Machine Learning Methods for HIV/AIDS Diagnostics and Therapy Planning

Inauguraldissertation

zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von

Sandhya Prabhakaran
aus Kerala, Indien

Basel, 2014

Attribution-NonCommercial-NoDerivatives 3.0 Switzerland
(CC BY-NC-ND 3.0 CH)

You are free: **to Share** — to copy, distribute and transmit the work

Under the following conditions:

Attribution — You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).

Noncommercial — You may not use this work for commercial purposes.

No Derivative Works — You may not alter, transform, or build upon this work.

With the understanding that:

- **Waiver —** Any of the above conditions can be **waived** if you get permission from the copyright holder.
- **Public Domain —** Where the work or any of its elements is in the **public domain** under applicable law, that status is in no way affected by the license.
- **Other Rights —** In no way are any of the following rights affected by the license:
    - Your fair dealing or **fair use** rights, or other applicable copyright exceptions and limitations;
    - The author's **moral** rights;
    - Rights other persons may have either in the work itself or in how the work is used, such as **publicity** or privacy rights.
- **Notice —** For any reuse or distribution, you must make clear to others the license terms of this work. The best way to do this is with a link to this web page.

Quelle: creativecommons.org/licenses/by-nc-nd/3.0/ch/deed.en Datum: 12.11.2013

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät

auf Antrag von

Prof. Dr. Volker Roth, Universität Basel, Dissertationsleiter

Prof. Dr. Thomas Vetter, Universität Basel, Korreferent

Basel, den 10. Dezember 2013

Prof. Dr. Jörg Schibler, Dekan

Machine Learning Methods for HIV/AIDS Diagnostics and Therapy Planning

Abstract

by Sandhya Prabhakaran

Submitted to the Faculty of Science at the University of Basel

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy in Computer Science

THE focus of the thesis is the development and application of Machine Learning methods to the domain of HIV/AIDS diagnostics and therapy planning. The thesis addresses this domain from two different perspectives:

Facet I. The first facet of the thesis analyses the genetically-diverse HIV populations present in an infected patient's blood samples. Understanding genetic diversity is crucial for further insights into the evolution of drug-resistant viral lineage within an infected host and for *personalised medication* where drugs are prescribed to a patient based on his/her viral lineage. With the help of recent sequencing technologies, one can generate shorter viral strains called *reads* from infected blood samples that are made use of in genetic-diversity studies. The puzzle is in matching every read to its parent strain or *haplotype*, which can be seen as a standard clustering task. Given error-prone reads with limited lengths, the main modelling challenge is that non-overlapping reads do not have any suitable *a priori* pairwise similarity measure; this leads to a *non-standard* clustering problem. None of the previous approaches have provided a convincing strategy to solve this issue. In this work we overcome this problem by introducing a propagating Dirichlet Process Mixture Model.

Facet II. The second facet of the thesis takes the first steps to identify similarity patterns between drugs used in HIV/AIDS therapy and active chemical compounds. Currently there exists only a frugal number of anti-HIV drugs available to prepare drug cocktails. When a viral lineage becomes resistant to a particular drug, it tends to show resistance to other drugs in the same drug category, a property called *cross-resistance*. This situation demands development of newer and resilient drugs and thus, an indepth understanding of similarities between the current drugs and active chemical compounds is necessary. This is done by examining a landscape of active chemical compounds that also contains the drugs. With respect to this, two models are developed:

Network structure learning. We present a fully probabilistic approach to inferring networks

from pairwise Euclidean distances obtained from kernel matrices of n objects. Traditional models (Lasso-type methods), are based on the central Wishart likelihood parametrised by the inverse covariance and sparsity of the latter is usually enforced by some penalty term. Assuming a central Wishart, however, is equivalent to assuming that the origin of the coordinate system is known. If these methods use on input only kernel matrices, then usually only the kernels' pairwise distance information is truly relevant. Since traditional methods rely on an assumed origin for any kernel, they might generate biased networks. The method we developed is specifically designed to work with pairwise distances since the likelihood depends only on these distances. Combining this likelihood with a prior suited for sparse network recovery, we are able to extract sparse networks using only pairwise distances. Now network inference can be carried out on any such distance matrix induced by a Mercer kernel on graphs, probability distributions or more complex structures. Given a set of chemical compounds which also includes anti-HIV drugs, we construct kernels using the *SMILES* string encodings of the compounds. The network extracted using the kernels can be used to read out cross-resistance properties shared amongst compounds from different chemical classes and drugs' functional groups.

Archetype analysis. Archetype analysis involves the identification of representative objects from amongst a set of multivariate data such that the observations can be expressed as a noisy convex combination of these representative objects. Conventional archetype analyses rely on residual sum of squares (RSS) decay curves for model selection, which in high-noise settings, tend to break down due to no sudden change in the decay. Another drawback is that these methods are sensitive to the initialisation of archetypes at the onset of the algorithm. This is crucial for a structured dataset, where these methods have difficulties in extracting the right archetypes. In the current work, we address these problems through a Group-Lasso formulation together with a well-defined criterion, Bayesian Information Criterion (BIC), for model selection. Further, the archetypes are initialised to all the observations desensitising our method to archetype initialisation. The usage of larger datasets requires efficient methods and we therefore use the Group-Lasso to enforce grouped sparsity. Since the Group-Lasso solution ensemble can be sampled at discrete steps using a fast active-set method, BIC can be computed stepwise for model selection, thereby effecting automatic archetype identification. The method is applied to extract archetypes from a set of active chemical compounds including anti-HIV drugs. From the resulting set of archetypal compounds, one can predict functional similarities that can be shared between drugs and archetypal compounds.

Acknowledgements

This thesis would not have materialised without the encouragement, support and insight of my thesis advisor, Prof. Dr. Volker Roth. I am greatly indebted to his passion, dedication and seemingly limitless clarity over concepts which are worth commending and certainly are sources of inspiration. Prof. Roth was ever ready to take forward a research idea. He was always available with his sound reasonings and constant encouragements during this research. I thank him for the several thorough reviews and feedback on my work and for instilling, over the years, the many academic-nuggets required for writing a paper, drafting a poster/presentation, drawing figures or running code. I am extremely humbled and honoured to have completed this thesis under his academic guidance.

I also am deeply grateful to my co-advisor, Prof. Dr. Thomas Vetter, for showing interest in the thesis and reviewing it.

It gives me great pleasure in acknowledging the collaborations I have had with Prof. Dr. Niko Beerenwinkel and Armin Töpfer (ETH Zürich, Basel), Dr. Karin J. Metzner, Dr. Huldrych Günthard and Francesca Di Giallonardo (University Hospital, Zürich), Dr. Osvaldo Zagordi (University of Zürich, Zürich) and Dr. Alexander Böhm (LOEWE Center for Synthetic Microbiology, Marburg, Germany). The many meetings and discussions are indeed cherishable and have proved to be an enriching experience.

I am indebted to my colleagues Sudhir Shankar Raman, Julia E. Vogt, Mélanie Rey, David Adametz, Behrouz Tajoddin and Dinu Kaufmann for providing a conducive research environment at Basel. I appreciate the discussions and brainstorming sessions we have had so far. Sincere thanks goes to Mélanie for sharing plenty of light moments, providing quick Math-clarifications, the morning Rhine-runs and the tiny yet strong chats on art, books, life and chocolates. In addition, I am extremely grateful to Mélanie and David for their multiple reviews over the entire thesis and helping me refine the text further.

Further, I would like to immensely thank the following people for having proofread parts of the thesis and for taking the time to communicate their suggestions via email or Skype: Nitya Hariharan (Max Planck Institute for Astrophysics, Germany), Manuela Mangold (BVB, Basel), Priyambada Sinha (SS&C PORTIA, Bangalore), Fengyuan Hu (Department of Haematology, University of Cambridge), Ranjith R (IBM Software Labs, Bangalore) and Shivashankar Subramanian (Ericsson Research, Chennai).

This thesis has been typeset in 10.5 point Helvetica font and I acknowledge modifying the PhD template available here: https://github.com/suchow/.

I am certainly thankful to the numerous people I met outside research. Special mention goes to Tasqiah Julianti and Anja Schramm (Department of Pharmaceutical Sciences, University of Basel), Tejal and Ajay (Infosys/Novartis Basel), Amandine Thomas (Trad8, Delémont), Dr. med. vet. Claudia Wenk (LABOKLIN, Basel), Katrin Westritschnig (NIBR, Basel), Silvia Balu (Psychiatry Clinic Königsfelden, Aargau), Suja Maria Thomas and Vidya Surendran (UST Global, US/India), Lekshmy Sasidharan (TCS, India/UK), Nitya and Priyambada for having shared time, be it through emails or the long enduring hikes and runs in Swiss/European terrain or the brewing debates over a cup of hot chocolate/tea or lunch/dinner. I truly appreciate the time spent together.

My stay at Katholische Universitätsgemeinde (KUG) has also been rewarding. I thank the staff for their commitment to ensure a safe and studious environment through the years. During my time here, I have come across several faces from around the globe and take this opportunity to thank each and every one of them. There was never a dearth in finding company for a spontaneous run, hike, tea, dinner or a stroll by the Rhine.

The journey I have embarked of being a student for many, many years was largely possible due to the love, understanding and encouragement of my family. I cannot thank my parents and little sister enough: my father who initiated me into education and reading and taught me the virtues of time and dedication, my mother who always had a good head over her shoulders and my sister who made me appreciate the small things in life. I am humbled by the trust they showed in me that allowed me to pursue my interests and for them being the thickest of company all throughout.

Thank you indeed, one and all.

Contents

List of figures

List of Tables

1 Introduction

THE principal focus of the thesis is the development and application of Machine Learning methods to the domain of HIV/AIDS diagnostics and therapy planning. The thesis is stereoscopic in that it looks at this domain from two different perspectives:

1. It analyses the genetically-diverse HIV populations present in an infected patient's blood samples. Understanding genetic diversity is crucial for further insights into the viral-host interactions, evolution of drug-resistant viral lineage within an infected host, progression of HIV infection and for *personalised medication* where drugs are prescribed to a patient based on his/her viral lineage.

2. It also takes the first steps to identify similarity patterns between drugs used in HIV/AIDS therapy and active chemical compounds. Currently there exists only a frugal number of anti-HIV drugs available to prepare drug cocktails. When a viral lineage becomes resistant to a particular drug, it tends to also become resistant to other drugs in the same drug category, a property called *cross-resistance*. This situation demands development of newer and resilient drugs and thus, an indepth understanding of similarities between the current drugs and active chemical compounds is necessary.

In what ensues, the underlying motivation for using the HIV/AIDS domain in this thesis is presented. This caters to understanding the complicated nature of HIV and the construing problems that offer a challenging and interesting ground for applying Machine Learning (ML) techniques.

1.1 Motivation - Why is HIV/AIDS domain challenging for ML techniques

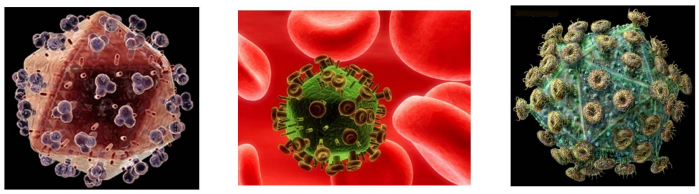


Figure 1.1.1: Appearances of HIV enveloped in the host cell membrane shown with the docking stations (seen as knobs) that aid the virus to connect onto host cells. [1]

Acquired Immune Deficiency syndrome (*AIDS*) is one of the most destructive pandemics in chronicled history and is caused by the presence of the human immunodeficiency virus (HIV) (see Figure 1.1.1) in an infected host body. The spread of HIV infection by demography, based on the UNAIDS Global Report 2006, is shown in Figure 1.1.2. The first cases of AIDS were reported in 1981 (Coffin and Varmus, 1997), and the discovery of its etiologic agent, a distinct subtype called human immunodeficiency virus type 1 (HIV-1), was identified in 1983 (Marmor et al. (2006)). Since then, there has been significant research towards understanding HIV-1 interactions at the host's cellular levels and in the development of effective antiretroviral therapy (Marmor et al. (2006)).

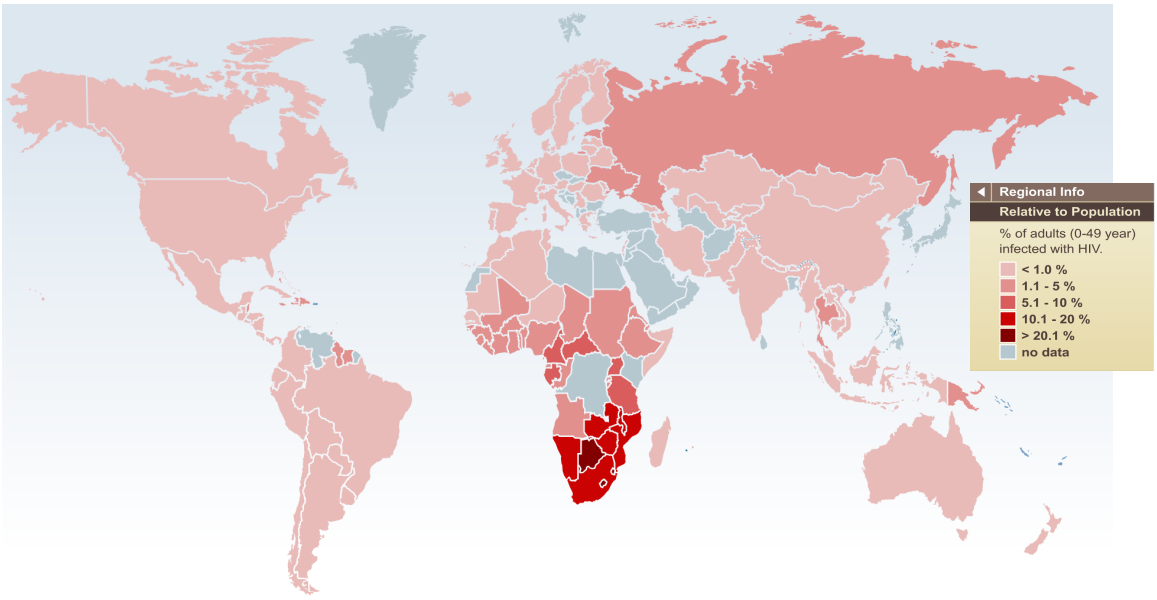


Figure 1.1.2: Spread of HIV infection. Figure courtesy: UNFPAs AIDS Clock (2006): http://www.unfpa.org/aids_clock/.

[1] Figure courtesy: http://www.rkm.com.au/VIRUS/HIV/HIV-virion-laevo.html, http://www.kurzweilai.net/protein-that-destroys-hiv-discovered, http://www.personal.psu.edu/afr3/blogs/SIOW/2011/10/good-news-for-hiv-victims.html.

1.1.1 Challenges in taming the HIV/AIDS infection

Although there have been plenty of scientific and clinical advancements in understanding and treating HIV infection now than at any point in history, the challenges to tame the HIV infection are still paramount (Marmor et al. (2006), USA (2010)). The challenges faced are primarily related to (Lamptey et al. (2006), CDC (2013), Schweighardt et al. (2010), NHS (2012)):

1. The nature of HIV: Specifically, HIV is a retrovirus and is highly mutagenic. Mutations can create newer viral strains that are likely to escape drug treatment and these drug-resistant lineages allow further disease progression.

2. Channels of HIV transmission: Transmission of HIV is primarily via a mucous membrane or bodily fluids like the blood stream or mother's milk. Utmost personal attention and stringent measures must be followed to curtail HIV transmission through these channels which, unfortunately, stand to be severely defaulted.

3. HIV/AIDS awareness of the general public: A high degree of complacency and general awareness are required.

4. Design, despatch and administering of antiretroviral drugs: Currently there exists only a meagre number of plausible drug combinations for anti-HIV therapy. When HIV becomes resistant to a particular drug, it also becomes cross-resistant to other drugs in the same drug family, eventually reducing the number of available drug combinations. Therefore, design of competent drugs is necessary to counteract the escape of drug-resistant viral mutants and to prevent further disease progression. Dissemination of drugs is also a problem in many AIDS-impacted societies due to poverty, higher medicine costs, transportation and shortage of medicines.

Each of the challenges spins further allied problems which require utmost attention. It is at this juncture, that the thesis pitches in with models tailored using ML techniques to quell some of these challenges primarily relating to the nature of HIV and aiding antiretroviral drug design.

1.2 Focus of the thesis

From the above discussions, it is clear that HIV/AIDS infection is an unabating menace. The perspective of this thesis lies in devising and applying ML techniques to analyse challenging problems that arise due to the virulent[2] nature of HIV. We mainly study two different facets of this virulency:

1. One facet deals with analysing the diverse virulent populations for identifying genetic diversity. Understanding the genetically-diverse population, throws light on drug-resistant strains - an impending aftermath caused by virulency - and the dynamics of the strains' reduced drug susceptibilities and associated therapy failure (Beerenwinkel, 2009). Further insights to the diverse populations also aid *personalised medicines*, i.e. drug concoctions tailored to the virus' genetic diversity given a specific patient.

[2]Virulency is the extremely infectious or harmful nature of a microorganism to cause disease (Vir).

2. Another facet looks for structural similarities between anti-HIV drugs and chemical compounds given the chemical compound landscape. This facilitates better understanding of drug cross-resistance and aids the design for new, potent and viable drugs.

Each facet is explained in detail below, together with the challenges posed and the thesis contribution in the area.

1.2.1 Facet One - Analysing Genetic Diversity/Identifying HIV haplotypes

The primary challenge in dealing with the HIV/AIDS infection is the inherent nature of the virus itself. HIV is a retrovirus and is known for its high mutagenecity (Mansky and Temin (1995)) which creates a diverse genetic population. This diversity facilitates the evolutionary escape of HIV from the host's immune response (Beerenwinkel et al. (2012a)) and leads to the possible emergence of drug-resistant viral strains (Meyer R Ph (2004)). The knowledge of genetic diversity is also essential for administering drugs tailored per patient that leads to the *personalised medication* model in drug treatment.

HIV is a retrovirus and is highly mutagenic. HIV is a retrovirus meaning that its genetic material is the RNA. Once it enters the cell, it copies its RNA onto the host DNA. This integration with the natural DNA results in the lifelong HIV infection (Meyer R Ph, 2004). Replication, mutation and recombination are vital for the HIV proliferation within the host (Negroni and Buc (2001), Mansky and Temin (1995)). HIV-1 has a mutation rate of 3.4×10^{-5} mutations per base pair (bp) per replication cycle (Mansky and Temin (1995)). This high mutation rate gives rise to a prolific viral load with variants that are medically termed as *viral quasispecies* or *viral population* or *mutant clouds* or *swarms* (Beerenwinkel et al. (2012a)). The mutated strains arise due to different selection pressures as the virus evolves within the host, between hosts and also depends on the infection stage (discussed in the next paragraph) (Rambaut et al. (2004), Yoshida et al. (2011), Beerenwinkel et al. (2012a)). The selection pressures are induced on the whole viral population and not on a single variant (Eigen and Schuster (1977), Eigen (1987)). Given HIV's high mutagenicity and drug-resistant nature, if left untreated, the infected person can host prolific genetically-diverse populations of HIV.

Stages in HIV infection and AIDS. The infection of HIV starts by targeting the *CD4+* cells (also known as the *T-cells*) of the human immune system which is the body's line of defense against illness and infection (Bushman et al. (2012), Mohammadi et al. (2013)). When the count of these cells falls gradually but continually, the immune system is systematically weakened. Refer Figure 1.2.1 for different stages of HIV progression.

The *acute infection phase* is the initial infection period (Klatt, 2013). This phase is marked by a higher concentration of less-diverse variants and renders the host highly contagious (Hollingsworth et al. (2008)). Over time, this results in an assemblage of symptoms and infections that the body cannot defend any further with the CD4+ cells gradually decreasing. During this *asymptomatic phase*, such a HIV-positive patient succumbs to various *opportunistic* diseases like Tuberculosis, Herpes, Hepatitis and measles (Fleming (1990)). This compounded state of infections within an

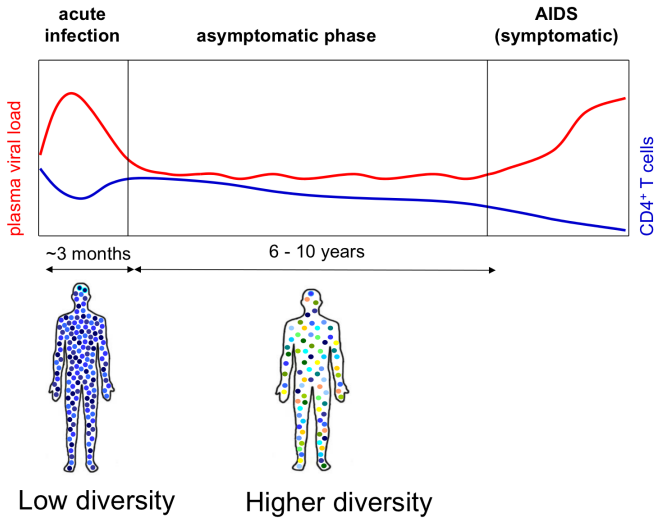


Figure 1.2.1: Various stages in HIV Viral progression (Hollingsworth et al. (2008)). Stage 1 is the acute infection phase where HIV has successfully entered the host cell and starts its replicative cycle. The host is highly contagious during this phase. Stage 2 is the asymptomatic phase where the host contracts *opportunistic* diseases due to severe weakening of the immune system. Stage 3 or the symptomatic phase is called AIDS where the infected body hosts highly-diverse HIV mutants and has a weak immune system (reduced CD4+ count). Figure courtesy: Philip Rieder and Karin J. Metzner, University Hospital, Zürich, Switzerland.

infected host eventually leads to the AIDS condition. As shown in Figure 1.2.1, stage 3 of the HIV infection or the *symptomatic phase* is called AIDS which is highlighted by a profound increase of highly-diverse viral variants and a sharply depleting CD4+ count (Vergis and Mellors (2000), Hollingsworth et al. (2008)).

Challenge in analysing genetic diversity. A HIV-infected person can host a highly-diverse viral population. Given such a diverse viral population, identifying the genetic diversity means inferring the constituent *haplotypes*, i.e. the original mutated viral strains. Input data for genetic diversity identification are the *sequenced reads* obtained from *Next-generation Sequencing* (NGS) techniques. The sequenced reads are shorter fragments of the constituent haplotypes. More details on NGS are discussed in Chapter 2. These sequencing techniques generate error-prone reads which further complicate the genetic diversity analysis: the reads that were already subject to biological diversity by way of mutations, replications and recombinations are further corrupted through machine errors (Beerenwinkel, 2009). The pursuit in identifying the genetic diversity now consists in being able to clearly separate machine error versus biological diversity to be able to effectively interpret the diverse viral load. The various methods employed for analysing genetic diversity are reviewed in Chapter 2.

Thesis contribution. The model, PredictHaplo, is introduced that infers the genetic diversity in HIV. This is a Bayesian nonparametric model and uses a propagating Dirichlet Process Mixture Model framework to infer the number of viral variants, their genetic makeup and corresponding frequencies. The model is elaborated in Chapter 3.

1.2.2 Facet Two - Aiding Antiretroviral drug design and therapy

Antiretroviral therapy (ART). Proper treatment is extremely crucial in arresting the further promulgation of viral multiplication (Klatt, 2013). As opposed to *monotherapy* where only a single drug was used to eradicate HIV, current anti-HIV drug therapy involves a mix of multiple drugs available for anti-HIV treatment (Lipsky (1996)). These drugs are known as *antiretroviral* (ART) or *anti-HIV* drugs. The aim of anti-HIV treatment, is to reduce the amount of copies or the 'viral load' to very low levels - an undetectable viral load which is below 20–50 copies per ml of blood (Martin-Blondel et al. (2012)).

Although there are *highly active antiretroviral treatments* (HAART) for AIDS and HIV to reduce the mortality and morbidity of the infection, there is no known cure (Mehanna (2003)). Apart from this, even availing or dispensing HAART promptly and economically is not possible. Further, the tussle of finding the right drug concoctions to counteract the drug-resistant variants is a lurking problem in anti-HIV therapy (Shafer and Schapiro (2008)).

To further understand the inherent challenges in HIV drug design and therapy, the different ART drugs are explored below.

Drug categories. There are currently 25 antiretroviral drugs in use for HAART (Johnson et al. (2010)). The functional crux of every drug is to prevent the immature HIV from becoming mature and infectious (Fou, 2013). Depending on the mode of HIV-host cellular interactions, these 25 drugs offer counter-attacks and can be classified into the following (Mehanna (2003), Johnson et al. (2010), AID (2012), NIA (2013), AID (2013)):

- *Entry Inhibitors*

  In order to enter the host's immune cell, HIV binds itself simultaneously onto receptor proteins (CD4 and CCR-5) present outside the immune cells. The *Entry Inhibitors* fuse to CCR-5 and block access to CCR-5, preventing entry of the virus into the host cell. Example drugs are *Maraviroc* and *Enfuvirtide*.

- *Integrase Inhibitors*

  These target the *Integrase* protein of the HIV (refer Figure 1.2.2 showing active drug target sites on the HIV genome) and prevent HIV from integrating onto the host DNA altogether. Example drugs are *Raltegravir* and *Elvitegravir*.

- *Protease Inhibitors*

  This set of anti-HIV drugs works once the HIV has infected the host cell. To initiate the viral replication within the host cell, HIV has to split into its functional constituents and uses a protein-cutting enzyme, *protease*, for this. The *Protease Inhibitors* prohibit the activity of the

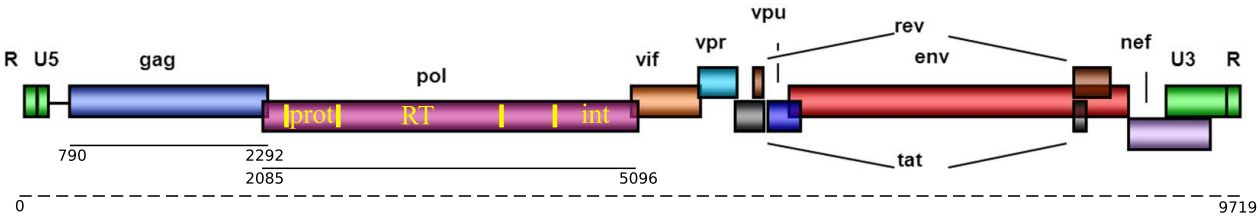


Figure 1.2.2: Landmarks of the HIV-1 genome showing the *pol* region (locations $2085 - 5096$) which constitutes the *Protease (prot)*, *Reverse Transcriptase (RT)* and *Integrase (int)* proteins; the active sites for ART drugs. Figure courtesy: Di Giallonardo Francesca, University Hospital, Zürich, Switzerland.

protease protein (refer Figure 1.2.2) by binding onto them thereby preventing further viral assemblage. Example drugs are *Darunavir* and *Tipranavir*.

- *(Non)/Nucleoside Reverse Transcriptase Inhibitors (NNRTIs/NRTIs)*

  After successful binding to the host's immune cells (CD4+ or T-cells), HIV infects the host cell by copying its genetic material present in the RNA onto the host's DNA. The copying from HIV viral RNA onto the host's cellular DNA occurs through the *Reverse-transcriptase* (RT) protein (Preston (1997)). The NNRTIs/NRTIs target the RT site of the HIV genome (refer Figure 1.2.2).

  1. NNRTIs bind to the RT enzyme itself, making the virus unable to replicate. Example drugs are *Nevirapine* and *Efavirenz*.
  2. NRTIs create faulty viral building blocks and binds to the RT protein. When HIV is in the process of copying genetic material, which in now the NRTI-generated faulty material, RT is unable to proceed with copying. Example drugs are *Stavudine*, *Lamividine* and *Tenofovir*.

  Both drug categories prevent the RT protein from copying the harmful viral RNA onto the host's DNA, prohibiting further infection.

Primary challenge in ART: Resistance to anti-HIV medications. HAART blends in an approriate drug mix from 3 different families of anti-HIV drugs available (Fou, 2013). It is important to maintain a steady level of anti-HIV drugs in the body, else HIV can proliferate quickly leading to resistance (Meyer R Ph, 2004). Resistance is the virus' ability to resist effects of the anti-HIV drugs in the body and still be able to mutate, reproduce and proliferate in the presence of antiretroviral drugs (see Figure 1.2.3) (Meyer R Ph (2004), WHO (2011)). Due to the selection pressure exerted by a particular drug, the virus sensitive to that drug (green blob) succumbs over time whereas the resistant virus (red blob) persists, proliferates and aids HIV progression. The other impending danger is that if the virus is resistant to a medication, it can become resistant to other medications in the same drug family; this is known as *cross-resistance*. For example, *NRTI*s are known to be highly cross-resistant (Johnson et al. (2010)). Further, drug concoctions for the infected person must make use of this knowledge and use drugs from other families where resistance has not yet been developed. If the cross-resistance aspect is not factored in during drug therapy, it can lead to dire consequences

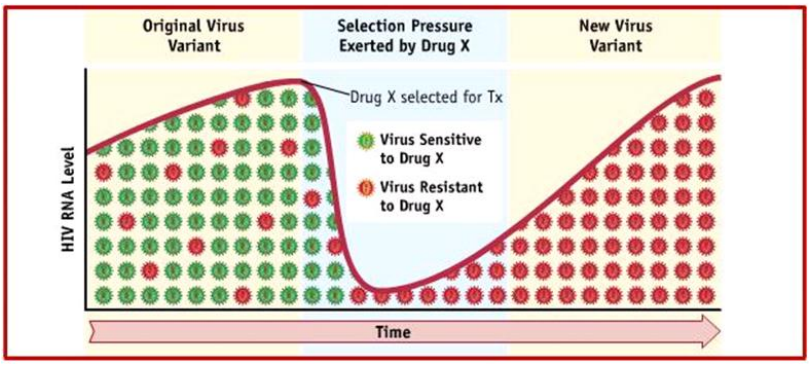


Figure 1.2.3: Temporal development of drug resistance: Selection pressure induced by ART drugs acts on the entire viral population. Over time certain strains (green blobs) succumb to the pressure whereas the drug-resistant variants (red blobs) escape, eventually progressing with their replicative cycles and are responsible for further disease intensification. Figure courtesy: Dr. Betty Chang, Microsoft Bing, U.S.

like treatment failure and spread of HIV-resistant strains which catalyse the AIDS infection (Stage 3 in Figure 1.2.1). This spins further problems in that new drugs are needed to combat the resistant strains and these drugs come at increased treatment costs. Current HIV treatments are classified as successful if they prevent initial infections (due to co-inhabiting with HIV patients) or ameliorate opportunistic diseases (Marmor et al. (2006)).

It is therefore the dire call of the hour that more and powerful anti-HIV drugs are needed to respond to HIV's drug-resistant nature since drug resistance is common and diverse, even among untreated patients (Hirsch et al. (2008), Metzner et al. (2013)). Currently there are only 25 commercialised drugs used in *ART* which fall into any of the 5 drug categories described above. Since certain drugs combinations are contraindicated (Macher et al., 2003), there remains only a handful of possible and allowed drug combinations between categories. It thus becomes necessary to investigate further the space of active chemical compounds and chalk out relations between the compounds and currently-administered ART drugs.

THESIS CONTRIBUTION. To understand HIV's behaviour towards drug selection pressures and how the drug concoctions perpetrate viral attacks, one promising strategy would be to explore structural similarities between the current 25 commercialised anti-HIV drugs and other active chemical compounds. Therefore, the other facet of the thesis explores the ART drugs where one could draw similarities of the existing anti-HIV drugs from amongst the landscape of other active chemical compounds. This would lead to better understanding of drug cross-resistance based on the structural similarities of drugs with the chemical compounds and could open doors to potent drug design and development.

Given a chemical compound landscape consisting of currently-used ART drugs and active compounds, ML techniques are developed in this thesis for:

- *Network structure recovery*: A fully-probabilistic model called *Translation-invariant Wishart network* (TiWnet) is presented that can extract a network of chemical compounds based on chemical structures. Details of the 25 ART drugs are superimposed onto the obtained network

to visualise structural similarities between the drugs and the active compounds and assists in sketching the drug cross-resistance profiles. This model is discussed in Chapter 5.

- *Automatic archetype analysis*: The model automatically identifies archetypal drugs from amongst the chemical compound landscape based on chemical structures of compounds. For any archetypal drug, an approximate convex set of active compounds is found that can be well explained by the archetype. Based on the location of an active compound in this set, predictions can be made as to how functionally similar they are to the archetype. This can be used as prior knowledge in the design and development of new anti-HIV drugs. The model is elaborated in Chapter 6.

