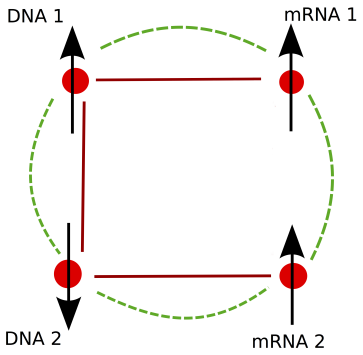


Figure 5.2: **A system of four Ising spins.** The spins correspond to the copy-numbers and expressions of two genes: DNA1, DNA2, mRNA1, mRNA2. The copy-numbers, DNA1 and DNA2, are coupled. On the other hand, mRNA1 is coupled to DNA1 and mRNA2 is coupled to DNA2. Therefore, even without regulatory interactions, i.e. a direct coupling (red lines), between mRNA1 and mRNA2, one still observes correlation (green arcs) between them.

We conclude that the correlations observed among gene expressions, apart from biological reasons, can also be induced by the correlations in copy-numbers of the genes. (Although the visual signal of the induction in figure 5.1 seems weak, we will show that

the effect is clear at the genome-wide scale.) In the following, based on the maximum entropy principle, we formulate a model that allows us to eliminate these induced correlations from the *observed correlations*, resulting in the so-called *bare correlations* in expression between genes.

5.2 Elimination of the copy-number-induced correlation

5.2.1 Maximum entropy model for induced-correlation elimination

With both copy-number and gene expression data available for N genes, a cancer sample is characterised by N pairs of variables (X_i, Y_i), where X_i stands for the copy-number and Y_i stands for the expression of gene i, with $i = \{1, 2, \ldots, N\}$. More precisely, X_i and Y_i are the signal intensities measured by microarrays subjected to preprocessing methods, taking on real values.[1] We note that this is different from chapter 4, where copy-numbers are integers obtained as the outputs of an impurity correcting algorithm which takes the measured signals as inputs. Here, it is more convenient to work with the real-valued copy-numbers before impurity-correction. This is also to be consistent with the expression data, where, to our knowledge, no impurity-correction is possible.

Following the general maximum entropy reasoning, in order to model the data, we need to specify the set of relevant observables, which constitutes the sufficient statistics. In this case, we choose the set of observables as

(i) the mean values of gene expressions, $\langle Y_i \rangle$;

(ii) the correlations among gene expressions, $\langle Y_i Y_j \rangle$;

(iii) the correlations between copy-numbers and gene expressions of the *same* genes, $\langle X_i Y_i \rangle$.

The corresponding maximum entropy model is then

$$\begin{aligned} H(\{x_i, y_i\}) &= H^X(\{x_i\}) - \sum_i \lambda_i x_i y_i + \\ &\quad - \sum_i \beta_i y_j + \frac{1}{2} \sum_{ij} A_{ij} y_i y_j, \end{aligned} \tag{5.1}$$

where $H^X(\{x_i\})$ models the copy-numbers in a way which we do not need to specify here (see also chapter 2). The corresponding partition function is

$$Z = \int \prod_i \mathrm{d}\, x_i \, \mathrm{d}\, y_i \exp[-H(\{x_i, y_i\})]. \tag{5.2}$$

[1] Raw data of gene expression were subjected to quantile normalisation and then log transformed. Raw copy-number data were segmentation smoothened as explained in chapter 4 and copy-numbers of genes were then derived from the segmented data.

The gene expression variables y_i in the partition function can be integrated out to yield

$$Z = \int \prod_i \mathrm{d}\, x_i e^{-\tilde{H}(\{x_i\})}, \tag{5.3}$$

where

$$\begin{aligned} \tilde{H}(\{x_i\}) &= H^X(\{x_i\}) - \frac{N}{2}\ln 2\pi + \frac{1}{2}\ln\det(A) + \\ &\quad -\frac{1}{2}\sum_{ij} B_{ij}(\beta_i + \lambda_i x_i)(\beta_j + \lambda_j x_j), \end{aligned} \tag{5.4}$$

with $B = A^{-1}$.

From the Hamiltonian (5.1), we see that the quantities B_{ij} are the correlations in gene expression *if* there are no copy-number alterations (i.e., $X_i = \text{const}$). These quantities B_{ij} are different from the *observed correlations* $C_{ij} = \langle Y_i Y_j \rangle - \langle Y_i \rangle \langle Y_j \rangle$, which are calculated in the presence of copy-number alterations (and which are observed in experiments). Therefore, we call B_{ij} *bare correlations*, which imply that B_{ij} are the correlations in gene expression after eliminating from C_{ij} the correlation induced by the copy-numbers. Bare correlations are supposed to be due to regulatory interactions exclusively.

From (5.3) and (5.4), and noting that

$$\langle X_i Y_i \rangle = \frac{1}{Z}\frac{\partial Z}{\partial \lambda_i}, \tag{5.5}$$

$$\langle Y_i \rangle = \frac{1}{Z}\frac{\partial Z}{\partial \beta_i}, \tag{5.6}$$

$$\langle Y_i Y_j \rangle = \frac{1}{Z}\frac{\partial^2 Z}{\partial \beta_i \partial \beta_j}, \tag{5.7}$$

it is straightforward to show that

$$C_{ij} = B_{ij} + \sum_{kl} B_{ik} B_{jl} D_{kl} \lambda_k \lambda_l, \tag{5.8}$$

$$R_i = \sum_j B_{ij} D_{ij} \lambda_j, \tag{5.9}$$

with $R_i = \langle X_i Y_i \rangle - \langle X_i \rangle \langle Y_i \rangle$ and $D_{ij} = \langle X_i X_j \rangle - \langle X_i \rangle \langle X_j \rangle$. The observed gene expression correlations C_{ij} and the gene expression–copy-number correlations R_i are identified with the corresponding estimated values from data by the requirement of maximum entropy models (see also chapter 2). If we further also use the experimental estimated values for D_{ij}, equations (5.8) and (5.9) can be solved for the bare correlations B_{ij} and the copy-number–gene expression couplings λ_i. However, a careful look at equation (5.9) reveals

that it suffers from an over-fitting problem. Indeed, the correlation matrices C_{ij} and D_{ij} are usually degenerate since the number of genes far exceeds the number of samples.[2] This means that there is no guarantee that the matrix $B_{ij}D_{ij}$ is invertible; in practice, it is always found to be degenerate. To get around this difficulty, we introduce the so-called *weak-correlation approximation*.

Weak-correlation approximation

The weak-correlation approximation assumes that the gene–gene bare correlations in expression are small, so that to the first approximation we can substitute B_{ij} by a diagonal matrix. This directly lifts the degeneracy of (5.9). In fact, both first approximations for B_{ii} and λ_i can then be found easily,

$$B_{ii} = C_{ii} - \frac{R_i^2}{D_{ii}}, \tag{5.10}$$

$$\lambda_i = \frac{R_i}{B_{ii}D_{ii}}. \tag{5.11}$$

In the second round of approximation, we use the solution of λ_i, and numerically solve equation (5.8) for B_{ij}. Using linear algebra notation, we rewrite (5.8) as a matrix equation

$$C = B + BEB, \tag{5.12}$$

where $E_{ij} = D_{ij}\lambda_i\lambda_j$ and we have used the symmetry of B, $B_{ij} = B_{ji}$. Since E is symmetric and positive definite, there exists F so that $E = F^2$. (The matrix F can be found by first using a similarity transformation to diagonalise E, taking the square-root of the eigenvalues, then back transforming the root with the reverse similarity transformation.) By multiplying the two sides of equation (5.12) from the left and the right by F, we have

$$\tilde{C} = \tilde{B} + \tilde{B}^2, \tag{5.13}$$

where $\tilde{C} = FCF$ and $\tilde{B} = FBF$. Equation (5.13) implies that the eigenvectors of $\tilde{B}$ are also the eigenvectors of $\tilde{C}$. Therefore, by a similarity transformation that brings $\tilde{C}$ into diagonal form, (5.13) becomes quadratic algebraic equations of eigenvalues, which can be solved easily.

Although the above procedure allows to solve for the bare correlation matrix B explicitly, the computational cost at the level of genome-wide scale is still not affordable in practice. Moreover, for very large system sizes, noise may be added up in such a global

[2] We have $\text{rank}(C) \leq \min\{N, M-1\}$, where N is the number of genes and M is the number of samples. The same is true for D.

solution. We therefore introduce a further approximation.

Isolated-pair approximation

The isolated-pair approximation implies considering each pair (i, j) independently from all the other genes. As a result, we have $N(N-1)/2$ versions of equation (5.8) for all pairs of genes, each has the B matrix of degree (2×2). Moreover, if we fix the diagonal elements B_{ii} and B_{jj} by the weak-correlation approximation (5.10), a quadratic equation for the only unknown B_{ij} can be obtained explicitly,

$$\begin{aligned} C_{ij} &= E_{ij}B_{ij}^2 + (1 + B_{ii}E_{ii} + B_{jj}E_{jj})B_{ij} + \\ &\quad + B_{ii}B_{jj}E_{ij}, \end{aligned} \tag{5.14}$$

where $E_{ij} = D_{ij}\lambda_i\lambda_j$. One of the two solutions of B_{ij} to the equation can be chosen by the condition that B_{ij} tends to C_{ij} as E_{ij} vanishes, namely

$$B_{ij} = \frac{-(1 + B_{ii}E_{ii} + B_{jj}E_{jj}) + \sqrt{\Delta}}{2E_{ij}}, \tag{5.15}$$

where $\Delta = (1 + B_{ii}E_{ii} + B_{jj}E_{jj})^2 - 4B_{ii}B_{jj}E_{ij}^2$.

Under these approximations, the calculation of B_{ij} is very fast. Moreover, the effective number of (intermediate) parameters is reduced from $O(N^2)$ in one problem to 1 parameter in $N(N-1)/2$ separated problems. Overfitting problems can therefore be avoided.

5.2.2 Analysing the copy-number and gene expression data of lung cancer patients

Test of the isolated-pair approximation

To study the quality of the isolated-pair approximation, we perform the inference for a restricted number of genes on chromosome 1 and compare the results with that of weak-correlation approximation alone (i.e., without isolated-pair approximation). The bare correlations inferred by the two methods after normalised as $\tilde{B}_{ij} = B_{ij}/\sqrt{B_{ii}B_{jj}}$ are plotted versus the normalised observed correlations $\tilde{C}_{ij} = C_{ij}/\sqrt{C_{ii}C_{jj}}$ in figure 5.3. We see that for many pairs of genes, while the observed correlations are large, the bare correlations inferred by either methods are rather small. Importantly, when the correlations are sufficiently strong, the results inferred by the isolated-pair approximation agree well with that of the weak-correlation approximation (region III in figure 5.3). This means it is safe using the isolated-pair approximation for pairs of genes that are correlated strongly enough. Note that it is also those correlated genes that we are most interested in. In the following, we will use isolated-pair approximation exclusively to eliminate the gene expression

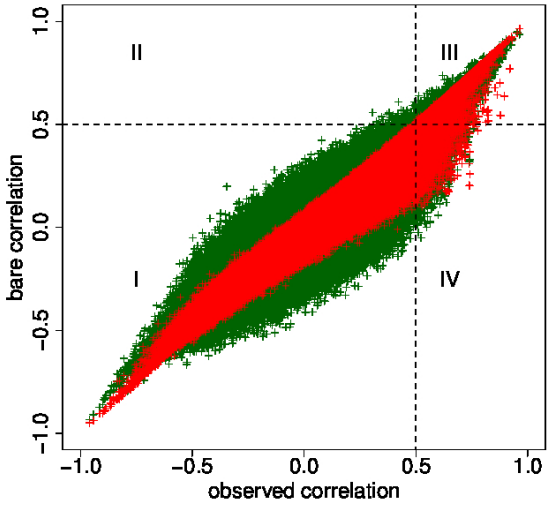


Figure 5.3: **Test of isolated-pair approximation.** The normalised bare correlations, $\tilde{B}_{ij} = B_{ij}/\sqrt{B_{ii}B_{jj}}$, inferred by weak-correlation approximation (green) and isolated-pair approximation (red) are plotted versus the normalised observed correlations $\tilde{C}_{ij} = C_{ij}/\sqrt{C_{ii}C_{jj}}$. The bare correlations inferred by both methods agree well for pairs that are strongly correlated (region III), which are of our central interest.

correlations induced by the correlations in copy-number at genome-wide scale.

Sliding windows and strong correlation cutoffs

The induction of correlation in gene expression by copy-number correlations is most important when the copy-number correlations D_{ij} are strong. Instead of scanning all pairs of genes over the whole genome, we restrict ourselves to those pairs of genes that are not further than W base-pairs (bp) away from each other, for which the copy-numbers are indeed highly correlated. Figure 5.4 shows a scatter plot between the inferred bare correlations, normalised as $\tilde{B}_{ij} = B_{ij}/\sqrt{B_{ii}B_{jj}}$, and the observed correlations, normalised as $\tilde{C}_{ij} = C_{ij}/\sqrt{C_{ii}C_{jj}}$, with $W = 500000$ bp. A threshold t is set on both axes to select for strong correlated pairs. The difference between the sets of pairs selected by $\tilde{B}_{ij} > t$ and by $\tilde{C}_{ij} > t$ is clear from the figure.

Connection between bare correlations and gene function

We study the connection between bare correlations and gene function using Gene Ontology (The Gene Ontology Consortium 2000). Gene Ontology (GO) is a system of terms to describe genes. The terms can be related to each other by different relations, which form a so-called semantic network. The GO consists of three of such networks (three fam-

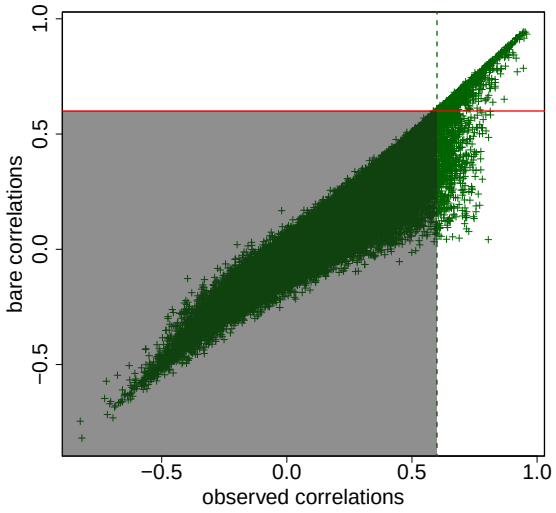


Figure 5.4: **Inferred bare correlations versus observed correlations.** Strong observed correlations can show very small bare correlations. Selection of strongly correlated pairs by setting the cutoff at $t = 0.6$ on the (normalised) bare correlations $\tilde{B}_{ij}$ (red, solid line) excludes a large number of pairs that are suggested by setting the same cutoff over the (normalised) observed correlation $\tilde{C}_{ij}$ (green, dashed line).

ilies): Molecular Function (MF), Biological Process (BP) and Cellular Component (CC). Each gene can be described by a set of terms in the networks. Here we are interested in the similarities between pairs of genes, which can be measured by different semantic similarity measures defined on the network. Specifically we use Wang's measure (J. Z. Wang et al. 2007), which is implemented in the R package GOSemSim (R Core Team 2012, Yu et al. 2010). Wang's similarity ranges from 0 to 1, with 1 for maximum similarity. A pair of genes will be called *GO-closely-related* if their GO similarity is larger than 0.95.

The fractions of GO-closely-related pairs in the sets of strongly correlated gene pairs selected using either $\tilde{C}_{ij} > t$ or $\tilde{B}_{ij} > t$ reflects how good the selections are in discovering genes that are biologically related. Figure 5.5 shows the fractions of GO-closely-related pairs in all the correlated pairs, selected either as $\tilde{C}_{ij} > t$ (green) or as $\tilde{B}_{ij} > t$ (red), when the threshold t varies from 0.5 to 1.0. Higher fractions of GO-closely-related pairs selected by using bare correlations, $\tilde{B}_{ij} > t$, than that by using observed correlations, $\tilde{C}_{ij} > t$, persist essentially for all thresholds t in the range (apart from the fluctuations happening when the thresholds are high). The differences are particularly clear in the case of Cellular Component GO family.

We then vary the window size W, keeping the threshold for correlation fixed at $t = 0.7$ (dashed black lines in figure 5.5). Because the fractions of GO-closely-related pairs in *all* considered pairs (i.e., all pairs of genes that are not farther than W bp from each other)

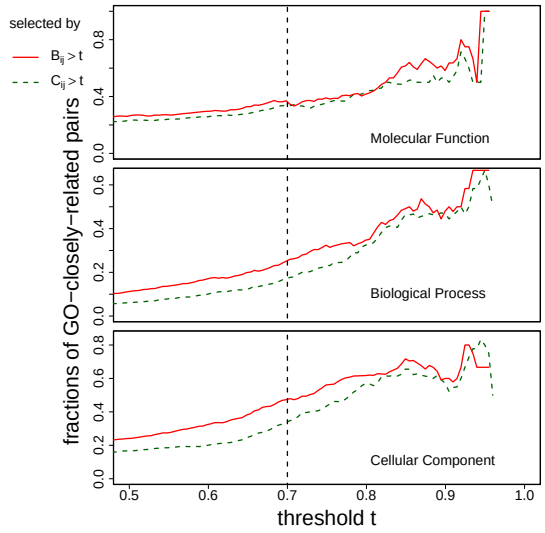


Figure 5.5: **Fractions of GO-closely-related gene pairs in the selected pairs versus the selection threshold t at $W = 1000$ kbp.** Fractions of GO-closely-related gene pair (for three families: Molecular Function, Biological Process, Cellular Component) in the gene pairs selected using bare correlations $\tilde{B}_{ij} > t$ (red, solid) are consistently higher than that selected by using observed correlations $\tilde{C}_{ij} > t$ (green, dashed). When the threshold is too high (more than 0.85), the fractions show fluctuations due to the small numbers of gene pairs selected.

decrease as the window sizes W get large, the fractions of GO-closely-related pairs in the *selected* pairs using $\tilde{B}_{ij} > t$ or using $\tilde{C}_{ij} > t$ both slowly decrease as W increases, see figure 5.6. Higher fractions of GO-closely-related selections using bare correlations $\tilde{B}_{ij}$ however persist for all three GO families in the full ranges of W up to 1000 kbp.

5.3 Conclusions and outlook

We have shown that the correlations in gene expression induced by the correlations in copy-number are significant in cancer cells with copy-number alterations. The effect is particularly strong for those genes that are close to each other in the genome. Based on the maximum entropy principle, we formulated a simple model to tell apart the effects from copy-number correlations and the effects from regulatory interactions in the gene expression correlation, introducing the concept of bare correlations which capture the effect of regulatory interactions alone. Gene pairs that have strong bare correlations indeed showed a closer similarity in terms of Gene Ontology than gene pairs that have strong observed correlations.

The copy-number-induced correlations in gene expressions studied in this chapter are closely related to the concept of indirect correlations introduced in the context of the inverse Ising problem in chapter 4. We consider the maximum entropy modelling as a gen-

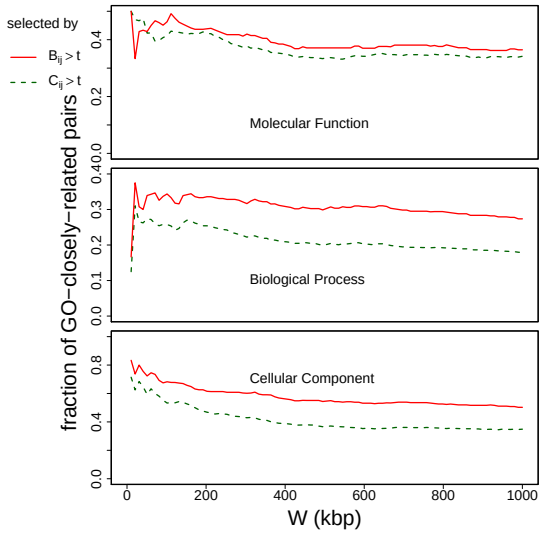


Figure 5.6: **Fractions of GO-closely-related gene pairs in the selected pairs versus window sizes W at selection threshold $t = 0.7$ (black dashed line in figure 5.5).** Varying the window size W up to 1000 kbp, we clearly observed that the fractions of GO-closely-related pairs using $\tilde{B}_{ij} > t$ are consistently higher than that of using $\tilde{C}_{ij} > t$ for all three families of Gene Ontology.

eral framework to address this problem, leading to simple and conceptually straightforward solutions.

References

Bonnet, E., T. Michoel, and Y. Van de Peer (2010). "Prediction of a gene regulatory network linked to prostate cancer from gene expression, microRNA and clinical data". *Bioinformatics* 26.18, pp. i638–44.

D'haeseleer, P., S. Liang, and R. Somogyi (2000). "Genetic network inference: from co-expression clustering to reverse engineering". *Bioinformatics* 16.8, pp. 707–726.

Golub, T. R. (1999). "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring". *Science* 286.5439, pp. 531–537.

Huang, J. et al. (2004). "Whole genome DNA copy number changes identified by high density oligonucleotide arrays". *Hum. Genomics* 1.4, pp. 287–99.

Huang, N., P. K. Shah, and C. Li (2012). "Lessons from a decade of integrating cancer copy-number alternations with gene expression profiles". *Brief. Bioinform.* 13.3, pp. 305–316.

Kirk, P., J. E. Griffin, R. S. Savage, Z. Ghahramani, and D. L. Wild (2012). "Bayesian correlated clustering to integrate multiple datasets". *Bioinformatics* 28.24, pp. 3290–3297.

LaFramboise, T. (2009). "Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances". *Nucleic Acids Res.* 37.13, pp. 4181–93.

Lashkari, D. A., J. L. DeRisi, J. H. McCusker, A. F. Namath, C. Gentile, S. Y. Hwang, P. O. Brown, and R. W. Davis (1997). "Yeast microarrays for genome wide parallel genetic and gene expression analysis". *Proc. Natl. Acad. Sci.* 94.24, pp. 13057–13062.

Masica, D. L. and R. Karchin (2011). "Correlation of somatic mutation and expression identifies genes important in human glioplastoma progression and survival". *Cancer Res.* 71.13, pp. 4550–4561.

Menezes, R. X., M. Boetzer, M. Sieswerda, G. B. van Ommen, and J. M. Boer (2009). "Integrated analysis of DNA copy number and gene expression microarray data using gene sets". *BMC Bioinformatics* 10.203.

Parmigiani, G. and E. S. Garett (2003). *The Analysis of Gene Expression Data: Methods and Software*. Springer.

Pinkel, D. et al. (1998). "High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays". *Nat. Genet.* 20.2, pp. 207–11.

R Core Team (2012). *R: A languagez and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Rhodes, D. R. et al. (2007). "Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles". *Neoplasia* 9.2, pp. 166–180.

Schena, M., D. Shalon, R. W. Davis, and P. O. Brown (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray". *Science* 270.5235, pp. 467–470.

Segal, E., N. Friedman, D. Koller, and A. Regev (2004). "A module map showing conditional activity of expression modules in cancer". *Nat. Genet.* 36.10, pp. 1090–8.

Segal, E., M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman (2003). "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data". *Nat. Genet.* 34.2, pp. 166–76.

The Gene Ontology Consortium (2000). "Gene Ontology: tool for the unification of biology". *Nat. Genet.* 25, pp. 25–29.

Wang, J. Z., Z. Du, R. Payattakool, P. S. Yu, and C. F. Chen (2007). "A new method to measure the semantic similarity of GO terms". *Bioinformatics* 23.10, pp. 1274–1281.

Wang, J., Y. Zhang, C. Marian, and H. W. Ressom (2012). "Identification of aberrant pathways and network activities from high-throughput data". *Brief. Bioinform.* 13.4, pp. 406–419.

Yu, G., F. Li, Y. Qin, X. Bo, Y. Wu, and S. Wang (2010). "GOSemSim: an R package for measuring semantic similarity among GO terms and gene products". *Bioinformatics* 26.7, pp. 976–978.

Chapter 6

Semi-supervised classification of lung cancers

Based on mutation profiles, cancer samples are classified into different subtypes for further study and for appropriate treatments. In this chapter, we introduce a simple method for cancer mutation profile classification using a mixture of several paramagnetic models with Ising spins (or mixture of paramagnetic Ising model, for simplicity). We further suggest a semi-supervised learning algorithm for the model, which is capable of both learning from unclassified (unlabelled) samples as well as correcting misclassified (mislabelled) samples. Comparison of the performance of the semi-supervised model with that of the supervised and unsupervised models on simulated datasets demonstrates its superior performance. We then apply the the model to a dataset of lung cancers to review the initial histological classifications of the samples.

6.1 Cancer classification based on mutation profiles

Today, cancer is viewed as a complex disease consisting of many different types of tumours with different behaviours. Traditionally, cancers are classified according to the tissues and cell-types from which they originated (see chapter 1). With deeper understanding of the molecular origin of cancer it is natural to think of classifying cancers according to their molecular markers (e.g., some particular mutated genes) (Lakhani and Ashworth 2001). Indeed, tumours with different molecular markers behave differently even when they originate from the same tissue and the same cell type. On the other hand, few markers are shared between cancers of different origins, indicating some common behaviours between them. More importantly, when cancers are classified according to molecular markers, their molecular defects are highlighted and personalised treatment strategy can be designed (Lakhani and Ashworth 2001).

Early molecular classifications of cancer were based on one or two molecular markers. With the development of genomic study, molecular signatures of cancer have been extended from few or single markers (e.g., single mutated genes) to genomic profile classi-

fication, which collectively involves a large number of markers. Early cancer classification methods based on genome-wide gene expression (which contain some thousands of genes at the time) were often simple applications of machine learning techniques of clustering, e.g., K-means clustering (Golub 1999). During the last decade, a diverse set of methods of classification of gene expression profiles, copy-number profiles and other kinds of mutation profiles have been proposed (Nutt et al. 2003, Pusztai et al. 2006, Sotiriou et al. 2003, Souto et al. 2008, Wang et al. 2006). However, classification at the level of genome-wide is a high dimensional problem, and is therefore subject to the *curse of dimensionality* as we also discussed in chapter 4 and chapter 5. A better approach is to concentrate on a smaller subset of markers, which may be still much larger than one or few selected markers, but not the whole genome. Since such a subset often requires other sources of knowledge, those methods are also *knowledge-driven*. To date, classification and molecular characterisation of cancer has been developed for many different types of cancers (The Cancer Genome Atlas Research Network 2012a,b,c, 2014, Viale 2012).

Working with collaborators from Department of Translational Medicine (University of Cologne), we are involved in the problem of classifying lung cancers (The Clinical Lung Cancer Genome Project (CLCGP) and Network Genomic Medicine (NGM) 2013). The collected dataset consists of 766 mutation profiles of lung cancer samples, each contains the binary states ('mutated' or 'not mutated') of 18 molecular markers, including gene copy-number alterations (amplification or deletion) or gene sequence mutations, as shown in the first panel of figure 6.1. The cancer samples are initially classified into five histological types by pathologists on the basis of visual appearance of cells under a microscope: adenocarcinomas (AD), squamous carcinomas (SQ), small cell lung carcinomas (SCLC), large cell carcinomas (LCC) and large cell neuroendocrine carcinomas (LCNEC), as shown in figure 6.1.

This dataset and the initial histological classification can be used to train a model for lung cancer classification. Many different statistical models can be used. Here, we focus on a simple one. We can think of the 18 molecular markers as 18 Ising spins. We start with associating each type of cancer with a paramagnetic Ising Hamiltonian characterised by a distinct set of fields acting on the 18 spins. Since there are 5 cancer types in the dataset, we have a system of 5 paramagnetic Ising Hamiltonians (called a *mixture model*). From the data, the fields for each Hamiltonian can be estimated (or *learned*). Now for each new sample, we can calculate the probability for it to come from a specific Hamiltonian. Based on the calculated probabilities, the sample is then classified into one of the 5 types.

The problem with such approach is: the initial classifications of the cancer samples used to train the model can be incorrect. These initial classifications were based on visual

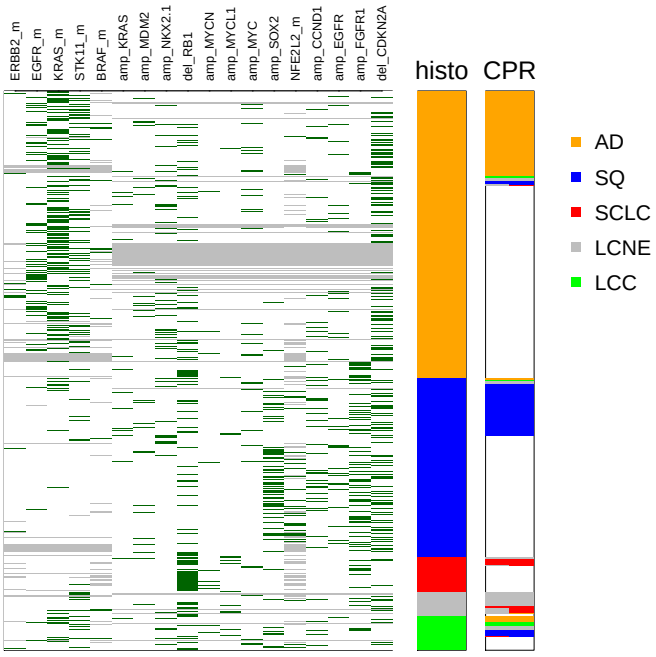


Figure 6.1: **Summary of the dataset.** The first panel shows the data matrix with loci on the columns and samples on the rows. A green line indicates that the locus is mutated. The grey lines are failures in measurements. The second panel (histo) shows the initial histological classifications of the samples into five types: adenocarcinomas (AD), squamous carcinomas (SQ), small cell lung carcinomas (SCLC), large cell lung carcinomas (LCC) and large cell neuroendocrine lung carcinomas (LCNEC). The last panel presents a secondary more reliable classification of the samples termed central pathological reviews (CPR).

inspection by pathologists and there can be errors: sometimes adenocarcinomas could be erroneously observed as squamous carcinomas, for example. Indeed, a more reliable secondary classification of a subset of the samples (will be referred to as central pathological reviews – CPR) reveals discrepancy, see the third panel of figure 6.1. Such misclassifications in the training dataset apparently affect our classification model: the fields of the paramagnetic Hamiltonian can be wrongly estimated. In the following, after detailing the framework of paramagnetic Ising mixture model, we design a semi-supervised learning algorithm that allows the model to correct the misclassified samples during learning.

There is another bonus to such a model: it can also learn from both classified (or labelled) and unclassified (or unlabelled) samples. This is important in many situations where the initial classifications of samples to train a supervised model are rare, while unclassified samples are abundant. Although unsupervised clustering methods are available for learning from unclassified samples, they waste available classified samples and the performance can be poor as a result. By learning from both classified samples (as a supervised model does) and unclassified samples (as an unsupervised model does), semi-supervised

models maximise the data available for learning, significantly improving the performance.

Curiously, although semi-supervised learning is widely discussed in machine learning literature (Zhu 2005, Zhu and Goldberg 2009), there are few explicit applications of them in cancer research to our knowledge. We also note that there is no single way of semi-supervising a model, many of them are rather arbitrary tricks. Here, we introduce a very simple one (The Clinical Lung Cancer Genome Project (CLCGP) and Network Genomic Medicine (NGM) 2013).

6.2 Semi-supervised model to classify cancer mutation profiles

6.2.1 Semi-supervising the mixture of paramagnetic Ising models

Let us recall the mixture of paramagnetic Ising models discussed in chapter 2. Instead of considering a mixture of separate paramagnetic Ising Hamiltonians, we can equivalently consider a system which consists of a Potts spin Φ coupled to N binary Ising spins $\{X_i\}$, $i = 1, 2, \ldots, N$, with Hamiltonian

$$H(\phi, \{x_i\}) = -\sum_t \sum_i b_i^t x_i \delta(\phi, t) - \sum_t a_t \delta(\phi, t), \tag{6.1}$$

where t denotes the state of the Potts spin and δ is the Kronecker-delta function. The coefficient a_t can be interpreted as a field acting on state t of the Potts spin and b_i^t is the coupling between state t of the Potts spin to Ising spin i. Recall that this is also the maximum entropy model with observables $\langle \delta(\Phi, t) X_i \rangle$ and $\langle \delta(\Phi, t) \rangle$ as sufficient statistics. Moreover, note that there is a constraint on the observables, $\sum_t \delta(\Phi, t) = 1$, thus a_t are underdetermined.