Both these models can be used to identify active chemical compounds analogous to the drugs based on similarities of their chemical structures.

1.3 Roadmap of the thesis

The thesis deals with novel ML techniques devised and applied for HIV/AIDS diagnostics and therapy planning.

Facet I is based on identifying the genetic diversity of HIV i.e. the swarm of intra-host viral quasispecies. Chapter 2 reviews genetic diversity, the various NGS techniques and currently-available computational methods for diversity analysis. PredictHaplo, the propagating DPMM for haplotype inference is presented in Chapter 3.

Facet II presents models that identify similarities between anti-HIV drugs and active chemical compounds. Similarities can be tethered by either extracting networks of drugs or identifying archetypal drugs, given a chemical compound landscape. Chapter 4 is an introduction to graphical models and in Chapter 5, the fully-probabilistic TiWnet network model is detailed. In Chapter 6 the model for automatic archetype analysis is presented. Conclusions of the thesis and future directions are outlined in Chapter 7.

An overview of the subject areas this thesis deals with is captured in Figure 1.3.1.

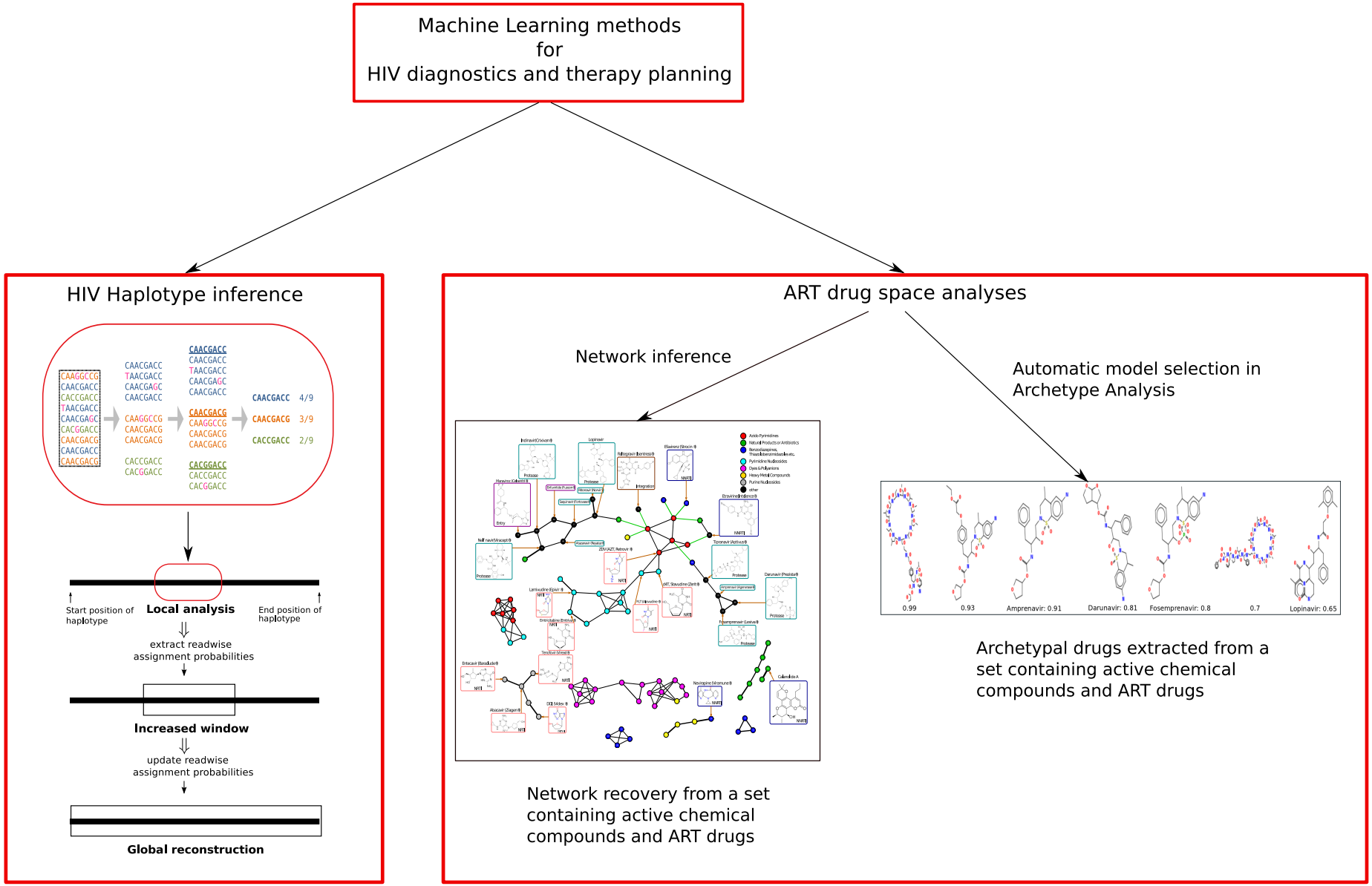


Figure 1.3.1: Graphical overview of the thesis.

1.4 List of publications

Following are the publications based on this thesis:

- "HIV Haplotype Inference using a propagating Dirichlet Process Mixture Model", Sandhya Prabhakaran, Melanie Rey, Osvaldo Zagordi, Niko Beerenwinkel and Volker Roth. IEEE/ACM Transactions on Computational Biology and Bioinformatics, [Epub ahead of print], 2013.

  An extended abstract was presented at the Machine Learning in Computational Biology (MLCB) Workshop, NIPS 2010.

- "Recovering networks from distance data", Sandhya Prabhakaran, David Adametz, Karin J. Metzner, Alexander Böhm and Volker Roth. Machine Learning Journal, volume 92:2-3, pages 251–283, 2013.

  A conference proceeding has appeared here: Asian Conference of Machine Learning (ACML'12), Journal of Machine Learning Research Workshop and Conference Proceedings, volume 25, pages 349–364, 2012.

- "Automatic Model Selection in Archetype Analysis", Sandhya Prabhakaran, Sudhir Raman, Julia E. Vogt and Volker Roth. 34^{th} DAGM/OAGM Symposium, volume 7476 of Lecture Notes in Computer Science, page 458–467, 2012.

Other publications also referred to in the thesis are:

- "Probabilistic Inference of Viral Quasispecies Subject to Recombination", Armin Töpfer, Osvaldo Zagordi, Sandhya Prabhakaran, Volker Roth, Eran Halperin and Niko Beerenwinkel. Journal of Computational Biology, volume 20:2, pages 113–123, 2013.

  A conference proceeding has appeared here: The 16th Annual International Conference on Research in Computational Molecular Biology (RECOMB'12), pages 342–354, 2012.

- "The Translation-invariant Wishart-Dirichlet Process for Clustering Distance Data", Julia E. Vogt, Sandhya Prabhakaran, Thomas J. Fuchs and Volker Roth. The 27th International Conference on Machine Learning (ICML'10), pages 1111–1118, 2010.

Facet I

Computational methods to infer HIV-1 Haplotypes using Next-generation Sequencing data

Goals in this part of the thesis:

- What is genetic diversity and why is it of significant interest.
- Understanding Next-generation sequencing techniques and sequenced reads used to model genetic diversity.
- Existing methods to infer HIV-1 haplotypes from NGS data

  SNV

  Local haplotype assembly

  Global haplotype assembly
- PredictHaplo - a propagating DPMM to infer global HIV-1 haplotypes

2

Computational Methods to infer HIV-1 Haplotypes using NGS data

THIS chapter reviews the various computational techniques available to reconstruct haplotypes and estimate genetic diversity present in an infected blood sample. Genetic diversity arises due to the presence of diverse HIV variants and is responsible for disease progression and hindrance of medical prognosis. We begin by understanding what constitutes a genetically-diverse sample and then take a tour of the NGS techniques available to sequence such samples, before reviewing the computational methods employed for haplotype assembly.

This chapter is structured based on Beerenwinkel et al. (2012a) and Beerenwinkel et al. (2012b).

2.1 Genetic diversity

HIV, which is a RNA virus, is known to be highly mutagenic thereby evolving into large viral populations that exhibit extreme genetic diversity (Mansky and Temin (1995), Preston and Dougherty (1996), Loeb et al. (1999)). They mutate at very high rates i.e. 3.4×10^{-5} mutations per base pair (bp) per replication cycle (Section 1.2.1 and Mansky and Temin (1995)). Such a replication cycle can entail error-prone steps that introduce base substitutions, frameshifts and recombinations creating genetic arrangements and leading to diverse mutants (Preston and Dougherty (1996)). The error rate of a viral mutation is 10^{-4} – 10^{-3} per base meaning that every replication of the HIV genome incurs $1-10$ errors (Nowak (1992)). An aftermath of such colossal evolutionary dynamics is that an HIV-infected patient harbours the virus as complex genetically-heterogeneous populations known as *mutant clouds*, *swarms* or *quasispecies* that are constantly evolving (Beerenwinkel et al. (2012a)).

The term *quasispecies* was coined in the works of Manfred Eigen (Eigen (1971), Eigen and Schuster (1977)) to describe a formal mathematical model explaining mutant populations. The quasispecies is defined as the equilibrium distribution of heterogeneous mutants generated by a mutation-selection process coupled with errors (Eigen (1987), Eigen and Winkler (1992), Nowak (1992), Eigen (1993), Eigen (1996)). The frequency of any mutant in a population depends on two entities: a) its ability to replicate error-free and b) the probability that it will arise by error-prone replications of other mutants in the distribution. Therefore, the resultant mutants are not independent but are coupled via mutations (Jenkins et al., 2001). These resultant mutants represent a cohesive structure that forms the target of evolutionary selection (Eigen (1987), Eigen et al. (1988) and Eigen and Winkler (1992)).

Domingo et al. (1978) was the first to suggest that RNA viruses might have quasispecies distributions, thus introducing the term *viral quasispecies*. This littany of diverse and proliferating viral mutants leads to disease progression mainly by the evolution of drug-resistant mutant variants (Koenig et al. (1995), Domingo and Holland (1997) and Domingo and Perales (2012)). This has shown to be the case with drug-resistant HIV mutants even before the onset of therapy (Bonhoeffer and Nowak (1997)). Such a prolific genetically-diverse quasispecies is crucial for the survival of HIV-1 to withstand various selection pressures both due to drug treatment (see Figure 1.2.3) and the patient's immune response (Rambaut et al. (2004), Yoshida et al. (2011)). Therefore, understanding and modelling the observed genetic diversity forms the heart to answering many medically-relevant questions related to disease prognosis and prophylaxis including drug design, development and delivery.

2.2 Next-generation Sequencing

At the helm of aiding the experimental studies for viral genetic diversity are *Next-generation Sequencing* (NGS) or 2^{nd} *generation Sequencing* techniques that present a whole gamut of methods to investigate viral populations for full-genome sequencing, metagenomics, epigenetics and transcriptomics (Metzker, 2010). Blood samples of an infected patient are used by NGS machines to output *reads* which are short fragments extracted from the viral mutants in the blood samples. The reads are further processed and used as input for diversity analysis. Previously, sampling viral populations was performed exclusively by the Sanger method (also known as the 1^{st} *generation sequencing* method) (Metzker, 2010) that was considered an epitome of high accuracy and had long read lengths of 800 bps but very low throughput ($\sim$ 6 Mega bps) (Kircher and Kelso (2010), English et al. (2012)), inferring only a consensus sequence of the sample (Zagordi et al., 2010b). The advent of NGS brought about a striking impact on later genomic research due to these quintessential features namely (Shendure and Ji (2008), Mardis (2008), Zhang et al. (2011)):

- Higher throughput: NGS allows massive production of millions of short sequenced fragments called *reads*. Figure 2.2.1 shows the advancements in throughput over the past 30 years (Stratton et al. (2009)).
- Efficient scalability: Ability to process a large number of infected blood samples.
- Greater speed.

- Reasonable resolution: NGS provides the choice to sequence from any bandwidth i.e. specific genomic sites to an entire genome.
- Cost-effective.

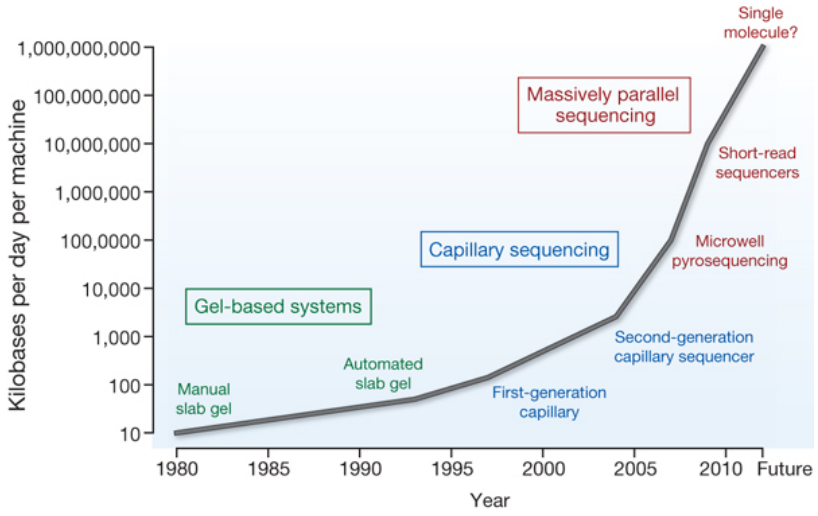


Figure 2.2.1: Advancements in DNA sequencing throughput over the preceeding 30 years from 1^{st} generation Sanger method to 2^{nd} generation highly-parallel systems to 3^{rd} generation single-molecule sequencing (Stratton et al. (2009)).

Due to these features, NGS-reads obtained from a single high-throughput NGS experiment are used in many applications in modern biology and genetics (Kircher and Kelso, 2010).

Figure 2.2.2: Current commercial NGS platforms[1].

Figure 2.2.2 shows the current commercial NGS platforms. Prominent players in NGS are 454/Roche, Illumina/Solexa, PacBio RS, Ion Torrent, HeliScope and ABI SOLiD. These sequencing platforms vary in runtime, per run costs, read lengths, machine-generated error and throughput (Metzker (2010), Zhang et al. (2011), Desai and Jere (2012)). In this thesis, work is mainly done with 454/Roche and Illumina/Solexa sequenced reads. 454/Roche GS Junior produces 100 K reads

[1]Figure courtesy: http://massgenomics.org/2010/03/next-gen-sequencing-in-2010.html.

per run with an average read length of 400 bps [2]. The more advanced 454/Roche GS FLX Titanium XL+ produces upto 1 million reads per run with an average read length of 700 bps [3]. Illumina MiSeq produces 30 million paired-end reads of 2×150 bps and 15 million single reads of 36 bps in length [4]. Illumina HiSeq 2500 produces 1.2 billion paired-end reads of 2×150 bps and 600 million single reads of 36 bps in length [5].

2.3 HIV-1 Haplotype assembly from NGS reads

As illustrated in the previous sections, NGS reads are used for estimating the viral genetic diversity. Starting with a sample containing a mixture of HIV-1 haplotypes, genetic diversity estimation or *haplotype assembly* aims at inferring:

1. The number of different haplotypes in the sample.
2. The respective frequencies of these different haplotypes.
3. The DNA sequence of each haplotype.

Figure 2.3.1 explains a typical workflow for genetic diversity analysis.

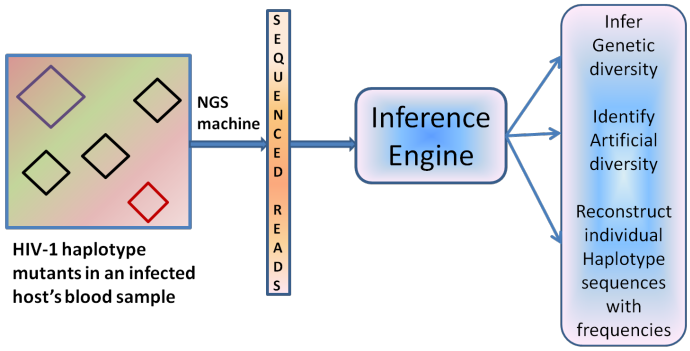


Figure 2.3.1: Snapshot of the workflow starting from wetlab experiments (extraction of infected blood samples and subsequent processing) proceeding to the sequencing machine to obtain sequenced reads. These reads are input to the haplotype inference mechanism where the reads first undergo filtering and alignment before the computational processing to infer haplotypes.

The data used for inference are the sequenced reads generated by a sequencing machine. Each read is a short base sequence which was read from one of the haplotypes by the machine. Since HIV is a *haploid* organism, it is valid to assume that a read emanates from any single haplotype. The haplotype from which a read was generated is referred to as the read's *parent* haplotype. Fig. 2.3.2 gives an illustration of reads and their corresponding haplotypes.

There are five major difficulties encountered whilst analysing these reads:

1. We do not know which of the haplotypes the reads were generated from.

[2] http://454.com/products/gs-junior-system/index.asp
[3] http://454.com/products/gs-flx-system/index.asp
[4] http://www.illumina.com/systems/miseq/performance_specifications.ilmn
[5] http://www.illumina.com/systems/hiseq_2500_1500/performance_specifications.ilmn

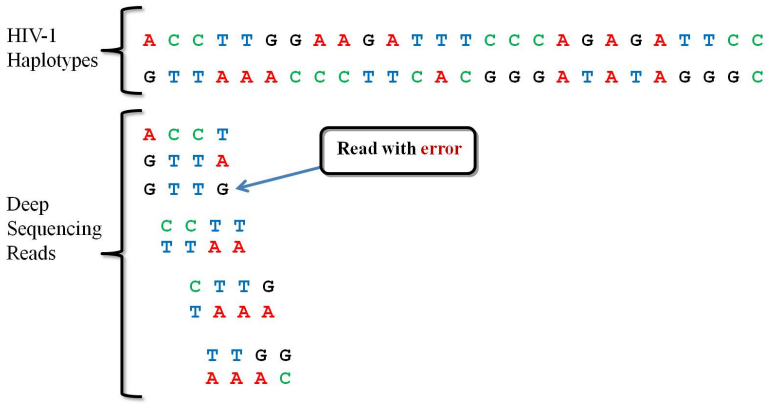


Figure 2.3.2: Error-prone reads sequenced from two different *parent* haplotypes.

2. The length of a read is much shorter than the haplotypes' length.

3. The reads are error-prone, thus a read string may not exactly match that part of the haplotype it was sequenced from. Platform-specific sequencing errors are introduced and it becomes crucial to alleviate these errors for a better statistical analysis of the data. 454/Roche introduces *indels* (insertions-deletions) in homopolymeric regions i.e. regions where the nucleotides are continuously repeated (Gilles et al. (2011)). Illumina/Solexa introduces artificial indels in the non-homopolymeric zones along with substitutions (Luo et al. (2012)). The errors are directly proportional to the read length, organism sequenced and genomic loci analysed (Gilles et al. (2011), Yang et al. (2012), Beerenwinkel et al. (2012a)).

4. The starting positions of the reads with respect to the haplotypes are unknown.

5. Non-uniform read coverage. Read coverage per position can be interpreted as the number of reads consisting any base at that given position on the genome.

For genetic diversity analysis, the ideal scenario would be having error-free reads and reads of uniform coverage i.e. uniform distribution of reads along the genomic stretch of interest. In practise, coverage is not uniform and complicates the situation further with the presence of error in reads (Zagordi et al. (2012a)). Therefore, the reads undergo rounds of preprocessing prior to using them in computational procedures for genetic diversity estimation. They are initially filtered based on their quality scores. The quality score, Q is a score based on the probability that a base is incorrectly called [6] and is calculated as $Q = -10 \times log_{10}(p)$ [7] where p is the estimated probability of the base call being wrong (Ewing and Green, 1998). Thus, a higher value of Q indicates a smaller error probability [7]. During filtering, the low-quality reads are removed. The rest of the reads are subject to alignment, where identifying the starting position of the reads can be efficiently done using

[6] Base calling is the process of matching one of the 4 bases to the chromatogram peaks, each peak taking one of the 4 allowed colours. A chromatogram is an output of a sequencing run and provides a visual depiction of a DNA genome with one coloured peak per genome position (Bio, Her).

[7] http://www.illumina.com/truseq/quality_101/quality_scores.ilmn

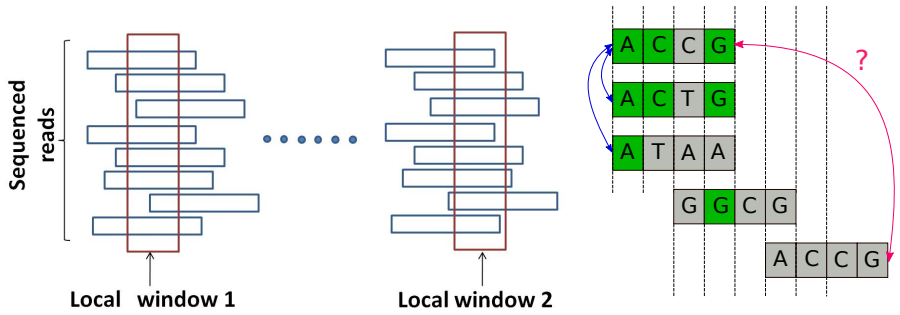


Figure 2.3.3: Clustering based on mixture models works for fully and partially-overlapping reads (left and central) but not for full-length reconstruction (right).

alignment techniques (Durbin et al. (1998)) with respect to a HIV reference sequence [8], thereby producing a set of *aligned reads*. For instance, reads are mapped individually to portions of the reference genome to arrive at a consensus alignment called *Multiple sequence alignment* (MSA) from all the pairwise alignments (Beerenwinkel et al., 2012a). An example of its usage is presented in Zagordi et al. (2011). Other aligners widely used are *Burrows-Wheeler Alignment Tool* (BWA) (Li (2012)), MOSAIK (Mos) and SEGEMEHL (Hoffmann et al. (2009)). Challenges in aligning relate to adequate handling of frameshifts, gap placements and substitutions. Sophisticated aligners tackling these problems are available (Langmead et al. (2009), Langmead and Salzberg (2012) and Li (2012)).

Once the reads are filtered and aligned, they are ready for haplotype assembly. Haplotype assembly refers to classifying the reads with respect to unknown haplotypes and is inherently a clustering problem: the data points to cluster being the aligned reads and the cluster centroids being the unknown haplotypes. This, however, is a *non-standard* clustering problem since the major difficulty posed is that there is no *a priori* 'natural' similarity measure defined for this complete read dataspace. The reason is that, since the reads are much shorter than the haplotypes and can start at any position along the haplotypes, two reads randomly chosen will generally not have overlapping positions. Figure 2.3.3 elucidates the non-standard clustering problem due to non-overlapping reads. Then the question encountered here is that of finding a similarity measure between two reads generated from different distant regions of the haplotypes. Since there is no direct coupling of non-overlapping reads, a potential similarity measure first needs to relate the reads to some intermediate object. This object is the set of haplotypes as will be shown in Section 3.4.

The absence of *a priori* pairwise relationships between these non-overlapping reads explains why obtaining a full-length or *global reconstruction* of the haplotypes constitutes a hard problem (Beerenwinkel and Zagordi (2011)). The problem becomes easier if we initially consider only a smaller region of the haplotypes i.e. a region small enough for every two reads to overlap and work within this local window (see left and central plots in Figure 2.3.3). The problem now translates to a standard clustering task where the goal is to differentiate between true mutations (sources of inter-cluster differences) and sequencing-machine errors (sources of intra-cluster differences). This local-window

[8]The reference sequence for HIV was obtained from http://www.hiv.lanl.gov/content/sequence/HIV/MAP/landmark.html.

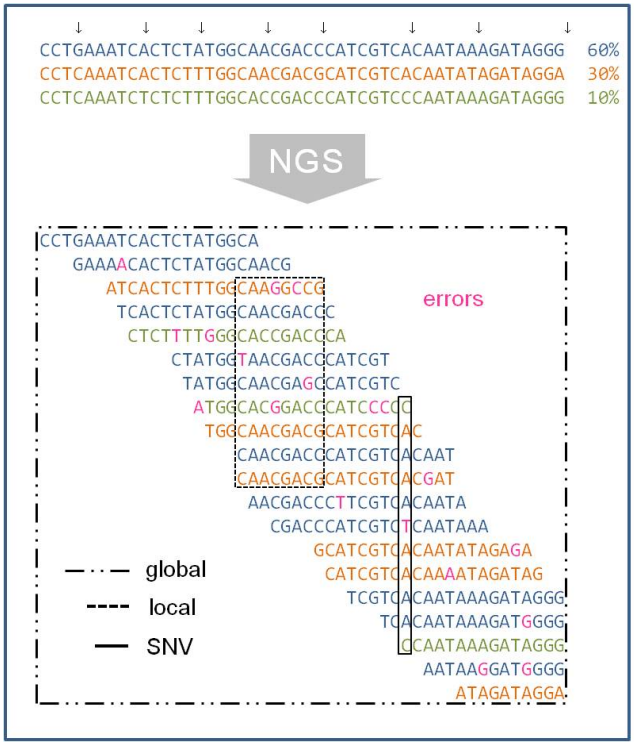


Figure 2.4.1: Spatial stratification used in diversity estimation. Three parental haplotypes of different proportions are sequenced to produce error-prone reads. Errors are shown in pink. For haplotype assembly, three spatial zones can be considered. SNV: single column analysis, local analysis: is for an aligned short stretch of the genome rather than the entire genome length, global analysis: where the whole genome is considered for haplotype reconstruction. Figure courtesy: Beerenwinkel et al. (2012a).

clustering solution leads to a *local reconstruction* of the haplotypes. Various computational approaches to haplotype reconstruction thereby generally start by solving local reconstructions before combining these local information to infer full-length haplotypes. Details of these approaches are presented in the forthcoming section.

2.4 Computational Approaches for HIV-1 haplotype assembly

This section reviews the various approaches available to infer the HIV-1 haplotypes given a set of (Roche or Illumina) sequenced reads. Depending on the length of the interest region for genomic diversity analysis, the approaches can be spatially stratified into 3 (Beerenwinkel et al., 2012a) namely:

- *Single-nucleotide variant* or SNV
- Local haplotype assembly
- Global haplotype assembly

Figure 2.4.1 elucidates these stratification zones used for haplotype assembly.

2.4.1 SNV

Single-nucleotide variant or SNV is a random abherration or defect occurring to a single nucleotide of a haplotype (Snp, 2009). Estimating genetic diversity by identifying SNV or what is referred to as *SNV calling* relates to statistical analyses based on count data which concurs to the per-position quantification of mutation prevalence (Beerenwinkel et al., 2012a). A naive approach to identify if the SNV is a true biological mutation or a sequencing-induced error, is to assume that 1) at every SNV-location the number of sequencing errors follows a fixed Poisson distribution, normally used to model count data, and that 2) true alleles are called based on a given error rate, when their frequency is higher than expected-by-chance alone (Wang et al. (2007)). Another method to model the allele counts per location is through a Binomial mixture model where the number of mixtures allowed or in other words, the number of SNVs called at that site, is fixed *a priori* (Crisan et al. (2012)). To improve accuracy in SNV-calling, control experiments are performed by sequencing the same viral samples. Then, count data analyses can be done between pairs of mixed and control samples simultaneously, for example, using Fisher's exact test for every allele (*Varscan 2* software by Koboldt et al. (2012), `http://varscan.sourceforge.net/`), or by assuming independent Poisson distributions to model sequencing errors per position and checking the number of mismatches in the observed alleles (*vipR* software by Altmann et al. (2011), `http://htsvipr.sourceforge.net/`). A refinement to these SNV-calling methods is to introduce a Beta-binomial distribution to model the per-site sequencing errors (*deepSNV* software by Gerstung et al. (2012), `http://www.bsse.ethz.ch/cbg/software/deepSNV`) or use hypothesis testing with (Bonferroni) correction for multiple testing (*Shimmer* software by Hansen et al. (2013)).

2.4.2 Local haplotype assembly

Extending the region of interest from a single genomic locus as that in SNV to a stretch of loci where the sequenced reads within this stretch tend to almost perfectly overlap, leads to *local* diversity analysis. Choice of the local window stretch is crucial as smaller windows, although have larger coverage, would contain fewer discriminating SNVs necessary for pairwise comparison of reads, whereas larger windows would have low coverage, but more SNVs (Beerenwinkel et al. (2012a)). Within the window, reads are subject to clustering with the assumption that similar reads have emanated from the same parent haplotype. This is valid only when the machine error is low with respect to the biological variation in the sample and the ability to identify true variants increases with coverage (Eriksson et al. (2008)).

A probabilistic flavour to clustering reads locally was imparted in Eriksson et al. (2008), Zagordi et al. (2009) and Zagordi et al. (2010a). Here, the error rate is estimated along with the number of mixture components. The predicted haplotypes are the cluster centres and the haplotype frequencies are given by the cluster weights. As depicted in Figure 2.4.2, reads are corrected locally by replacing all read alleles with its cluster centre. This method of local error correction reduces per-base error, decreases false positives of local haplotype inference and improves the haplotype frequency estimates (Zagordi et al. (2010b), Beerenwinkel et al. (2012a)).

Diversity estimation based on flowgram clustering instead of the sequenced reads is devised as *AmpliconNoise* (Quince et al. (2009), Quince et al. (2011)). A flowgram is a bar graph of light

intensities generated in the NGS chambers (454, 2007). The signal intensity is directly proportional to the number of nucleotides per position (454 (1996), 454 (2007)) (see Figure 2.4.3). The sequences read out from the flowgrams are assumed to resemble the true sequences but are subject to machine errors. Clustering error-prone flowgrams leads to estimating a mixture model for the proposed true sequences (Beerenwinkel (2009), Beerenwinkel et al. (2012a)).

Eventually, clustering sequences or flowgrams should result in differentiating biological variations as inter-cluster variations and machine-induced errors as intra-cluster variations (Beerenwinkel et al., 2012a).

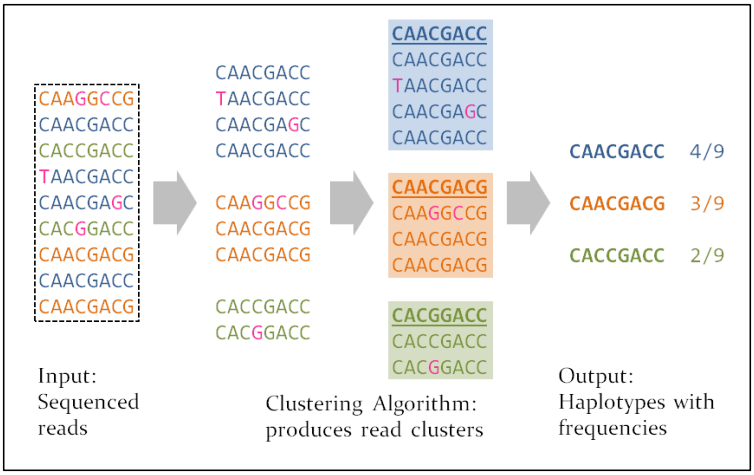


Figure 2.4.2: Local haplotype assembly. Given a set of error-prone sequenced reads, local analysis clusters the reads. All intra-cluster reads are then locally corrected based on the cluster centroid. The cluster centroids are the reconstructed haplotypes and the cluster weights are the haplotype proportions. Figure modified from Beerenwinkel et al. (2012a).

2.4.3 Global haplotype assembly

Global haplotype assembly is effected by increasing the spatial window from a local stretch bounded by the average read length to a window that spans the entire length of the genome. The aim is in finding whole-length haplotypes that represent the viral population's genetic makeup or the *quasispecies* irrespective of the sequencing machine's specifications like average read length or read coverage (Beerenwinkel et al., 2012a). It gives a broader picture over the entire viral population being sequenced. The assembling approaches for genome-wide haplotypes fall in roughly 3 main genres namely (Beerenwinkel et al. (2012a), Beerenwinkel et al. (2012b)):

- Graph-based combinatorial assembly
- Probabilistic assembly using mixture models
- *De novo* methods: these do not assume any reference genome but construct one instead from the sequenced reads (Narzisi and Mishra (2011), Finotello et al. (2012)).