The Boltzmann distribution corresponding to this Hamiltonian is

$$p(\phi, \{x_i\}) = \frac{1}{Z} e^{-H(\phi, \{x_i\})}, \tag{6.2}$$

where

$$Z = \sum_\phi e^{a_\phi} \prod_i 2 \operatorname{ch}(b_i^\phi). \tag{6.3}$$

In considering our discussion in chapter 2, we define new parameters

$$\pi_t = \frac{e^{a_t} \prod_i 2 \operatorname{ch}(b_i^t)}{\sum_{t'} e^{a_{t'}} \prod_i 2 \operatorname{ch}(b_i^{t'})}, \tag{6.4}$$

and rewrite the distribution in a somewhat simpler form,

$$p(\phi, \{x_i\}) = \prod_t [\pi_t p(\{x_i\}|t)]^{\delta(\phi,t)}, \tag{6.5}$$

where $p(\{x_i\}|t) = e^{\sum_i b_i^t x_i}/Z_t$ with $Z_t = \prod_i 2\,\mathrm{ch}\, b_i^t$. Note that while a_t are underdetermined, π_t satisfy $\sum_t \pi_t = 1$ and are completely determined.

In the context of cancer classification, the Potts spin Φ represents K different types of cancers, and the N spins represent N binary states of N loci (mutated or not mutated) – the mutation profile. The classification is the problem of inferring the state of the Potts spin Φ based on the observed Ising spins $\{x_i\}$. The problem can also be considered as labelling the Ising spin configurations with the states of the Potts spins; therefore the type of a cancer sample is also called *label*. In order to do so, we have to estimate the parameters $\{a_t, b_i^t\}$, or equivalently $\{\pi_t, b_i^t\}$.

Supervised learning

In the standard *supervised learning*, we need a training dataset of M samples, $D = \{\phi^\mu, x_i^\mu\}$, where all mutation profiles $\{x^\mu\}$ and the types $\{\phi^\mu\}$ of the cancer samples are known. The log-likelihood for the parameters $\{\pi_t, b_i^t\}$ can be written down easily,

$$\ln L(\{\pi_t, b_i^t\}) = \sum_\mu \sum_t \delta(\phi^\mu, t) \left[\ln \pi_t + \sum_i \left(h_i x_i^\mu - \ln 2\,\mathrm{ch}\, b_i^t \right) \right], \tag{6.6}$$

which is to be maximised, taking into account the normalisation constraint $\sum_t \pi_t = 1$. To incorporate the normalisation constraint, we introduce a Lagrange multiplier λ, and modify the log-likelihood function to

$$\tilde{l}(\{\pi_t, b_i^t\}) = \ln L(\{\pi_t, b_i^t\}) - \lambda(\sum_t \pi_t - 1). \tag{6.7}$$

Taking the derivatives of $\tilde{l}$ with respect to π_t and setting them to zero, we can solve for π_t,

$$\pi_t = \frac{1}{M} \sum_\mu \delta(\phi^\mu, t), \tag{6.8}$$

where we have used the constraint $\sum_t \pi_t = 1$ to find that $\lambda = M$.

Taking the derivatives of $\tilde{l}$ with respect to m_i^t and setting them to zero, we can solve for b_i^t,

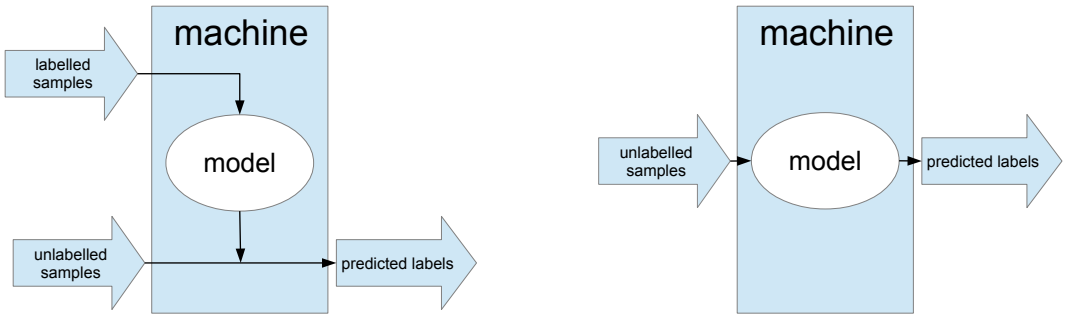
$$\mathrm{th}\, b_i^t = \frac{\sum_\mu x_i^\mu \delta(\phi^\mu, t)}{\sum_\mu \delta(\phi^\mu, t)}. \tag{6.9}$$

Note that $\langle \delta(\Phi,t) \rangle = \pi_t$, $\langle \delta(\Phi,t) x_i \rangle = \text{th}\, b_i^t$, $\langle \delta(\Phi,t) \rangle^D = 1/M \sum_\mu \delta(\phi^\mu, t)$, $\langle \delta(\Phi,t) x_i \rangle^D = 1/M \sum_\mu x_i^\mu \delta(\phi^\mu, t)$, equations (6.8) and (6.9) are identical to (2.100) and (2.101),

$$\langle \delta(\Phi,t) \rangle = \langle \delta(\Phi,t) \rangle^D , \tag{6.10}$$

$$\frac{\langle \delta(\Phi,t) X_i \rangle}{\langle \delta(\Phi,t) \rangle} = \frac{\langle \delta(\Phi,t) X_i \rangle^D}{\langle \delta(\Phi,t) \rangle^D} . \tag{6.11}$$

In practice, training datasets, for which the types of all the cancer samples are known (labelled samples), are usually small. On the other hand, samples to be classified, of which the types are not known (unlabelled samples), are abundant. Let us discuss the *unsupervised learning* (or *clustering*), which is capable of learning from unlabelled samples (see figure 6.2).



(a) (b)

Figure 6.2: **The workflows of supervised and unsupervised models.** (a) A supervised learning algorithm learns from labelled samples and classifies unlabelled samples. (b) An unsupervised learning algorithm takes all samples as unlabelled and classifies them using the clustering effect.

Unsupervised learning: EM algorithm

The presence of different discrete levels of Φ, which are coupled to the Ising spins, causes the configurations of the Ising spins to group into clusters in their configuration space. (Each cluster corresponds to a single state of the Potts spin, and therefore a single paramagnetic Ising Hamiltonian.) So, even though the states of Φ are not observed, the clustering structure of the Ising spins gives us some information about the presence of these states. In fact, this clustering effect allows us to infer these hidden states of Φ up to a permutation of the states.

Suppose the dataset consists of M configurations of N Ising spins $D = \{x_i^\mu\}$ and ϕ^μ are completely unobserved. The log-likelihood function for the parameters $\{\pi_t, b_i^t\}$ in this

case needs to be marginalised over the unobserved variable Φ,

$$\ln L(\{\pi_t, b_i^t\}|D) = \ln \prod_\mu \sum_\phi p(\phi, \{s_i^\mu\}), \tag{6.12}$$

or explicitly,

$$\begin{aligned} \ln L(\{\pi_t, b_i^t\}) &= \sum_\mu \ln \left\{ \sum_\phi \prod_t \left[\pi_t p(x_i^\mu|t)\right]^{\delta(\phi,t)} \right\}. \\ &= \sum_\mu \ln \left\{ \sum_t \left[\pi_t p(x_i^\mu|t)\right] \right\}. \end{aligned} \tag{6.13}$$

To implement the constraint $\sum_t \pi_t = 1$, we introduce a Lagrange multiplier λ and modify the likelihood function accordingly,

$$\tilde{l}(\{\pi_t, b_i^t\}) = \ln L(\{\pi_t, b_i^t\}) - \lambda(\sum_t \pi_t - 1). \tag{6.14}$$

Taking the derivatives of $\tilde{l}$ with respect to π_t and setting them to zero, we find

$$\pi_t = \sum_\mu \frac{\pi_t p(\{x_i^\mu\}|t)}{\sum_t \pi_t p(\{x_i^\mu\}|t)}, \tag{6.15}$$

where we have used the constraint $\sum_t \pi_t = 1$ to find that $\lambda = M$. Defining

$$\gamma_\mu^t = \frac{\pi_t p(\{x_i^\mu\}|t)}{\sum_t \pi_t p(\{x_i^\mu\}|t)}, \tag{6.16}$$

then

$$\pi_t = \sum_\mu \gamma_\mu^t. \tag{6.17}$$

Similarly, taking the derivatives of $\tilde{l}$ with respect to b_i^t and setting them to zero, we find

$$\text{th}\, b_i^t = \frac{\sum_\mu \gamma_\mu^t x_i^\mu}{\sum_\mu \gamma_\mu^t}. \tag{6.18}$$

From (6.16), we see that γ_μ^t is the posterior probability for sample μ to belong to type t, given the observed configurations $\{x_i^\mu\}$. This posterior probability is dependent on the unknown parameters π_t, b_i^t. Therefore, equations (6.15) and (6.18) together with the definition of γ_μ^t in (6.16) form a system of self-consistent equations to be solved for $\{\pi_t, b_i^t\}$. The coupled system of equations can be solved by initiating the values for γ_μ^t to calculate π_t using (6.15) and b_i^t using (6.18); and then the results are used to update γ_μ^t using (6.16).

Such a procedure can be proved to converge to a (local) maximum of the likelihood function. In fact, it belongs to a general categories of algorithms, the expectation–maximisation (EM) algorithms, see Bishop 2006, for example.

Although the EM algorithm allows us to find the maximum of the likelihood function, the solutions are sometimes not good in validation. This is because the maxima of the likelihood function are not sharp, and the likelihood maximisers are rather uncertain. In addition, by the nature of unsupervised learning, the likelihood function is always invariant under a permutation of the K types, its maximum is therefore also $K!$-fold degenerate. The reason for these limitations is because unsupervised models waste labelled samples in training datasets; these all will be overcome by *semi-supervised models*.

Semi-supervised learning

It is easier to introduce the semi-supervised learning algorithm by discussing the problem of correcting the errors in the known types of the training data as mentioned when opening this chapter. Let us think about how a pathologist does the classification for the training dataset: the classifications are normally based on the colours, the volumes or the shapes of the cancer cells. In any case, histological classification is always based on some sort of *observation*, much in the same way that we use the mutation profiles to classify the samples. From that perspective, it is rather obvious that we should somehow treat the classifications of the pathologist as noisy observations. This was not the case for the supervised model, where initial classifications were considered exact (while the unsupervised model did not consider these observations).

The difficulty in the implementation of this idea is that normally the observations made by the pathologists are based on human judgement, rather complicated and only qualitative. Nevertheless, we can model them in a simplified way as follows: we think of the initial classification of the pathologist not as the true state of Φ (the cancer type), but as the state of a Potts spin Φ' which is coupled to Φ by a finite coupling, as illustrated in figure 6.3. Since the coupling is finite, the observed types Φ' can be different from the hidden true types Φ. A sample μ now consists of the observed state of ϕ'^{μ} and the mutation states of N loci, x_i^{μ}. The joint probability distribution for Φ, Φ' and $\{X_i\}$ can be modelled as

$$p(\phi, \phi', \{x_i\}) = \prod_t \left[\pi_t p(\phi'|t) p(\{x_i\}|t)\right]^{\delta(\phi,t)}, \tag{6.19}$$

where we also assumed that given the true type $\Phi = t$, Φ' and $\{X_i\}$ are independent. In comparison to (6.5), the key modification in the joint distribution is $p(\phi'|t)$, which depends on the nature of the initial classifications and can vary from sample to sample. In the case

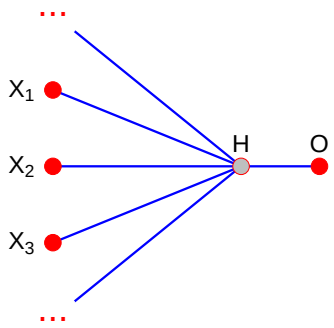


Figure 6.3: **Graphical presentation of the semi-supervised model.** The semi-supervised model considers the initial classifications as an observation of a variable Φ', denoted by O (observed) in the figure, which is coupled by a finite coupling to the true type Φ, denoted by H in the figure (hidden) . Since the coupling is finite, the states of Φ' (O) and Φ (H) can be generally different.

where the initial classification is certain, $p(\phi'^{\mu}|t)$ can be modelled by a Kronecker delta function, $p(\phi'^{\mu}|t) = \delta(\phi'^{\mu}, t)$. In the case where classification is unknown, $p(\phi'^{\mu}|t)$ can be modelled by a flat distribution, $p(\phi'^{\mu}|t) = 1/K$, where K is the number of cancer types. More generally, for the cases where the initial classifications are available, but not always correct, we can model $p(\phi'^{\mu}|t)$ by some finite *confidence* in the initial classifications,

$$p(\phi'^{\mu}|t) = \begin{cases} f & \text{if } t = \phi'^{\mu}, \\ (1-f)/(K-1) & \text{if } t \neq \phi'^{\mu}, \end{cases} \tag{6.20}$$

with the confidence f taking values between 0 and 1. When $f = 1$, we completely trust the initial classifications and when $f = 0$, we disregard the information of the initial classifications. A finite value of confidence f allows the posterior values of Φ given Φ', $\{X_i\}$ to be different from Φ'; errors in the initial classification can be therefore corrected.

By considering the initial classifications ϕ'^{μ} as noisy observations to infer the hidden true types ϕ^{μ}, the inference problem is very similar to the procedure discussed in unsupervised learning equations (6.17) and (6.18). The only modification is the definition of γ_{μ}^{t}, now includes $p(\phi'^{\mu}|t)$ explicitly

$$\gamma_{\mu}^{t} = \frac{\pi_t p(\phi'^{\mu}|t) p(\{x_i^{\mu}\}|t)}{\sum_t \pi_t p(\phi'^{\mu}|t) p(\{x_i^{\mu}\}|t)}. \tag{6.21}$$

The confidence in the initial classification f can vary from sample to sample. This provides a way for the model to learn from both labelled (classified) and unlabelled (un-

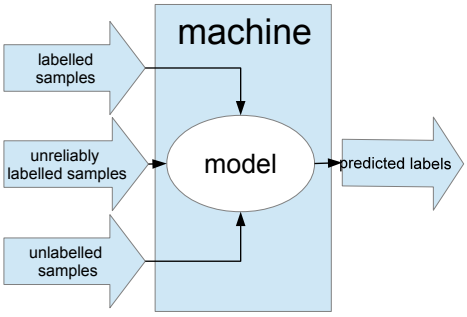


Figure 6.4: **The workflow of semi-supervised model.** Semi-supervised model learns from all reliably labelled, un-reliably labelled and unlabelled samples and reclassify them by incorporating the genetic data.

classified) samples, see figure 6.4, where $f = 0$ for every labelled samples and $f \neq 0$ for every labelled samples, in particular $f = 1$ for reliably labelled samples.

To compare the performance of the semi-supervised model with the supervised and unsupervised model, we first use a simulated dataset. To this end, we simulate a mixture of paramagnetic Ising models, which consists of 2 paramagnetic Hamiltonians of 30 Ising spins, each with a distinct set of fields. The two components of the Hamiltonian can be regarded as two types of cancer, say AD – adenocarcinomas, traditionally coloured orange, and SQ – squamous carcinomas, traditionally coloured blue. The dataset consists of 200 configurations for each Hamiltonian, as shown in figure 6.5a. We then study the models in two aspects: learning from both labelled and unlabelled samples, and correct erroneous labels (misclassifications).

Simulated data: learning from both labelled and unlabelled samples

We delete a fraction α of the labels in the dataset with α ranging from 0.01 to 0.99, as shown in the right panel in figure 6.5b, second panel. We then use the supervised learning algorithm and the semi-supervised learning algorithm to train the mixture of Ising models to predict the deleted labels. Note that for $\alpha \to 1$, the semi-supervised model becomes unsupervised. The initial classifications, when being known (labelled samples), are taken as perfect ($f = 1$). Unlabelled samples can be considered as samples with arbitrary initial labels but without any confidence ($f = 0$). The results are summarised in figure 6.5b. The first panel plots the fraction of labels correctly predicted for both types by the model training with the supervised learning algorithm and the semi-supervised learning algorithm. The figure clearly shows that the fractions of correct predictions of the semi-supervised model are consistently higher than that of supervised model. While the second panel

shows the deleted labels in the dataset, the third panel and the fourth panel detail the predictions of the supervised model and the semi-supervised model. It is easy to see by eyes that the semi-supervised model gives better prediction even when a large number of labels are deleted. The classifications by unsupervised model, which correspond to that of the semi-supervised model at $\alpha \to 1$, are also rather poor.