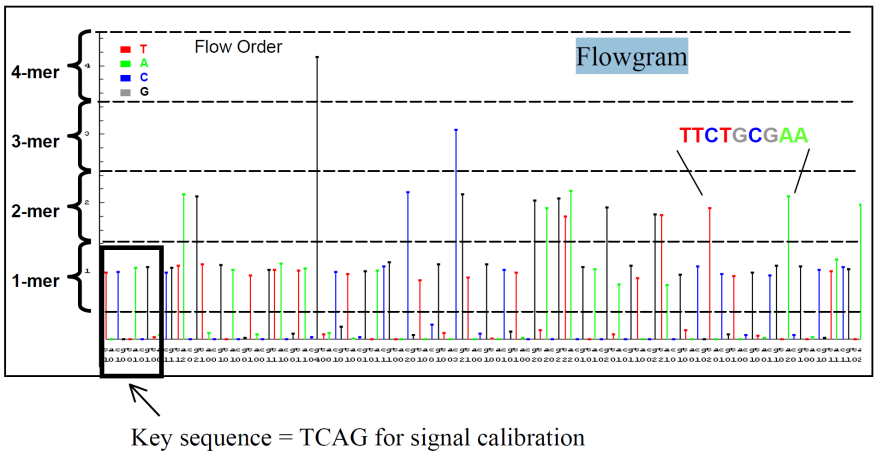


Figure 2.4.3: Flowgram: A flowgram is a bar graph of light intensities generated in the NGS chambers (454, 2007). The signal intensity is directly proportional to the number of nucleotides per position (454 (1996), 454 (2007)). In the figure, the light intensities for sequence *TCAG* is shown. The height of the signal strength (y-axis) per sequence position (x-axis) shows the number of same nucleotides incorporated at that position. Figure courtesy: Droege and Hill (2008).

Graph-based Combinatorial assembly. The local methods mentioned above (Section 2.4.2) can also be rendered to address the global assembly problem. Here, the reads are first locally corrected for machine error. Next, to obtain a global reconstruction of the haplotypes, these corrected reads are used to create a *read graph* from which a minimal set of haplotypes best explaining the read graph – the minimal path cover – is derived. The read graph construction is shown in Figure 2.4.4. Nodes in the graph represent locally error-corrected reads and the directed edges indicate the reads' alignment order with respect to a reference genome (Beerenwinkel et al., 2012a). Redundant nodes are removed if they overlap exactly. To complete this finite automata, a universal source and sink node are provided from which all plausible paths start and to which all such paths terminate, respectively. Only those edges are retained between nodes such that they are the only *informative* connecting links between nodes; a property called *transitive reduction* (Westbrooks et al. (2008)). In Figure 2.4.5, the transitive reduction property is shown for 3 reads. The path from node u to node w always overlaps node v entirely, therefore the *direct* path between node u and node w is removed.

In terms of the read graph, a haplotype is defined as a path from source to sink. The quest of identifying quasispecies' constituents using the read graph can be formulated as finding that set of source-sink paths that explain the locally error-corrected reads well (Beerenwinkel et al., 2012a).

In Eriksson et al. (2008), the path cover concept is synonymous to that of read graph. The nodes of the graph are error-corrected reads where the correction procedure is done in several steps comprising statistical tests and *k-means* clustering of the reads. An example of a path cover for 20 reads is shown in Figure 2.4.6. Quasispecies assembly refers to finding the minimal path cover over all the reads or finding the minimal set of haplotypes. The minimal path cover can be computed in

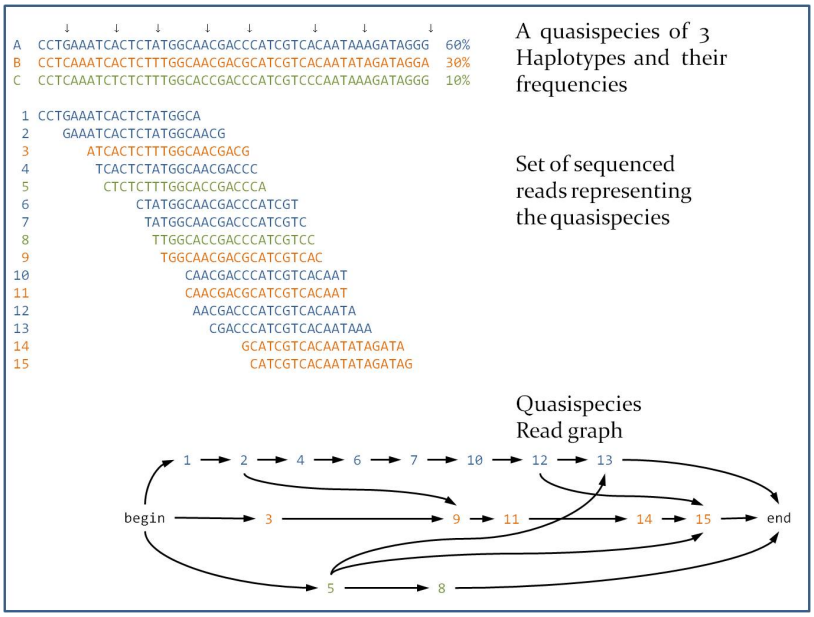


Figure 2.4.4: Quasispecies Read graph: Shown are 3 haplotypes with their corresponding frequencies and a set of 15 reads sequenced from the haplotypes. Given these reads, a read graph is constructed having one read per node and edges between 2 nodes imply an allowed overlap between the read pairs. A haplotype can be read out from the read graph as a path from 'begin' to 'end'. The graph must explain a minimal set of haplotypes by sketching the minimal path cover over all reads. Figure modified from Beerenwinkel et al. (2012a).

$O(N^3)$ where N is the total number of reads (Eriksson et al. (2008), Beerenwinkel et al. (2012b)). The same general approach is followed by Zagordi et al. (2009), Zagordi et al. (2010a) but with a different method to locally correct the reads: a Dirichlet process mixture model (DPMM) is used to distinguish true mutations from machine errors. Their software implementation is available as *ShoRAH* (Zagordi et al. (2011)).

Another read graph-based assembly is cast as a network-flow problem in Westbrooks et al. (2008). In the network-flow problem, one views each path (i.e. a haplotype) as a connection from source to sink. A flow f through the read is the number of haplotypes that contains the read (Westbrooks et al. (2008), Beerenwinkel et al. (2012b)). The main idea for global assembly is to state the optimisation problem as one that minimises f leading to the most parsimonious quasispecies assembly subject to the constraint that each read is part of at least one haplotype (Westbrooks et al. (2008)). A more general variant of this approach is implemented in the software *ViSpA* (Astrovskaya et al. (2011)) where sequencing errors are taken into account by allowing mismatches in the overlap between reads and then constructing a haplotype as a weighted consensus sequence over all reads.

In all of the above combinatorial assemblies, the relative frequencies of these explaining haplotypes are estimated using the *Expectation-Maximisation* (EM) algorithm.

Another haplotype reconstruction method is presented in Prosperi et al. (2011) and uses a graph approach slightly different from read graph. An *overlap graph* is constructed, using greedy path sampling, to obtain a set of haplotypes which minimises the number of false variants, called *in-silico*

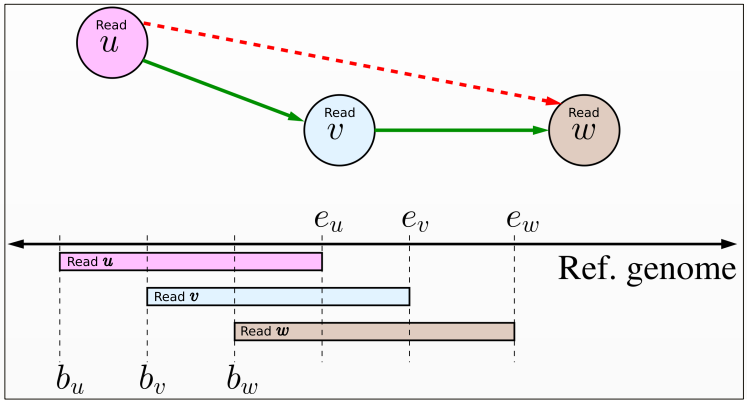


Figure 2.4.5: Transitive Reduction: Consider 3 reads u, v and w with their 'begin' b and 'end' e positions obtained after aligning to the reference genome. The reads constitute nodes in the read graph. The path from node u to node w always overlaps entirely node v, therefore the *direct* path between node u and node w is removed as the path from $u \rightarrow w$ possesses the transitive property by going from $u \rightarrow v$ and from $v \rightarrow w$. Figure modified from Beerenwinkel et al. (2012b), Westbrooks et al. (2008).

recombinants. This is synonymous to evading redundant paths in the read graph (Beerenwinkel et al., 2012a). The software of Prosperi et al. (2011) is provided as *QuRe* (Prosperi and Salemi (2012)). *QuRe* makes use of the amplicon-based structure [9] seen in sequenced reads for haplotype reconstruction (Beerenwinkel et al., 2012a) but is however designed for error-free reads and does not provide any mechanism for handling sequencing errors.

A similar approach to using locally error-corrected reads and the read graph method for global haplotype assembly is devised as *AmpMCF* and *ShotMCF* that cater to amplicon-based reads or NGS reads, respectively (Skums et al. (2013)).

Another haplotype reconstruction problem, again based on graphs, and defined as a vertex-colouring problem called *QColors* is dealt with in Huang et al. (2011). There are two complementary graphs used whose nodes consist of overlapping reads: an *overlap graph* where edges are present between reads that have non-conflicting overlaps and a *conflict graph* where egdes are present between reads that have conflicting overlaps (Huang et al., 2011). The reconstruction problem is posed as finding that partition of reads satisfying a minimum number of non-conflicting subsets, akin to a vertex-colouring problem. A drawback of *QColors* is that it could be sensitive to machine errors and erroneous alignments (Beerenwinkel et al. (2012a)).

However, most of these combinatorial graph-based approaches have some potential drawbacks. Local error correction is potentially misleading since it is impossible to revise this step in a global context. Such a correction step also necessarily removes the stochastic nature of the error-prone reads which precludes a proper modelling of the uncertainty in the haplotypes. This strategy of locally correcting the reads turns the haplotype assembly problem into becoming deterministic in nature (Beerenwinkel et al. (2012a)). Further, reconstructing global haplotypes using a minimal path cover of the read graph is a linear programming problem and therefore computationally expensive

[9] An amplicon is a nucleic acid (DNA or RNA) strand that gets amplified into multiple copies (Amp, 2013). Primers are attached to ends of the amplicon that serve as starting points for the DNA synthesis (Pri, 2013). Reads amplified using amplicon-based strategies tend to show the demarcation corresponding to these primers.

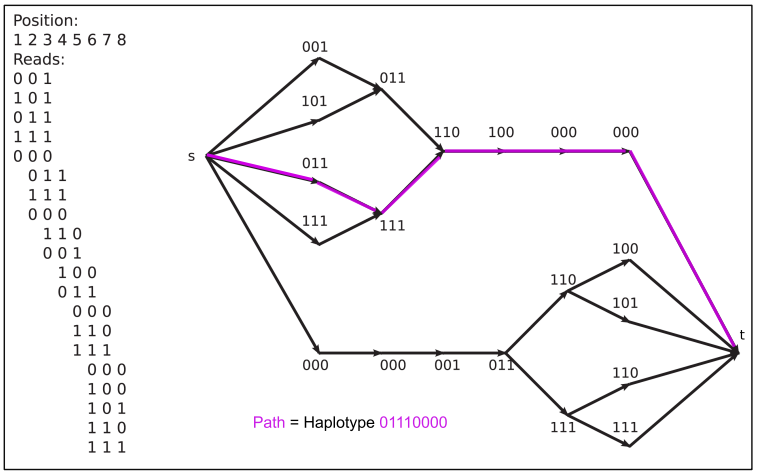


Figure 2.4.6: Path Cover: A set of 20 locally error-corrected reads is given. Every read is a node in the graph and a haplotype is a path from start node s to end node t that explains the reads well. As an example, the purple path designates haplotype 01110000. Figure modified from Eriksson et al. (2008).

and ambiguous, since the corresponding optimisation problem does not offer a unique solution, in general.

PROBABILISTIC ASSEMBLY. The read graph-based methods imparted a deterministic nature to haplotype assembly due to the reads' premature error removal step, that removed all the stochasticity from the set of observed reads. With a view to take into account the inherent stochasticity and model the error accordingly, probabilistic approaches have been put forward. Rather than pursuing an optimal parsimonious solution as in graph-based methods, probabilistic approaches resort to specifying probability distributions that encode some *a priori* information of the sequencing machines. Generally, the complexity of probabilistic models is given by a single parameter that is used to tune the number of haplotypes. With this simple parametric form, one can verify the reconstructed haplotypes based on the model assumptions and also segregate between false and true haplotypes (Beerenwinkel (2009)).

A probabilistic hierarchical model reproducing the reads' generative stochastic process was devised in Jojic et al. (2008). Parameters and hidden variables include the parent haplotype, starting position and error transformation and are estimated with maximum likelihood estimation using EM. This model however assumes that the number of haplotypes is known, and does not describe any formal method to estimate this number. In practise, this estimation constitutes a major challenge for haplotype reconstruction.

Another probabilistic approach for haplotype assembly that also caters to automatically inferring the number of haplotypes is *PredictHaplo*. This is the new model developed and introduced in the thesis and is the theme of Chapter 3. Here, every haplotype is represented as a set of location-specific probability tables over the four nucleotides (refer Figure 3.4.1). The underlying generative model assumes that reads are sampled from a mixture model, where every mixture component represents one haplotype and the component's mixing proportion estimates the haplotype frequency in the

given viral population. Since the number of haplotypes are not known up-front, it is assumed that there are infinitely-many different haplotypes from which reads are generated. Figure 2.4.7 explains the assumed stochastic process for read generation. The probability of choosing any haplotype h follows a multinomial distribution and given h whose total length is L, a location l is randomly chosen on h. A read is then read out starting from l till l_{end} where $l < l_{end} \leq L$ and terminates at the end state. *PredictHaplo* uses a truncated version of the *Infinite Mixture Model (IMM)* also known as *Dirichlet Process Mixture Model (DPMM)* (Ewens (1972), Ferguson (1973), Rasmussen (2000)) that adaptively chooses the number of haplotypes. Entire working details of *PredictHaplo* are presented in Chapter 3.

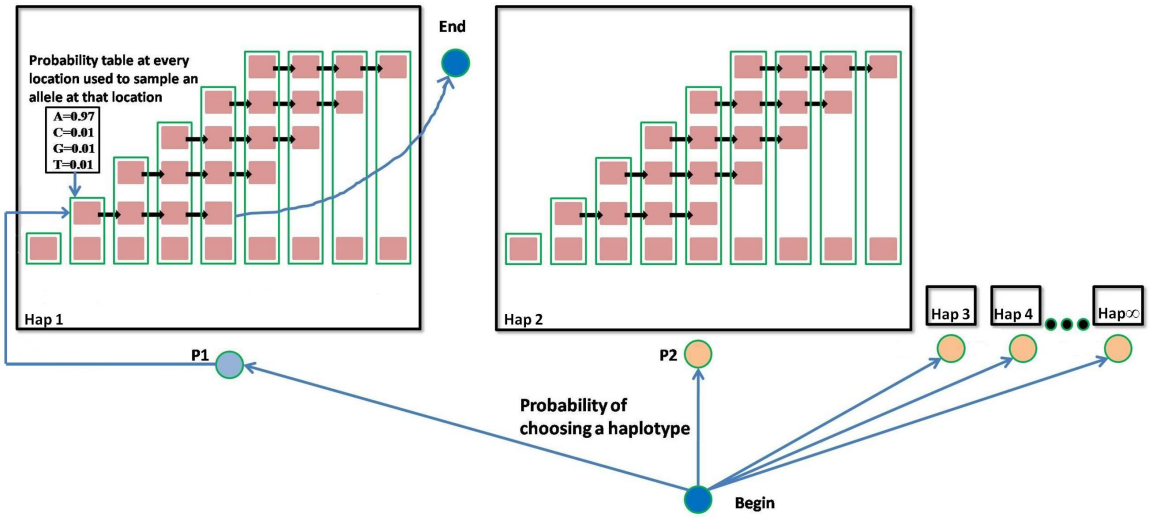


Figure 2.4.7: Infinite-state automata explaining the assumed stochastic process for read generation: Consider infinitely-many haplotypes from which reads can be generated. One read emanates from only one haplotype. The probability of choosing any haplotype h follows a multinomial distribution and given h, a location l is randomly chosen. A read is then read out starting from l till the end of h or before and terminates at the end state.

QuasiRecomb is an approach to infer the *distribution* of *generators*, the set of sequences that mutate and recombine and are responsible for quasispecies creation (refer Zagordi et al. (2012b), Töpfer et al. (2013)). *QuasiRecomb* relies on the fact that HIV is highly recombinant in nature and that recombination is amongst the prime factors for maintaining diversity (Beerenwinkel et al. (2012a)). The generative model for haplotypes is based on the presumption that they emanate from a small set of *generators* by virtue of mutation and recombination (Zagordi et al. (2012b)). Figure 2.4.8 depicts the underlying haplotype and read generation process given a set of generators. The model is designed using a *jumping Hidden Markov Model* (jHMM) framework that makes use of HMM switching to be able to switch or jump between potential generators (Zagordi et al. (2012b)).

De novo METHODS. The third genre of global diversity analysis consists of *de novo* methods. *De novo* read assembly is the construction of longer sequences called *contigs* from shorter sequenced reads without *a priori* information of the read order or reference genome but solely relies on the pairwise overlaps between reads (MacLean et al. (2009)). Figure 2.4.9 shows the workflow of how a

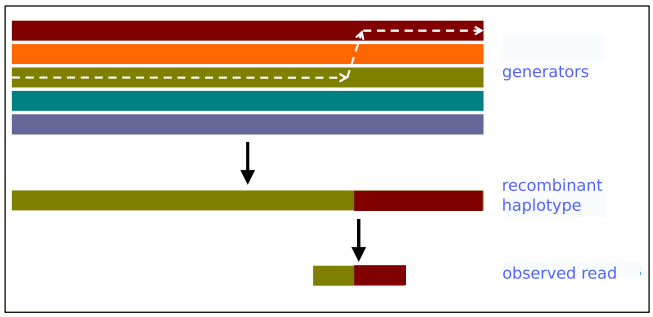


Figure 2.4.8: Modelling recombinants using *QuasiRecomb* (Zagordi et al. (2012b), Töpfer et al. (2013)). From a set of 5 generators, a *recombinant haplotype* is created using the 1^{st} and 3^{rd} generator. A read is sampled at random from this recombinant haplotype which possesses parts of both the 1^{st} and 3^{rd} generator. Modelling recombinant reads is done using a jumping HMM that makes use of the switching property of HMMs. Figure modified from Beerenwinkel et al. (2012b).

contig is constructed from a set of sequenced reads. In the *de Bruijn graph*, the read is represented as a sequence of k-mers (short fragments k-alleles long). For the overall overlap consensus in the *de novo* assembly, only those overlaps between reads explained by the *de Bruijn graph* are used, thus removing redundant paths (see MacLean et al. (2009)). There are many *de novo* assemblers that are platform-specific (Narzisi and Mishra (2011)) and those that work on reads by mixing platforms (see Aury et al. (2008), MacLean et al. (2009)). Although in *de novo* methods only a single contig is reconstructed, this can be seen as a read-pre-processing step to the many approaches for quasispecies assembly mentioned before (Ramakrishnan et al. (2009), Beerenwinkel et al. (2012a)).

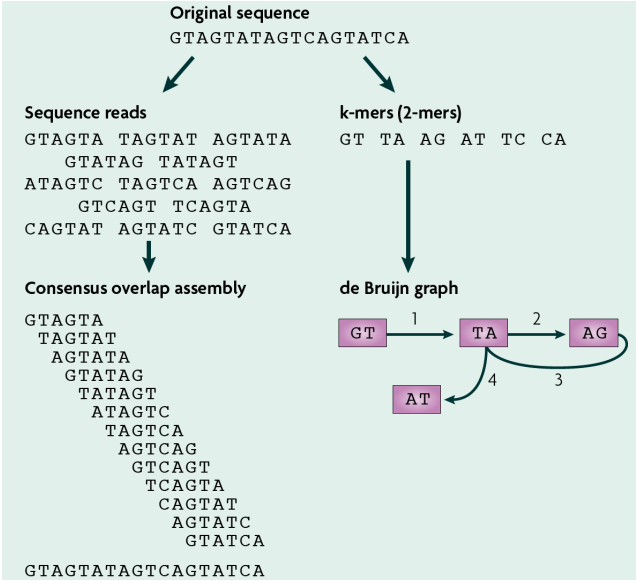


Figure 2.4.9: *De novo* read assembly: Sequenced reads are used to construct a contig based on overlaps explained by each read's *de Bruijn graph*. The *de novo* method does not rely on any reference genome or read order for contig construction (MacLean et al. (2009)). Figure courtesy: MacLean et al. (2009).

2.5 Conclusion

With the preceeding discussions on the many computational methods developed for genetic diversity estimation using NGS reads, it is clear that diversity estimation is a crucial link for the better understanding of HIV-host interactions, evolution of drug-resistant viral variants, disease prognosis/prophylaxis and provides the necessary leeway for new ART drug/vaccine design and development. Next-generation sequencing has made it possible to look at diversity problems based on different lengths – from single positions to whole stretches – of the genome and thereby open research questions relevant to this spatial stratification. There is still plenty of scope for improving genetic diversity estimation and this is discussed in Chapter 7.

3

HIV Haplotype Inference using a propagating Dirichlet Process Mixture Model

3.1 Introduction

THIS chapter presents a new computational technique for the identification of genetically-diverse HIV (Human Immunodeficiency virus) haplotypes present in an infected blood sample. HIV is a retrovirus that causes the widespread, life-threatening AIDS (Acquired Immunodeficiency Syndrome), attacking the human immune system (for more details, refer Chapter 1). Since HIV mutates fast, upto the order of 10^{-5} mutations per bp per replication cycle (Mansky and Temin (1995)) with every mutation having an error rate of the order of 10^{-4} per base (Nowak (1992)), a patient generally hosts many different virus mutants (Mansky (1998)). The particular DNA sequence which constitutes the genetic material of a mutant is called a *haplotype*. The variety of mutants present in the patient pose a major issue in HIV treatment because many of these mutants can be resistant to different drugs (Perrin and Telenti (1998)). Therefore, identifying the haplotypes present in a particular patient enables adapting the treatment to the specific patient, paving the way for *personalised medication* to administer the most efficient drug concoction.

One of the latest technological innovations enabling the extraction of information about the haplotypes present in a sample is *deep sequencing*/NGS. More details on NGS can be found in Section 2.2. Here, the mixture of haplotypes is processed through a series of chemical manipulations carried out by a NGS machine. The resulting data are short base [1] sequences which have been generated (or read) by the machine from a random part of *any* haplotype present in the sample.

[1] The bases are A, C, T and G for the four nucleotides which are the basic building blocks of DNA.

These sequences are called *reads*. Recent advances in deep sequencing technologies, for example *454/Roche* or *Illumina/Solexa* have made it possible to generate vast amounts of reads while reducing sequencing costs. However, the limited lengths of reads and their non-negligible sequencing errors pose new statistical challenges as will be explained in Section 3.2.

From a statistical point of view, identifying haplotypes is a clustering problem: matching the reads to unknown haplotypes actually means classifying the reads with respect to unknown cluster centroids. This, however, is a *non-standard* clustering problem since the non-overlapping nature of the reads precludes an *a priori* definition of a suitable similarity measure for the entire read dataspace.

Outline of the chapter. In Section 3.3, an introduction to the well-established mixture modelling framework is given. We present our haplotype inference model in Section 3.4. Experimental results based on simulated and real clinical data are discussed in Section 3.5. Finally, the chapter concludes in Section 3.6.

3.2 Computational Approaches to Haplotype Reconstruction

For a complete review of the problem of haplotype reconstruction and various previous combinatorial and probabilitistic approaches designed to address the problem, refer Chapter 2.

To ensure clarity, here is a quick recapitulation of the problem we have at hand. Starting with a sample containing a mixture of HIV-1 haplotypes, we want to infer the genetic diversity of the sample i.e.

1. The number of different haplotypes in the sample.
2. The respective frequencies of these different haplotypes.
3. The DNA sequence of each haplotype.

The data we use for inference are the reads generated by a NGS machine. Each read is a short base sequence which was read from one of the haplotypes by the machine. Refer Figure 2.3.2 for an illustration of sequenced error-prone reads and their corresponding haplotypes.

The challenges faced while analysing these reads for genetic diversity estimation are that:

1. We do not know which of the haplotypes the reads were generated from.
2. The length of a read is much shorter than the haplotypes' length.
3. The reads are error-prone, thus a read string may not exactly match that part of the haplotype it was sequenced from.
4. The starting positions of the reads with respect to the haplotypes are unknown.
5. Uneven coverage of reads. Read coverage per position can be interpreted as the number of reads consisting any base at a given location on the genome.

Before we present details of our probabilistic haplotype inference model, *PredictHaplo* that is based on infinite mixture models, we briefly describe the mixture model framework in the succeeding section.

3.3 Primer to Mixture Models

Sources for this section. This section is primarily based on Yu (2006a). Others sources include Teh (2010), Frigyik et al. (2010), Ghosal (2010) and Bartle and Sherbert (2000).

3.3.1 Mixture Models

Mixture models are techniques used to model processes whose output comes from several different underlying distributions (Everitt and Hand (1981)). Given a set of n i.i.d. observations $\{x_i\}_{i=1}^n$ where $x_i = (x_{i1}, \cdots, x_{id}) \in \mathbb{R}^d$, clustering using mixture models resorts to

1. Estimating parameters θ_k for each k^{th} component distribution of the mixture. Each component represents a *cluster* or group. Observations within a cluster are deemed similar.
2. Inferring the unknown group assignment, k, of an observation.

For clustering using mixture models, one first defines the generative model, then derives the likelihood of these observations, specifies model parameters and for the Bayesian formulation, assigns prior distributions to the model parameters. Then through inference, the best model parameters are learnt.

3.3.2 Finite Mixture Model

For a Finite Mixture model (FMM), the observations can be modelled using a mixture of finite (say K) distributions.

The generative model is:

1. Select one of the K clusters with probabilities $\pi = \{\pi_1, \cdots, \pi_K\}$ where $\sum_{k=1}^K \pi_k = 1$.
2. Sample an observation x from the probability distribution of the selected cluster.
3. Repeat steps 1 and 2 n times to sample n i.i.d. observations.

This is pictorially depicted in Figure 3.3.1.

We introduce the class assignment variable, c_i which denotes the class to which the i^{th} observation belongs. The likelihood of x_i can be written as:

$$P(x_i|\pi, \Theta) = \sum_{k=1}^{K} P(c_i = k|\pi) P(x_i|\theta_k) \tag{3.1}$$

with $\Theta = \{\theta_1, \cdots, \theta_k\}$ being the parameters for all the mixture component distributions. The model parameters are π and Θ.

For a Bayesian FMM, we assign prior distributions to the parameters π and Θ, namely a Dirichlet prior over π and a conjugate-family prior over each θ_k. The hyperparameters for these distributions are α and G_0, respectively.

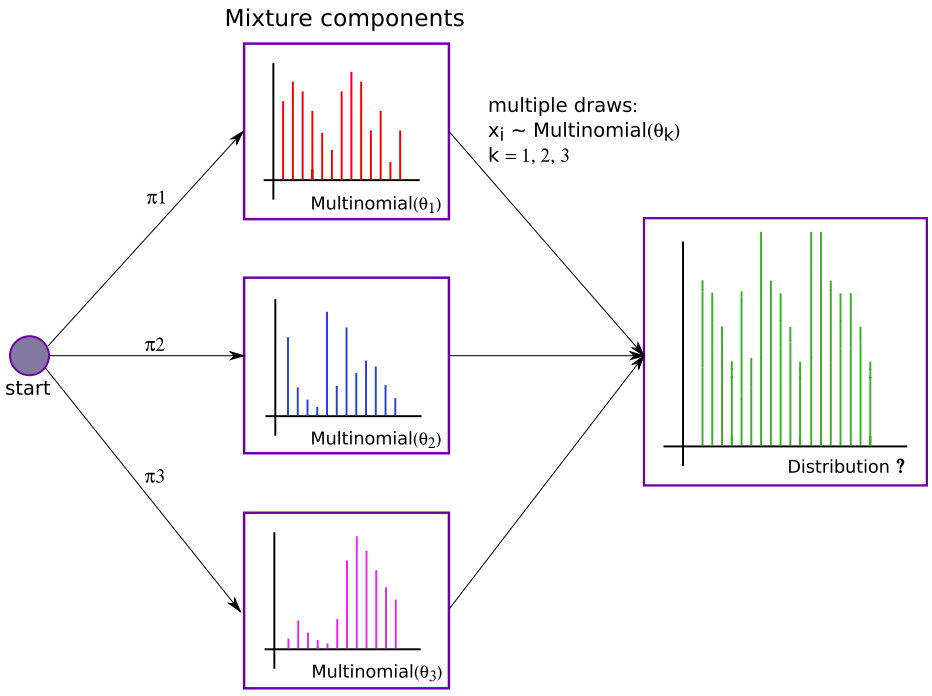


Figure 3.3.1: Generative process for a Finite Mixture model: Given are 3 multinomial distributions shown as histograms. They are chosen randomly according to π_k, $k = 1, 2, 3$ and depending on π_k, an observation is drawn from the corresponding component distribution i.e. $x_i \sim \text{Multinomial}(\theta_k)$.

The FMM is equivalent to the following distributions:

$$\begin{aligned} x_i|\Theta, (k = c_i) &\overset{\text{ind}}{\sim} P(x_i|\theta_k), \quad i = 1, \cdots, n \\ c_i|\pi &\overset{\text{iid}}{\sim} \text{Multinomial}(c_i|\pi), \quad i = 1, \cdots, n \\ \theta_k &\overset{\text{iid}}{\sim} \mathsf{G}_0, \quad k = 1, \cdots, K \\ \pi|\alpha, K &\sim \text{Dirichlet}(\pi|\tfrac{\alpha}{K}, \cdots, \tfrac{\alpha}{K}) \end{aligned} \tag{3.2}$$

The corresponding plate model for the FMM is shown in the left panel of Figure 3.3.2.

Equation 3.1 can be equivalently written as

$$\mathrm{P}(x_i|\pi, \Theta) = \int_\theta P(x_i|\theta)\mathsf{G}_K(\theta)\mathrm{d}\theta \tag{3.3}$$

where $\mathsf{G}_K(\theta) := P(\theta|\pi, \Theta) = \sum_{k=1}^K \pi_k \delta_{\theta_k}(\theta)$. Here, $\delta_{\theta_k}(\theta)$ is a point mass distribution located at θ_k taking the value 1 for $\theta = \theta_k$ and 0 otherwise. $\mathsf{G}_K(\theta)$ can be seen as a discrete prior distribution over θ meaning that θ can choose from only amongst the K values from Θ weighted by π. α and G_0 act as parameters to $\mathsf{G}_K(\theta)$.

By integrating out π and Θ, the likelihood for x_i can be written as:

$$P(x_i|\alpha, \mathsf{G}_0) = \int_{\mathsf{G}_K} P(\mathsf{G}_K|\alpha, \mathsf{G}_0) \left(\int_\theta P(x_i|\theta)\mathsf{G}_K(\theta)\mathrm{d}\theta \right) \mathrm{d}\mathsf{G}_K. \tag{3.4}$$

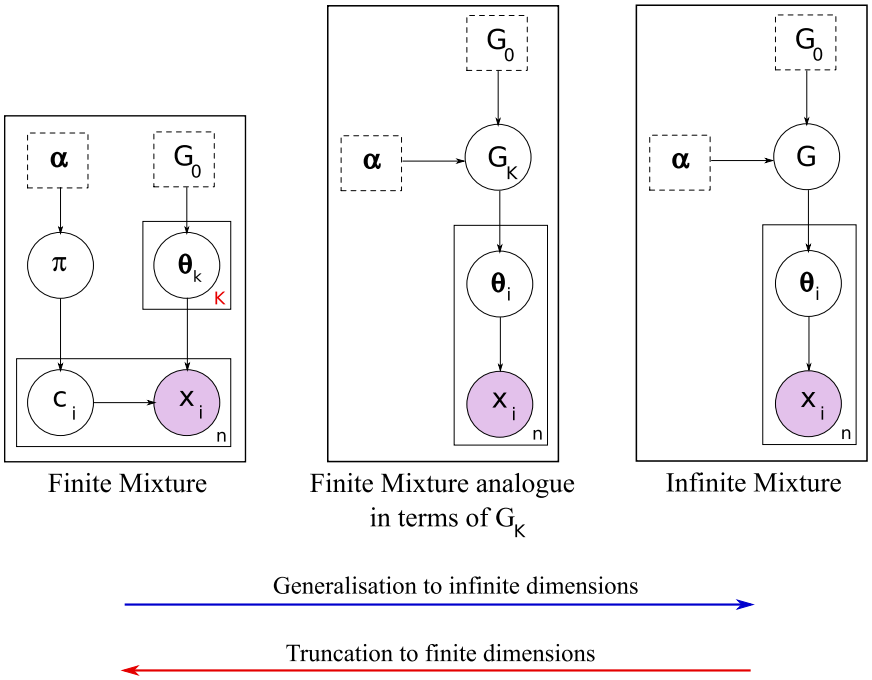


Figure 3.3.2: Plate model: Finite Mixture model with class assignments, its random-measure analogue in G_K (discrete prior over θ) and the Infinite Mixture model in $\mathsf{G} \sim DP(\mathsf{G}_0, \alpha)$. x_i is an observation, white circles denote latent variables of interest, squares indicate replications where the replicative factor is at the bottom right corner and dotted squares are predefined parameters. Figure modified from Yu (2006b).

The plate model denoting the FMM with respect to $\mathsf{G}_K(\theta)$ is given in the central panel of Figure 3.3.2 and the corresponding set of equations to sample x_i is:

$$\begin{aligned} x_i|\theta_i &\overset{\text{ind}}{\sim} P(x_i|\theta_i), \quad i = 1, \cdots, n \\ \theta_i &\overset{\text{iid}}{\sim} \mathsf{G}_K, \quad i = 1, \cdots, n \\ \mathsf{G}_K(\theta) &= \sum_{k=1}^{K} \pi_k \delta_{\theta_k}(\theta) \end{aligned} \tag{3.5}$$

3.3.3 Infinite Mixture Model

Instead of the K *a priori* fixed clusters used in the FMM, the FMM can be generalised to the infinite case i.e. $K \to \infty$ clusters. The nonparametric mixture model this leads to is called the *infinite mixture model* (IMM). Figure 3.3.3 gives the generative stochastic process for observations x_i using the IMM. $\mathsf{G}_K(\theta) = \sum_{k=1}^{K} \pi_k \delta_{\theta_k}(\theta)$ which was a finite sum of weighted point mass functions for FMM is now extended in the IMM to an infinite sum of weighted point mass functions i.e. $\mathsf{G}(\theta) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\theta)$. When $K \to \infty$, the discrete prior G_K tends to be a realisation from a *Dirichlet Process* (DP) i.e. $\mathsf{G} \sim DP(\alpha, \mathsf{G}_0)$ where α is the concentration parameter and G_0 is the base distribution of the DP. The DP defines a distribution for random distributions (Freedman (1963), Ferguson (1973) and Ferguson (1974)).

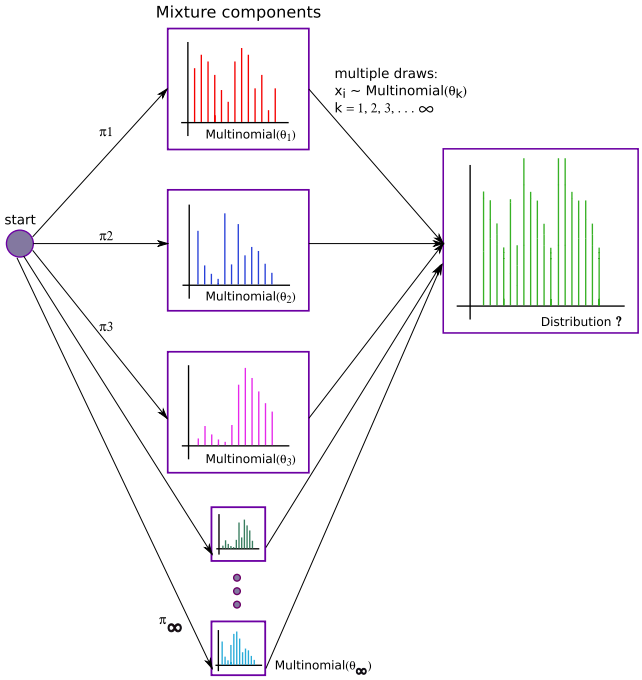


Figure 3.3.3: Generative process for an Infinite Mixture model: Consider an ∞ number of component distributions, here they are multinomial distributions shown as histograms. These are chosen according to π_k, $k = 1, \cdots, \infty$ and an observation is drawn from the corresponding component distribution i.e. $x_i \sim \text{Multinomial}(\theta_k)$.

Since the IMM uses the DP as conjugate prior over the class parameters Θ, the IMM is also known as *Dirichlet Process Mixture Model* (DPMM). A DPMM can be written as:

$$\begin{aligned} x_i|\theta_i &\overset{\text{ind}}{\sim} P(x_i|\theta_i), \quad i = 1, \cdots, n \\ \theta_i &\overset{\text{iid}}{\sim} \mathsf{G}, \quad i = 1, \cdots, n \\ \mathsf{G}(\theta) &= \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\theta) \end{aligned} \tag{3.6}$$

The likelihood for x_i in the DPMM is given as:

$$P(x_i|\alpha, \mathsf{G}_0) = \int_{\mathsf{G}} P(\mathsf{G}|\alpha, \mathsf{G}_0) \left(\int_{\theta} P(x_i|\theta) \mathsf{G}(\theta) \mathrm{d}\theta \right) \mathrm{d}\mathsf{G}. \tag{3.7}$$

and it can be seen that this is similar to the likelihood in Equation 3.4 for a FMM. The corresponding plate model is depicted in the right panel of Figure 3.3.2.

Sample generation from a DP. There are 3 different ways to generate samples from a DP. These are:

- *Chinese Restaurant Process* (Pitman (2006))
- *Polya Urn Process* (Blackwell and Macqueen (1973))

- *Stick-breaking process* (Sethuraman (1994), Ishwaran and James (2001) and Ishwaran and Zarepour (2002))

Of the three, we make use of the Chinese Restaurant Process (*CRP*) in our current haplotype inference model and therefore describe it below.

Chinese Restaurant Process. Assume an empty restaurant with endless table capacity and endless seating capacity per table. The first customer x_i arrives and chooses an empty table k (equivalent to *selecting a class* in mixture modelling framework) and orders food (equivalent to the class parameters, $\theta_k \sim \mathsf{G}_0$). Subsequent customers joining him will be limited to this table's food i.e. they all share the same parameters θ_k. Any customer has the choice to either join an already existing table k with probability $\propto n_k$ (the number of people already seated at table k) or resort to a new table altogether with probability $\propto \alpha$. α is the dispersion parameter controlling the number of newer cluster formations. A larger α leads to a larger number of clusters. Therefore, the class assignments (tables) c_i define partitions over the finite-numbered observations x_i (customers). This can be seen as a sample from a *DP* and the resulting conditional prior distribution over class assignments is called the *CRP* prior. The *CRP* prior is depicted in Figure 3.3.4 and the corresponding IMM using the *CRP* prior is:

$$
\begin{aligned}
x_i|\theta_k, k = c_i &\overset{\text{ind}}{\sim} P(x_i|\theta_k) \quad \text{(customer)} \\
\theta_k &\overset{\text{iid}}{\sim} \mathsf{G}_0 \qquad \text{(food at table } k\text{; group-level parameter)} \\
c_i|c_{-i}, \alpha \sim CRP(\alpha) &= \begin{cases} \frac{n_k}{n+\alpha-1} & \text{(for an existing class } k\text{)} \\ \frac{\alpha}{n+\alpha-1} & \text{(for a new class)} \end{cases} \qquad \text{(table assignment)}
\end{aligned}
\tag{3.8}
$$

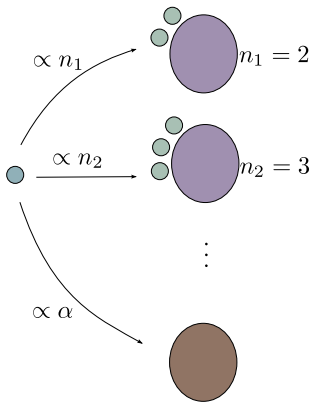


Figure 3.3.4: Chinese Restaurant Process: The conditional probability distribution for assigning the i^{th} customer to a populated table $k \propto n_k$ (number of customers at table k) or to a new table $\propto \alpha$.

Next, we present our probabilistic haplotype inference method, *PredictHaplo*, that uses a propagating DPMM for global haplotype assembly.

3.4 Haplotype Reconstruction using a propagating Dirichlet Process Mixture Model

3.4.1 The Haplotype Representation

In our model, a haplotype is represented using a set of probability tables (θ), one at every location i as in Fig. 3.4.1. This representation is designed to model the uncertainty of the inferred haplotypes. In terms of inference, our aim would then be to infer these tables over all the haplotypes' locations. Since the haplotype reconstruction is being carried out for a mixed sample of haplotypes, the problem involves finding the set of unknown haplotypes which can best explain the sequencing reads. Our probabilistic approach assumes that these reads are sampled from a Bayesian multinomial mixture model where each mixture component represents a haplotype. The mixing proportions of the different components then provide an estimate of the corresponding haplotypes' frequencies. To account for the uncertainty in the number of haplotypes we use an infinite mixture model which does not require to *a priori* fix this number. In the infinite mixture formulation, a Dirichlet process prior is used in place of the standard conjugate Dirichlet distribution prior. For optimising the computational efficiency, we implement a truncated approximation of this Dirichlet process in the Markov Chain Monte Carlo sampling scheme used for inferring the haplotypes.

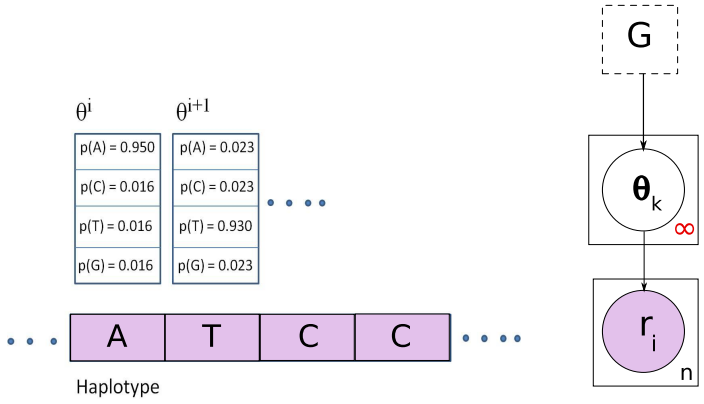


Figure 3.4.1: Left: A haplotype represented as a set of position-wise multinomial probability tables θ. **Right**: Assumed stochastic generative process for reads. The i^{th} read, r_i is generated from the k^{th} haplotype, $\theta_k = (\theta_k^1, \cdots, \theta_k^L)$, $k = 1, \cdots, \infty$. The prior distribution over θ is the Dirichlet Process, G.

3.4.2 Likelihood and Prior

The n reads of the data set are denoted by $r_1, \ldots, r_n$. Each read r_j has L components $r_j = \left(r_j^1, \ldots, r_j^L\right)$ corresponding to the positions $loc_j = \left(l_j^1, \ldots, l_j^L\right)$ of that read. Every component r_j^i is a categorical vector of length 4 specifying the base found at position i for read j: $r_j^i = (r_{j1}^i, \ldots, r_{j4}^i)$,

where only one of the 4 values is 1 and the others are equal to 0. To simplify the notation we denote by L the length of every read but in practise the reads have different lengths. We align the reads using the *Burrows-Wheeler Aligner* (BWA) (Li and Durbin (2010)).

The reads are modelled as i.i.d. samples of a Dirichlet process mixture model (DPMM) and the density is given by:

$$P\left(r_j|\alpha, \mathsf{G}_0\right) = \int_{\mathsf{G}} P(\mathsf{G}|\alpha, \mathsf{G}_0) \left(\int_{\theta_j} P(r_j|\theta_j)\mathsf{G}(\theta)\mathrm{d}\theta \right) \mathrm{d}\mathsf{G}, \tag{3.9}$$

where θ_j represents the probability table for read j, G is the DP prior placed over θ_j and $\mathsf{G} \sim DP(\alpha, \mathsf{G}_0)$ where α denotes the concentration parameter and G_0 is the base distribution (Ferguson (1973)). Figure 3.4.1 shows the haplotype representation as a set of probability tables θ and Figure 3.4.2 gives the stochastic generative process modelling reads.

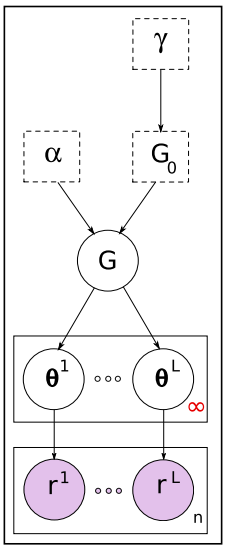


Figure 3.4.2: Detailed plate model for read generation using DPMM: Consider n observed reads r of length L. They can possibly arise from one of the infinitely-many haplotypes θ of length L. The prior distribution of haplotype (or class) assignment is according to the *Polya Urn Scheme* or *Chinese Restaurant Process* (see Section 3.3.3). For the i^{th} read r_i, draw its haplotype parameters θ_i from already-seen values $\{\theta_1, \cdots, \theta_{i-1}\}$ i.e. $\theta_i = \theta_k$ with a probability $\propto n_k$, the number of reads belonging to haplotype k or a new value from the base distribution G_0 with probability $\propto \alpha$.

Since we assume independence between the locations, the mixture components are modelled using a product of independent multinomial distributions:

$$r_j|\theta_j \sim \prod_{i \in loc_j} \text{Multinomial}\left(r_j^i|\theta_j^i\right), \tag{3.10}$$

where θ_j^i represents the probability table entries for read j at location i. To obtain conjugacy in the model, G_0 is chosen as a product of independent Dirichlet distributions. The latent variables of the observations' class assignments are denoted by c_j where $j = 1, \ldots, n$. If we suppose that classes $k = 1, \ldots, K$ are already populated by $n_1, \ldots, n_K$ reads then, using the *CRP* sampling scheme

for DPMM (see Section 3.3.3), the conditional prior distribution for the class assignment has the following form:

1. Probability of assignment to an already populated class k is: $P(c_j = k|c_{-j}, \alpha) \propto \frac{n_k}{n+\alpha+1}$.
2. Probability of assignment to a new class is: $P(c_j \neq 1, \ldots, K|c_{-j}, \alpha) \propto \frac{\alpha}{n+\alpha+1}$.

3.4.3 Including Prior Information from previous local analyses

As was explained in Section 3.2, to be able to extend our model to solve the global reconstruction problem we need to address the issue of missing direct coupling between non-overlapping reads. To introduce indirect relationships between these reads we use prior information extracted from local reconstructions.

Truncated DPMM

A computational challenge in implementing the DPMM is in handling the infinite mixture (Ishwaran and James (2001), Gelfand and Kottas (2002)). The principled workaround to this is to truncate the DP to approximate the full process by choosing a sufficiently-large positive integer K such that $\mathsf{G} = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\theta) \approx \sum_{k=1}^{K} \pi_k \delta_{\theta_k}(\theta)$ and $\pi_k = 0 \ \forall k > K$ (Blei and Jordan (2004), Ohlssen et al. (2007)). Thus, by choosing a large K, G becomes a finite sum of weighted point-mass funtions and using such a truncated DP leads to the truncated DPMM which acts as a good approximation to the DPMM.

Hence, for inference we implement a truncated DPMM as shown in Fig. 3.4.3. A positive integer K large enough for our problem is chosen to be the maximum number of clusters allowed in the model. This approximation which considerably improves the running time enables our model to handle larger datasets of several hundred thousands of reads. A practical strategy to set the value for K, is to check if the DPMM returns K-fully populated clusters. If this is the case, one significantly raises the value for K. On the other hand, if the DPMM returns many empty clusters, it is an indication that the bound K was set high enough. Another point to be relied upon while setting the value for K is that there are frequency detection limits imposed by the sequencing error rate. For a given coverage, say we assume that all the haplotypes are of frequencies below 0.1%. Then we expect to detect at most $K = 1000$ different haplotypes which, with the current available methods, would be unrealistic to reliably detect.

Since the truncated DPMM is formally identical to a finite mixture model (Section 3.3.2), G can be rewritten as a finite sum of point-mass measures with weights $\pi = (\pi_1, \ldots, \pi_K)$ and the distribution of the reads is now:

$$P\left(r_j|\pi, \phi\right) = \sum_{k=1}^{K} \pi_k P\left(r_j|\phi_k\right), \tag{3.11}$$

where $\phi = (\phi_1, \ldots, \phi_K)$ and ϕ_k relates to the k^{th} haplotype's probability tables. We again denote by $c_j, j = 1, \ldots, n$ the variables taking values in $\{1, \ldots, K\}$ which represent the cluster assignments for every read. The conditional distribution of a read given its cluster assignments is then given

analoguously to Equation 3.10:

$$r_j|c_j, \phi \sim \prod_{i \in loc_j} \text{Multinomial}\left(r_j^i|\phi_{c_j}^i\right), \tag{3.12}$$

and the variables c_j are distributed according to:

$$c_j|\pi \sim \text{Multinomial}\left(c_j|\pi\right). \tag{3.13}$$

To complete the model specification we still need to give the prior distributions of ϕ and π. The prior distribution of the parameters ϕ_k^i for every location i is a Dirichlet of order 4:

$$\phi_k^i|\gamma \sim \text{Dirichlet}(\tfrac{\gamma}{4}, \ldots, \tfrac{\gamma}{4}), \quad k = 1, \ldots, K, \tag{3.14}$$

where $\gamma \in \mathbb{R}$ and is divided by the number of possible nucleotides (A, C, T, G). The prior distribution of the mixture proportions π for our truncated DPMM with K components is a Dirichlet of order K:

$$\pi|\alpha \sim \text{Dirichlet}(\tfrac{\alpha}{K}, \ldots, \tfrac{\alpha}{K}), \tag{3.15}$$

with $\alpha \in \mathbb{R}$.

Inference – Gibbs sampling

Inference is obtained via Gibbs sampling based on the Chinese Restaurant Process (*CRP*) (*CRP* is discussed in Section 3.3.3). The Gibbs sampler is a Markov chain Monte Carlo (MCMC) method used for sampling from the posterior probability distribution over variables given the probabilistic model (Geman and Geman (1984), Casella and George (1992)). This involves iterations where one draws variables of interest from the corresponding conditional posterior distributions alternately, keeping the others fixed, while repeatedly sweeping through all observations. The variables of interest are c_j, ϕ and π. Iterative Gibbs sampling results in a sequence of samples for these variables that form a Markov chain which converge to samples from the joint distribution $P(\{c_j\}_{j=1}^n, \{\phi_k\}_{i=1}^K, \pi|\{r_j\}_{j=1}^n)$.

The conditional posterior distributions required by the Gibbs sampler based on the *CRP* are:

1. For the latent class assignment variable c_j:

$$\begin{aligned} P(c_j|\boldsymbol{c}_{-j}, \{r_j\}_{j=1}^n, \{\phi_{c_j}\}_{c_j=1}^K, \gamma, \pi, \alpha) &= P(c_j|r_j, \{\phi_{c_j}\}_{c_j=1}^K, \pi) \\ &\propto \underbrace{P(c_j|\{\phi_{c_j}\}_{c_j=1}^K, \pi)}_{\text{prior distribution}} \underbrace{P(r_j|c_j, \{\phi_{c_j}\}_{c_j=1}^K, \pi)}_{\text{likelihood}} \\ &\propto P(c_j|\pi)P(r_j|\phi_{c_j}) \end{aligned} \tag{3.16}$$

   where the right hand side of the equation is known and the normalisation constant can be determined by simply adding over the possible values for c_j.

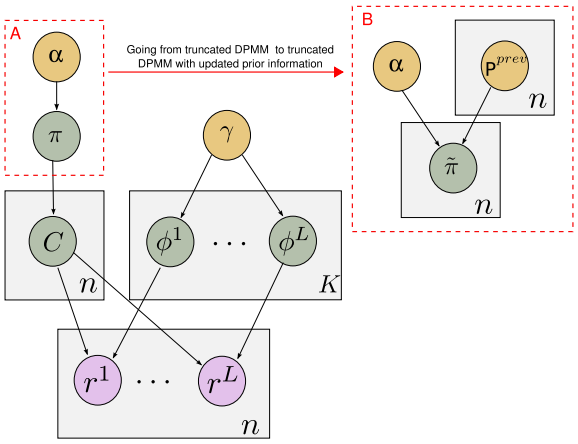


Figure 3.4.3: Plate model for the truncated Dirichlet prior mixture model (A). When using updated prior information the mixing proportions π are replaced by $\tilde{\pi}$ (B). We can see that when information from the previous window is available in the form of P^{prev} the class assignment probabilities $\tilde{\pi}$ are differentiated for every read.

2. For the k^{th}-haplotype's probability table, ϕ_k:

$$\begin{aligned} P(\phi_k^i|\{r_j\}_{j=1}^n, c, \alpha, \pi, \gamma) &= P(\phi_k^i|\{r_j^i\}_{\forall j:c_j=k}, \gamma) \\ &\propto \underbrace{P(\phi_k^i|\gamma)}_{Dirichlet\ \text{prior}} \underbrace{\prod_{j:c_j=k} P(r_j^i|\phi_k^i)}_{\text{product of } Multinomials} \qquad \text{(using Bayes' Theorem)} \\ &= \text{Dirichlet}\left(\frac{\gamma}{4} + \sum_{j:c_j=k} r_{j1}^i, \ldots, \frac{\gamma}{4} + \sum_{j:c_j=k} r_{j4}^i\right) \qquad \text{(due to conjugacy)} \end{aligned} \tag{3.17}$$

3. For the component proportions π, using Bayes' theorem and conjugacy of the Dirichlet prior:

$$\begin{aligned} P(\pi|c, \{r_j\}_{i=1}^n, \{\phi_k\}_{k=1}^K, \alpha) &= P(\pi|c, \alpha) \\ &\propto p(\pi|\alpha)p(c|\pi) \\ &= \text{Dirichlet}(\frac{\alpha}{K} + \sum_{j=1}^n \delta_1(c_j), \ldots, \frac{\alpha}{K} + \sum_{j=1}^n \delta_K(c_j)) \end{aligned} \tag{3.18}$$

where $\delta_k(c_j)$ is the Kronecker delta.

Truncated DPMM with updated prior information

We split the global problem into a sequel of several reconstruction tasks of increasing difficulty. We start with a local reconstruction initiated at the region of maximum coverage [2] (see Fig. 3.4.4). Then we progressively increase the window currently analysed until that window covers the entire

[2] The coverage at location i is the number of reads covering location i.

haplotypes' length. For every window, we perform Gibbs sampling and obtain a clustering of the reads which fall in that window. Using this clustering we can update the prior information about clusters probabilities which will then be an input for the subsequent window. Fig. 3.4.5 depicts our model's workflow.

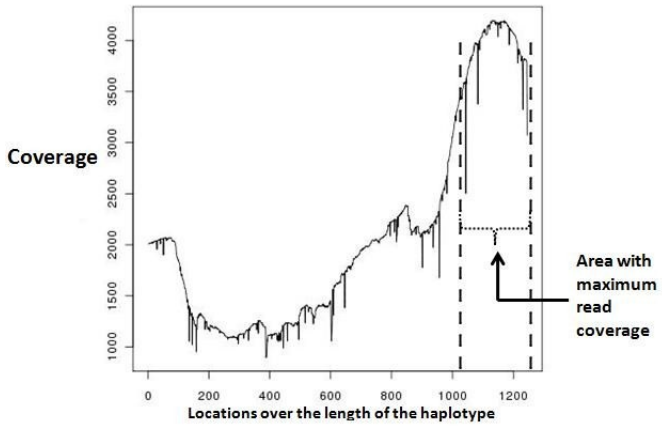


Figure 3.4.4: Uneven coverage landscape of non-PCR real data. Coverage is the number of reads covered per haplotype position.

After every window analysis, we update the prior of the class assignment probabilities (π) of the reads. The key idea is that we allow this prior to be different for every read. As a consequence, the class assignment probabilities also become different for every read and we use $\tilde{\pi}_j = (\tilde{\pi}_{j1}, \ldots, \tilde{\pi}_{jK})$ for $j = 1, \ldots, n$ (see Fig. 3.4.3). The reads are now modelled as independent samples of several mixture distributions sharing a common parameter ϕ:

$$P\left(r_j|\tilde{\pi}_j, \phi\right) = \sum_{k=1}^{K} \tilde{\pi}_{jk} P\left(r_j|\phi_k\right), \quad j = 1, \ldots, n, \tag{3.19}$$

and Equation 3.15 is replaced by:

$$\tilde{\pi}_j|\alpha, P_j^{prev} \overset{\text{ind}}{\sim} \text{Dirichlet}(\alpha P_{j1}^{prev}, \ldots, \alpha P_{jK}^{prev}), \quad j = 1, \ldots, n, \tag{3.20}$$

where $P_j^{prev} = (P_{j1}^{prev}, \ldots, P_{jK}^{prev})$. Here P_{jk}^{prev} is the posterior probability of the event $\{c_j^{prev} = k\}$ given $r, \phi^{prev}, \tilde{\pi}^{prev}$ obtained during the sampling performed for the previous window. It gives a measure of how likely the assignment of read j was in the previous window. We finally obtain a clustering of all the reads where each read j is assigned to a haplotype represented by its estimated probability table ϕ_{c_j}. The respective proportions of the haplotypes can be simply estimated by counting the number of reads assigned to each cluster and dividing it by the total number of reads.

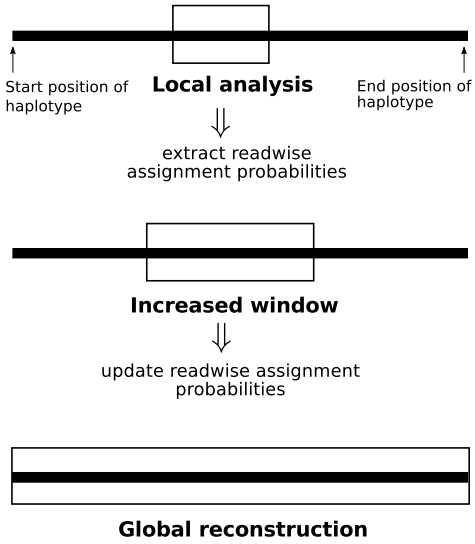


Figure 3.4.5: Model workflow for global haplotype reconstruction: Local analysis for haplotype reconstruction is initially performed on a smaller section of the entire sequence. Here the reads have significant overlap to enable clustering. The cluster centroids form intermediate haplotypes. The locally-available class assignment probability of every read is extracted and aids as prior information to the DPMM for an enhanced assembly of the intermediate haplotypes in subsequent larger sections of the sequence. This process of updating the prior information continues over the entire length of the sequence thus reconstructing the whole haplotype.

3.5 Results

We tested our model on both simulated and real sequencing reads. Simulations were done on the *gagpol* region of 4306 bases and real data focussed on the *pol* gene region of 1245 bases. Choosing a longer region for simulations was done to verify our model's stability while inferring haplotypes over longer genomic stretches. Further, the *gagpol* region is known to be medically relevant for drug resistance. Our software *PredictHaplo* implementing the model is available as Open-source software from `http://bmda.cs.unibas.ch/HivHaploTyper/`.

3.5.1 Simulated Reads

Simulation setup

Ten haplotypes were simulated as mutants from the HIV-1 reference genome with proportions ranging from 50% to 0.1% following a geometric-decay series and thereafter reads were simulated from these haplotypes. For each haplotype the mutation sites were drawn independently using a mutation probability per position of 0.5% with respect to the reference genome. The new base substituting the original one was then drawn with each of the three possible replacements having equal probability. All haplotypes are simulated as mutants from the *gagpol* region starting from 790 till 5096 which is 4306 bases long. We used MetaSim (Richter et al. (2008)) to generate 200,000 reads from these haplotypes. MetaSim explicitly models the light intensity emitted by the *454/Roche* sequencing machines along with the possible resulting sequencing errors. The error model used in Metasim was the 454 error model and the parameters were chosen as recommended in Richter et al. (2008).

We generated MetaSim-simulated reads for two different read lengths: 340 and 700 bases. These two read lengths correspond to the average read length of the real data we analysed (see section 3.5.2 and Zagordi et al. (2010a)) and to the typical read length for the *454/Roche* GS FLX Titanium XL+ sequencing machine, respectively. These different lengths also capture the evolution of *454/Roche* sequencing machines.

Another set of MetaSim-simulated reads was generated for the same two read lengths with haplotypes having a mutation probability of 1.5% with respect to the reference genome.

Performance

To compare the performance between different MetaSim-simulated reads, 15 simulation runs each were carried out on a particular combination of read length and mutation probability. Figure 3.5.1 shows the comparative performance between read lengths of 340 and 700 and mutation probabilities of 0.5% and 1.5%. Each boxplot denotes the F-scores over these 15 runs. For read length of 340 bases with a mutation probability of 0.5% i.e. an average of 1% diversity between the haplotypes, all the haplotypes present with proportions higher than 12.5% could be reconstructed and exactly matched the original sequences of bases. For the same mutation probability but longer reads of 700 bases, all haplotypes above 3.125% were reconstructed without errors. The former setting detected more false positives (on average 3) whereas in the latter only 1–2 haplotypes were signalled as false positives. For read length of 340 bases with a higher mutation probability of 1.5%, haplotypes above 1.6% were reconstructed without errors whereas for longer reads of 700 bases, haplotypes above 0.8% were reconstructed error-free. The number of false positives for shorter reads was between 1–2 whereas for longer reads, there were 0–1 false positives. We can conclude that longer read lengths together with higher mutation rates show an improved performance since both factors contribute in bridging the gaps between identical regions in different haplotypes.

In general, these simulations depict that haplotype reconstruction is a harder problem when the read lengths are shorter. This also illustrates the benefits with longer reads as brought about by the evolution in *454/Roche* sequencing machines. Further increasing the reconstruction window, when with shorter reads, reconstructs only a fewer number of haplotypes with more false positives. This can be attributed to the fact that as the window spans, shorter reads would not constitute adequate overlapping positions necessary to identify the appropriate haplotypes. This leads to forfeiting potential haplotypes in the inference process and also increasing the number of mismatches in the resulting inferred haplotypes. Thus longer read lengths (700 bases or more) as provided by the latest *454/Roche* GS FLX Titanium XL+ lead to better results. Further, for a fixed read length increased diversity renders a higher number of true positives as the diversity aids in distinguishing true haplotypes at the local-window levels itself.

Comparison with previous methods

We compare our method with *ShoRAH* (which also implements the read-graph approach of Eriksson et al. (2008)) and *QuRe*. *ViSpA* was not included in the comparison for reasons explained further below.

An initial comparison experiment was based on the sets of simulated reads described in Section

3.5.1, however none of these competing softwares was able to handle the 200,000 simulated reads. Therefore, this lead us to conduct comparisons on a smaller scale, reproducing the charasteristics of our real data sets, see Section 3.5.2. We simulated 10,000 reads of mean length 340 bases, again using the MetaSim 454 error model. The reads were simulated from 10 different haplotypes having a mutation probability of 1.5% with respect to the reference genome, and the haplotypes constitute the same decreasing proportions as described in Section 3.5.1. Since reconstructing a region of 4306 bases long is not possible with only 10,000 reads, the haplotypes now considered have a decreased length of 1321, corresponding to *Reverse Transcriptase* (RT) region of the *pol* gene. Figure 3.5.2 depicts the number of correctly reconstructed haplotypes and the number of false positives as a function of the number of mismatches tolerated between the reconstructed haplotypes and the ground truth. We can see that *PredictHaplo* perfectly reconstructs 6 of the 10 haplotypes. This performance can also be attained by *ShoRAH* if we allow a maximum of 4 mismatches to the ground truth. *ShoRAH*, however, suffers from a high number of false positives. *QuRe* performed poorly on this simulated data set reconstructing only 1 haplotype with less than 5 mismatches, which could be attributed to the specificities of the inbuilt MetaSim 454 error model that introduces a lot of insertions (representing approximately 80% of the simulated errors) within simulated reads. Insertions are treated differently in *QuRe* than in *PredictHaplo* and *ShoRAH*. Therefore, to ensure a fair comparison, we present in Fig. 3.5.2 results for *QuRe* obtained with reads previously corrected using *ShoRAH* as was recommended for *ViSpA* in Astrovskaya et al. (2011). We see that *QuRe* can reconstruct up to 3 haplotypes for a mismatch tolerance of 2 with the number of false positives lower than *ShoRAH* but considerably higher than *PredictHaplo*. This performance of *QuRe* is in accordance with how it performs on real reads (discussed in Section 3.5.2).

We did not include *ViSpA* in the comparison because it suffered from instability problems on simulated reads, preventing us from reproducing results and presenting definite conclusions. Memory leak issues appeared for runs with more than 20,000 reads whilst the coverage obtained with 10,000 reads seemed insufficient to recover haplotypes from error-prone reads. *ViSpA* results were extremely sensitive to the parameters used, delivering from 1 to 430 haplotypes without being able to reconstruct more than 1 haplotype that always corresponded to the most frequent one. Even using *ShoRAH*-corrected reads did not improve the results. The only setup for which we obtained an improvement of reconstructing 4 haplotypes exactly and 2 haplotypes with less than 2 mismatches along with 10 false positives, was when using MetaSim simulated error-free reads. Problems using *ViSpA* were previously also reported in Schirmer et al. (2012).