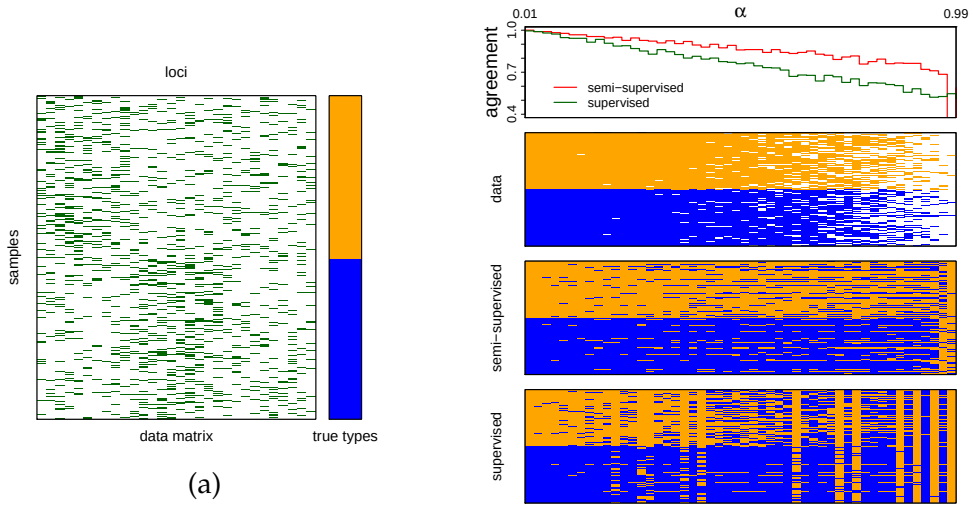


Figure 6.5: **Prediction of missing labels in a simulated dataset.** (a) A simulated data matrix of a mixture of two paramagnetic Ising models. Each component of the mixture is a paramagnetic model of 30 spins with specific pre-determined fields. Each row of the first panel is a configuration of the spins, with green bars for $+1$, and white bars for -1. The second panel indicates the types of the samples, AD (orange) or SQ (blue). (b) Prediction of the missing labels. The top panel summaries the overlaps of the predictions by the supervised model (green) and semi-supervised model (red) with the true types known from the original simulated dataset. The second panel describes the dataset after randomly deleting labels (white bars) with varying fractions α from 0.01 to 0.99. The third panel shows the predictions of the semi-supervised model and the fourth panel shows the predictions of the supervised model. Note that the classifications by the unsupervised model correspond to the semi-supervised model when $\alpha \to 1$.

Simulated data: reclassification of the erroneous initial classifications

Next we examine the ability of correcting misclassifications of the semi-supervised model. From the original dataset, we turn 20% of AD labels into SQ, and 20% of SQ labels into AD. We then run the semi-supervised learning algorithm with varying confidence in the initial classification f. The results are presented in figure 6.6. Starting from $f = 1$, meaning we completely trust the initial classification, the model gives the outputs identical to the initial classifications. And therefore, at $f = 1$, only 80% of the reclassifications for both types are correct. As f decreases from 1 to 0, we lift our confidence on the initial classification and misclassifications gradually get corrected, see the lower panel of figure 6.6. Corre-

spondingly, the fractions of correct classifications of both types increase up to 90% when $f \approx 0.7$, see the upper panel of figure 6.6. These fractions then decrease as the confidence in the initial classifications gets lower; this is because of the decrease in the supervising information at small f; in fact at $f = 0$, the model becomes unsupervised. We know that unsupervised models, which are based on clustering effects, cannot distinguish permutations of labels in the results. Interestingly, we observed accordingly a label-permutation of the two types, signalled by a colour-swapping in the lower panel of figure 6.6 at $f = 0$.

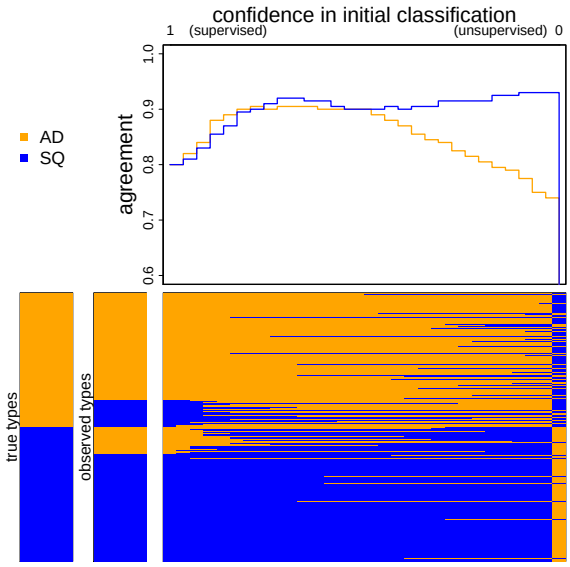


Figure 6.6: **Reclassifications of misclassified samples.** In the simulated dataset, 20% of the labels are swapped (lower, left and middle panels). The semi-supervised model then reclassifies the samples with varying confidence in the initial classification f from 1 to 0. The fractions of correct reclassifications when comparing with the true types known from the samples in the original dataset as functions of f are plotted in the upper panel for the two types. The lower panel details the predicted types for each samples at different values of the confidence f.

6.2.2 Analysing the lung cancer mutation profiles

We now come back to the dataset of mutation profiles of lung cancer patients, where the problem is the potential misclassifications in the supervising data (panels labelled Histo in figure 6.1 and figure 6.7). We run the semi-supervised learning algorithm for the model with varying confidence in the initial classification f. The results are shown in figure 6.7 in the same spirit of figure 6.6. We take the secondary more reliable classifications, namely, the central pathological reviews (CPR), as the validation data. In contrast to the appearance of the data in figure 6.1, we reorder the samples in types labelled by the central pathological reviews (considered as true types) instead of the initial histological classifications by visual signatures (considered as noisy observations). The fractions of corrected types

classified by the model are plotted as a function of the confidence in the initial classification f separately for different types. The lower panel details the predictions by the model, and the data matrix is shown for a reference.

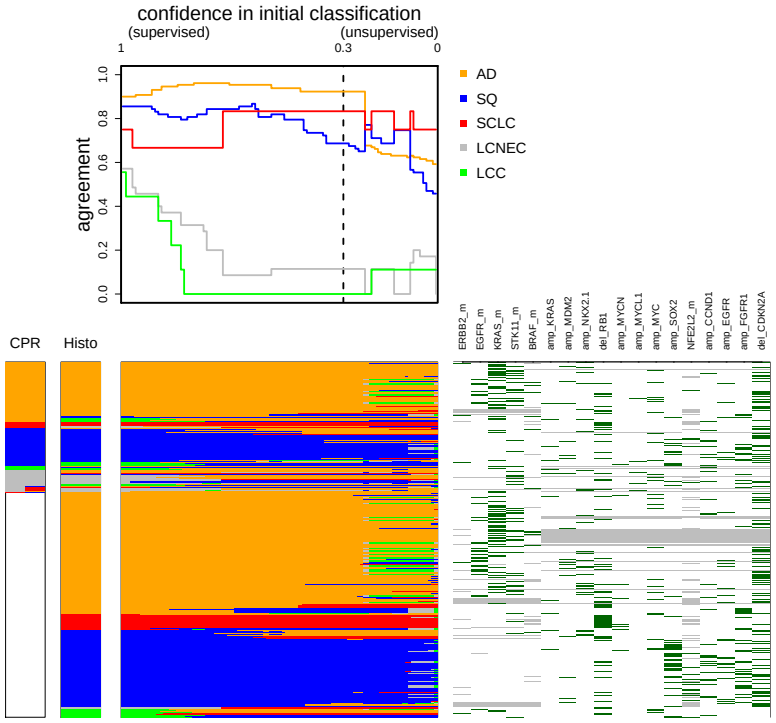


Figure 6.7: **Reclassification of mutation profiles of lung cancer patients**. The central pathological reviews and the initial histological classifications are shown on the left. The middle panel gives the semi-supervised reclassification based on mutations and copy number calls from 18 marker genes (on the right genetic alterations are indicated per sample as green lines). For each sample, the prediction is given as a function of the confidence in the initial histological classification (x-axis: free parameter 1 to 0.3). The right end of this plot describes an unsupervised classification of the samples, disregarding of the initial classification but only taking genetic alterations as the basis for prediction. Prediction tends to agree with the pathological review for AD (orange), SCLC (red), and SQ (blue) (upper panel). On the contrary, LC and LCNEC samples are quickly reclassified into other types on the basis of the genetic data, traceable through the drop of the lines in the top panel.

We observed that adenocarcinomas (AD) and squamous carcinomas (SQ) gradually get corrected. Small cell lung carcinomas (SCLC) are known to be very well characterised; nevertheless, few of them are also corrected. Interestingly, large cell carcinomas (LCC) and large cell neuroendocrine carcinomas (LCNEC) are all reclassified into other types, causing the agreement curves going down very early when f decreases. Conservatively, the result suggests that LCC and LCNEC are not characterised by the mutation at the inspected loci. Comparing the initial classifications with the central pathological reviews, we observe that many LCC and LCNEC samples are also reclassified. This suggests that LCC and LCNEC

are also not well characterised in histological inspection or CPR (The Clinical Lung Cancer Genome Project (CLCGP) and Network Genomic Medicine (NGM) 2013).

Interestingly, at $f < 0.3$, we also observe a colour-swapping in the figure, where the green colour of large cells (LCC) occupies a subset of samples in adenocarcinomas (AD). Inspection of the mutation data shows that the subset of samples is a subtype of adenocarcinomas, which is well characterised by mutated EGFR.

6.3 Conclusions and outlook

In this chapter we discussed a simple semi-supervised learning algorithm for a mixture of paramagnetic Ising model to classify cancer mutation profiles. We run the model on a simulated dataset to study its performance. The semi-supervised model showed a superior performance in comparison to supervised and unsupervised learning algorithms. The advantage of the semi-supervised model is two-fold. On one hand, it maximises the training data by learning from both classified (labelled) and unclassified (unlabelled) samples. On the other hand, the semi-supervised model is capable of correcting the misclassifications in the training data. Both problems are often met in practice. We then applied the model to classify the mutation profiles of lung cancer samples. The model reclassifications agree better with secondary more reliable classifications of the samples, except for large cell carcinomas. The result suggests that large cell carcinomas are not well characterised, at least by the mutations in the proposed set of loci.

References

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Golub, T. R. (1999). "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring". *Science* 286.5439, pp. 531–537.

Lakhani, S. R. and A. Ashworth (2001). "Microarray and histopathological analysis of tumours: the future and the past?" *Nat. Rev. Cancer* 1.2, pp. 151–7.

Nutt, C. L. et al. (2003). "Gene expression-based classification of malignant gliomas correlates better with survival than histological classification". *Cancer Res* 63.7, pp. 1602–1607.

Pusztai, L., C. Mazouni, K. Anderson, Y. Wu, and W. F. Symmans (2006). "Molecular classification of breast cancer: limitations and potential". *Oncologist* 11.8, pp. 868–77.

Sotiriou, C. et al. (2003). "Breast cancer classification and prognosis based on gene expression profiles from a population-based study." *Proc. Natl. Acad. Sci. U. S. A.* 100.18, pp. 10393–8.

Souto, M. de, I. Costa, D. de Araujo, T. Ludermir, and A. Schliep (2008). "Clustering cancer gene expression data: a comparative study". *BMC Bioinformatics* 9.1, p. 497.

The Cancer Genome Atlas Research Network (2012a). "Comprehensive molecular characterization of human colon and rectal cancer". *Nature* 487.7407, pp. 330–7.

The Cancer Genome Atlas Research Network (2012b). "Comprehensive genomic characterization of squamous cell lung cancers". *Nature* 489.7417, pp. 519–25.

The Cancer Genome Atlas Research Network (2012c). "Comprehensive molecular portraits of human breast tumours". *Nature* 490.7418, pp. 61–70.

The Cancer Genome Atlas Research Network (2014). "Comprehensive molecular characterization of gastric adenocarcinoma". *Nature* 513.7517, pp. 202–9.

The Clinical Lung Cancer Genome Project (CLCGP) and Network Genomic Medicine (NGM) (2013). "A genomics-based classification of human lung tumors". *Science Transl. Med.* 5.209, 209ra153.

Viale, G. (2012). "The current state of breast cancer classification". *Ann. Oncol.* 23.suppl 10, pp. x207–10.

Wang, Y., F. Makedon, and J. Pearlman (2006). "Tumor classification based on DNA copy number aberrations determined using SNP arrays". *Oncol. Rep.* 15 Spec no, pp. 1057–9.

Zhu, X. (2005). *Semi-Supervised Learning Literature Survey*. Tech. rep. 1530. Computer Sciences, University of Wisconsin-Madison.

Zhu, X. and A. B. Goldberg (2009). *Introduction to semi-supervised learning*. Ed. by R. Brachman. Morgan & Claypool Publishers.

Appendix A

Drug sensitivity of cancer cells

In this appendix, we are interested in the association of drug sensitivity and genetic features of cancer cells. The traditional method for this problem is to use regression models with appropriate regularisation. This approach is however not the optimal way to exploit the information from the data since the dependence between drugs is ignored. Although a vector regression model takes this dependence into account, it requires that the training dataset contains complete vectors of observations. This is however an important limitation, since training datasets very often contain missing values. Therefore, the problem should be considered as a data-recovery problem, that is, the problem of estimating the values of missing values in a dataset and parameters are estimated consistently. Interestingly, this turns out to be closely related to the semi-supervised model we discussed in chapter 6. From that point of view, it is clear that the purpose of the data-recovery procedure here is not only to predict the missing values in the dataset, rather, is to optimally exploit the information in the dataset. We illustrate that the data-recovery model has lower variations in the estimated parameters in comparison to the regression model. The trade-off is that it contains significantly more parameters than the regression model.

A.1 The drug responses of cancer cells

Targeted therapy is a very important advance in cancer treatment of the last one or two decades. Targeted therapy directly fix the defects in the molecular machine of cancer cells (Huang et al. 2014). Each targeted drug is often designed to interfere with some specific biomolecules in the biochemical networks of the cancer cells (their *targets*). Therefore, targeted therapy is usually more cancer-specific and less toxic than cytotoxic chemotherapy, potentially reducing side-effects for long-term cancer treatments. Developing targeted therapies requires detailed understanding of the molecular biology of cancer cells, which is in most cases patient-specific. Targeted therapy therefore also came with the concept of *personalised medicine*.

A broad range of compounds is used in targeted therapy. Some of them are complex proteins found in nature. Some others are mono-clonal antibodies. A large class of drug molecules are small molecules engineered in laboratories. While the detailed biology of different types of cancers is still to be explored, small molecules are constantly generated in laboratories by small chemical modifications of the other compounds, forming a so-called *drug library*[1]. In many cases, the targets of those molecules are only poorly characterised; many of them have unknown targets and may show unexpected effects.

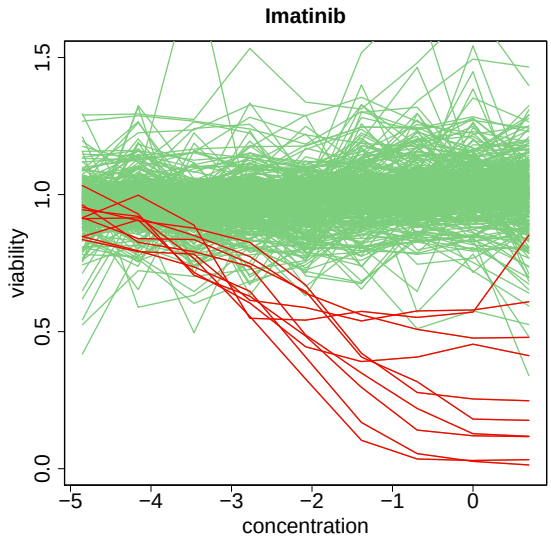


Figure A.1: **The response curves of imatinib (data from Garnett et al. 2012).** Imatinib is screened on 359 cell-lines at 9 concentrations (we use the logarithm of concentration measured in µM), resulting in a family of response curves. The response curves that show a signature of drug efficacy are coloured red, while the others are coloured cyan.

The very first step to benchmark the anti-cancer effects of these drugs is to experimentally screen them on cancer cells in laboratories. The aim is to find out how much a drug inhibits the growth of a population of cancer at different concentrations. Conventionally, the inhibition activity is measured by the ratio of cell population treated with the drug and that of a control population, both taken after 96 hours. We refer to this ratio as relative *cell viability*. The outcome of many screening experiments at different concentrations of the drug is an inhibitory curve for a specific drug and a specific cancer cell-line. An example of inhibitory curves of imatinib on 359 cell-lines is plotted in figure A.1. From such curves, one obtains an idea if a particular cell-line is *sensitive* to a particular drug or not. However, more often than not, distinguishing sensitive cell-lines from insensitive cell-lines on the basis of their response curves is less straightforward than in the case pre-

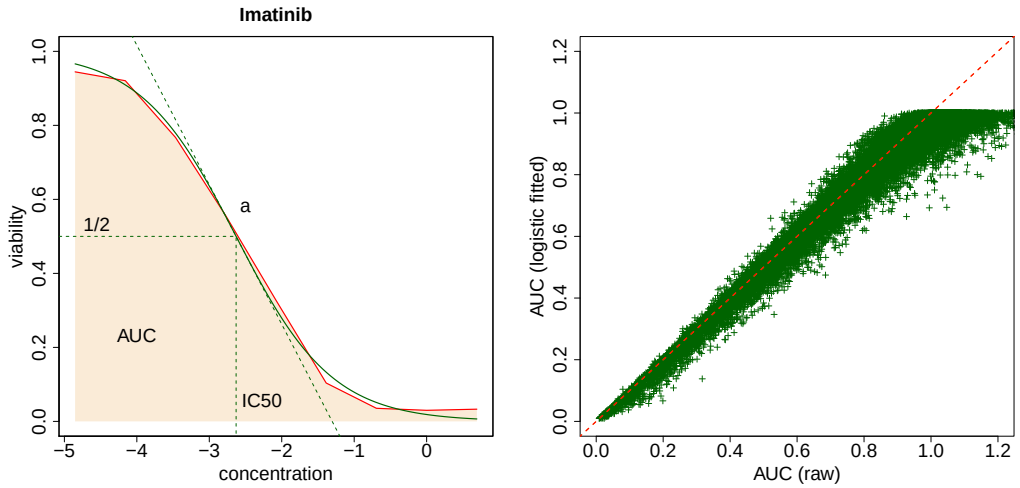
[1] In principle, these small molecules are called *compounds*. They are only called drugs after they have passed certain tests for their efficacy. However, in this chapter, we refer to them as drugs for the sake of simplicity.

sented in figure A.1. Instead of assigning a binary value to a response curve ('sensitive' or 'insensitive'), we may think of quantitatively summarising the response curve by some continuous quantity.

Often, a response curve is fitted with a logistic function,

$$y = \frac{1}{1 + e^{A(x-B)}}, \tag{A.1}$$

where x is the drug concentration (often in logarithmic scale), y is cell viability and α and β are parameters of the logistic function. There emerge three quantities that can be use to characterise the sensitivity of a cell-line to a drug: the half-inhibitory concentration IC50, the tangent of the response curve at IC50 and the area under the response curve, see figure A.2a.



(a) (b)

Figure A.2: **The definitions of IC50, tangent at IC50 and AUC.** (a) A response curve is fitted with a logistic function $y = 1/(1 + e^{A(x-B)})$. The raw data (Garnett et al. 2012) is shown in red, while the fitted curve is shown in green. The IC50 values at which the viability equals 0.5 is indicated on the figure. The tangent of the response curve at IC50 is shown in a dashed green lines. The AUC is coloured light-brown. (b) The AUC calculated from raw data is well agreed with the AUC calculated from the fitted curve when the fitting is converged.

The first is called the *half-inhibitory concentration*, IC50. This is the concentration at which the inhibition activity reaches 0.5, in terms of the logistic fitting function IC50 $= B$. However, it cannot be used to compare different drugs.

Another quantity is the *tangent of the response curve at IC50*. This quantity is interesting in that it is dimensionless if the x-axis is presented on a logarithmic scale. Therefore, in principle, it is possible to compare the tangents at IC50 of different drugs. A steeper tangent at IC50 indicates that the cell-line may be more sensitive to the drug, and vice

versa.

The third quantity is the *area under the curve* (AUC), or the *acting area*. As the name indicates, it is the area under the response curve, spanning the minimum concentration to the maximum concentration (in log scale). The AUC is often normalised to the area of the curve $y = 1$ (no response) spanning on the same range. A normalised AUC of 1 means no response, and a small AUC means strong response of the cell-line to the drug. As with IC50, it is not possible to compare AUC of different drugs. However AUC has an important advantage: it can be calculated from the raw data instead of from the fitted curve, therefore it does not suffer from ambiguities in the fitting procedure. On the other hand, the IC50 and the tangent of a response curve at IC50 can only be determined from fitted curves, and are sensitive to the choice of fitting function. In practice, the AUC of the raw curves are in agreement with the AUC of the fitted curves when the fitting convergences, as shown in figure A.2b. In the following, we will choose to use AUC to summarise the information of a drug-cell-lines response curve (where the we use the logarithm of concentration measured in µM). In fact, we often use the logarithm of AUC in our analysis, but also referred to as AUC for simplicity.

Assuming it can be determined that a particular cell-line is sensitive to a particular drug from the response curves (or at least an estimate of an AUC), the next step is to determine which genetic features make a particular cell-line sensitive to a particular drug. This requires the study of the genetics of the cell-lines. Sequencing or gene expression profiling the cell-lines are then necessary. By comparing the genetic contents of the drug-sensitive cell-lines and the drug-insensitive cell-lines, we can in principle statistically pinpoint which genetic features characterise their difference.

For this reason, the development of targeted therapy from small molecules demands not only for the drug–cell-line screening data, but also the genetic data of the cell-lines. Recently, two large databases of drug-sensitivity together with genetic data have been compiled, which call for detailed analysis (Barretina et al. 2012, Garnett et al. 2012). In the analysis presented in this appendix, we will use the dataset from Welcome Trust Sanger Institute as an example (Garnett et al. 2012, Yang et al. 2013). This dataset consists of 707 genetically annotated cell-lines of 16 cancer types classified according to the tissues where they emerged, see figure A.3. Information on mutations at 64 loci known to be involved in cancer is also given for the cell-lines. This includes both sequence mutations and copy-numbers alterations. After filtering those loci with too low frequencies of mutation, we retain 61 loci, see the middle panel of figure A.3. We also restrict drugs to those that have more than 60% of AUC measured, see the right panel of figure A.3, which shows the data before the restriction. The AUC of all drug–cell-lines in the database of Sanger Institute

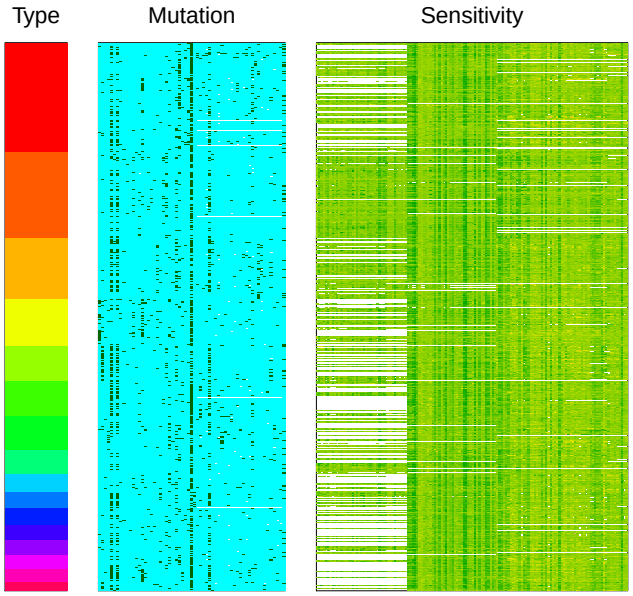


Figure A.3: **Data summary.** 16 types of the cancer cell-lines are shown in the 16 colours in the left panel. The genetic information of the cell-lines at 64 loci is summarised in the middle panel (green bars: mutated, cyan bars: white type, white bars: no information). In the right panel, the AUCs of the cell-lines with respect to 139 drugs are presented in terrain scale (green-like colours: small; white-like colours: large).

are extracted, which result in a data matrix of AUC with cell-lines on the row and drugs on the columns, as shown in the right panel of figure A.3.

We are interested in the dependence between the drug sensitivity data and the genetic data. A standard way to study this dependency is to use linear regression with drug responses as response variables and genetic information as signals (see chapter 2). However, we will show that this is not the optimal way of exploiting the available information. This is because the dependency between the drugs is not taken into account in linear regression. Considering the dependency between the drugs, the problem turns from a regression problem to a data-recovery problem. Interestingly, this turns out to be also formally related to the formalism of semi-supervised learning methods discussed in chapter 6.

A.2 Turning the regression problem to a data-recovery problem

A.2.1 Drug response as a regression problem

The association of genetic information with response to a drug can be viewed as a simple regression problem, where the response variable is the AUC values of the drugs and the signal is the mutation profile of the cell-line (Garnett et al. 2012). In practice, it turns out

that besides the mutation profiles, the cell types also play an important role in predicting the response of a cell-line to drugs. We begin with the regression model as discussed in chapter 2, Y as a function of $(\Phi, \{X_i\})$, where Y stands for the response of a sample to some particular drug, Φ is the tissue type of the sample and $\{X_i\}$ is its mutation profile. Note that Φ is categorical (see chapter 2). The Hamiltonian can be written as

$$H(y;\phi,\{x_i\}) = H(\phi,\{x_i\}) - \sum_t a_\phi \delta(\phi,t) y - \sum_i b_i x_i y + \frac{c}{2} y^2, \tag{A.2}$$

or in terms of conditional probability,

$$p(y|\phi,\{x_i\}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{\left(y - \sum_i \beta_i x_i - \alpha_\phi\right)^2}{\sigma^2} \right\}, \tag{A.3}$$

where a_ϕ, b_i, α_i, β_i, c and σ are model parameters. The coefficients β_i or b_i characterise the effect of the genetic mutation on the response of the cell-line to the drug. Note that this corresponds to the maximum entropy model with sufficient statistics $\langle Y \rangle$ and $\langle Y X_i \rangle$ and $\langle Y \delta(\Phi, t) \rangle$ and an implicit model for $(\Phi, \{X_i\})$. In analogy to the Ising model, we also refer to β_i as couplings between the genetic mutations and the drug response, and refer to α_t as couplings between the cancer types and the drug response.

Vector regression

The regression model described above ignores the dependence between the drugs. In practice, as many drugs are different only in small chemical modifications, this dependency is important (Garnett et al. 2012). How do we model such dependency? We can think of predicting a vector of drug responses instead of single-drug regression. From the maximum entropy reasoning perspective, the solution is simply adding the observables $\langle Y_k Y_l \rangle$ into the sufficient statistics of the regression model. The model becomes a vector regression problem, of which the conditional probability of observing the response vector $\{y_k\}$ given the genetic data $\{x_i\}$ is given by

$$p(\{y_k\}|\phi,\{x_i\}) \propto \exp\left\{ -\frac{1}{2} \sum_{kl} A_{kl} (y_k - \sum_i \beta_i^k x_i - \alpha_k^\phi)(y_l - \sum_j \beta_j^l x_j - \alpha_l^\phi) \right\}. \tag{A.4}$$

In this model, the response of one drug can tell us about the response of the other drug via the interacting coefficient A_{kl}. As a simple application, this vector regression problem can be used to reconstruct missing values of AUC of the drug–cell-line response curve. Suppose we know the mutation profile $(\phi, \{x_i\})$ and the response y_a to a set of drugs a of a particular cell-line, and we want to find the (mean) responses to the other set of drugs

$\bar{y}_{b|a}$. From (A.4), we find that this is a conditional Gaussian inference problem (Bishop 2006), which leads to

$$\bar{y}_{b|a} = \bar{y}_b + B_{ba}(B_{aa})^{-1} y_a, \tag{A.5}$$

where $B = A^{-1}$ and $\bar{y}_b$ is the mean AUC predicted by the genetic information alone,

$$\bar{y}_b = \beta^b x + \alpha_b^\phi. \tag{A.6}$$

Note that α_b^ϕ, y_a, $\bar{y}_b$ and $\bar{y}_{b|a}$ are partitioned vectors and β^b, B_{ba} and B_{aa} are partitioned matrices with respect to the sets of drugs a and b.

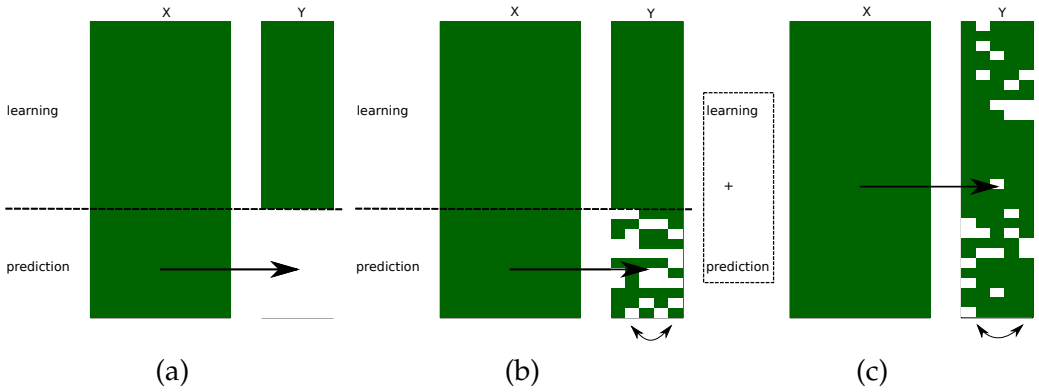


Figure A.4: **From regression problem to data-recovery problem.** (a) Single drug regression does not use the information of the dependency between different drugs. (b) Vector regression does use the dependence between different drugs, but needs to learn from a complete training dataset to predict an incomplete dataset. (c) Turning the regression to a data-recovery problem, we do not distinguish between learning and predicting. The model consistently learns the incomplete data and predicts the missing values.

From regression to data-recovery problem and semi-supervised problem

Note that to train a vector regression model, one needs a large complete dataset, which is not always available. A better model is to consistently learn from an incomplete dataset and predict the missing values. This is done in the framework of *data-recovery* instead of regression. The missing data in the data-recovery problem can be considered as unobserved data, which can be estimated by maximum likelihood estimation via EM-algorithm (see chapter 6). Instead of writing the procedure formally, we give here a practical description.

1. *Initiation:* for each drug, we run the single drug regression algorithm (A.3) based on observed data and fill in the missing data with the corresponding predicted values by the regression model.

2. *Maximisation:* use the newly (fully) filled data to learn the parameters for the vector regression model (A.4).

3. *Expectation:* use the vector regression (A.5) to (re)infer new values for the missing data. If the differences between the new values and the old ones are smaller than some predefined small value ϵ, then exit and return the recovered data matrix, otherwise, return to step 2.

Note that in such a model, there is no distinguishing between training data and data for prediction anymore, see figure A.4. Let us recall from chapter 6, where we used the semi-supervised model to predict missing labels (classifications) in a simulated dataset of lung cancer classification. This is nothing but a data-recovery problem. This unification of the data-recovery problem and the semi-supervised model brings an interesting perspective, normally not considered in data-recovery problems. We suggest: the aim of the data-recovery procedure is *not only* to infer the missing values in the dataset, rather, it should be considered as a way of improving the analysis by maximising the use of the information in the training dataset. In the same way, the semi-supervised classification model discussed in chapter 6 is not only to predict missing labels but also to exploit the information in the training dataset efficiently. We will show that the parameters in the data-recovery model are estimated *more reliably* than that in the regression model.

A.2.2 Analysis of the drug-response library

In the following, we will apply this data-recovery procedure to the drug sensitivity data from Welcome Trust Sanger Institute (Garnett et al. 2012, Yang et al. 2013) and compare the results with the single-drug regression model.

To verify the predictive power of the different procedures, we randomly delete 30% of the measured AUC values (in figure A.3, panel c), saving them for cross validation. The resulting data is then used to run both the data-recovery algorithm and single regression. The cross validation results are shown in figure A.5. The data recovery algorithm gives a higher predictive power than the regression, sometimes much higher as in the case of gemcitabine. This is to be expected: the data-recovery procedure exploits the correlation between drugs for the predictions while the single-drug regression is entirely based on the genetic data.