Significance of read length for haplotype reconstruction

To test the importance of read length addressing the haplotype reconstruction problem, we run simulations using different read lengths ranging from 36 bp to 350 bps. This range emulates the Illumina/Solexa read length of 36 bps to 454/Roche (GS FLX Titanium) read length of 450 bps. From the haplotypes reconstructed using different read lengths, we deduce that read length is a significant factor to answer the *global* haplotype reconstruction problem and also a criterion when it comes to choosing between sequencing platforms, i.e. one platform generating smaller number of

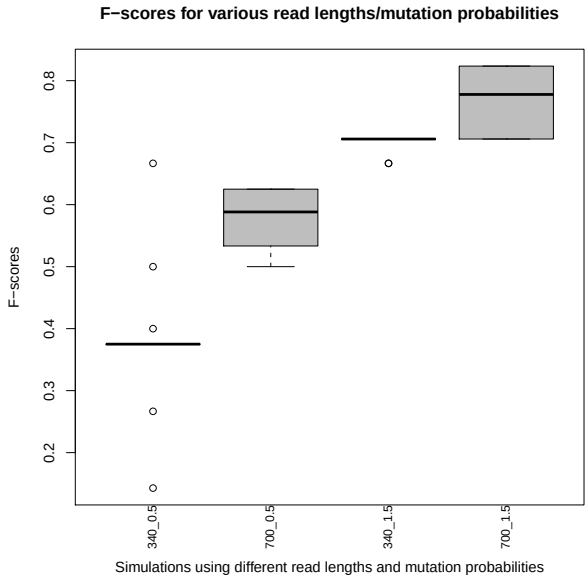


Figure 3.5.1: Boxplots denoting F-scores over 15 runs each for different combinations of read lengths (340 and 700 bases) and mutation probabilites (0.5% and 1.5%). Longer reads, irrespective of the mutation probabilities, turns in a higher number of true positives with lesser false positives. The results obtained for longer reads portray the benefits brought about by the latest *454/Roche* technology. Increased diversity also aids in inferring more number of true haplotypes.

longer reads (454/Roche) versus another with larger number of shorter reads (Illumina/Solexa). If the goal is large-scale global reconstruction of highly abundant haplotypes, we recommend using longer reads from 454/Roche, whereas if it is local reconstruction of scarcer haplotypes, one should opt for Illumina/Solexa by virtue of their deeper coverage. This further reinforces the results of Zagordi et al. (2012a). Figure 3.5.3 compares the significance of platform-dependent read lengths specifically addressing the global haplotype reconstruction problem.

3.5.2 Real Reads

Sequencing data description

A genetically diverse sample was prepared by mixing 10 haplotype clones of length 1245, in known proportions ranging from 38.3% to 0.02% (refer Table 3.5.1 for actual proportions and see Zagordi et al. (2010a)). These 10 clones were previously isolated from the plasma of HIV infected patients and the clones consist of the protease and a part of the reverse transcriptase portion of the *pol* gene. One aliquot of this sample underwent *polymerase chain reaction* (PCR) amplification to access the viability of utilising amplified samples for haplotype reconstruction. Both samples were sequenced using a *454/Roche* sequencing machine with an average read length of 340 bases. The PCR dataset contained 25,716 reads whereas the non-PCR dataset had 10,174 reads.

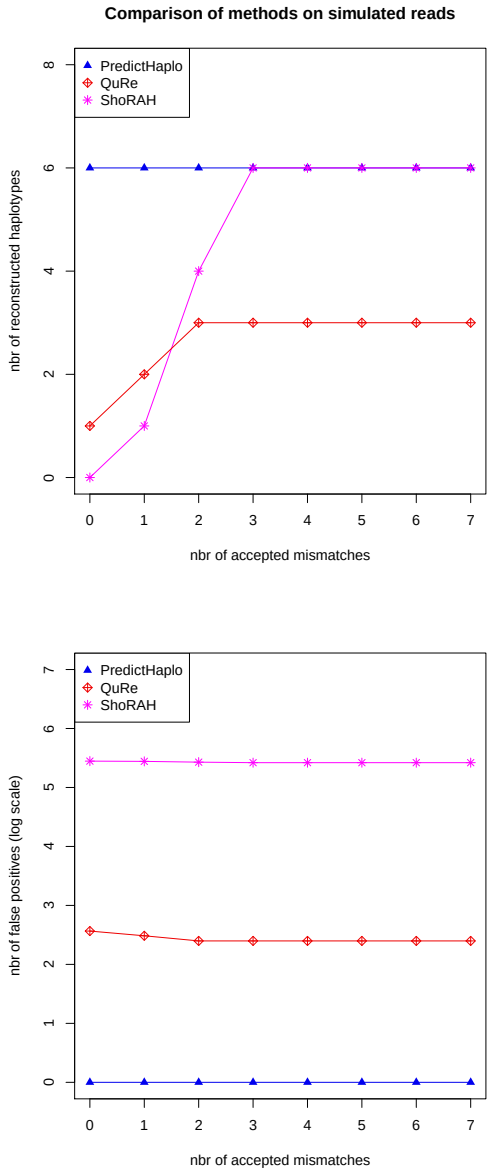


Figure 3.5.2: Top: Number of correctly reconstructed haplotypes. Bottom: number of false positives on a log scale (the numbers shown are exactly $\log(fp + 1)$ where fp is the number of false positives) as a function of the tolerated number of mismatches for simulated reads.

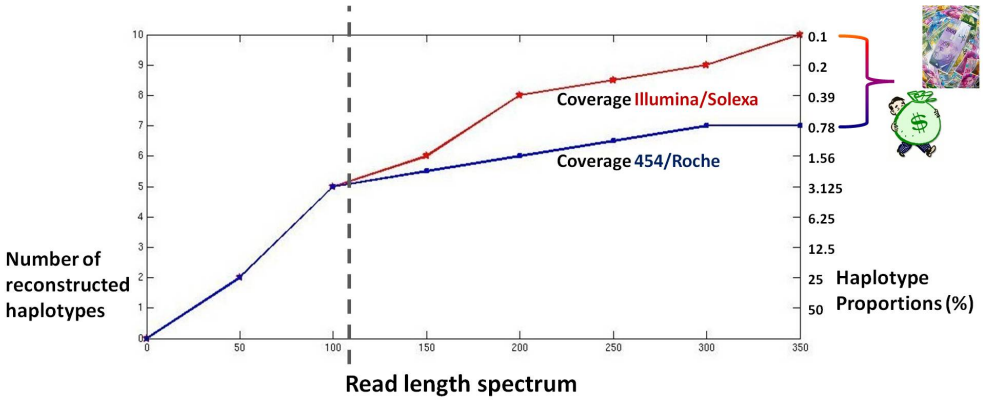


Figure 3.5.3: Number of reconstructed haplotypes (left axis) and the corresponding haplotype frequencies (right axis) against read length for simulated data. Higher the read length, more are the number of globally-reconstructed haplotypes.

Results on real reads

The true haplotypes are in general either correctly reconstructed (meaning that the inferred bases matched exactly with the true ones at every location) or not detected by the model. The inferred proportions are very close to the true values as can be seen in Table 3.5.1. For the PCR amplified reads the model was able to correctly reconstruct all the haplotypes present with proportions between 38.3% and 5.6% and for non-PCR amplified reads between 29.3% and 6.2%.

Comparison with previous methods

We compare our experimental results from *PredictHaplo* to those obtained from *ShoRAH*, *QuRe* and *ViSpA*. Multiple runs for each of these methods are performed by varying the available parameters to obtain precision and recall values. For *ShoRAH*, the Dirichlet process rate is varied that controls the number of reconstructed haplotypes. With *ViSpA*, reads are first corrected with *ShoRAH*, as recommended in Astrovskaya et al. (2011). We then varied the number of mismatches allowed to cluster reads around super reads and the mutation-based range to obtain different numbers of reconstructed haplotypes. For *QuRe*, we changed both the homopolymeric and non-homopolymeric error rates.

The upper plot of Fig. 3.5.4 shows the best F-score obtained for different methods on the non-PCR reads [3] whereas the bottom plot depicts the highest precision and recall values as well as the best F-score (harmonic mean of the precision and recall) obtained for these different methods on PCR amplified reads. The reader should note here that the highest precision value coincided with its best F-score value. All values are given for a tolerated number of mismatches of less than 5 with respect to the ground truth.

From Fig. 3.5.4, it is evident that all methods perform equally well in terms of recall since at most 5 to 6 haplotypes could be reconstructed for conducive parameter settings. However, *PredictHaplo* is

[3]Best F-score, highest precision and highest recall values coincided for all models except for *ViSpA*, which attained a maximal recall of 0.6.

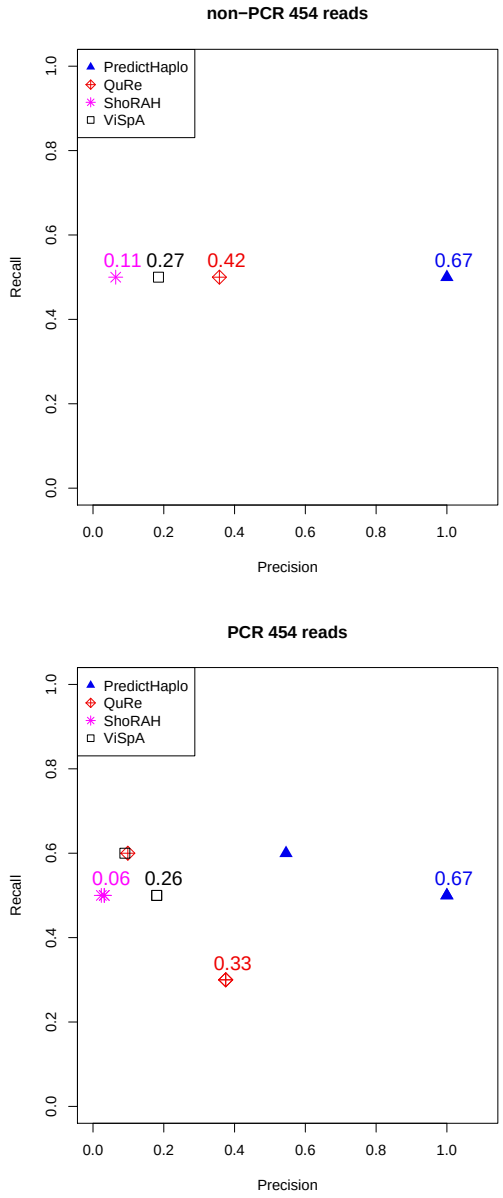


Figure 3.5.4: Plots showing best F-score / highest precision and recall values obtained for *PredictHaplo, QuRe, ShoRAH* and *ViSpA* using 454 sequencing reads. For each method, several sets of parameters were varied (see Section 3.5.2) leading to corresponding precision-recall points from which the highest precision and recall values were chosen. The best F-score (harmonic mean of precision and recall) for each method is also shown. Top: Best F-score values shown for the 4 methods on non-PCR data. Bottom: Highest precision and recall along with best F-score values plotted for the 4 methods on PCR-amplified data. Note here that the highest precision values coincide with the corresponding best F-scores. It can be seen from both the plots that *PredictHaplo* clearly stands out in terms of better precision and thereby has a higher F-score as compared to the previous methods.

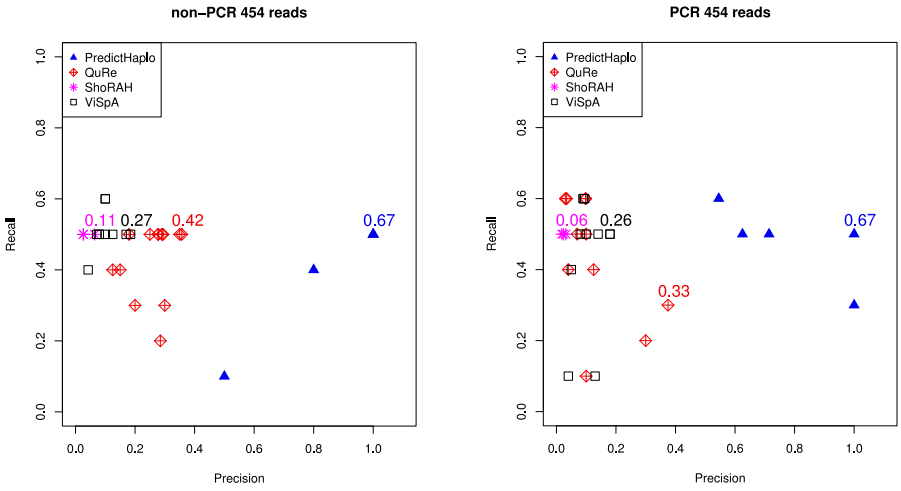


Figure 3.5.5: Plots showing precision and recall values obtained for *PredictHaplo, QuRe, ShoRAH* and *ViSpA* using 454 sequencing reads. For each method, several sets of parameters were varied (see Section 4.2.3) leading to corresponding precision-recall points. The number of obtained points can vary between the methods since different parameter values can still lead to the same performance. For each method we also show the value of the best F-score attained. **Top:** Values shown for the 4 methods on non-PCR data. **Bottom:** Values plotted for the 4 methods on PCR-amplified data. It can be seen from both the plots that *PredictHaplo* clearly stands out in terms of better precision.

significantly ahead of its competitors in terms of offering better precision values. We run *PredictHaplo* with values for the Dirichlet prior over probability tables, γ, between 0.01 and 50 and the Dirichlet prior over the mixtures, α, between 0.1 and 10 (although changing α did not influence the model's performance significantly). For appropriate parameter settings our method can find 5 haplotypes for non-PCR and PCR reads, without detecting any false positives. By tuning the only available parameter for *ShoRAH*, the Dirichlet process rate, over a large range (1e-5 to 1e3), false positives could not be reduced below 73 (present with frequencies between 0.02% and 3.1%) and 152 (between 0.01% and 6.4%) on the non-PCR and PCR data respectively. A more detailed version of Fig. 3.5.4 containing all points obtained for the different sets of parameters is given in Figure 3.5.5.

These experimental results confirm that our probabilistic model performs remarkably better than Astrovskaya et al. (2011), Prosperi et al. (2011) and the four-step approach detailed in Eriksson et al. (2008) and Zagordi et al. (2009).

Table 3.5.1: Actual and reconstructed haplotypes proportions obtained using *PredictHaplo* for non-PCR and PCR *454/Roche* reads. All values are in %. X denotes 'undetected haplotype'.

Non-PCR	Actual	29.3	28.6	22.1	10.4	6.2	2.3	0.7	0.26	0.16	0.04
	Reconstructed	32.1	26.0	23.7	10.5	7.7	X	X	X	X	X
PCR	Actual	38.3	35.4	10.1	9.5	5.6	0.46	0.32	0.08	0.06	0.02
	Reconstructed	37.6	34.9	11.8	10.1	5.7	X	X	X	X	X

Table 3.5.2: Links to softwares used in comparison experiments.

PredictHaplo	http://bmda.cs.unibas.ch/HivHaploTyper/
ShoRAH	http://www.bsse.ethz.ch/cbg/software/shorah
QuRe	https://sourceforge.net/projects/qure/
ViSpA	http://alla.cs.gsu.edu/~software/VISPA/vispa.html

3.5.3 Datasets used and links to competing softwares

To ensure reproducibility of results presented in Sections 3.5.1 and 3.5.2, we provide the links to datasets used as well as softwares used for comparison purposes. Data used for comparison experiments based on simulated data (in Section 3.5.1) can be found at http://bmda.cs.unibas.ch/HivHaploTyper/ and that used for comparison experiments based on real reads (in Section 3.5.2) is published at https://wiki-bsse.ethz.ch/display/ShoRAH/Data under *454 Data* section.

Links to the various softwares deployed in this chapter are provided in Table 3.5.2.

3.6 Conclusions

This research deals with analysing the different haplotypes obtained from deep-sequencing data. With the advent of these powerful sequencing technologies, there is a plenitude of reads being curated but their limited lengths and machine-induced errors pose challenges in identifying the exact haplotypes' diversity within the sample. Being able to identify the genetic diversity is an important step in administering personalised medication.

The main modelling challenge here arises due to non-overlapping reads not having any suitable *a priori* similarity measure defined between them. None of the previous approaches have provided a convincing strategy to solve this issue. In this work we successfully overcome this problem by introducing a propagating DPMM. Our model does not *a priori* need to know the number of genetically-diverse HIV haplotypes present in the sample. A Gibbs sampler is used for inferring the unknown haplotypes from the error-prone reads. Through this the full posterior distribution of model parameters is inferred and by using Gibbs sampling we eliminate potential local minima issues arising with EM. Other added values of this Bayesian model are that it is computationally efficient and requires only a few input parameters. From our results based on simulated reads we can see that the model's performance is stable under simulations conducted with varying diversities. Experiments with real data also confirmed the model's performance.

Facet II

Computational methods for structure recovery in antiretroviral drug space

Goals in this part of the thesis:

- Graphical models – An introduction.
- TiWnet – a Bayesian method for network structure recovery using distance data.
- Automatic Archetype Analysis.

Introduction to Facet 2

HIV is highly virulent spawning mutants that escape drugs over time, thereby becoming a menace difficult to eradicate. Preventive measures like ART drugs must be powerful enough to avert aggressive HIV proliferation that eventually lead to the AIDS infection. The ART drugs prescribed are a concoction from amongst 25 commercialised drugs falling in 5 categories as discussed in Section 1.2.2. When HIV becomes drug-resistant to a drug of a particular category, it in turn becomes resistant to other drugs within the same category, giving rise to what is known as *cross-resistance*. This effectively reduces the number of possible drug combinations available for prescription. It is therefore important to analyse for similarities between ART drugs and other available chemically-active compounds for an effective ART.

The 2^{nd} facet of the thesis looks into extracting similarities between drugs and chemical compounds, based on their chemical structures. This is done by examining a landscape of active chemical compounds, also encompassing the drugs:

1. for extracting networks amongst the active chemical compounds. This helps in understanding relations between chemical compounds and drugs. The fully-probabilistic model, TiWnet, is developed and deployed for this purpose. This is discussed in Chapter 5.

2. for identifying archetypal compounds. A Group-Lasso based approach along with an efficient model selection criterion to identify archetypes is developed. The method is applied to extract archetypes from a set of chemical compounds including drugs. From the resulting archetypal drugs, one can draw deeper insights into the functional similarities (for example, the cross-resistant nature of HIV strains towards drugs) that can possibly be shared within convex sets of archetypal drugs and compounds. This is further discussed in Chapter 6.

4

Graphical Models

Graphical models constitute a well-studied research field and have been used in many application domains including natural language processing, analysis of biological networks, speech recognition and image processing. In this chapter, we introduce the basic concepts used in graphical models before discussing our contribution to this field. This chapter follows the content provided in Lauritzen (1996) and Whittaker (1990).

4.1 Introduction

A network or graph is a blueprint deciphering the connectedness between a set of objects. The study of graphs is called *graph theory* where one learns the structure of the graph (or network). The structure represents the pairwise relationships between objects in the graph. A graph $G = (V, E)$ is composed of a vertex set V and an edge set $E \subseteq V \times V$. The elements of V are called *vertices* and the number of vertices or *vertex cardinality* is denoted by $|V|$. The elements of E are called *edges* of the graph and the size of the graph is $|E|$. A graph is *directed* if all the pairwise edges have directional information and it is *undirected*, if the directional information is absent. The focus of the current chapter and Chapter 5 is of discovering structures of the underlying *undirected* graph given a set of objects.

Undirected graphs have amassed prolific interest in recent years due to its intuitive mechanism of representing and visualising complex connectedness between objects. More specifically, they provide a rigid formalism to represent high-dimensional distributions of random variables (objects). Given a $n \times d$-dimensional random matrix X with n objects and d i.i.d. measurements (observations), an *undirected graphical model* for X is the family of probability density functions that represent the

conditional dependencies amongst these n objects. Fitting such a graphical model to X is called *graphical modelling*.

Next, the definitions of (conditional) independence that play a central role in graphical modelling are presented.

Consider a random vector $Y = (Y_1, \cdots, Y_n)^t$ with Y_i being a continuous random variable for $i = 1, \cdots, n$.

Definition 4.1 (Independence) *Two continuous random variables Y_i and Y_j with marginal densities $f(y_i)$ and $f(y_j)$ respectively are* independent*, denoted as $Y_i \perp\!\!\!\perp Y_j$, if and only if their joint probability density function (pdf) factorises as a product of their marginal densities i.e. $f(y_i, y_j) = f(y_i)f(y_j)$ $\forall (y_i, y_j)$.*

Definition 4.2 (Conditional Independence) *Two continuous random vectors Y_i and Y_j are* conditionally independent *on Y_k, denoted as $Y_i \perp\!\!\!\perp Y_j | Y_k$, if and only if the conditional pdf $f(y_i, y_j|y_k) = f(y_i|y_k)f(y_j|y_k)$ $\forall (y_i, y_j, y_k)$ satisfying $f(y_k) > 0$.*

Definition 4.3 (Markov Chain) *A (discrete time)* Markov Chain *with discrete state space $Y_n \in \{0, 1, 2, \ldots\}$ is a sequence of random variables $Y_0, Y_1, Y_2, \ldots$ such that for all states $i_{n+1}, i_n, i_{n-1}, \cdots, i_0$ and all discrete time points $n = 0, 1, 2, \ldots$, the Markov property is satisfied i.e.*

$$f(y_{n+1}|y_n, y_{n-1}, \ldots, y_0) = f(y_{n+1}|y_n).$$

From Definition 4.3, it can be said that the future observation Y_{n+1} is conditionally independent of past observations $\{Y_{\setminus n}\}$ [1] given the current observation Y_n i.e. $(Y_{n+1} \perp\!\!\!\perp Y_{n-1}, \ldots, Y_0 | Y_n)$.

4.1.1 Relation between network structure estimation & inverse covariance matrix and conditional independence of a corresponding probability distribution

To make the relation between the probability distribution explaining the network structure and its inverse covariance matrix clearer, let us assume 1) a graph $G = (V, E)$ with $|V| = n$ and 2) a n-dimensional random vector $Y = (Y_1, \cdots, Y_n)^t$, $Y \in \mathbb{R}^n$ and Y_i is a continuous random variable having a marginal density $f(y_i)$ for $i = 1, \cdots, n$. Each Y_i forms the i^{th} node of G.

Consider the joint probability distribution $P(Y)$ on the random variables Y_i. The conditional independence in a Markov chain (Definition 4.3) can be written reflecting Y_i in G. These Markov properties (or conditional independence properties) that P might have with respect to G are:

1. **Definition 4.4** *(Pairwise Markov property)*

   For any given non-adjacent pair of vertices (Y_i, Y_j) where $Y_i \neq Y_j$ and $(i, j) = 1, \cdots, n$, Y_i is conditionally independent of Y_j given the rest of the variables i.e. $Y_i \perp\!\!\!\perp Y_j | Y_{V \setminus \{i,j\}}$.

2. **Definition 4.5** *(Local Markov property)*

   For any given vertex $Y_i \in V$, Y_i is conditionally independent of the rest given its neighbours $Y_{ne(i)}$ i.e. $Y_i \perp\!\!\!\perp Y_{V \setminus (ne(i) \cup \{i\})} | Y_{ne(i)}$.

[1] $\{Y_{\setminus n}\}$ denotes all past observations $Y_0, \cdots, Y_{n-1}$ excluding the current observation Y_n.

3. **Definition 4.6** *(Global Markov property)*

For any triple (Y_A, Y_B, Y_C) of disjoint subsets of V such that Y_C separates Y_A from Y_B in G, subsets Y_A and Y_B are conditionally independent given Y_C i.e. $Y_A \perp\!\!\!\perp Y_B | Y_C$.

Proposition 4.1 *(Lauritzen (1996) Proposition 3.4)*

For any undirected graph G and any probability distribution $P(Y)$, if $P(Y)$ satisfies the global Markov property with respect to G then it also satisfies the pairwise Markov property of G.

Definition 4.7 *(Factorisation property)*

A joint probability distribution $P(Y)$ is said to possess the factorisation *property with respect to a given undirected graph G if it can be written as the product of non-negative functions $\phi_r(Y_r)$ such that $P(Y) = \prod_{r \subseteq \mathbb{C}} \phi_r(Y_r)$ where $\mathbb{C}$ is the set containing fully-connected subgraphs or* cliques *of G and Y_r is the set of nodes in clique r.*

Proposition 4.2 *(Lauritzen (1996) Proposition 3.8)*

For any undirected graph G and any probability distribution $P(Y)$, if $P(Y)$ satisfies the factorisation property with respect to G then it also satisfies the pairwise Markov property of G.

The converse of Proposition 4.2 does not hold for all distributions except for strictly positive distributions $P(Y)$ i.e. $P(Y) > 0$. The necessary and sufficient conditions under which a strictly positive probability distribution has its pairwise Markov property equivalent to its factorisation property is given by the Hammersley-Clifford theorem (Shen, 2011).

Theorem 4.3 *(Hammersley and Clifford) (Lauritzen (1996) Theorem 3.9)*

For any undirected graph G and a probability distribution $P(Y)$ with respect to G, $P(Y)$ satisfies the pairwise Markov property with respect to G if and only if it factorises according to G.

According to Proposition 4.2, if $P(Y)$ factorises according to G then it also satisfies the pairwise Markov property. The converse needs to be shown i.e. if $P(Y)$ satisfies the pairwise Markov property then it factorises according to that particular G. This proof can be found in Lauritzen (1996).

Let Y follow a multivariate Gaussian distribution with mean ζ and covariance matrix Σ. $\Sigma^{-1} = \Psi$ is the inverse covariance matrix of the distribution $P(Y)$. Given these, the corollary to Theorem 4.3 can be stated as follows:

Corollary 4.4 *The zeroes in Ψ of the multivariate Gaussian distribution of Y correspond to missing edges in the network G.*

Proposition 4.5 *(Lauritzen (1996) Proposition 5.2)*

Assume $Y \sim \mathcal{N}_n(\zeta, \Sigma)$ where ζ is the mean vector and Σ is invertible. For $(i, j) = 1, \cdots, n$ and $i \neq j$, it holds that the pairwise Markov property viz. $Y_i \perp\!\!\!\perp Y_j | Y_{V \setminus \{i,j\}} \iff \psi_{ij} = 0$ where $\Psi = \{\psi_{ij}\} = \Sigma^{-1}$.

Thus conditional independence in the multivariate Gaussian distribution is captured in Ψ as zero entries. Lauritzen (1996) uses the pairwise Markov property between two vertices Y_i and Y_j

to further show the relationship between the inverse covariance matrix Ψ and the elements of the partial correlation matrix as follows:

$$\rho_{ij|V\setminus\{i,j\}} = -\frac{\psi_{ij}}{\sqrt{\psi_{ii}\psi_{jj}}} \quad \forall i \neq j \tag{4.1}$$

where $\rho_{ij|V\setminus\{i,j\}}$ is the partial correlation coefficient between variables Y_i and Y_j given the rest of the variables $Y_{V\setminus\{i,j\}}$. Since Y is multivariate Gaussian distributed, if Y_i and Y_j satisfy the pairwise Markov property, then $\rho_{ij|V\setminus\{i,j\}} = 0$. Therefore $\rho_{ij|V\setminus\{i,j\}} = 0$ is synonymous to $\psi_{ij} = 0$ and conditional independence can be asserted between nodes Y_i and Y_j.

As seen in Equation 4.1, the partial correlations measure the strength of pairwise direct interactions only and since Ψ contains scaled partial correlations, identifying zeroes either in the pairwise partial correlations or Ψ forms the crux to network structure recovery. This lays the basis for a procedure called *covariance selection* introduced by Dempster (1972) where graph structure recovery is made possible by estimating Ψ of the underlying Gaussian distribution. Then it suffices to read out the zero-entry indices from Ψ and construct a graph where the corresponding indices have missing edges. Since the underlying distribution considered is multivariate Gaussian, covariance selection models are also called *Gaussian graphical models* (GGMs) (Lauritzen (1996)).

4.2 Challenges related to structure recovery

Identifying networks – estimating dependencies between objects and thereby determining their underlying graph structure – is a challenging problem. In the simplest case where the number of measurements is greater than the number of objects ($d \gg n$), the standard estimation of partial correlations involves either the inversion of the sample covariance matrix, or the estimation of n least squares regression problems. The problem is more pronounced in high-dimensional settings i.e. when the number of objects n is far larger than the measurements d themselves ($n \gg d$) and then the sample covariance matrix becomes non-invertible (see Dykstra (1970), Stifanelli et al. (2011)). Having to learn the unknown network structure from noisy observed measurements further aggravates the structure recovery problem. Another statistical challenge one faces is that the number of possible undirected networks is exponential in the number of objects n (Erds and Rényi (1960), Stifanelli et al. (2011)). Example works for structure recovery based on solving n regularised neighbourhood regression problems was dealt with in Meinhausen and Bühlmann (2006). Based on the nonnegative garrote (Breiman (1995)) and Lasso (Tibshirani (1996)) for the linear regression, Yuan and Lin (2007) introduced a regularised estimation of the Ψ using a ℓ_1-type penalty on the entries of Ψ when maximizing the multivariate Gaussian log-likelihood. The ℓ_1 norm forces certain entries of the estimated Ψ to be exactly zero. Similar sparsity-enforcing techniques on the entries of Ψ have been dealt with in Banerjee et al. (2008), Friedman et al. (2007) and d'Aspremont et al. (2008).

Apart from dealing with high-dimensional data, another problem-inflicting aspect to traditional network inference models is that they depend on geometric translations of the data which require knowledge of the underlying geometric coordinates. In many real-world scenarios, especially those dealing with non-vectorial objects like strings, graphs etc, one rarely has access to the objects' underlying vectorial representations but only to their pairwise distances implying that the geometric

translations are entirely lost. Therefore, it becomes pertinent to devise a network inference procedure that looks from the angle of pairwise distances, hence being devoid of any vectorial representations of the objects. This forms the goal of the next chapter.

A novel sparse network inference mechanism designed to work solely with *pairwise distances* of the data X is introduced where X is a $n \times d$-dimensional random matrix with n objects and d i.i.d. measurements (observations). To deal with network structure recovery in high-dimensional settings, the construction of *module networks* using the pairwise distance representation is also described.

4.3 Graphical abstract

For clarity, a graphical abstract (Figure 4.3.1) is provided that captures the purview of network inference. The top panel shows the classical operational regime for GGMs that uses the vectorial representation of an object for network recovery. These vectors are present in the observed $X_{n\times d}$ matrix where n is the number of objects and d the measurements. The bottom panel sketches the regime of our work which deals with the non-vectorial representations of objects. These objects can be those having a structure like graphs, strings, probability distributions etc. For such objects, it is natural to look into their pairwise representations and therefore for network recovery, their pairwise representations assembled in a pairwise distance matrix $D_{n\times n}$ is made use of. The model is detailed in the subsequent chapter.