However, we want to emphasise again that it may not be the sensitivity prediction the most important point, rather the better estimate for the parameters may be more important. This can be observed in the plot of the standard deviations of the model parameters inferred by data-recovery procedure and by single-drug regression in figure A.6. The standard deviations are estimated by bootstrapping 300 times on the data with 10% of the measured AUC deleted. The figure clearly shows that while the means of the coefficients agree well between the two methods, the standard deviations of the coefficients inferred by

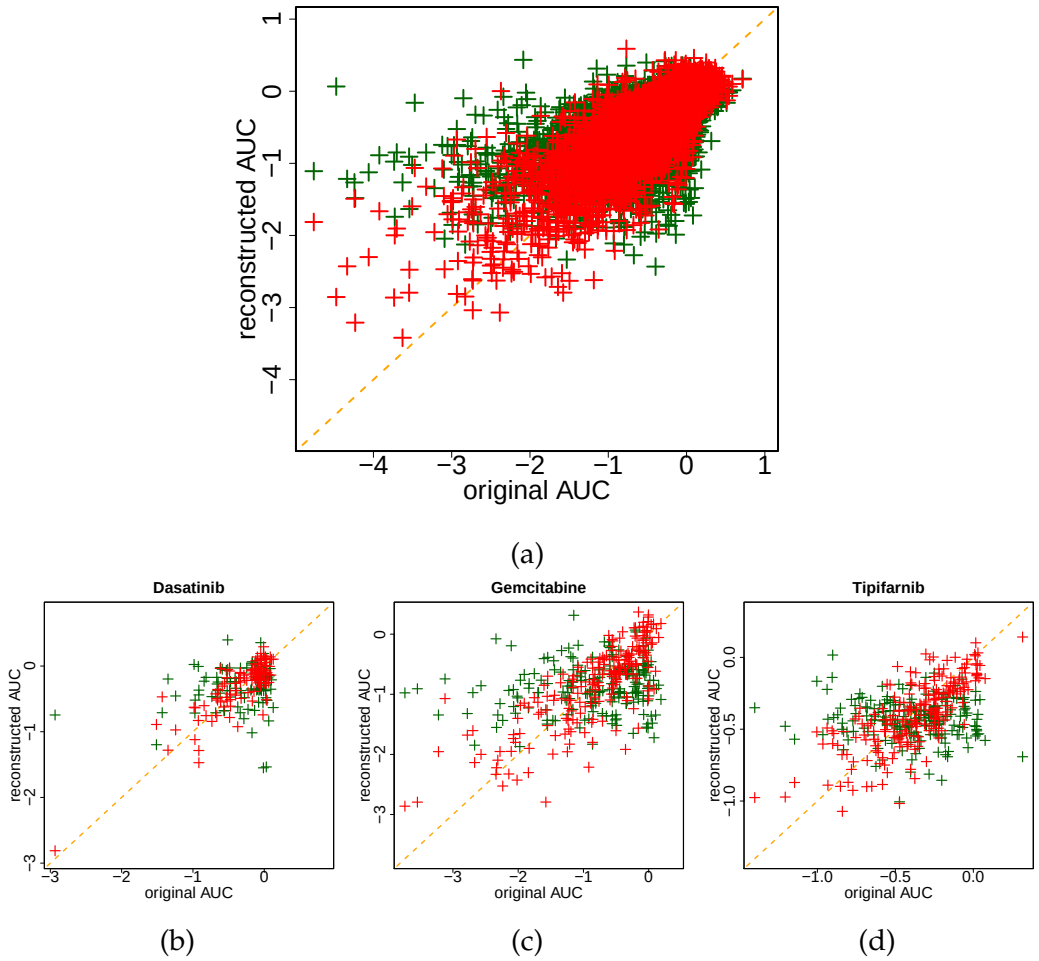


Figure A.5: **Scatter plot between values of deleted AUC against reconstructed AUC.** (a) For all drugs, (b) dasatinib, (c) gemcitabine, and (d) tipifarnib. The predictions by single-drug-regression are coloured green, while the predictions by the data-recovery procedure are coloured red. One sees that the data-recovery model gives better predictions than the regression model does.

the data-recovery procedure is consistently lower than that of the single-drug regression model.

There is however a drawback of the data-recovery model: its has significantly more parameters than the regression model. This causes the failure of the data-recovery model when there are too many missing values in the dataset indicating by the divergence of the model parameters. Note that, however, this is the problem of this particular model (Gaussian model); in general, the number of parameters of data-recovery model is not necessarily much larger than that of the corresponding regression problem.

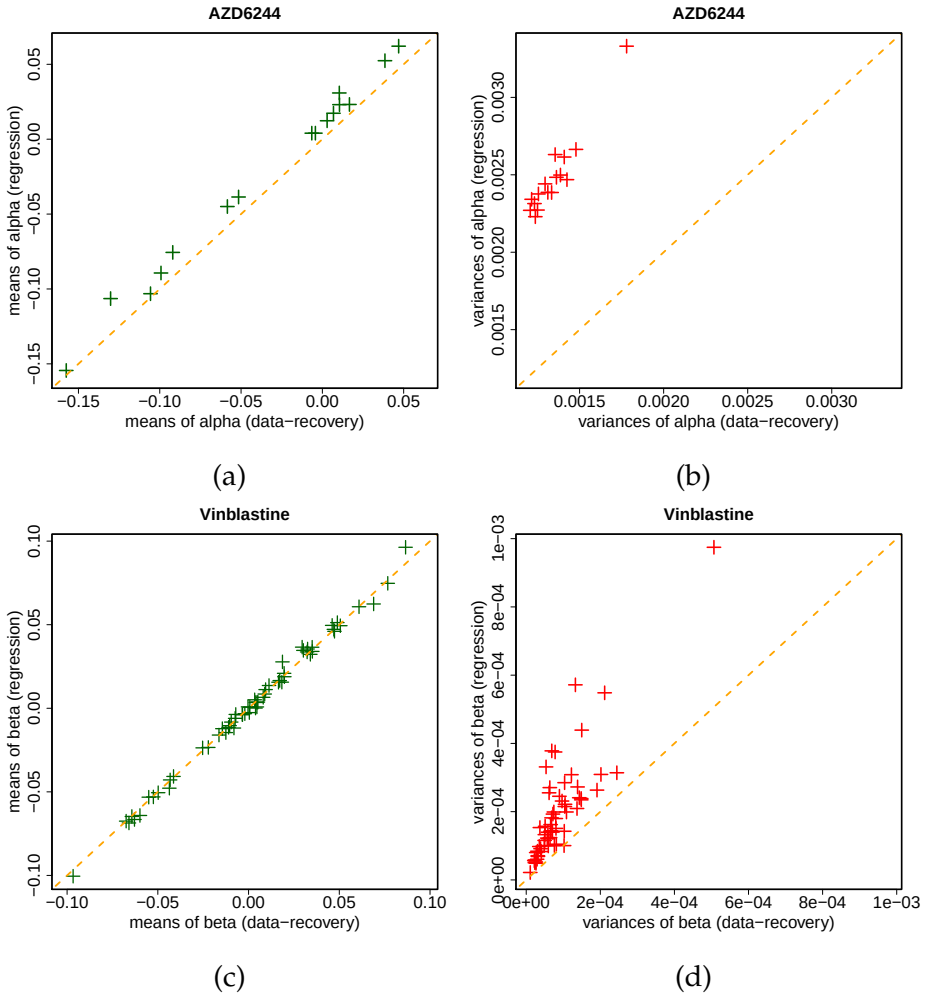


Figure A.6: **Scatter plot to compare the model parameters estimated by data-recovery procedure and single-drug regression.** (a) Comparison of the means of couplings α between the types of cancer and the AUC (see equations (A.3) and (A.4)). (b) Comparison of the variances of α. (a) Comparison of the means of the couplings β between the mutation states and the AUC (see equations (A.3) and (A.4)). (b) Comparison of the variances of β.

A.3 Conclusions and outlook

In this chapter, we examined the drug sensitivity data library from Welcome Trust Sanger Institute using the method of linear regression and data-recovery. We suggested that the problem should be viewed as a data-recovery problem, instead of a regression problem. We brought a unified view on data-recovery and the semi-supervised learning algorithm developed in the previous chapter. Our discussions are a proof of principle. Downstream analysis and biological validation of the outputs of the data-recovery procedure are still to be developed. Incorporating the data-recovery procedure with different regularisations, in particular subset-selective regularisations (Hastie et al. 2009), can also be considered.

References

Barretina, J. et al. (2012). "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity". *Nature* 483.7391, pp. 603–7.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Garnett, M. J. et al. (2012). "Systematic identification of genomic markers of drug sensitivity in cancer cells". *Nature* 483.7391, pp. 570–5.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning*. Springer.

Huang, M., A. Shen, J. Ding, and M. Geng (2014). "Molecularly targeted cancer therapy: some lessons from the past decade". *Trends Pharmacol. Sci.* 35.1, pp. 41–50.

Yang, W. et al. (2013). "Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells". *Nucleic Acids Res.* 41, pp. D955–61.

Appendix B

The emergence of drug resistance in expanding tumours

We are to study the evolution of drug resistance in cancer. While therapy resistant clones often exist as minority in a stable tumour, we observe that under specific circumstances, resistant clones can be found more frequently in expanding tumours. We illustrate this in a logistic model of tumour growth and numerically determine the range of parameters that allow for such a phenomenon. The mathematical basis to analyse the model is the theory of non-homogeneous birth-death processes with arrivals (or migrations), which will also be discussed in detail.

B.1 Evolution of drug resistance in cancer

Despite great advances during the last decades, cancer treatments are hindered by therapy resistance. Indeed, most of cancers respond either partly or completely at least to some particular therapy, but almost invariably relapse within a few months or a few years with resistance to the therapy (Gottesman 2002, Holohan et al. 2013). Further treatments may be applied, but again fail at some point. This short appendix discusses several aspect of the emergence of drug-resistance.

Within the evolution model of cancer, it is believed that cancer resistance has its main root in the fast evolving nature of the population of cancer cells. Under the selective pressure exerted by therapy, cells with genetic (or epigenetic) features that allow them to live in the therapeutical environment have tremendous reproductive advantage and spread in the population. Do these resistant cells pre-exist in the population or do they emerge (de novo) during therapy? This is of course not a well-defined question, unless we give it a probabilistic nature: are these resistant cells more likely to pre-exist in the cell population or more likely to emerge during therapy?[1]

In many cases, it is observed the majority of the cells in a cancer cell population carry

[1] This is different from the question: does the mutation rate rise under therapy (hypermutation), which is well-defined biologically and of non-probabilistic nature.

the genetic contents that can be directly targeted by (some) drug. If the resistant cells pre-exist, they only constitute a minority in the population. This reflects that in therapy-free environments, drug-sensitive cells often grow faster than resistant cells. In some case, resistance arises because of the development of pump proteins on the surfaces of cancer cells, which pump the drug out of the cells. As this pumping activity requires energy, the cells have less resource for division (Gottesman 2002).

Imagine a picture of a tumour freely expanding from a single cell, which is sensitive to a particular therapy. This gives rises to an exponentially expanding clone of sensitive cells. Errors in DNA replications create genetic variation in the population, and resistant cells may occur during the divisions of sensitive cells. Over time, the resistant cells accumulate. They also divide forming sub-clones of resistant cells in the tumour. But if the resistant cells divide sufficiently slower than the sensitive cells, they remain at small fractions in the population. This model of accumulation of resistant cells in a freely expanding population was discussed by Luria and Delbrück and then by Iwasa and Michor, see Iwasa et al. 2006. According to this model, the resistant cells are almost sure to pre-exist in tumours at a detectable size (Iwasa et al. 2006).

One may be unsatisfied with this answer. Indeed, the picture of an exponentially expanding population of cancer cells is unrealistic: it essentially assumes that cancer cells grow and divide limited only by their intrinsic ability to grow and divide (division rate). In reality, cancer cell growth and division are heavily dependent on the resources they compete for, which are usually limited in the internal environment of the body (Greaves and Maley 2012). We also learned that tumour cells may also compete for space, direct inhibit each other by mechanical stress, see chapter 1.

The cell-cell competition changes the picture of tumour expansion and resistant cell accumulation radically. When the tumour expands to some large size, cell-cell competition becomes apparent; as a result, cell divisions decrease and cell deaths may also increase. The whole population suffers from this slowing down in growth, but resistant cells may suffer more because of their disadvantage in competition. This means, at some point in the expanding phase of the tumour, it may happen that resistant cells do not accumulate in the tumour but actually are lost from the tumour due to the competition with sensitive cells. The consequence is that there may be an intermediate maximum of the number of resistant cells in the population. We will call this phenomena *boosting of resistance* during tumour expansion.

In the following, we illustrate quantitatively the possibility of such boosting in a simple model of tumour expansion when the competition between cells is considered. We start with a discussion of the birth-death processes with arrivals, which will serve as the basis

for our model.

B.2 The mergence of a mutated clone in an expanding population

B.2.1 Birth-death process with arrivals

Definition (birth-death processes)**.** A (linear) birth-death stochastic process $X(t)$ with arrivals taking values in $\{0, 1, \ldots\}$ is defined by the transition rates

$$\begin{aligned} q_{k\to k+1} &= b(t)k + r(t), && \text{(B.1)} \\ q_{k\to k-1} &= d(t)k, && \text{(B.2)} \\ q_{k\to k} &= 1 - q_{k\to k+1} - q_{k\to k-1}, && \text{(B.3)} \\ q_{k\to p} &= 0 \text{ for } p \neq k, k-1, k+1, && \text{(B.4)} \end{aligned}$$

where $b(t)$, $d(t)$ and $r(t)$ are continuous functions of t almost everywhere, which are called the *birth rate*, the *death rate* and the *arrival rate* respectively. If $r(t) = 0$, the process is called *linear birth-death process without arrivals* or *simple birth-death process*. If $b(t)$, $d(t)$ and $r(t)$ are constant over time, the process is called *homogeneous*.

In practice, we usually think of $X(t)$ as the number of cells or number of particles at time t. Note that, for the process to be fully defined, we need to specify the initial condition, for example, $X(0) = k$.

The master equation for the distribution p_k for such a birth-death process with arrivals, defined as $p_k = \mathrm{P}\{X(t) = k\}$, can be written down explicitly,

$$\frac{\mathrm{d}\, p_k}{\mathrm{d}\, t} = [(k-1)b(t) + r(t)]p_{k-1} + (k+1)d(t)p_{k+1} - [r(t) + kb(t) + kd(t)]p_k. \tag{B.5}$$

Multiplying the two sides of (B.5) with z^k, and performing the summation over k, we obtain the equation for probability generating function,

$$\frac{\partial G(z,t)}{\partial t} = [b(t)z - d(t)](z-1)\frac{\partial G(z,t)}{\partial z} + r(t)(z-1)G(z,t), \tag{B.6}$$

where

$$G(z,t) = \sum_{k=0}^{+\infty} p_k(t) z^k. \tag{B.7}$$

From the equation for the probability generating function, the equation for the mean $m(t) = \langle X(t)\rangle$ and the variance $s(t) = \langle X(t)^2\rangle - \langle X(t)\rangle^2$ of X can be found easily. Indeed, noting

that

$$m(t) = \left.\frac{\partial G(z,t)}{\partial z}\right|_{z=1}, \tag{B.8}$$

by taking the derivative of (B.6) with respect to z, we obtain

$$\frac{\mathrm{d}\, m}{\mathrm{d}\, t} = [b(t) - d(t)]m(t) + r(t), \tag{B.9}$$

which can be solved to yield

$$m(t) = e^{w(t)}\left[m_0 + \int_0^t \mathrm{d}\, t' e^{-w(t')} r(t')\right], \tag{B.10}$$

with $w(t) = \int_0^t \mathrm{d}\, t'[b(t') - d(t')]$ and m_0 is the initial value of m at $t = 0$.

Suppose the process is initiated at t_1 with $X(t_1) = 1$ and we are interested in the probability of the process to be at extinction,

$$P_0(t_1, t_2) = \mathrm{P}\{X(t_2) = 0\}. \tag{B.11}$$

If the arrival rate is non-zero, this extinction is temporal since at some point a new arrival will set a new clone. If the process is a birth-death process without arrivals, the extinction is permanent and we use another notation

$$P_{\mathrm{ext}}(t_1, t_2) = \mathrm{P}\{X(t_2) = 0\}. \tag{B.12}$$

Note that $\mathrm{P}\{X(t) = 0\} = G(0, t)$; in principle, equation (B.6) can be solved, and $P_0(t_1, t_2)$ can be found (see Stirzaker 2005 and Foo and Michor 2010, for example). We however use another route, which elucidates some important properties of birth-death processes.

Lemma. *The probability of (permanent) extinction by time t_2 of a non-homogeneous birth-death process without arrivals $X(t)$ with birth rate $b(t)$ and death rate $d(t)$, which is initiated at $t = t_1$ as $X(t_1) = 1$, is given by*

$$P_{\mathrm{ext}}(t_1, t_2) = 1 - \frac{1}{e^{-w(t_1,t_2)} + \int_{t_1}^{t_2} \mathrm{d}\, t e^{-w(t_1,t)} b(t)}, \tag{B.13}$$

where $w(t_1, t_2) = \int_{t_1}^{t_2} \mathrm{d}\, t[b(t) - d(t)]$.

Proof. The lemma can be proved by the standard *first step analysis* of the process (Stirzaker 2005). Think of $P_{\mathrm{ext}}(t_1, t_2)$ as a function of t_1 when t_2 is fixed. At time $t = t_1 + h$, to the first order of h, the process can be at $X(t_1 + h) = 0$, $X(t_1 + h) = 1$ or $X(t_1 + h) = 2$. We

have the decomposition of $P_{\text{ext}}(t_1, t_2)$ with respect to these three events,

$$\begin{aligned} P_{\text{ext}}(t_1, t_2) &= P_{\text{ext}|X(t_1+h)=0}(t_1+h, t_2)\,\mathrm{P}\{X(t_1+h)=0\} + \\ &\quad + P_{\text{ext}|X(t_1+h)=1}(t_1+h, t_2)\,\mathrm{P}\{X(t_1+h)=1\} + \\ &\quad + P_{\text{ext}|X(t_1+h)=2}(t_1+h, t_2)\,\mathrm{P}\{X(t_1+h)=2\} + O(h^2), \end{aligned} \tag{B.14}$$

where, by the definition of the process

$$\mathrm{P}\{X(t_1+h)=0\} = d(t_1)h + O(h^2), \tag{B.15}$$

$$\mathrm{P}\{X(t_1+h)=2\} = b(t_1)h + O(h^2), \tag{B.16}$$

$$\mathrm{P}\{X(t_1+h)=1\} = 1 - [b(t_1)+d(t_1)]h + O(h^2), \tag{B.17}$$

and

$$P_{\text{ext}|X(t_1+h)=0} = 1, \tag{B.18}$$

$$P_{\text{ext}|X(t_1+h)=1} = P_{\text{ext}}(t_1+h, t_2), \tag{B.19}$$

$$P_{\text{ext}|X(t_1+h)=2} = P^2_{\text{ext}}(t_1+h, t_2), \tag{B.20}$$

where, in the last expression we have used the fact that conditioned on $X(t_1+h)=2$, the two cells at t_1+h initiate two *independent* birth-death sub-processes with the same birth rate $b(t)$ and death rate $d(t)$, for the process $X(t)$ to go extinct by t_2, both of these sub-processes should go extinct. Now, expanding the decomposition (B.14) to the first order in h then taking the limit $h \to 0$, we have a differential equation for P_{ext} as a function of t_1,

$$\partial_{t_1} P_{\text{ext}} = -P^2_{\text{ext}} b(t_1) + P_{\text{ext}}[b(t_1)+d(t_1)] - d(t_1). \tag{B.21}$$

Solving this differential equation[2] with the ending condition $P_{\text{ext}}(t_2, t_2) = 0$, we have the solution (B.13). □

Theorem. *Consider a non-homogeneous birth-death process with arrivals $X(t)$ with birth rate $b(t)$, death rate $d(t)$ and arrival rate $r(t)$, which is initiated at $t = t_1$ as $X(t_1) = 0$. The probability for the process is at temporal extinction at t_2 is then give by*

$$P_0(t_1, t_2) = \exp\left\{ \int_{t_1}^{t_2} \mathrm{d}t \frac{e^{-w(t_1,t)} r(t)}{K(t_1,t) - K(t_1,t_2) - e^{-w(t_1,t_2)}} \right\}, \tag{B.22}$$

where $w(t_1, t_2) = \int_{t_1}^{t_2} \mathrm{d}\, t[b(t) - d(t)]$ and $K(t_1, t_2) = \int_{t_1}^{t_2} \mathrm{d}t e^{-w(t_1,t)} b(t)$.

[2]The equation can be solved by Kendall's transformation $P_{\text{ext}}(t_1, t_2) = 1/u(t_1, t_2) + 1$

Proof. The idea is that in a birth-death process with arrivals, each arrival (which comes at rate $r(t)$) will initiate a simple birth-death process (without arrivals) *independently* from each other. This new process starts a new clone. The probability that the process has no cell at time t_2, $\mathrm{P}\{X(t_2) = 0\}$, is the probability that all such clones initiated at any time $t < t_2$ all have gone extinct by t_2. We implement this idea in the first step analysis. Let us consider the decomposition of $P_0(t_1, t_2)$ into the conditions $\{X(t_1 + h) = 0\}$ (no clone initiated) and $\{X(t_1 + h) = 1\}$ (a clone initiated) to the first order in h, where h is a small time interval,

$$\begin{aligned} P_0(t_1, t_2) &= P_{0|X(t_1+h)=0}(t_1, t_2)\,\mathrm{P}\{X(t_1 + h) = 0)\} + \\ &\quad + P_{0|X(t_1+h)=1}(t_1, t_2)\,\mathrm{P}\{X(t_1 + h) = 1)\} + O(h^2). \end{aligned} \tag{B.23}$$

Note that at $t = t_1$, $X(t_1) = 0$, we have

$$\mathrm{P}\{X(t_1 + h) = 0)\} = 1 - r(t_1)h + O(h^2), \tag{B.24}$$

$$\mathrm{P}\{X(t_1 + h) = 1)\} = r(t_1)h + O(h^2), \tag{B.25}$$

and

$$P_{0|X(t_1+h)=0} = P_0(t_1 + h, t_2), \tag{B.26}$$

$$P_{0|X(t_1+h)=1} = P_0(t_1 + h, t_2)P_{\mathrm{ext}}(t_1 + h, t_2), \tag{B.27}$$

where in the last equation, conditioned on $X(t_1 + h) = 1$, there is a clone initiated by this single cell, and $P_{\mathrm{ext}}(t_1 + h, t_2)$ is the probability for this clone to die out by t_2, given by (B.13). Expanding the decomposition (B.23) to the first order in h and taking the limit $h \to 0$, we have a differential equation for $P_0(t_1, t_2)$ as a function of t_1,

$$\partial_{t_1} P_0(t_1, t_2) = r(t_1)[1 - P_{\mathrm{ext}}(t_1, t_2)]P_0(t_1, t_2), \tag{B.28}$$

which can be solved with the boundary condition $P_0(t_2, t_2) = 0$ to gives

$$P_0(t_1, t_2) = \exp\left\{-\int_{t_1}^{t_2} \mathrm{d}\, t_1' r(t_1')[1 - P_{\mathrm{ext}}(t_1', t_2)]\right\}. \tag{B.29}$$

With $P_{\mathrm{ext}}(t_1, t_2)$ calculated by (B.13), we obtain equation (B.22). □

This proof is complementary to the proof by Foo and Michor 2010, where an integral argument was used instead of the first-step differential equation. The following corollary follows:

Corollary. *Consider a non-homogeneous birth-death process with arrivals $X(t)$ with birth rate $b(t)$, death rate $d(t)$ and arrival rate $r(t)$, which is initiated at $t = t_1$ as $X(t_1) = k_0$. The probability for the process to be at temporal extinction at t_2 is give by*

$$P_0(t_1, t_2; k_0) = [P_{\text{ext}}(t_1, t_2)]^{k_0} P_0(t_1, t_2), \tag{B.30}$$

where $P_0(t_1, t_2)$ is the probability for the process starting with no cell to be at temporal extinction at t_2 given by (B.22), and P_{ext} is the probability of extinction of a birth-death process with birth rate $b(t)$ and death-rate $d(t)$ but without arrivals and starting with a single cell, given by (B.13).

We now analyse the birth death processes with arrivals at equilibrium. Let $b_\infty = \lim_{t\to+\infty} b(t)$, $d_\infty = \lim_{t\to+\infty} d(t)$, $r_\infty = \lim_{t\to+\infty} r(t)$ and assuming $r_\infty > 0$. It is clear that if $b_\infty \geq d_\infty$, the mean number of cell keeps increasing and the process escapes to infinity and there is no stationary distribution. We are concerned with the case

$$b_\infty < d_\infty, \tag{B.31}$$

where the process is trapped above extinction. In this case, there exists a stationary distribution of the number of cells. The equation for the probability generating function at equilibrium, $G(z) = \lim_{t\to+\infty} G(z, t)$, can be found by letting $t \to +\infty$ in (B.6), which gives

$$(b_\infty z - d_\infty)G'(z) + r_\infty G(z) = 0. \tag{B.32}$$

Solving this differential equation with the condition $G(1) = 1$, we have

$$G(z) = \left(\frac{d_\infty/b_\infty - z}{d_\infty/b_\infty - 1}\right)^{-r_\infty/b_\infty}. \tag{B.33}$$

It follows that the mean number of cells at equilibrium $m_\infty = \lim_{t\to+\infty} m(t)$ can be found as

$$\begin{aligned} m_\infty &= G'(1) \\ &= \frac{r_\infty}{d_\infty - b_\infty}, \end{aligned} \tag{B.34}$$

and the probability for the process to be at temporal extinction at equilibrium, $P_\infty = \lim_{t_2\to+\infty} P_0(t_1, t_2)$ (which does not depend on t_1), can be found as

$$\begin{aligned} P_\infty &= G(0) \\ &= \left(\frac{d_\infty/b_\infty}{d_\infty/b_\infty - 1}\right)^{-r_\infty/b_\infty}. \end{aligned} \tag{B.35}$$

Monte Carlo simulation

The non-homogeneous birth-death processes can be simulated via the *rejection* (or *projection*) method (Lewis and Shedler 1978). This is a general method to simulate any Markov process with time-dependent transition matrix. Consider such a time-dependent Markov process in a space $\mathscr{F}$, defined by the time dependent transition matrix $W_{j\to i}(t)$. Suppose at time $t = 0$, the system is at state j, we want to simulate the next jump of the process according to the time-dependent rate $W_{j\to i}(t)$. We determine the jump in two steps: (a) decide when the jump happens and (b) decide where the jump ends at.

(a) The next event happens with a time-dependent rate,

$$\lambda(t) = \sum_{i\neq j} W_{j\to i}(t). \tag{B.36}$$

The probability for the waiting time τ to be larger than $t + h$ is the joint probability for τ to be larger than t and no event happens in $[t, t + h]$, that is

$$\mathrm{P}(\tau > t + h) = \mathrm{P}(\tau > t)(1 - \lambda(t)h) + O(h^2). \tag{B.37}$$

Therefore

$$\frac{\mathrm{d}\,\mathrm{P}(\tau > t)}{\mathrm{d}\,t} = -\,\mathrm{P}(\tau > t)\lambda(t), \tag{B.38}$$

which can be solved with the initial condition $\mathrm{P}(\tau > 0) = 1$ to give

$$\mathrm{P}(\tau > t) = \exp\left\{-\int_0^t \mathrm{d}\,t'\lambda(t')\right\}. \tag{B.39}$$

The following lemma allows us to simulate the waiting time according to this accumulated distribution.

Lemma (accumulated rejection)**.** *Consider a Poisson process with time-dependent rate $\lambda(t)$. Starting at $t = 0$, the waiting time of the next event to happen is given by (B.39). Let $\lambda_{\max} = \sup\{\lambda(t)\}$. We simulate a process, starting at $i = 1$ and $T = 0$:*

(i) Generate a waiting time T_i exponentially distributed with inverse mean $\lambda_{\max}$.

(ii) Let $T = T + T_i$. Generate a uniformly distributed random variable p between 0 and 1. If $p < \lambda(T)/\lambda_{\max}$, accept the event and exit; otherwise increase i by 1 and come back to (i).

The outcome of this accumulated rejection procedure is a value of the random number T, which is the expected waiting time with accumulated distribution (B.39).

We skip the simple proof, interested readers are redirected to Lewis and Shedler 1978.

(b) After the waiting time τ is simulated, the actual state for the transition to happen

is chosen with probability

$$p_i = \frac{W_{j\to i}(t)}{\lambda(t)}. \tag{B.40}$$

A series of two-step simulation allows us to simulate any Markov process with arbitrary time-dependent rates.

B.2.2 A model of tumour expansion and the emergence of resistant cells

Consider the expanding of a population of tumour cells starting from a single cell which is assumed to be sensitive to a particular drug. To model the competition between cells, we assume the logistic growth of the tumour in a deterministic description (Otto and Day 2007),

$$\frac{\mathrm{d}\,N(t)}{\mathrm{d}\,t} = b_0[1 - N(t)/C]N(t) - d_0N(t), \tag{B.41}$$

where b_0 is the competition-free birth rate and d_0 is the death rate of the sensitive cells and C is the environment carrying capacity. The environment carrying capacity C models the competition of the cells for resources by linearly modifying the actual birth rate b_0 to $b_0(1 - N/C)$ in the present of N cells in the population (Otto and Day 2007). The differential equation can be solved easily to yield

$$N(t) = \frac{N_{\max}}{1 + (N_{\max}/N_0 - 1)e^{-(b_0-d_0)t}}, \tag{B.42}$$

where $N_{\max} = C(1 - d_0/b_0)$ is the maximum number of cells the environment with carrying capacity C can carry and N_0 is the initial number of cells at $t = 0$, here $N_0 = 1$. The population of sensitive cells is usually very large, constituting the most part of a tumour, therefore such a deterministic description is plausibly sufficient. Such the growth of the population of sensitive cells is sketched in figure B.1.

Cells that are resistant to the drug emerge as mutant clones from the sensitive cells. Since they are normally present in minority in the population in a therapy-free environment, a stochastic description will be necessary. We assume that during a replication of a sensitive cell, a resistant cell arises at rate μ due to replication errors. Since the number of divisions per time unit of sensitive cells is $b_0[1 - N(t)/C]N(t)$, the rate of emergence of a resistant cell per time unit is

$$\nu(t) = \mu b_0[1 - N(t)/C]N(t). \tag{B.43}$$

Let $K(t)$ be the stochastic number of resistant cells, the transition rates between different

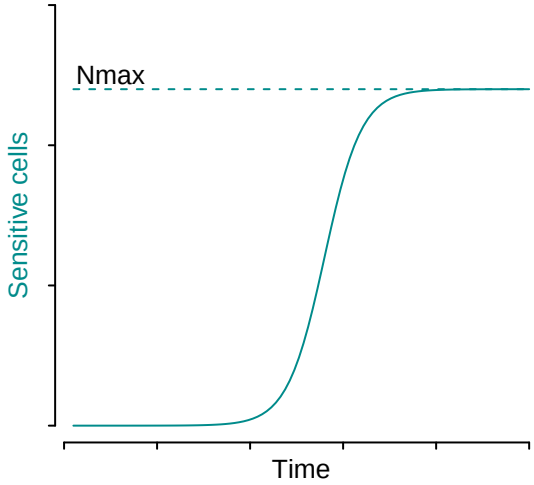


Figure B.1: **A sketch of the logistic growth of the population of sensitive cells.**

state of K are

$$q_{k\to k+1} = b[1-N(t)/C]k+\nu(t), \tag{B.44}$$
$$q_{k\to k-1} = dk, \tag{B.45}$$
$$q_{k\to k} = 1-q_{k\to k+1}-q_{k\to k-1}, \tag{B.46}$$
$$q_{k\to p} = 0 \text{ for } p\neq k, k-1, k+1, \tag{B.47}$$

where b is the competition-free birth rate of resistant cells, which is modified to $b[1-N(t)/C]$ in the present of competition, and d is the death rate of resistant cells. Note that we ignored the number of resistant cells in the competition, $b[1-(N(t)+K(t))/C] \approx b[1-N(t)/C]$, and ignored the reduction of the birth rate of sensitive cells due to mutation to resistant cell, $1-\mu \approx 1$.