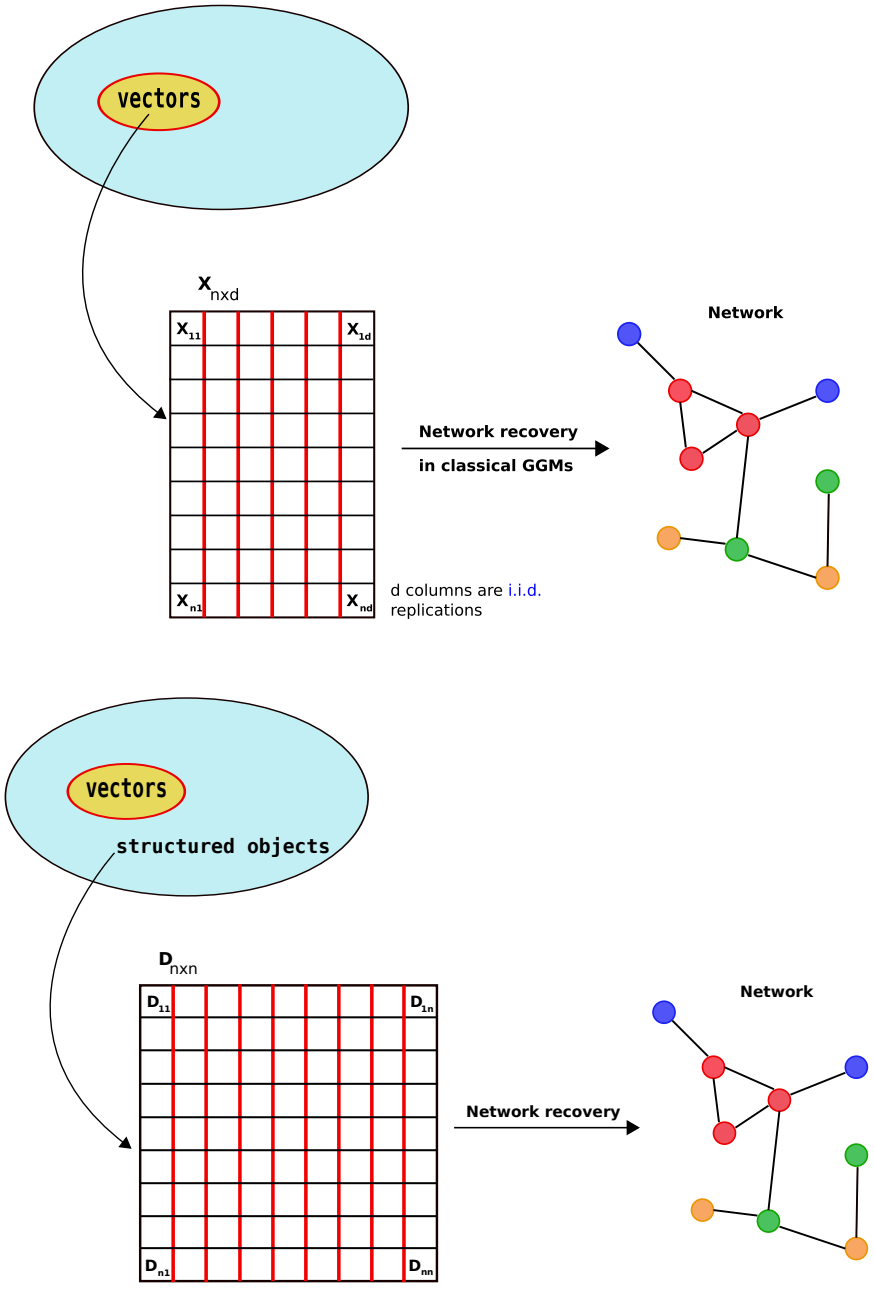


Figure 4.3.1: Graphical abstract. Consider the space of objects having a vectorial or non-vectorial representation. (**Top**) Classical GGMs operate in a vectorial regime where networks are extracted from objects represented as vectors in an observed $X_{n\times d}$ matrix with n objects of interest and d observations. (**Bottom**) Current focus of this work deals with objects possessing a non-vectorial representation i.e. these objects have a structure like a string or graph. For such objects, it is natural to consider their pairwise representations rather than vectorial representations. To enable network extraction for such structured objects, their pairwise representations collected in a pairwise distance matrix $D_{n\times n}$ is made use of.

5

Recovering Networks from Distance Data

5.1 Introduction

In the current chapter, we introduce a novel sparse network inference mechanism called the *Translation-invariant Wishart Network* (TiWnet) model that is designed solely to work on pairwise distances. This applicability to situations in which we can only observe distance information constitutes the strength of this new model over similar approaches involving the matrix-valued Gaussian likelihood (Allen and Tibshirani, 2010). We denote by $D_{n\times n}$, the matrix that contains the pairwise distances between n objects. To the best of our knowledge this is the first work that deals with network structure discovery in situations where no vectorial representation of objects is available and only pairwise distances are observed. Additionally, the presence of certain objects having a relatively higher confluence of edges gives rise to central *hub* regions. Extracting the network structure from amongst hubs given noisy measurements makes it, in general, difficult to summarise the entire network succinctly. To handle this, we present the construction of *module networks* where networks are learned on groups of variables called *modules*, thereby effectively reducing n to the number of modules.

Outline of the chapter. In Section 5.2, we explain the classical setting for GGMs. The underlying problems with existing methods are elaborated in Section 5.3. In Section 5.4, we discuss the solution to these problems and further explain how our model, TiWnet, caters to this solution. Section 5.5 details the TiWnet network inference model. We describe module networks in Section 5.6. Comparison experiments on simulated data along with three real-world application areas are demonstrated in Section 5.7. In Section 5.8, we discuss TiWD (Vogt et al., 2010) that uses the same

likelihood as TiWnet and TiWD's incapability to extract networks. The contributions of TiWnet are highlighted in Section 5.9 and in Section 5.10, we conclude the chapter.

5.2 Classical GGMs

To set the stage, we begin with a description of the classical framework for estimating sparse GGMs. One usually starts with a $n \times d$ observed data matrix X^o (the superscript ^{o} means "original" and is used here only for notational consistency), its d columns interpreted as the outcome of a measuring procedure in which some property of the n objects of interest is measured. In a biological setting, for instance, the objects could be n genes and one set of measurements (one column) could be gene expression values from one microarray. All d columns in X^o are assumed to be i.i.d. according to $\mathcal{N}(\mathbf{0}, \Sigma)$. Then, the inner product matrix $S^o = \frac{1}{d} X^o (X^o)^t$ follows a central Wishart distribution $\mathcal{W}_d(\Sigma)$ in d degrees of freedom [1] (Muirhead, 1982) (if $d \geq n$ otherwise S^o is pseudo-Wishart [2]), and its likelihood as a function of the inverse covariance $\Psi := \Sigma^{-1}$ is

$$\mathcal{L}(\Psi) \propto |\Psi|^{\frac{d}{2}} \cdot \exp\left[-\tfrac{d}{2} \operatorname{tr}(\Psi S^o)\right]. \tag{5.1}$$

The corresponding generative model is sketched in Figure 5.2.1. Every algorithm for network reconstruction relies on some potentially interesting sparsity structure garnered within the inverse covariance matrix $\Psi := \Sigma^{-1}$. Ψ contains the (scaled) partial correlations between the n random variables forming the nodes in the network: a zero entry in Ψ_{ij} concurs to no edge prevailing between the pair of random variables (i, j) in the network.

Related work. There exists a plethora of literature on network structure estimation using i.i.d. samples. To infer the underlying network, it is straightforward (at least from a methodological viewpoint) to maximise the Wishart likelihood while ensuring that Ψ is sparse. This is exactly the approach followed in Yuan and Lin (2007), Banerjee et al. (2008), d'Aspremont et al. (2008) and *graph lasso* (Friedman et al. (2007)) where a ℓ_1 sparsity constraint on Ψ is used:

$$\log \mathcal{L}(\Psi) \propto \tfrac{d}{2} \log |\Psi| - \tfrac{d}{2} \operatorname{tr}(\Psi S^o) - \lambda ||\Psi||_1 \tag{5.2}$$

where λ controls the amount of penalisation and $||\Psi||_1 = \sum_i |\Psi_i|$, the ℓ_1 norm which is the sum of absolute values of the elements in Ψ. A methodologically similar, but simplified approach that decouples this joint estimation problem into n independent neighbourhood-selection problems is dealt in Meinhausen and Bühlmann (2006). The neighbourhood selection problem is cast into a standard regression problem and is solved efficiently using a ℓ_1 penalty. The model presented in Kolar et al. (2010a) deals with conditional covariance selection where the neighbourhoods of nodes are conditioned on a random variable that holds information about the associations between nodes. They

[1] The central standard Wishart distribution is defined for $S^o = X^o (X^o)^t$. Throughout the chapter, we use $S^o = \frac{1}{d} X^o (X^o)^t$ so that d appears in the central Wishart distribution and can be later used as an annealing parameter in the inference procedure.

[2] The names of the Wishart distribution are inconsistent in the literature. We use the notation in Díaz-García et al. (1997).

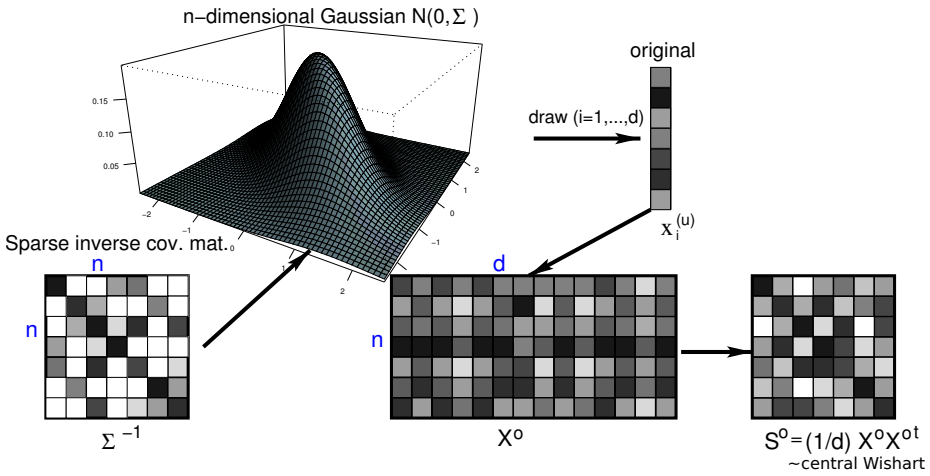


Figure 5.2.1: Assumed underlying generative process in classical GGMs. Black arrows indicate the workflow when drawing samples from this model; *n*, *d*: matrix dimensions. Every d^{th} draw from the n-dimensional Gaussian is an i.i.d. replication and stacked as a column of X^o. A draw represents a set of observations and a row denotes an object of interest.

employ a logistic regression model with a $\ell_{1,2}$ penalty for the neighbourhood-selection problem while additionally assuming this conditioning variable which steers sparsity of edges. Another method to extract networks called *walk-summable graphs* is introduced in Johnson et al. (2005b) where a neighbourhood is constructed based on *walks* accumulated by every node in the graph and weighted as a function of the edgewise partial correlations present in Ψ.

5.3 Underlying Problems with Existing Methods

The above papers and related approaches, however, have been built on an assumption that the d columns in X^o are i.i.d. This particular assumption of considering columns to be *identically* distributed might be too restrictive: even if the underlying Gaussian generative process is a valid model, different column-wise bias terms are common in practice. In the above biological example, there might be global expression differences between the d microarrays. It is therefore indispensable to model these unknown shifts (biases) for valid network inference. An ensuing consequence of modeling these biases is that the column i.i.d. assumption gets relaxed i.e. one ends up working with just independent data since the columns now come from different distributions.

Employing non-i.i.d. data for network recovery has been dealt with in the past, primarily in the area of time-varying data. Here, the data are no longer identically distributed since observations are taken at d discrete time points. In this case, the time-varying GGMs aim in capturing the longitudinal relational structure between objects. Examples of such work that deal with transient non-i.i.d. data due to discrete time points can be found in Kolar et al. (2010b), Zhou et al. (2010) and Carvalho and West (2007). In these references, it must be noted that every observation assumes to have been generated from either a common-mean discrete-distribution Ising model (Kolar et al. (2010b)) or zero-mean multivariate normal distribution (Zhou et al. (2010) and Carvalho and West (2007)). At this juncture, our work differs from this fraternity in that although we also deal with

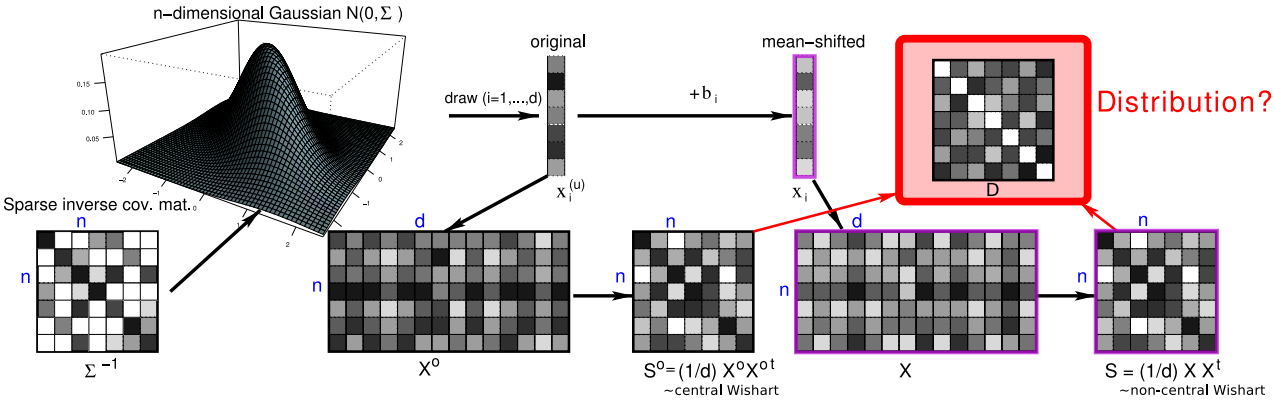


Figure 5.3.1: Assumed underlying generative process. Black arrows indicate the workflow when drawing samples from this model; *n*, *d*: matrix dimensions. The red arrows highlight the same distance matrix D produced from either the "original data" X^o (consisting of i.i.d. samples) or the "mean-shifted" data X (purple-outlined boxes).

non-i.i.d. data, the non-i.i.d. nature arises not due to the time component but due to admitting different column-wise biases.

To model these column-wise biases in TiWnet, they are included in the generative model by introducing a shifting operation in which scalar bias terms $b_{(i=1,\dots,d)}$ are added to the "original" column vectors $\boldsymbol{x}_i^o$, which results in a mean-shifted vector $\boldsymbol{x}_i$, forming the i-th column in X, cf. Figure 5.3.1 (purple-outlined boxes). Hence the columns come from *different* distributions i.e. they cease to be *identically distributed*. In the classical case of not considering column biases, X^o is distributed as $\mathcal{N}(\mathbf{0}, \Sigma)$, but in TiWnet which now accommodates these column biases, the joint distribution of all matrix elements is expressed, that here is matrix normal $X \sim \mathcal{N}(M, \Omega)$ with mean matrix $M := \mathbf{1}_n \boldsymbol{b}_d^t$ and covariance tensor $\Omega := \Sigma_{n\times n} \otimes I_d$. This model implies that $S = \frac{1}{d} X X^t$ follows a *non-central* Wishart distribution $S \sim \mathcal{W}_d(\Sigma, \Theta)$ with non-centrality matrix $\Theta := \Sigma^{-1} M M^t$ (Gupta and Nagar, 1999). Practical use of the non-central Wishart for network inference, however, is severely hampered by its complicated form and more so, the problem of estimating the unknown non-centrality matrix Θ based on only one observation of S which is problematically analogous to identifying the mean of any distribution given only a single data point.

It is, thus, desirable to use a simpler distribution. One possible way of handling such column biases is to "center" the columns by subtracting the empirical column means $\hat{b}_i$, and using the matrix $S_C = \frac{1}{d}(X - \mathbf{1}\hat{\boldsymbol{b}}^t)(X - \mathbf{1}\hat{\boldsymbol{b}}^t)^t$ in the standard central Wishart model. Since the entries in the i-th column, $\{x_{1i}, \cdots, x_{\mathrm{n}i}\}$, are not independent but coupled via the Σ-part in Ω, this centering, however, brings about undesired side effects; apart from removing the additive shift, the original columns are modified with the resulting column-centered matrix S_C being rank deficient. As a consequence, $S_C \not\sim \mathcal{W}(\Sigma)$ i.e. S_C is not central Wishart distributed. Instead, S_C follows the more complicated *translation invariant* Wishart distribution, see Equation 5.12 below.

Figure 5.3.2 exemplifies these problems where we depict the performance of *graph lasso* (Friedman et al., 2007) based on (*i*) the original unshifted data generated using Figure 5.2.1 (GL.o), (*ii*) mean-shifted data generated using Figure 5.3.1 (GL.s) and (*iii*) column-centered data (GL.C). *Graph lasso*

maximises the Wishart likelihood using a ℓ_1 sparsity constraint (see Equation 5.2) and works best in case (i) where the model assumptions are met. The boxplots in Figure 5.3.2 confirm that the presence of column-wise biases (case ii) significantly deteriorates the performance of *graph lasso* and even column-centering (case iii) does not augment the performance. Thus column-biases are not only a theoretical problem of model mismatch but also a severe practical problem for inferring the underlying network.

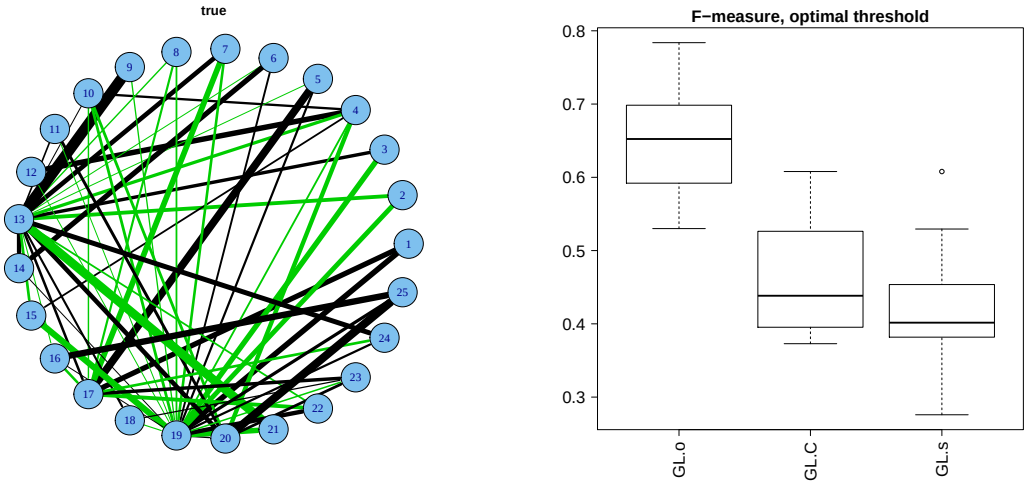


Figure 5.3.2: Left: Example network, artificially created from a data generator. **Right**: performance of edge recovery for the *graph lasso (GL)* method which maximises the standard Wishart likelihood with a ℓ_1 sparsity penalty. The leftmost boxplot refers to the original (unshifted) data (GL.o), meaning that the model assumptions are correct, the rightmost boxplot refers to data with column shifts (GL.s), and the middle boxplot refers to empirically centering the columns (GL.C). Refer section 5.7 for details on sample generation, methods, model selection and evaluation criteria.

Another problem-arising situation is where even observing $X_{n\times d}$ is not valid, instead one assumes access to a measuring procedure which directly returns pairwise relationships between n objects. Two variants are considered: either a positive definite similarity matrix identified with the matrix S is measured, or pairwise squared distances arranged in a matrix D is measured, defined componentwise as $D_{ij} = S_{ii} + S_{jj} - 2S_{ij}$. In the first case with S or in the second case with D, column-centering is still possible by the usual "centering" operation in kernel PCA (Schölkopf et al., 1998): $S_C = QSQ^t = -(1/2)QDQ^t$, with $Q_{ij} = \delta_{ij} - \frac{1}{n}$. However, using this column-centered matrix S_C in the standard Wishart model induces obviously the same problems related to model mismatch as in the vectorial case above (Figure 5.3.2).

5.4 Novel Solution to Network Inference

To overcome the above intertwined problems of having to work with column-wise biases and the complicated non-central Wishart we need to rely on a model that makes use of only pairwise distances. Figure 5.4.1 shows how one can move from $X \mapsto S \mapsto D$ and the information loss involved therein. When one moves from X to S, the rotational information is lost and when one moves from S to D, the translational information is lost. Once in D, we are devoid of any relevant geometric

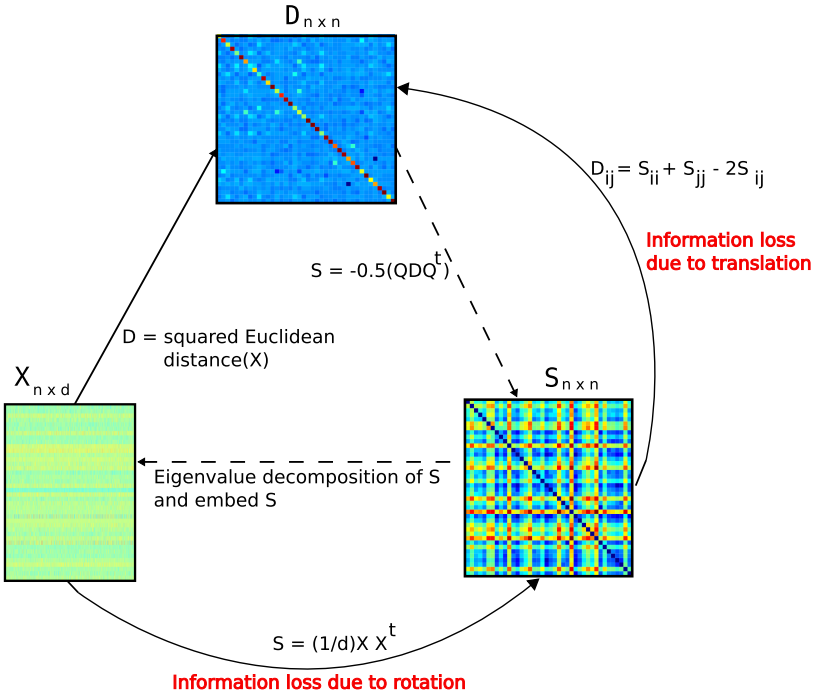


Figure 5.4.1: Relationship between data matrix X, similarity matrix S and pairwise distance matrix D and the *information loss* procured by moving between them. The straight lines from $X \mapsto S$, $X \mapsto D$ and $S \mapsto D$ show a unique mapping whereas the dotted lines from $D \mapsto S$ and $S \mapsto X$ show a non-unique mapping. Since we deal with squared-Euclidean pairwise distances throughout, the distances are preserved. It is the non-uniqueness that poses the real problem which requires attention.

information i.e. D is both translation and rotation invariant. Since we consider D to contain the squared-Euclidean pairwise distances, the distances are preserved throughout. On the other hand, the mappings from $D \mapsto S$ and $S \mapsto X$ are not unique and this non-uniqueness is the problem that requires careful handling. We explain more on this non-uniqueness and how we handle it in the following.

Since by assumption D contains squared Euclidean distances, there is a set of inner product matrices S that fulfill $D_{ij} = S_{ii} + S_{jj} - 2S_{ij}$ (McCullagh, 2009). If S_* is one (any) such matrix, the equivalence class of these matrices mapping to a single D is formally described as set $\mathbb{S}(D) = \{S | S = S_* + \mathbf{1}\boldsymbol{v}^t + \boldsymbol{v}\mathbf{1}^t, S \succeq 0, \boldsymbol{v} \in \mathbb{R}^n\}$. The elements in $\mathbb{S}(D)$ can be seen as Mercer kernels that represent many objects ranging from graphs to probability distributions to strings etc. Mercer kernels are kernels that satisfy Mercer's theorem conditions (Vapnik (1998) and Cristianini and Shawe-Taylor (2000)). These kernels are viewed as similarity measures between structured objects that have no direct vectorial representation [3]. For example, Figure 5.4.2 represents a structured object like a graph for which different Mercer kernels S_1 and S_2 can be constructed wherein $S_1, S_2 \in \mathbb{S}(D)$ and therefore map to the same D. This $\mathbb{S}$ is exactly the set of inner product matrices that can be constructed by

[3]This does not necessarily imply that it is *meaningful* to use any Mercer kernel for reconstructing a Gaussian graphical model. The main focus here is not on kernels as a means for mapping input vectors to high-dimensional feature spaces in order to exploit nonlinearity in the input space but as similarity measures.

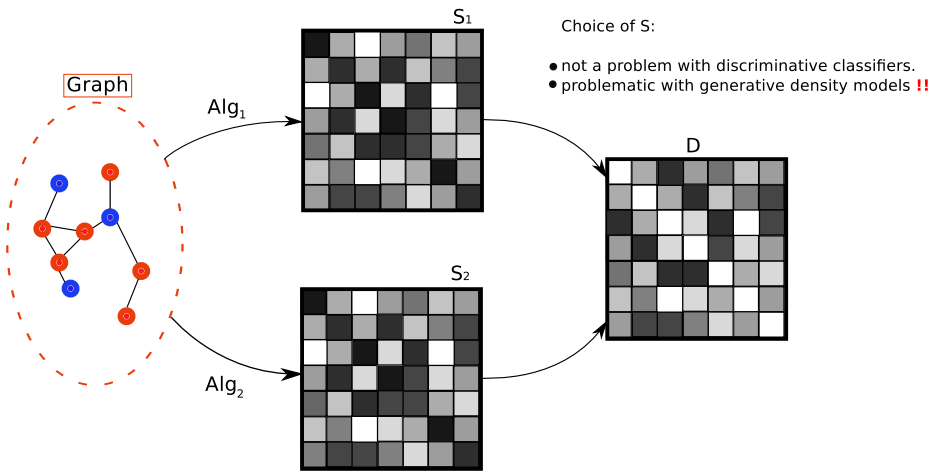


Figure 5.4.2: A structured object like a graph for which two different similarity matrices (Mercer kernels) S_1 and S_2 exist that give rise to the same D. In this case, the choice of S for usage in a probabilistic setup is irrelevant whereas if they did not map to the same D, then the choice of S is critical for the probabilistic model. In a discriminative framework, the choice of S is irrelevant.

arbitrarily biasing the column vectors in $X_{n\times d}$. Shifting the viewpoint from column to row vectors, this invariance means that the density does not depend on the origin of the coordinate system in which the n objects are represented as vectors containing d different measurements. Column-wise biases referred to before reduce in this view to simple shifts of the origin of an underlying coordinate system.

Most of the methods used for constructing kernels have no information about the origin of the kernel's underlying space meaning that we have no knowledge whether the probability distribution of either S_1 or S_2 is that of S_C i.e. the S having zero-column shifts. This indicates that as long as the kernels belong to set $\mathbb{S}(D)$, the exact form of the kernel matrix is irrelevant. On the other hand, were S_1 or $S_2 \notin \mathbb{S}(D)$, then the choice of S is critical in the framework of probabilistic models whereas for discriminative classifiers, the choice of S does not pose a problem. Most supervised kernel methods like SVMs are invariant against choosing different representatives in $\mathbb{S}$, and in common unsupervised kernel methods like kernel PCA (Schölkopf et al., 1998) the *rows* of X are considered i.i.d. implying that subtracting the empirical column means (leading to S_C) is the desired centering procedure for selecting a candidate in $\mathbb{S}(D)$. However, the sampling model for GGMs is not invariant against choosing $S \in \mathbb{S}$. If one adopts column centering, then this reduces to selecting one specific representative S_C from the set of all possible $S \in \mathbb{S}(D)$, namely the one whose origin is at the sample mean. This leads to implicitly assuming the underlying vectorial space. Such column centering, however, destroys the central Wishart property of S (assuming it was a Wishart matrix before) as discussed in Section 5.3. The strategy is therefore to avoid the selection of a representative $S \in \mathbb{S}$ altogether.

Instead, the proposed solution is to use a probabilistic model for squared Euclidean distances D. We use a likelihood model in TiWnet that depends only on D where these distances are not affected by any column-wise shifts (translations), cf. the red arrows in Figure 5.3.1. The likelihood model invariant to shifts has been studied before in the *Translational-invariant Wishart Dirichlet*

(TiWD) cluster process (Vogt et al., 2010). In Section 5.8, we discuss further the TiWD model and its unsuitability for network extraction.

5.5 The TiWnet Model

In this section, we discuss the likelihood model common to both TiWD and TiWnet, the prior construction we use suitable for network inference and the network inference mechanism.

5.5.1 Likelihood model

One starts with an observed matrix D containing pairwise squared distances between row vectors of an unobserved matrix $X \sim \mathcal{N}(M, \Omega)$. This means that in addition to the classical framework for GGMs, arbitrary column biases $b_{(i=1,\dots,d)}$ are now allowed which "shift" the columns in X but leave the pairwise distances unaffected.

As elaborated in Section 5.4 and depicted in Figure 5.4.2, there exists $\mathbb{S}(D)$, the set of kernel matrices mapping to the same D. We can now work with either D or with any $S \in \mathbb{S}(D)$ i.e. a specific S is not required. Since there exists no convenient expression for the distribution of D, the likelihood in terms of D can be computed based on the distribution of S (McCullagh, 2009). Here, it is shown that the distribution of an arbitrary $S \in \mathbb{S}$ can be derived analytically as a singular Wishart distribution with a rank-deficient covariance matrix. The likelihood is developed through the concept of marginal likelihood (Harville, 1974, Patterson and Thompson, 1971). Below, we explain the constructs for marginal likelihood and then define it in terms of D.

Marginal likelihood. The term *marginal likelihood* is not consistently used in the literature. What is sometimes called the "classical" marginal likelihood, (Harville, 1974, Patterson and Thompson, 1971), is a decomposition of the likelihood into one part which depends on the parameters of interest and a second one depending only on "nuisance" parameters. The "Bayesian" marginal likelihood, on the other hand, is computed by integrating out the nuisance parameters after placing prior distributions on them. In the following we will use the first definition, which involves a partition of the likelihood into an "interesting" part and a "nuisance" part. In some cases, this classical marginal likelihood coincides with the *profile* likelihood, which is obtained by replacing the nuisance parameters with their maximum likelihood (ML)-estimates. This interpretation indeed holds true in our case, implying that here the intuitive idea of plugging-in the ML estimates leads to a valid likelihood function (which is not always true for profile likelihoods). Further technical details on this equivalence between profile- and marginal likelihood are given in Section 5.11, and a discussion of these likelihood concepts from a Bayesian viewpoint can be found in Berger et al. (1999).

Let the data matrix X be distributed according to $p(X|\alpha, \theta)$, where the distribution is parametrised by the interest parameter α and the nuisance parameter θ. Assume there exists a statistic $t(X)$ whose distribution depends only on α. Then $p(X|\alpha, \theta)$ can be decomposed as follows:

$$\begin{aligned} p(X|\alpha,\theta) &= p(t(X), X|\alpha,\theta) \\ &= \underbrace{p(t(X)|\alpha)}_{\text{ML of interest}} \quad p(X|t(X),\alpha,\theta). \end{aligned} \tag{5.3}$$

We base our inference on $p(t(X)|\alpha)$ which is the "classical" marginal likelihood based on the interest parameter alone. We notate $p(t(X)|\alpha)$ as $\mathcal{L}(\alpha; t(X))$ where $t(X) = \frac{(X - \mathbf{1}_n \hat{\boldsymbol{b}}^t)}{||X - \mathbf{1}_n \hat{\boldsymbol{b}}^t||}$ is the standardised statistic and the interest parameter $\alpha = \Psi$. The nuisance parameters θ consist of bias estimates $\hat{\boldsymbol{b}}$ and scale factor τ. Note that this specific statistic $t(X)$ is constant on the set of all X and S matrices that map to the same D. Therefore $t(X)$ can be seen as a function that depends only on the scaled version of D i.e. $f(\frac{D}{||D||})$.

Proposition 5.1 *McCullagh (2009)*

Consider the standardised statistic $t(X) = \frac{(X - \mathbf{1}_n \hat{\boldsymbol{b}}^t)}{||X - \mathbf{1}_n \hat{\boldsymbol{b}}^t||}$ *where* $t(X)$ *is a function* $f(\frac{D}{||D||})$ *depending only on (scaled)* D*. The interest parameter is* Ψ*. The shift- and scale- invariant likelihood in terms of* D *is:*

$$\mathcal{L}(\Psi; \tfrac{D}{||D||}) \propto \det\left(\widetilde{\Psi}\right)^{\frac{d}{2}} \operatorname{tr}(-\tfrac{1}{2}\widetilde{\Psi} D)^{-\frac{(n-1)d}{2}} \tag{5.4}$$

where $\widetilde{\Psi} = f(\Psi) = \Psi - (\mathbf{1}_n^t \Psi \mathbf{1}_n)^{-1} \Psi \mathbf{1}_n \mathbf{1}_n^t \Psi$.

The proof of Proposition 5.1 is given in Section 5.11.

Thus, there is a valid probabilistic model underlying Equation 5.4, and with a suitable prior Bayesian inference for Ψ is well-defined.

The reader should notice that Equation 5.4 can be computed either from the distances D, or from *any* inner product matrix $S \in \mathbb{S}(D)$. Rather than choosing any S and implicitly fixing the underlying coordinate system, our solution is to make the distribution invariant to the choice of any S (refer Section 5.4). This is achieved by working directly with D whereby any $S \in \mathbb{S}(D)$ can be used. The practical advantage of this property is that one can now make use of the large "zoo" of Mercer kernels that represent structured objects whose vectorial representations are generally unknown. With TiWnet based on D, we make no assumption of the underlying coordinate system and can now use these Mercer kernels for reconstructing GGMs without being dependent on the choice of $S \in \mathbb{S}$.

5.5.2 Prior construction

For network inference in a Bayesian framework, we complement the likelihood (Equation 5.4) with a prior over Ψ. We develop a new prior construction that enables network inference. This prior is similar to the spike and slab model introduced in Mitchell and Beauchamp (1988). In principle, any distribution over symmetric positive definite matrices can be used. The likelihood has a simple functional form in $\widetilde{\Psi}$, but our main interest is in Ψ, since zeros in Ψ determine the topology. Unfortunately, the likelihood in Ψ is not in standard form making it plausible to use a MCMC sampler. For any two Σ matrices, Σ_1 and Σ_2 that are related by $\Sigma_2 = \Sigma_1 + \mathbf{1}\boldsymbol{v}^t + \boldsymbol{v}\mathbf{1}^t$, the likelihood is the same for Σ_1 and Σ_2 (McCullagh, 2009). This means that Ψ is non-identifiable and a sampler will have problems with such constant likelihood regions by continuing to visit them unless a prior is used that breaks this symmetry.

To deal with this problem, we quantise the space of possible Ψ-matrices such that any two candidates have different likelihood. This is achieved with a two-component prior: $P_1(\Psi)$ is uniform over the discrete set $\mathcal{A}$ of symmetric diagonally-dominant matrices with off-diagonal entries in $\{-1, +1, 0\}$, and diagonal entries are deterministic, conditioned on the off-diagonal elements

i.e. $\Psi_{ii} = \sum_{j \neq i} |\Psi_{ij}| + \epsilon$ where ϵ is a positive constant added to ensure full rank of Ψ. Thus $\mathcal{A} = \{\Psi | \Psi_{ij} \in \{-1, +1, 0\}, \Psi_{ji} = \Psi_{ij}, \Psi_{ii} = \sum_{j \neq i} |\Psi_{ij}| + \epsilon\}$. Note that we treat only the off-diagonal entries as random variables. Enforcing such a diagonally-dominant matrix construction ensures that the matrix will be positive definite. The usage of diagonally-dominant matrices for network reconstruction is further justified since these matrices form a strict subclass of GGMs that are walk summable (Johnson et al., 2005a) and in Anandkumar et al. (2011) theoretical guarantees are provided establishing that walk-summable graphs make consistent sparse structure estimation possible. It is clear that such a three-level quantisation of the prior which differentiates only between positive, negative and zero partial correlations encodes a strong prior belief about the expected range of the partial correlations. However, it is straightforward to use more quantisation levels, or even switch to continuous priors like the ones introduced in Daniels and Pourahmadi (2009), Joe (1996) which parametrise the "semi-partial" correlations. On the other hand, our simulation experiments below suggest that the simple three-level prior performs very well in terms of structure recovery.

The second component of the prior is a sparsity-inducing prior $P_2(\Psi)$. This corresponds to a Laplacian prior over the number of edges for each node and is given by $P_2(\Psi|\lambda) \propto \exp(-\lambda \sum_{i=1}^{n} (\Psi_{ii} - \epsilon))$ where $(\Psi_{ii} - \epsilon)$ denotes the number of edges for the i^{th} node and λ is equivalent to the regularisation parameter controlling the sparsity of the connecting edges.

5.5.3 Inference in TiWnet

To enable Bayesian inference in our model, we make use of the likelihood given in Equation 5.4 and the two-component prior described in Section 5.5.2. For inference we devise a Metropolis-within-Gibbs sampler where the Metropolis-Hastings step proposes an appropriate Ψ matrix by iteratively sample one row/column in the upper triangle part of Ψ, conditioning on the rest, and the Gibbs iteration involves repeating the Metropolis-Hastings step for every node.

The proposal distribution defines a symmetric random walk on the row/column vector taking values in $\{-1, +1, 0\}$ by randomly selecting one value and resampling it with identical probability to the two other possible values. After updating the i-th row/column in the upper triangle matrix and copying the values to the lower triangle, the corresponding diagonal element is imputed deterministically as $\Psi_{ii} = \sum_{j \neq i} |\Psi_{ij}| + \epsilon$. This creates $\widetilde{\Psi}_{\text{proposed}}$ which is then accepted according to the usual Metropolis-Hastings equations based on the posterior ratio $P(\widetilde{\Psi}_{\text{proposed}}|\bullet)/P(\widetilde{\Psi}_{\text{old}}|\bullet)$. The acceptance threshold is given by just the posterior ratio since we implement a symmetric random walk Metropolis sampling. The entire Metropolis-within-Gibbs sampler is described in Algorithm 1. For a pictorial representation, see Figure 5.5.1.

Since the proposal distribution, $\widetilde{\Psi}_{\text{proposed}}$, defines a symmetric random walk on set $\mathcal{A}$ consisting of diagonally-dominant matrices, one can reach any other element in $\mathcal{A}$ with non-zero probability after a sufficient number of $\frac{n(n-1)}{2}$ steps that account for number of elements in the upper triangle of Ψ. This construction ensures ergodicity in the Markov chain.

Note that the (usually unknown) degrees of freedom d in the shift- and scale-invariant likelihood (Equation 5.4) appears only in the exponents and, thus, has the formal role of an annealing parameter. In the annealing framework, the likelihood equation is seen as the energy function with d as the annealing temperature. We use this property of d during the burn-in period, where d is slowly

Algorithm 1 Metropolis-within-Gibbs sampler

in i^{th} row/column vector in upper triangle of Ψ
1: Uniformly select index k, $k \in \{1, \cdots, i-1, i+1, \cdots, n\}$
2: Resample value at Ψ_{ik} by drawing with equal probability from $\{-1, +1, 0\}$
3: Set $\Psi_{ki} = \Psi_{ik}$ and update Ψ_{ii} and Ψ_{kk} (to ensure diagonal dominance). This is $\widetilde{\Psi}_{\text{proposed}}$
4: Compute $P(\tilde{\Psi}|\bullet) \propto \mathcal{L}(\widetilde{\Psi})P_1(\Psi)P_2(\Psi)$
5: Calculate the acceptance threshold $\mathbf{a} = \min\,(1, \frac{P(\widetilde{\Psi}_{\text{proposed}}|\bullet)}{P(\widetilde{\Psi}_{\text{old}}|\bullet)})$
6. Sample $\mathbf{u} \sim \text{Unif}(0, 1)$
7: **if** $(\mathbf{u} < \mathbf{a})$ accept $\widetilde{\Psi}_{\text{proposed}}$, **else** reject.

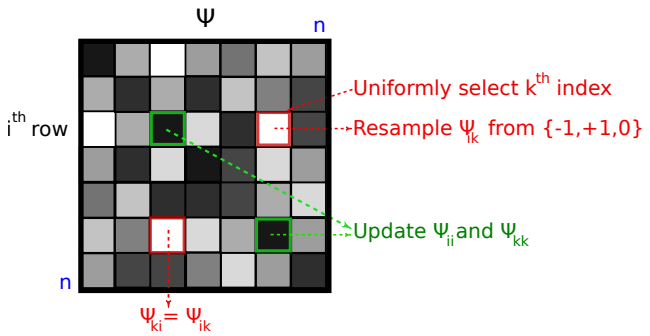


Figure 5.5.1: A Ψ proposal using the Metropolis-within-Gibbs sampler.

increased to "anneal" the system until the acceptance probability reaches below a certain threshold, and then the sampled Ψ-matrices are averaged to approximate the posterior expectation. If a truly sparse solution is desired, the annealing is continued until a network is "frozen".

Implementation & complexity analysis. Presumably the most efficient way to recompute $P(\widetilde{\Psi}|\bullet)$ after a row/column update of Ψ is through the identity: $\det(\widetilde{\Psi}) = (\det(\Psi)/\mathbf{1}^t\Psi\mathbf{1}) \cdot n$ (McCullagh, 2009). Assume now we have a QR factorisation of Ψ_{old} before the update. Then the new $\Psi = \Psi_{\text{old}} + \boldsymbol{v}_i\boldsymbol{v}_i^t + \boldsymbol{v}_j\boldsymbol{v}_j^t$ where i,j are the row/column indices of Ψ_{old} to be updated along with the corresponding diagonal elements and this accounts for two rank-one updates. Thus the QR factorisation of the new $\widetilde{\Psi}$ can also be computed in $O(n^2)$ time and $\det(\widetilde{\Psi})$ is then derived as $\prod_i R_{ii}$. The trace $\text{tr}(\widetilde{\Psi}D)$ is also computed in $O(n^2)$ time, as it is the sum of the *element-wise* products of the entries in $\widetilde{\Psi}$ and D. It is clear that this scaling behavior is prohibitive for very large matrices, but matrices of size in the hundreds can be easily handled, and for larger matrices with a "complex" inverse covariance structure the statistical significance of the reconstructed networks is questionable anyway, unless a really huge number of measurements is available. Moreover there is an elegant way of avoiding such large matrices by reconstructing *module networks* as outlined in the next section.

5.6 Inferring Module Networks

A particularly interesting property of TiWnet is its applicability to learning module networks. We define a module as a completely-connected subgraph, forming nodes in a module network. As a motivating example we refer to our gene-expression example of $X_{n\times d}$ where the measurements con-

sist of d microarrays for n genes. In usual situations having far more objects than measurements, one should not be too optimistic to reconstruct a meaningful network, in particular if the measurements are noisy and if the underlying network has "hubs"– nodes with high degrees. Generally when the node neighbourhoods are small, networks can be learned well whereas when the neighbourhoods tend to grow larger as in the case with hubs, learning networks gets difficult due to the higher-order dependencies existing between nodes. Unfortunately, both high noise and existence of hubs are common in such data. To address these issues, we present the computationally-attractive method of initially creating clusters of objects, that we connote as modules, over which networks are learned. Considering the gene-expression example, there are usually groups of genes which have highly correlated expressions and can often be jointly represented by one cluster without losing too much relevant information, due to high noise. To create clusters, we begin with the d-dimensional expression profile vectors, $\boldsymbol{x} \in \mathbb{R}^d$, of the n genes and use a mixture model to cluster these expression vectors into "modules", reducing n to the effective number of modules. The mixture model density is given by $p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k p_k(\boldsymbol{x})$ where π_k is the mixing coefficient and $p_k(\boldsymbol{x})$ is the component distribution of the k^{th} module. Partition matrices can be viewed as block-diagonal covariances (see McCullagh and Yang (2008), Vogt et al. (2010)), and in the terminology of GGMs the blocks define independent subgraphs with completely connected nodes, which is what we have defined as modules.

The link to learn networks on top of these modules goes via kernels defined on probability distributions. We can use kernels like *Bhattacharyya kernel* (Jebara et al., 2004):

$$K_B(k,j) = \int (\sqrt{p_k(\boldsymbol{x})}\sqrt{p_j(\boldsymbol{x})})\, d\boldsymbol{x} \tag{5.5}$$

or the *Jensen-Shannon kernel* (Martins et al., 2008):

$$K_{JS}(k,j) = \ln(2) - \mathcal{H}(\frac{p_k(\boldsymbol{x}) + p_j(\boldsymbol{x})}{2}) + \frac{\mathcal{H}(p_k(\boldsymbol{x})) + \mathcal{H}(p_j(\boldsymbol{x}))}{2} \tag{5.6}$$

(where $\mathcal{H}$ is the Shannon entropy) over the component distributions of the modules to compute an inner-product matrix of the modules. Network inference is then performed using this resulting inner-product matrix.

Usually, there is no information available about the origin of the underlying space, and by reconstructing networks from such kernels we heavily rely on the geometric invariance encoded in the TiWnet model. This elegant solution for inferring module networks overcomes statistical problems, and is also a principled way of applying the TiWnet to large problem instances. An example of this strategy is presented in Section 5.7.

5.7 Experiments

5.7.1 Toy Examples

The TiWnet is compared with the *graph lasso* method (Friedman et al., 2007) and with its non-invariant counterpart *Wnet* on artificial data. The *graph lasso* maximises the standard Wishart likelihood under a sparsity penalty on the inverse covariance matrix, see Equation 5.2. *Wnet* replaces

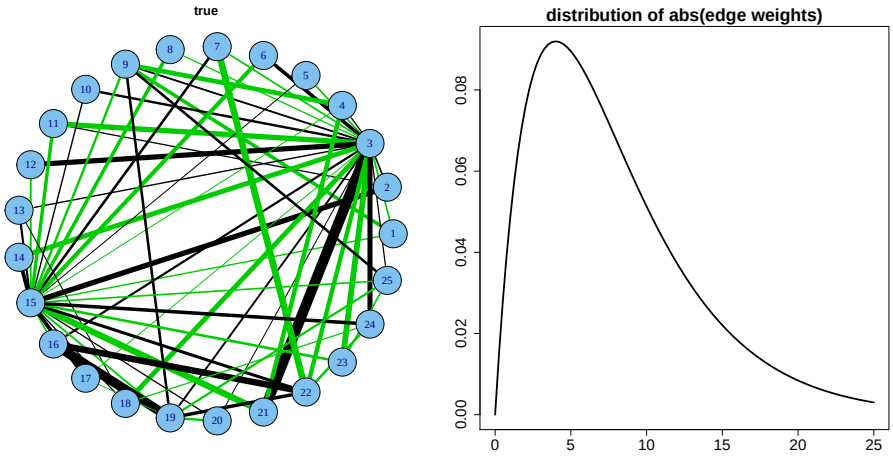


Figure 5.7.1: Left: Example network drawn from the data generator. **Right**: generative distribution of the edge weights.

the invariant Wishart used in TiWnet with the standard Wishart (Equation 5.1), but uses otherwise exactly the same MCMC code.

SAMPLE GENERATION. For these experiments we implemented a data generator that mimics the assumed generative model as shown in Figure 5.3.1. First, a sparse inverse covariance matrix $\Psi \in \mathbb{R}^{n\times n}$ with $n = 25$ is sampled. Networks with uniformly sampled node degrees are relatively easy to reconstruct for most methods, while networks with "hubs" are better suited for showing differences. Hubs are nodes with high degrees that appear naturally in many real networks since they often are scale-free i.e. their node degrees follow a power law. We simulate such networks by drawing node degrees from a Pareto$(7 \times 10^{-5}, 0.5)$-distribution and use these values as parameters in a binomial model for sampling 0/1 entries in the rows/columns of Ψ. The sign of these entries is randomly flipped, and scaled with samples from a Gamma- or uniform distribution (see below for a precise description of the distribution of the edge weights). The diagonal elements are imputed as the row-sums of absolute values plus some small constant $\epsilon\,(= 0.1)$ to ensure full rank. We draw d vectors $\boldsymbol{x}_i^o \in \mathbb{R}^n$ from $\mathcal{N}(\mathbf{0}_n, \Psi)$, and arrange them as columns in X^o. $S^o = \frac{1}{d}X^o(X^o)^t$ is then a central Wishart matrix. To study the effect of biased measurements, we randomly generate biases $b_{(i=1,\ldots,d)}$, resulting in the mean-shifted vectors $\boldsymbol{x}_i$ in Figure 5.3.1. The resulting matrix S is non-central Wishart with non-centrality matrix $\Theta = \Sigma^{-1}MM^t$, and $M = \mathbf{1}\boldsymbol{b}^t$. In fact, we always sample two i.i.d. replicates of the matrices S^o and S, and we use the second ones as a test set to tune all model parameters of the respective methods (the ℓ_1 regularisation parameter in *graph lasso* and the corresponding λ-parameter in the prior $P_2(\Psi)$ of TiWnet and Wnet) by maximising the predictive likelihood on this test set. In order to separate the effects of parameter tuning from the "true" differences in the models themselves, we additionally compared all models by tuning them to the same sparsity level. Figure 5.7.1 shows an example network drawn from our data generator together with a Gamma(2,4)-distribution of the absolute values of the edge weights.

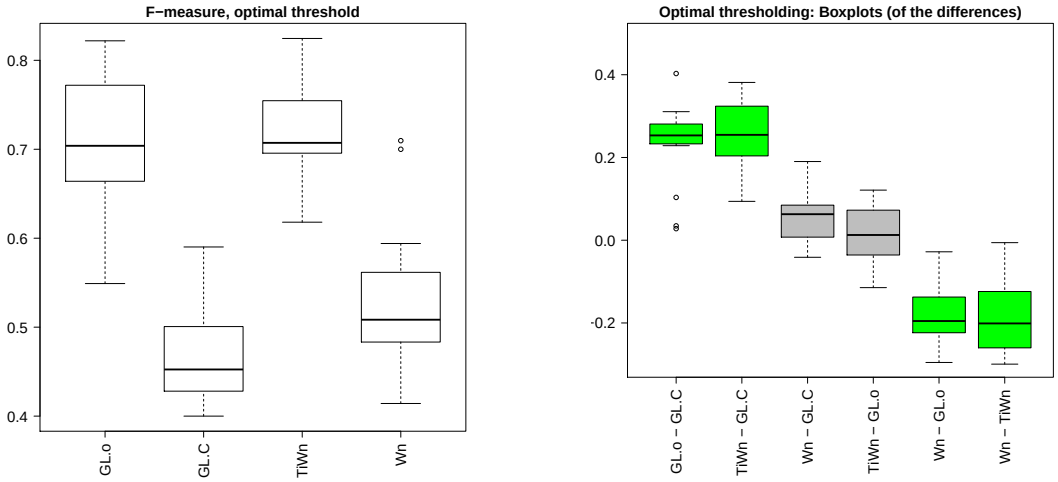


Figure 5.7.2: Left: Boxplots of F-scores obtained in 20 experiments with randomly generated Ψ-matrices for *graph lasso (GL)*: GL.o uses original S^o and GL.C uses column-centered S, *TiWnet*, and *Wnet*. **Right**: Boxplot of the pairwise differences together with color-coded significance (green, if multiple-testing-corrected $p < 0.05$) computed by a non-parametric Friedman test with post-hoc analysis (Wilcoxon-Nemenyi-McDonald-Thompson test, see Hollander and Wolfe (1999)).

SIMULATIONS. In a **first experiment**, we compare the performance of TiWnet with *graph lasso* and *Wnet*. The quality of the reconstructed networks is measured as follows: A binary vector $\boldsymbol{l}$ of size $n(n-1)/2$ encoding the presence of an edge in the upper triangle matrix of Ψ is treated as "true" edge labels, and this vector is compared with a vector $\hat{\boldsymbol{l}}$ containing the absolute values of elements in the reconstructed $\hat{\Psi}$ after zeroing those elements in $\hat{\boldsymbol{l}}$ which are not sign-consistent with the non-zero entries in Ψ (meaning that sign-inconsistent estimates will always be counted as errors). The agreement of $\boldsymbol{l}$ and $\hat{\boldsymbol{l}}$ is measured with the F-measure, i.e. the highest harmonic mean of precision and recall under thresholding the elements in $\hat{\boldsymbol{l}}$. The left panel in Figure 5.7.2 shows boxplots of F-scores obtained in 20 experiments with randomly generated Ψ-matrices for *graph lasso*, *TiWnet*, and *Wnet*. For *graph lasso*, a series of $\hat{\Psi}$ estimates with increasing ℓ_1 penalty parameter is computed using the *glassopath* function from the *glasso* R package [4]. For the MCMC-based methods TiWnet and Wnet, $\hat{\Psi}$ is computed as the sample average of networks drawn from the Gibbs samples after a certain burn-in period. The right panel shows the outcome of a Friedman test (i.e. non-parametric ANOVA) with post-hoc analysis for assessing the significance of the differences, see figure caption for further details. From the results we conclude that for the methods relying on the standard Wishart distribution (i.e. *graph lasso* and Wnet), column centering does not overcome the problem of model mismatch due to column biases. Further, TiWnet using only the pairwise distances D performs as well as *graph lasso* on the original (*not* shifted) data. Note that for the original S^o, *graph lasso* might indeed serve as a "gold standard", since the model assumptions are exactly met. And last but not least, the invariance properties of the likelihood used in TiWnet are indeed essential for its good performance, since its non-invariant counterpart Wnet uses exactly the same MCMC code (apart from using the standard Wishart likelihood, of course).

[4] `http://www-stat.stanford.edu/~tibs/glasso`

The left column of Figure 5.7.3 shows the networks reconstructed by the different methods (networks with highest predictive likelihood for *graph lasso* and sample average in the case of TiWnet and Wnet). The right column depicts the thresholded networks according to the best F-score with respect to the known ground truth. Analysing the reconstructed networks in the left column of Figure 5.7.3, it is obvious that the *graph lasso* networks are very dense, and that thresholding the edge weights is essential for a high F-score. Note, however, that such thresholding is only possible if the ground truth is known. The average TiWnet/Wnet result is also dense, since it represents the empirical distribution of networks sampled during the MCMC iterations. Thresholding the edges is also essential here, but for the MCMC models we can easily compute a truly sparse network by annealing the Markov chain *without* having access to the ground truth. Further studying this effect leads us to a **second experiment**, where we directly compare the lasso-type networks reconstructed using a sequence of ℓ_1 regularisation parameters with the "frozen" TiWnet after annealing. In this comparison, however we do *not* allow for further thresholding the edge weights when computing the F-score (i.e. we replace the entries in $\hat{\boldsymbol{l}}$ by their sign). The left panel in Figure 5.7.4 shows that TiWnet clearly outperforms all other methods. We conclude that model selection in the lasso methods does not work satisfactorily, probably because the ℓ_1 penalty not only sparsifies the solution, but also globally shrinks the parameters. As a result, truly sparse solutions have a relatively small predictive likelihood. Further, it is obvious that in the case of TiWnet, the annealing mechanism in our MCMC sampler produces very sparse networks of very high quality. The direct comparison with the non-invariant Wnet model shows that the invariance in the Wishart likelihood is indeed the essential ingredient of TiWnet.

It is clear that the results of the previous experiment crucially depend on the model selection step. To exclude differences caused by model selection, in a **third experiment** we additionally investigated the performance of the models after tuning all of them to the *same sparsity level* as the annealed network obtained by TiWnet. The results are presented in Figure 5.7.5. It is obvious that TiWnet clearly outperforms its competitors. Inspecting the recovered networks for the *graph lasso*, we see that under these restrictive sparsity constraints, the lasso selection has particular problems to recover the edges connecting *hubs* in the network.

We test the dependency of these results on the validity of the model assumptions, in a **fourth experiment**. The TiWnet in its simplest form uses only three levels for edge weights: $0, +1, -1$. It is clear that this simple model will have problems recovering networks with a very high dynamic range of edge weights (the generalisation to more than 3 levels, however, is straight forward). Since the edge weight distribution in the previous experiments was relatively concentrated around the mode of the gamma distribution (see Figure 5.7.1), we changed the distribution to a uniform distribution over the interval $[0.2, 20]$. This choice implies a uniform dynamic range over two decades. The performance of TiWnet measured in terms of the F-score, however, did not change significantly, see the top row in Figure 5.7.6 in comparison to Figure 5.7.2.

In order to further test the robustness under model mismatches, in a **fifth experiment**, we substituted the Gaussian to produce X^o with a Student-t distribution in our data generator. The resulting plot of F-scores (Figure 5.7.6, bottom row) has the same overall-structure as in Figure

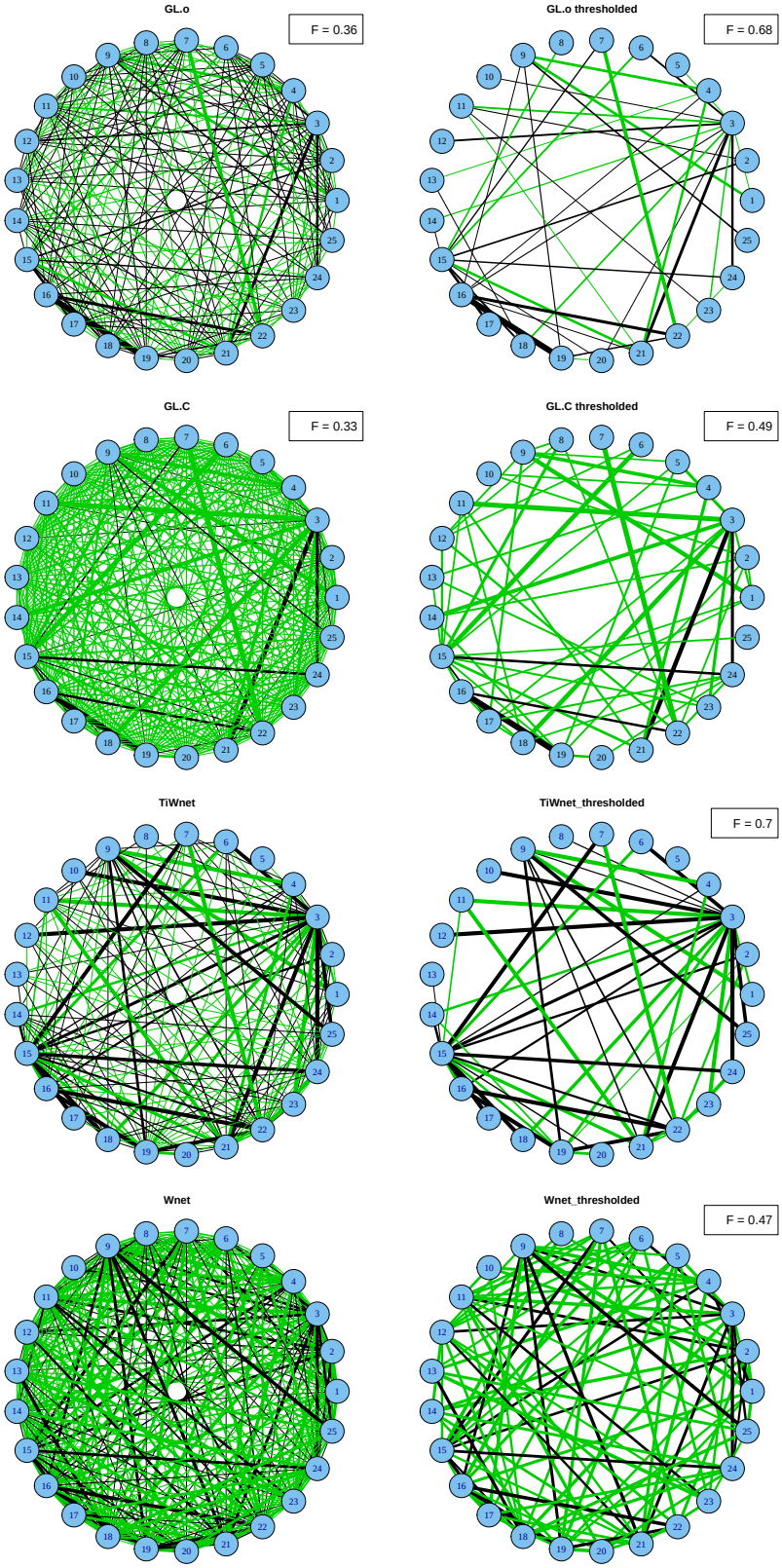


Figure 5.7.3: Left column: Networks with highest predictive likelihood for *graph lasso (GL)*: GL.o uses original S^o and GL.C uses column-centered S and sample averages for *TiWnet*, and *Wnet*. **Right column**: Optimally thresholded networks according to the best F-score with respect to the known ground truth. The underlying ground truth network is the one depicted in Figure 5.7.1.

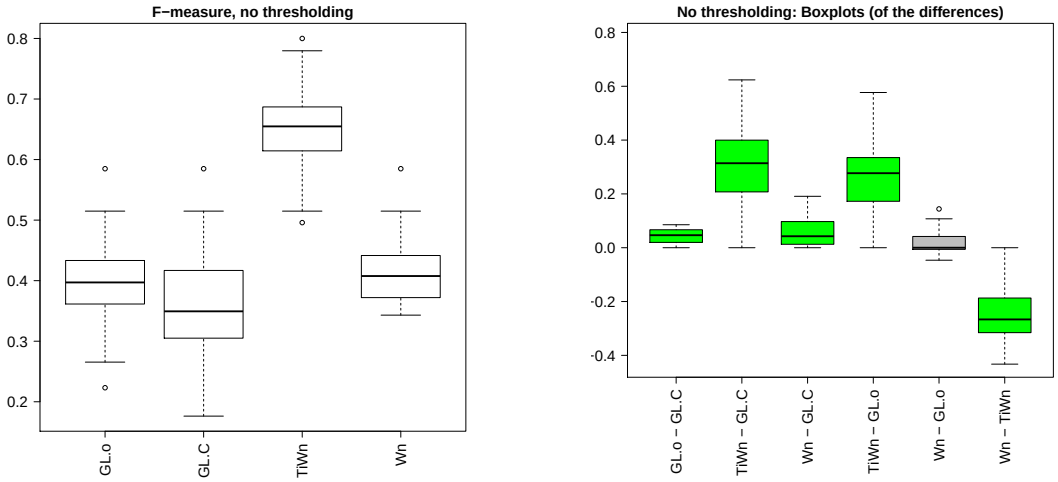


Figure 5.7.4: Left: F-scores without additional thresholding for *graph lasso (GL)*: GL.o uses original S^o and GL.C uses column-centered S, (model selected according to best predictive likelihood) and *TiWnet*/*Wnet* (also selected according to predictive likelihood, then annealed). **Right**: Corresponding boxplots of the pairwise differences.

5.7.2, which shows that TiWnet is relatively robust under such model mismatches. In summary, we conclude from these experiments that TiWnet significantly outperforms its competitors, and that the main reason for this good performance is indeed attributed to the invariant Wishart likelihood.

5.7.2 Real-world Examples

A Module network of *Escherichia coli* genes. For inferring module networks in a biological context, we applied the TiWnet to a published dataset of promoter activity data from ≈ 1100 *Escherichia coli* operons (Zaslaver et al., 2006). The promoter activities were recorded with high temporal resolution as the bacteria progressed through a classical growth curve experiment experiencing a "diauxic shift". Certain groups of genes are induced or repressed during specific stages of this growth curve. Cluster analysis of the promoter activity data was performed using a spherical Gaussian mixture model with shared variance σ: $p(x) = \sum_k \pi_k \mathcal{N}(x|\mu_k, \sigma)$ along with a Dirichlet-process prior to automatically select the number of clusters. This revealed the presence of 14 distinct gene clusters (see expression profiles of nodes in Figure 5.7.7). Network inference with TiWnet was carried out on a Bhattacharyya kernel K_B computed over the Gaussian clusters where $K_B(k,j) = \exp^{-||\mu_k - \mu_j||^2/8\sigma^2}$ (see Jebara et al. (2004)). When the clusters were analysed, genes known to be co-regulated were predominantly found in the same or nearby clusters with positive partial correlations. For example, during the diauxic shift experiment, the transcriptional activator *CRP* induces a certain set of genes in a specific growth phase (Keseler et al., 2011). Strikingly, of the 72 known *CRP* regulated operons in the dataset, 43 genes are found in cluster 6 or the four neighbouring clusters (3,9,11,13). Likewise, genes involved in specific molecular functions (those coding for proteins involved in amino acid biosynthesis pathways) were found in close proximity in the network, for example in nodes 1 and 2 (Figure 5.7.7). Physiologically, this co-regulation makes sense since protein biosynthesis (carried out by the ribosome) depends on a constant supply of synthesised amino acids. Thus TiWnet can successfully identify connections between genes co-regulated

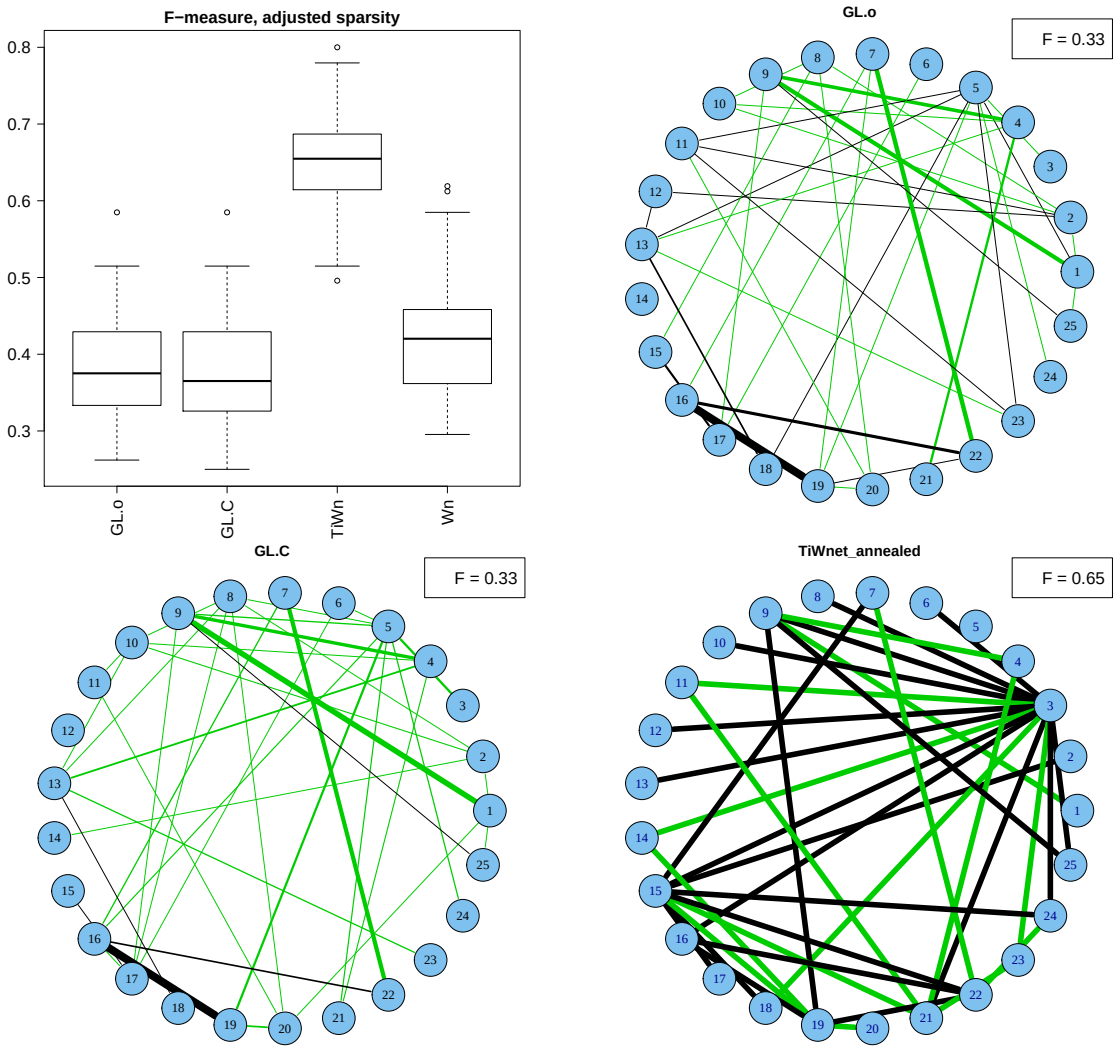


Figure 5.7.5: F-scores obtained by tuning the models to (roughly) the same sparsity level as the annealed TiWnet, averaged over 20 randomly drawn networks (top left). Other panels: networks recovered by *graph lasso (GL)*: GL.o uses original S^o and GL.C uses column-centered S and TiWnet in one of the 20 experiments. The underlying ground truth network is again the one depicted in Figure 5.7.1.