This process is nothing but a non-homogeneous birth-death process with arrivals with birth rate $\tilde{b}(t) = b[1-N(t)/C]$, death rate $\tilde{d}(t) = d$ and arrival rate $\tilde{r}(t) = \nu(t)$; note that $\tilde{b}(\infty) = bd_0/b_0$, $\tilde{d}(\infty) = d$, $\tilde{r}(\infty) = \mu d_0 N_{\max}$. The condition for the number of resistant cells to be small in the population, which is expressed in (B.31), implies

$$\frac{d}{d_0} > \frac{b}{b_0}. \tag{B.48}$$

This equation can be loosely interpreted as resistant cells having lower fitness than sensitive cells. In fact, our approximation, $K \ll N$, implies $m(\infty) \ll N_{\max}$, which means

$$\frac{d}{d_0} - \frac{b}{b_0} \gg \mu. \tag{B.49}$$

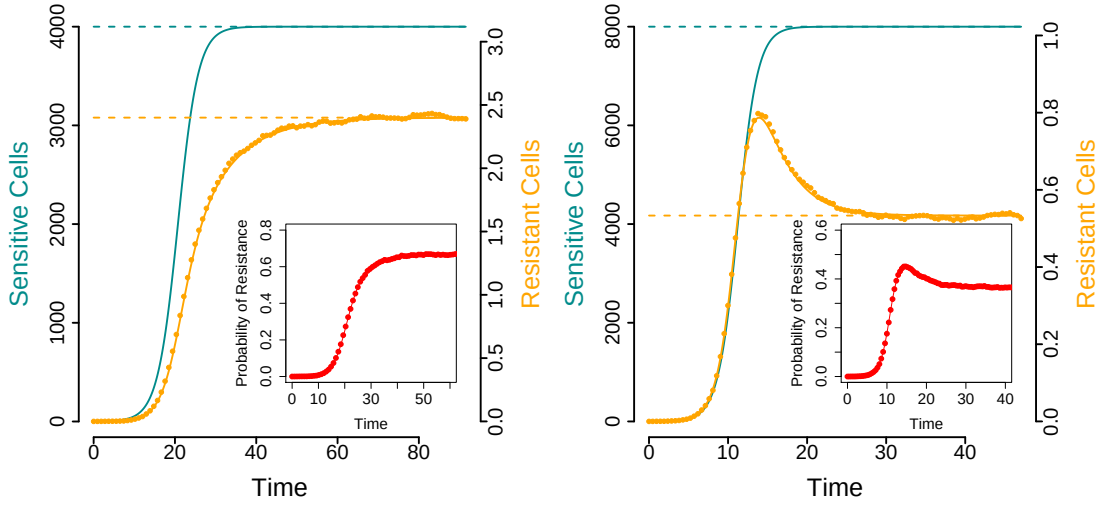


Figure B.2: The number of sensitive cells (cyan), the mean number of resistant cells (orange) and the probability of having at least one resistant cell (red, insets) as functions of time (solid lines: numerical calculations, dashed lines: asymptotic values, circles: Monte Carlo simulation). We use $N_{\max} = 10^4$, $\mu = 10^{-4}$, and the initial conditions $N(0) = 1$ and $K(0) = 0$. (a) $b_0 = 1$, $b = 0.5$, $d_0 = 0.6$, $d = 0.4$ (in an arbitrary inverse time unit), the mean number of resistant cells and the probability of resistance increase monotonically to stable values (type I–*monotonic*); (b) $b_0 = 1$, $b = 0.5$, $d_0 = 0.2$, $d = 0.4$ (in an arbitrary inverse time unit), the mean number of resistant cells and the probability of resistance increase to maxima and then decrease to stable values (type II–*non-monotonic*).

Assuming that the tumour starts from a single sensitive cell, therefore, there is no resistant cell at the beginning ($K(0) = 0$). The mean number of resistant cells $m(t)$ and the probability that the population has no resistant cells $P_S(t)$ (or conversely, the probability of having at least one resistant cell $P_R(t) = 1 - P_S(t)$, or simply *the probability of resistance*) can be calculated numerically via (B.10) and (B.22), respectively. The results are shown in figure B.2 for two different sets of parameters. We observed that both the number of resistant cells and the probability of resistance in the expanding course of sensitive cells are of two types: type I–monotonic or type II–non-monotonic. For $b_0 = 1$, $b = 0.5$, $d_0 = 0.6$, $d = 0.4$ (in arbitrary inverse time units), the mean number of resistant cells and the probability of resistance increase monotonically to stable values (type I, figure B.2 a). Instead, for $b_0 = 1$, $b = 0.5$, $d_0 = 0.2$, $d = 0.4$ (in arbitrary inverse time units), the mean number of resistant cells and the probability of resistance increase very fast to maxima before decreasing towards stable values (type II, figure B.2 b).

The next question is which parameters determine the two different types of emergence

of the resistant clone? We have six parameters: μ, $N_{\max}$, b_0, d_0, b and d. From (B.10) and (B.22), we see that $m(t) \propto \mu$ and $\ln P_0(t) \propto \mu$, the mutation rate therefore does not affect the monotonicity of $m(t)$ and $P_R(t)$. Likewise, $N_{\max}$ can only weakly affect the monotonicity of $m(t)$ and $P_R(t)$. This can be seen by observing that in (B.42), the phenomena should not change drastically by changing from $N_0 = 1$ to $N_0 = 1/f$ for some arbitrary f relatively large. Therefore, if we increase $N_{\max}$ by some factor f, and decrease N_0 by the same factor f, the function $N(t)$ simply scales by f but $N(t)/N_{\max}$ remains invariant. This leads to the rescaling by a factor of f of $m(t)$ and $\ln P_R(t)$ without changing their monotonicity. Therefore we are left with four rates, b_0, d_0, b and d, from which, only three dimensionless ratios can be derived. We then explore the phenomena numerically.

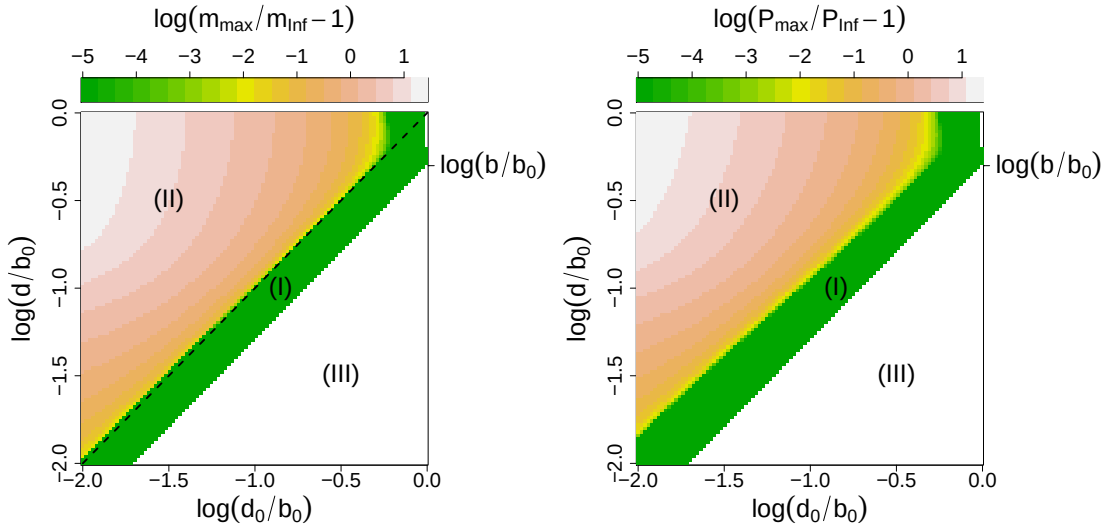


Figure B.3: The "order parameters" l_m (a) and l_P (b) as functions of $(d_0/b_0, d/b_0)$ in log scale. The diagram defines three different regimes where the emergence of the resistant clone has different characters: (I) monotonic; (II) non-monotonic; (III) resistant cells are fitter than sensitive cells. ($N_{\max} = 10^4$, $\mu = 10^{-4}$; $b_0 = 1$; effective zero: $\epsilon = 10^{-5}$.)

To distinguish the two types of emergence of the resistant clones, we define an "order parameter" for the mean number of resistant cells,

$$l_m = \frac{\sup\{m(t)\}}{m_\infty} - 1, \tag{B.50}$$

and another "order parameter" for the probability of resistance,

$$l_P = \frac{\sup\{P(t)\}}{P_\infty} - 1. \tag{B.51}$$

If the emergence of the resistant clone is of type I, $l_m = 0$ and/or $l_P = 0$; on the other hand,

if the emergence of the resistance clone is of type II, $l_0 \neq 0$ and $l_P \neq 0$. Figure B.3 plots the terrain maps of $\ln l_m$ and $\ln l_P$ as functions of d_0/b_0 and d/b_0. The terrain map of $\ln l_m$ defines a clear boundary between the two types (figure B.3 a). Further analytical analysis (not shown) confirms this numerical results, and a "phase boundary" can be drawn (dashed line, figure B.3 a). For the terrain map of $\ln l_P$, the boundary of the two types are very similar, but do not coincide with the that of $\ln l_m$ (figure B.3 b). Analytical analysis for the boundary on the terrain map of $\ln l_P$ has not been possible.

B.3 Conclusions and outlook

Our minimal model illustrates a very particular effect of cell-cell competition, namely the possibility of boosting of a resistant clone in the course of expansion of a sensitive clone. Current models of evolution of drug resistance in cancer do not explicitly consider clone competition (Iwasa et al. 2006) or treat the competition insufficiently (Bozic et al. 2012). The biological nature of clone competition in cancer is also not well explored, despite strong evidences (Greaves and Maley 2012). Interestingly, the situation is rather different for epidemic diseases. For example, in the research of malaria drug resistance, different mechanisms of clone competition between clones in the population have been investigated in detail (Mideo 2009, Read and Taylor 2001). Nevertheless, recently clone competition in the context of cancer has been recognised as more and more important (Chmielecki et al. 2011, Gatenby 2009). One of the innovative concepts is the adaptive therapy, introduced by Gatenby et al. 2009, which also leads to the development of optimally controlling a tumour (Fischer et al. 2014). In the next decades, we are waiting for more and more knowledge of clone competition in the context of cancer research, which may radically change our view on cancer treatment (Gatenby 2009).

References

Bozic, I., B. Allen, and M. A. Nowak (2012). "Dynamics of targeted cancer therapy". *Trends Mol. Med.* 18.6, pp. 311–6.

Chmielecki, J. et al. (2011). "Optimization of dosing for EGFR-mutant non-small cell lung cancer with evolutionary cancer modeling". *Sci. Transl. Med.* 3.90, 90ra59.

Fischer, A., I. Vazquez-Garcia, and V. Mustonen (2014). *The value of monitoring for controlling evolving populations*. arXiv:1406.6957.

Foo, J. and F. Michor (2010). "Evolution of resistance to anti-cancer therapy during general dosing schedules". *J. Theor. Biol.* 263.2, pp. 179–88.

Gatenby, R. A. (2009). "A change of strategy in the war on cancer". *Nature* 459.7246, pp. 508–9.

Gatenby, R. A., A. S. Silva, R. J. Gillies, and B. R. Frieden (2009). "Adaptive therapy". *Cancer Res.* 69.11, pp. 4894–903.

Gottesman, M. M. (2002). "Mechanisms of cancer drug resistance". *Annu. Rev. Med.* 53.1, pp. 615–627.

Greaves, M. and C. C. Maley (2012). "Clonal evolution in cancer". *Nat. Rev. Cancer* 481, pp. 306–313.

Holohan, C., S. Van Schaeybroeck, D. B. Longley, and P. G. Johnston (2013). "Cancer drug resistance: an evolving paradigm". *Nat Rev Cancer* 13.10, pp. 714–726.

Iwasa, Y., M. A. Nowak, and F. Michor (2006). "Evolution of resistance during clonal expansion". *Genetics* 172.4, pp. 2557–66.

Lewis, P. A. W. and G. S. Shedler (1978). *Simulation of nonhomogeneous poisson processes by thinning*. Tech. rep. IBM, p. 2286.

Mideo, N. (2009). "Parasite adaptations to within-host competition". *Trends Parasitol.* 25.6, pp. 261–268.

Otto, S. P. and T. Day (2007). *Mathematical modeling in ecology and evolution*. Princeton University Press.

Read, A. F. and L. H. Taylor (2001). "The ecology of genetically diverse infection". *Science (80-.).* 292, pp. 1099–1101.

Stirzaker, D. (2005). *Stochastic Processes and Models*. Oxford University Press.

Erklärung

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie - abgesehen von unten angegebenen Teilpublikationen - noch nicht veröffentlicht worden ist, sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Johannes Berg betreut worden.

Köln, November 2014

Hai Chau Nguyen

Teilpublikationen:

1. H. C. Nguyen and J. Berg (2012a). "Bethe–Peierls approximation and the inverse Ising problem". *J. Stat. Mech.* P03004
2. H. C. Nguyen and J. Berg (2012b). "Mean-field theory for the inverse Ising problem at low temperatures". *Phys. Rev. Lett.* 109, p. 50602
3. The Clinical Lung Cancer Genome Project (CLCGP) and Network Genomic Medicine (NGM) (2013). "A genomics-based classification of human lung tumors". *Science Transl. Med.* 5.209, 209ra153

Lebenslauf

Hai Chau Nguyen
Geburtsdatum: 28.08.1986
Staatsangehörigkeit: vietnamesisch
E-Mail: cnguyen@thp.uni-koeln.de

Studium

2010–heute: Doktorand. Institut für Theoretische Physik, Universität zu Köln (Köln, Deutschland). Betreuer: Prof. Johannes Berg.

2009–2010: Diplom in Festkörperphysik. The Abdus Salam International Center for Theoretical Physics (Trieste, Italy). Abschlussarbeit: *"Collision-dominated spin transport in graphene"*. Betreuer: Prof. Markus Müller.

2008–2009: Wissenschaftlicher Mitarbeiter im Bereich Festkörperphysik. Vietnam Institute of Physics (Hanoi, Vietnam). Betreuer: Prof. V. Lien Nguyen.

2004–2008: Bachelor of Science in Festkörperphysik. Hanoi University of Science (Hanoi, Vietnam). Abschlussarbeit: *"Quasi-gebundene Zustände von Dirac-Fermionen in Graphen"* (in vietnamesisch). Betreuer: Prof. V. Lien Nguyen.

Forschungsinteressen

Statistische Physik, statistische Inferenz, Informationstheorie, Nichtgleichgewichtsphänomene, Biophysik.

Publikationen

1. H. Chau Nguyen in the Clinical Lung Cancer Genome Project (CLCGP) and Network Genomic Medicine (NGM), *"A genomics-based classification of human lung tumours"*, Science Transl. Med. **5** 209ra153 (2013).
2. H. Chau Nguyen and Johannes Berg, *"Mean-field theory for the inverse Ising problem at low temperatures"*, Phys. Rev. Lett. **109** 050602 (2012).
3. H. Chau Nguyen and Johannes Berg, *"Bethe-Peierls approximation and the inverse Ising problem"*, J. Stat. Mech. P03004 (2012).
4. Markus Müller and H. Chau Nguyen, *"Collision-dominated spin transport in grahene and Fermi liquids"*, New J. Phys. **13** 035009 (2011).

5. C. Huy Pham, H. Chau Nguyen, V. Lien Nguyen, *"Massless Dirac fermions in a graphene superlattice: a T-matrix approach"* , J. Phys.: Condens. Matter **22** 425501 (2010).

6. H. Chau Nguyen and V. Lien Nguyen, *"Tunneling of Dirac electrons through one-dimensional potentials in graphene: a T-matrix approach"*, J. Phys.: Condens. Matter **21** 045305 (2009).

7. H. Chau Nguyen, M. Tien Hoang and V. Lien Nguyen, *"Quasi-bound states induced by one-dimensional potentials in graphene"*, Phys. Rev. B **79** 035411 (2009).

Lehrveranstaltungen

- Herbst 2014: Entwicklung von Übungsaufgaben und Beaufsichtigung von Tutoren für den Kurs *Statistische Physik*
- Herbst 2013: Entwicklung von Übungsaufgaben und Unterrichten einer Übungsgruppe für den Kurs *Fortgeschrittene Statistische Physik*
- Herbst 2012: Entwicklung von Übungsaufgaben und Unterrichten einer Übungsgruppe für den Kurs *Statistische Physik und Informationstheorie*

Ehrungen

- Stipendium der Bonn–Cologne Graduate School (BCGS) und Mitgliedschaft im "honours branch"der BCGS
- Motivationspreis für junge Forscher des Vietnam Institute of Physics 2008
- Silbermedaille in der internationalen Physikolympiade 2004 (Pohang, Korea)
- Bronzemedaille in der asiatischen Physikolympiade 2004 (Hanoi, Vietnam)
- Zweiter Preis in der nationalen Physikolympiade 2004 (Hanoi, Vietnam)
- Zweiter Preis in der nationalen Physikolympiade 2003 (Hanoi, Vietnam)