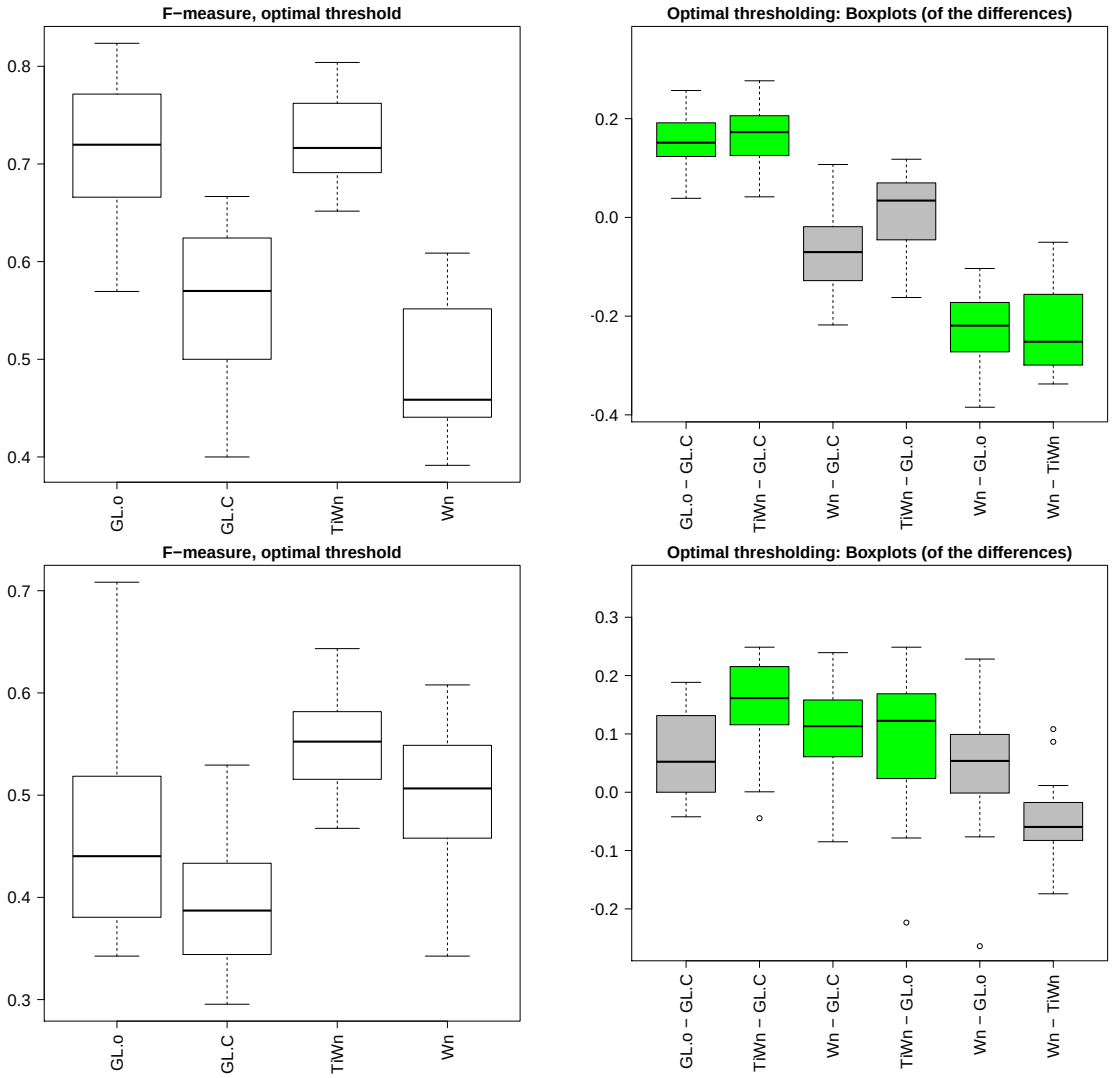


Figure 5.7.6: Top row: Testing the quality of the three-level prior on the elements in the inverse covariance matrix by simulating edge-weights with a uniform distribution on the interval $[0.2, 20]$ for *graph lasso (GL)* (GL.o uses original S^o and GL.C uses column-centered S) and *TiWnet*/*Wnet*. **Bottom row**: Results using a multivariate Student-t distribution in three degrees of freedom instead of a normal distribution to generate the columns in X^o.

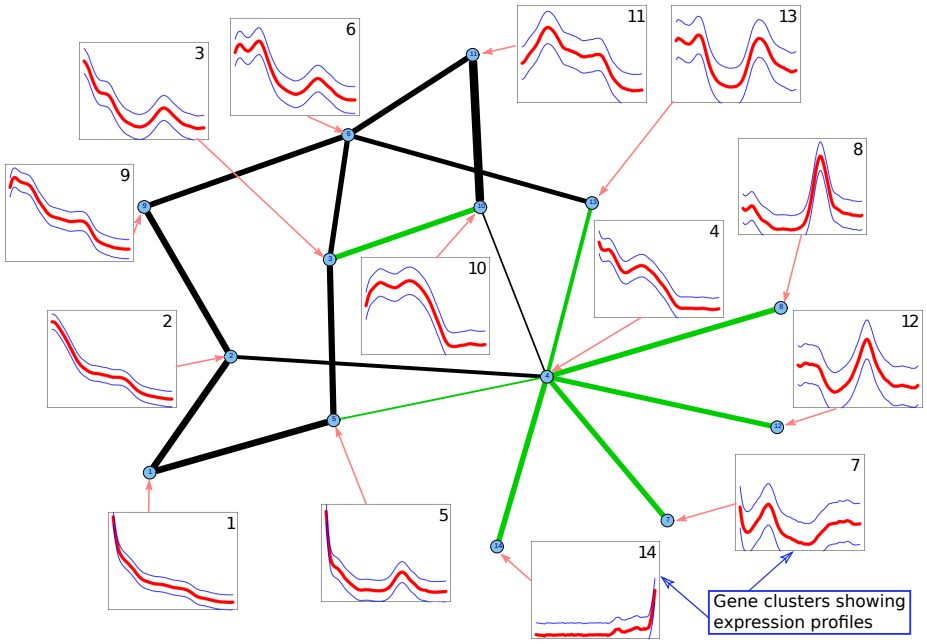


Figure 5.7.7: Module Network of *Escherichia coli* Genes. Black/green edges = positive/negative partial correlation.

by the same molecular factor, or are involved in interlinked molecular processes.

"Landscape" of chemical compounds with *in vitro* activity against HIV-1. As a second real-world example TiWnet is used to reconstruct a network of chemical compounds. We enriched a small list of compounds identified in an AIDS antiviral screen by NCI/NIH available at `http://dtp.nci.nih.gov/docs/aids/searches/list.html` with all currently available anti-HIV drugs, yielding a set of 86 compounds. *Chemical hashed fingerprints* were computed from the chemical structure of the compounds that was encoded in SMILES strings (Weininger, 1988). The *Tanimoto* kernel, a similarity matrix S of inner-product type, is constructed by the pairwise Tanimoto association scores (Rogers and Tanimoto, 1960) between the compounds. Since the geometric position of the underlying Euclidean space is unclear, we again relied heavily on the geometric invariance inherent in TiWnet. The resulting network (Figure 5.7.8) shows several disconnected components which nicely correspond to chemical classes (the node colors). Currently available anti-HIV drugs are indicated by their chemical and commercial names alongside their 2D-structures depicting the chemical similarity underlying this network. These drugs belong to the functional groups "Nucleoside reverse transcriptase inhibitors (NRTI)", "Non-nucleoside reverse transcriptase inhibitors (NNRTI)", "Protease inhibitors", "Integrase inhibitors", or "Entry inhibitors", and most compounds of a certain functional type cluster together in the graph. Medically, this network can be very useful to predict "cross-resistance" between resistant HIV-1 variants and drugs and is especially distinctive for NRTIs. The pairs *lamivudine-emtricitabine*, *tenofovir-abacavir*, and *d4T-zidovudine(ZDV)* show almost the same resistance profiles (Johnson et al., 2010). This similarity is very well reflected by our network

where these pairs are in close proximity.

It is worth noting that *graph lasso* has similar difficulties on this dataset as in the toy examples. When following the solution path by varying the penalty parameter, it is difficult to find a good compromise between sparsity and connectivity: either the obtained graphs are very dense being difficult to plot and harder to interpret, or are increasingly sparse in which, however, several interesting structural connections are lost since many singleton nodes are created. For a graphical depiction, refer Figures in Appendix 8.1. The *R* and *C++* source code for this experiment using TiWnet is available at http://bmda.cs.unibas.ch/TiWnet.

THE "LANDSCAPE" OF GLYCOSIDASE ENZYMES OF *Escherichia coli.* In yet another real-world experiment, we use TiWnet to extract the network of Glycosidase enzymes of *Escherichia coli.* Every enzyme is represented by its vectorised *contact map* computed from their *PDB* (Protein Data Bank) files. A contact map is a compact representation of the topological information of the 3D protein structure, present in the *PDB* file, into a symmetric, binary 2D matrix consisting of pairwise, inter-residue contacts.

For a protein with R amino acid residues, the contact map (see Figure 5.7.9) would be a $R \times R$ binary matrix CM where $CM_{ij} = 1$ if residues i and j are similar or 0 otherwise. The starting point for TiWnet is the contact map representation of an enzyme whose row-wise vectors serve as strings. To obtain the pairwise distances between strings in these contact maps, we compute the *Normalised Compression Distance (NCD)* (Li et al. (2004)) which is an approximation to the *Normalised Information Distance (NID).* The *NID* (Li et al. (2004)) is a distance metric minimising any admissible metric between objects. Given strings x and y, *NID* is proportional to the length of the shortest program that computes $x|y$ as well as $y|x$ and is defined as

$$NID(x,y) = \frac{\max\{K(x|y), K(y|x)\}}{\max\{K(x), K(y)\}} = \frac{K(xy) - \min\{K(x), K(y)\}}{\max\{K(x), K(y)\}}$$

where $K(x)$ is the Kolmogorov complexity of the string x. The real-world approximated version of *NID* is given by *NCD* and is calculated as follows:

$$NCD(x,y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}$$

where $C(xy)$ represents the size of the file obtained by compressing the concatenation of x and y. We use the *ProCKSI-Server* (Barthel et al. (2007), Krasnogor and Pelta (2004)) to compute $NCD(x,y)$.

The network extracted by TiWnet from the NCD values is shown in Figure 5.7.10. The network shows a clear formation of subnets of enzymes given by node colors. To further analyse the obtained subnets, we look at their corresponding Gene Ontology (GO) annotations. The GO annotations are part of a Directed Acyclic Graph (DAG), covering three orthogonal taxonomies: molecular function, biological process and cellular component (Ashburner et al., 2000). For two subnets (shown in dotted circles in Figure 5.7.10), we inspect the GO subgraphs that are subsets of the entire GO graph. The three taxonomic components of the GO subgraphs explain the proteins in these subnets

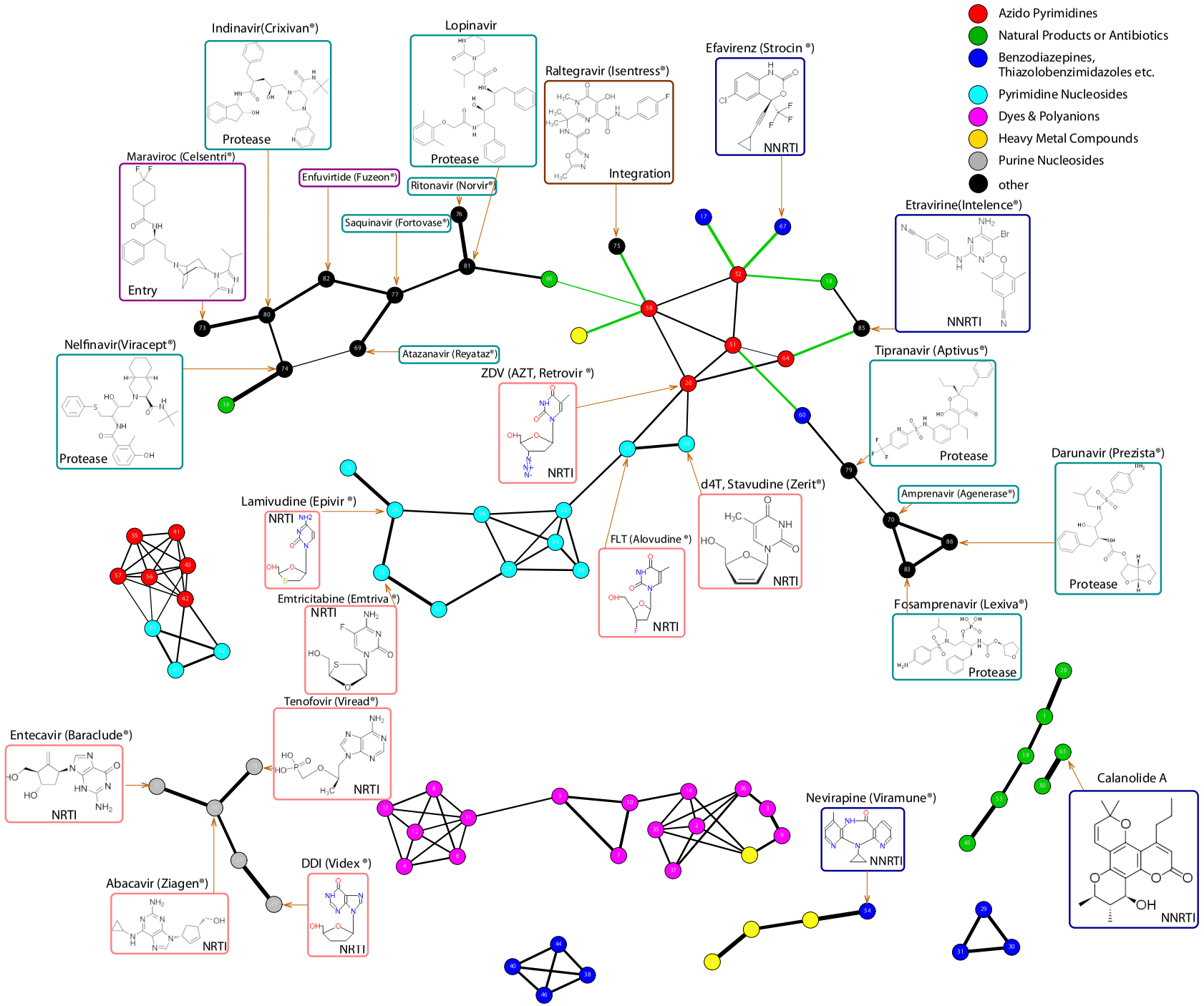


Figure 5.7.8: "Landscape" of Chemical Compounds with *In Vitro* Activity against HIV-1. Black/green edges = positive/negative partial correlation.

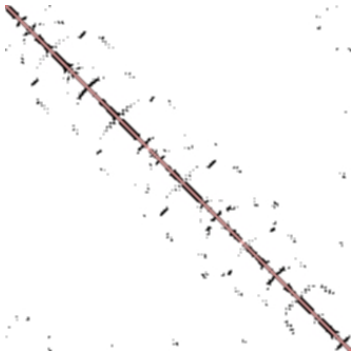


Figure 5.7.9: A contact map which is the vectorised 2D matrix capturing the 3D representation of a protein.

and show the relevance of these proteins through the color-scaling scheme where red accounts for highly-frequent enzymes. As depicted, the GO subgraphs plotted for the two subnets consist of many highly-significant enzymes thus emphasising that the subnets so obtained using TiWnet are not random, but instead consist of groups of enzymes having shared annotations. Subnets of this kind are beneficial to identify the most important GO domains for a given set of enzymes and also suggest biological areas for further exploration.

5.8 TiWD versus TiWnet

In this section, we describe the *Translational-invariant Wishart Dirichlet* (TiWD) cluster process (Vogt et al., 2010) (previously mentioned in Section 5.4) and explain why it is unsuited for extracting networks. TiWD is a fully-probabilistic model for clustering and is specifically devised to work with pairwise Euclidean distances by suitably encoding the translational and rotational invariances. Although the TiWD clustering model and TiWnet use identical likelihoods, the priors in both models are different.

The TiWD clustering model uses a Dirichlet-Multinomial type prior over clusters with the priors being restricted to block-diagonal form. This kind of prior construction is incompetent for network inference since if such a prior is used, all networks would always decompose into separated clusters which are maximal cliques i.e. fully connected within themselves. Therefore, to enable network recovery an enhanced prior construction is necessary and to this end, TiWnet uses a prior that relaxes the block-diagonal form. The two-component TiWnet prior (Section 5.5.2) is designed that, along with the invariance encoded in the likelihood, leads to sparse network recovery. The resulting Ψ is constructed to be a sparse diagonally-dominant matrix.

We illustrate the difference between the TiWnet and TiWD prior constructions in Figure 5.8.1. The top panel of Figure 5.8.1 depicts the original network generated using Ψ (no longer block-diagonal) meant for network inference and the inferred network using TiWnet. The black/green edges depict the positive/negative partial correlations between the nodes. The bottom panel of Figure 5.8.1 shows the inferred block-diagonal Ψ (left) obtained from TiWD clustering that uses a block-diagonal prior and different views of the network obtained using this Ψ: the center plot shows that the network is densely connected bearing no resemblance to the original network and the right

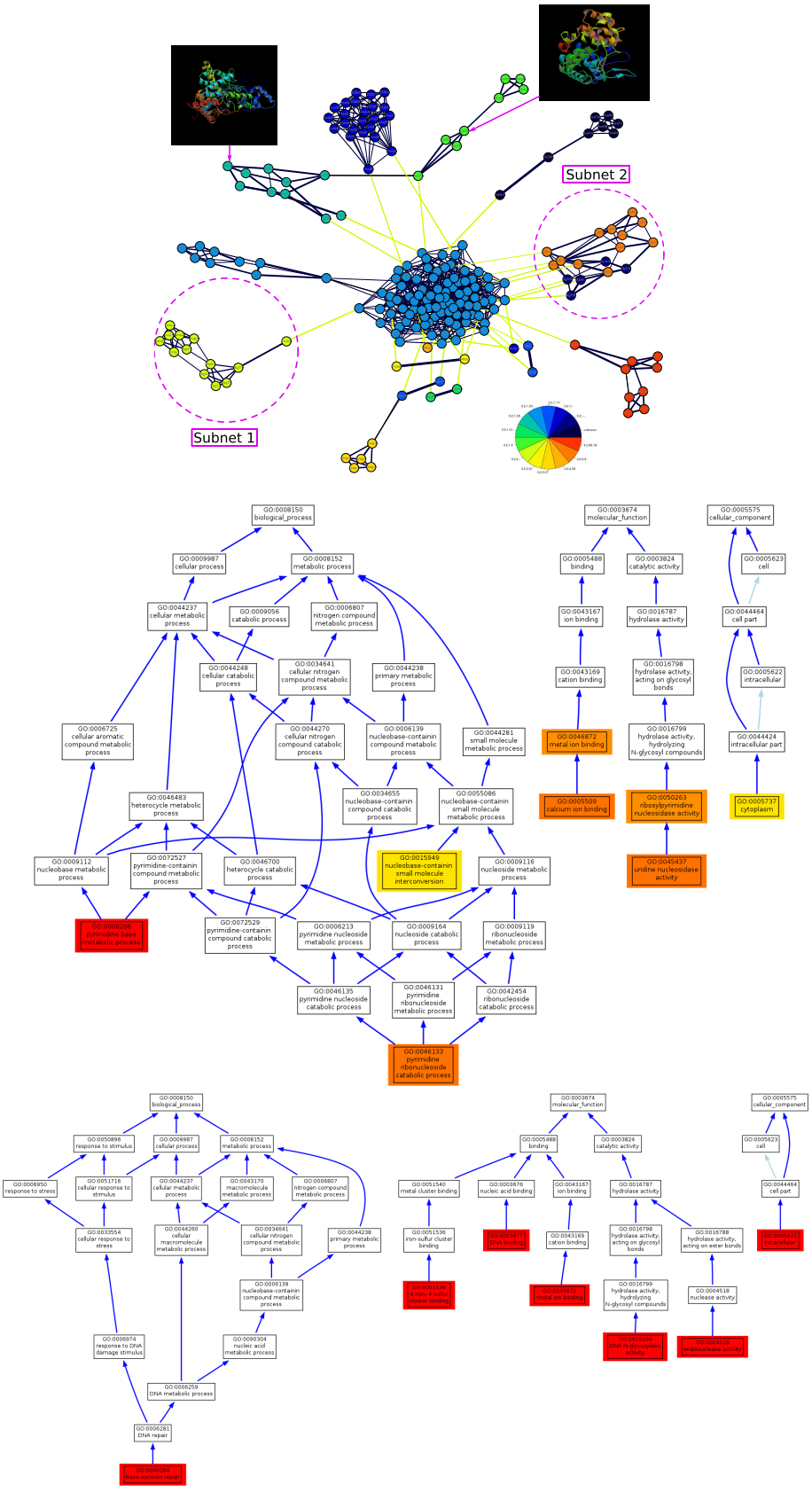


Figure 5.7.10: Top: "Landscape" of Glycosidase enzymes of *Escherichia coli*. Black/green edges = positive/negative partial correlation. For two subnets, Subnet 1 and 2 (encircled by dots), the corresponding Gene Ontology (GO) subgraphs (**centre** and **bottom**) are given that explain the enzymes present in the subnet. The multiple red/orange-hued boxes in the GO subgraph signal highly-frequent enzymes thus showing that the subnets extracted by TiWnet are not random but instead contain groups of enzymes having shared annotations.

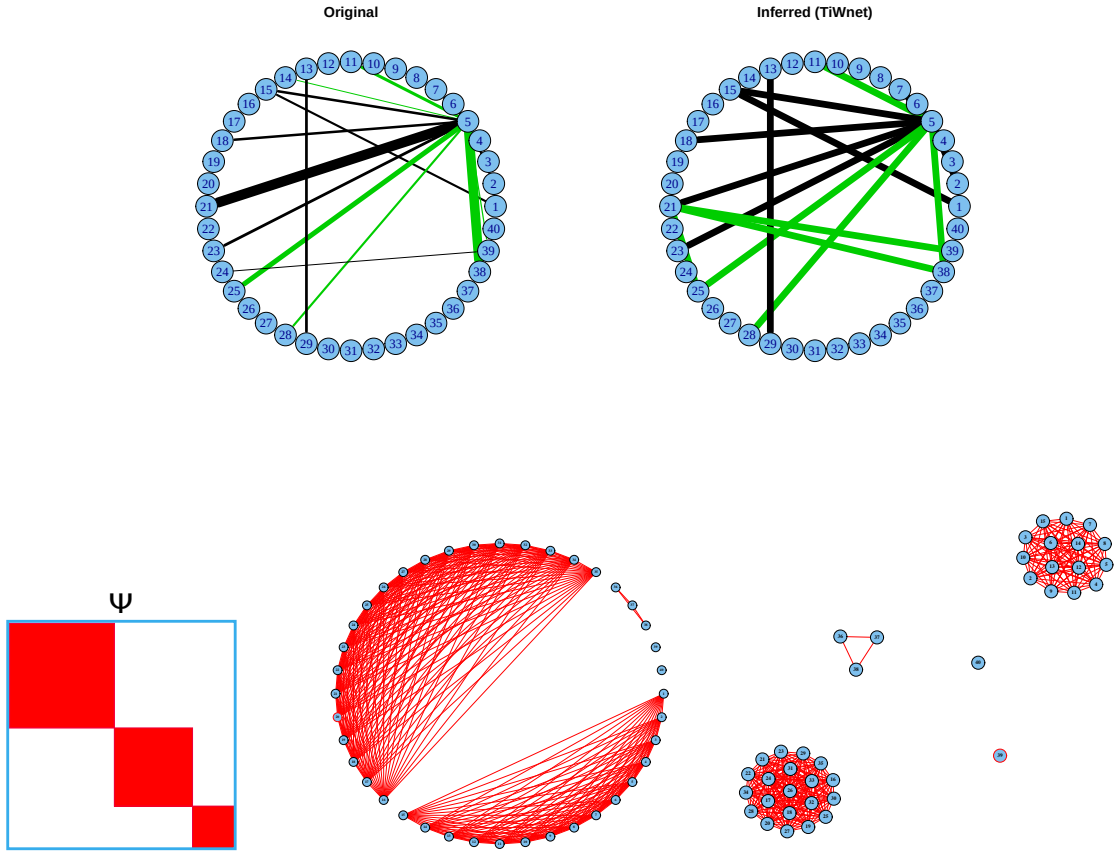


Figure 5.8.1: Illustration of the difference between TiWnet and TiWD clustering (Vogt et al., 2010) using data generated from Ψ (no longer block-diagonal) designed for network inference. **Top**: Left: Original network. Right: Inferred sparse network using TiWnet. The black/green edges denote positive/negative partial correlations between nodes. **Bottom**: Left: Inferred Ψ using TiWD clustering that has a block-diagonal structure which leads to fully-connected clusters (maximal cliques). Center: Densely-connected network obtained using this block-diagonal Ψ. The edges do not differentiate between positive/negative partial correlations. Right: The same network as in the center now showing that the network decomposes into separate maximal cliques. Here the network decomposes into 5 clusters viz. 3 fully-connected and 2 singletons.

plot highlights that the network gets decomposed into separate fully-connected clusters (maximal cliques). Moreover, the network fails to capture the positive/negative partial correlations between the nodes since the inferred Σ in the case of TiWD clustering only contains information regarding the cluster structure but without signs.

From the above discussion, it is obvious that clustering is a specialised case of network inference and that general networks cannot be recovered using the TiWD clustering model of Vogt et al. (2010). Thus the prior designed for use in TiWnet is not of the block-diagonal form thereby allowing any possible internodal interaction. Combining this enhanced prior suitable for network reconstruction with the likelihood, we are able to perform Bayesian network inference in TiWnet. We refer the reader to the Section 5.5 for complete details of our inference mechanism.

5.9 Contributions of TiWnet

TiWnet deals with distance data and is therefore, shift invariant. Classical GGMs extract networks from vectorial representations of objects and are based on the standard (central) Wishart likelihood model. The central Wishart model is only justified for *zero* column-shifts (i.i.d. data). These methods have solely relied on the i.i.d. assumption and not catered to the inherent column-shifts, thereby possibly generating biased networks. *Graph lasso*'s performance on column-shifts (Figure 5.3.2) and our extensive comparison experiments in Section 5.7 validate that not handling the column-wise biases is detrimental to network extraction. Instead, TiWnet based on D is shift-invariant and can therefore handle non-i.i.d. data (non-vectorial data). We show that in practical applications this shift invariance is an essential ingredient for recovering correct networks. Due to this, network reconstruction is possible using any D induced by a Mercer kernel that represents objects with structures for which the underlying vectorial space is unknown.

Generate module networks. Being able to derive networks from such complex objects, for example graphs and probability distributions, further leads to the development of module networks which addresses the high-dimensionality problem setting. A module connotes a cluster of homogeneous objects, thereby reducing the number of objects to that of the overall clusters, where each module is now represented by a probabilistic distribution or a graph over which a Mercer kernel can be constructed and used for network discovery.

TiWnet provides a distribution over networks. *Graph lasso* was devised for estimating a truly sparse network from the data. Since TiWnet is fully probabilistic, on output we not only obtain a single network but a distribution of networks explaining the data. For many cases in reality, this is more meaningful since one has access to possible structural variations of the extracted networks.

TiWnet provides an annealed network. Further, if required, our method has the flexibility to yield a single MAP-estimate network by simulated annealing and this is possible even without knowing the underlying ground truth. On the contrary, obtaining such an equivalent sparse network with *graph lasso* would require thresholding the edge weights and this too is only possible if the ground truth is known. The *graph lasso's* sparse networks obtained by the highest predictive likelihood are comparatively less better than TiWnet's (Figure 5.7.4). This could probably be to the improper model selection in the lasso-based models in the presence of column-shifts in the data.

TiWnet can extract hub nodes. Comparing TiWnet with *graph lasso* and *Wnet* based on the

same sparsity level, we see that *graph lasso* clearly fails in recovering *hub* nodes (Figure 5.7.5). TiWnet still returns a sparse annealed network with these desirable properties that seem difficult to be achieved by *graph lasso*. Thus, the experiments justify TiWnet's superior performance against lasso-based non-invariant models and the reason can be clearly attributed to the translation-invariance encoded in the Wishart likelihood.

5.10 Conclusion

The TiWnet model is a fully probabilistic approach to inferring GGMs from pairwise Euclidean distances obtained from inner-product similarity matrices (i.e. kernels) of n objects. Traditional models for reconstructing GGMs, for example lasso-type methods, are based on the central Wishart likelihood parametrised by the inverse covariance, and sparsity of the latter is usually enforced by some penalty term. Assuming a central Wishart, however, is equivalent to assuming that the origin of the coordinate system is known. If these methods use on input only kernel matrices, then usually only the kernels' pairwise distance information is truly relevant. Since traditional methods solely rely on the origin implicitly encoded in any such kernel, they might generate biased networks. Our TiWnet method is specifically designed to work with pairwise distances since the likelihood used in inference depends only on these distances. Combining this likelihood with a prior suited for sparse network recovery, we are able to extract sparse networks using only pairwise distances. This property opens up a huge new application field for GGMs, because network inference can now be carried out on any such distance matrix induced by a Mercer kernel on graphs, probability distributions or more complex structures. We also present an efficient MCMC sampler for TiWnet making it applicable to medium-size instances, and the possibly remaining scaling issues may be overcome by inferring module networks using kernels defined on probability distributions over groups of nodes. Comparisons with competing methods demonstrate the high quality of networks obtained from TiWnet, evoking the effectiveness of working with pairwise distances. TiWnet is also robust to model mismatches unlike existing methods. The three real-world examples provide an insight into the huge variety of possible applications.

5.11 Proof of Proposition 5.1

The marginal likelihood in terms of D, $\mathcal{L}(\Psi; t(X))$, is developed indirectly through the distribution of S. Here, $t(X) = \frac{(X - \mathbf{1}_n \hat{\boldsymbol{b}}^t)}{||X - \mathbf{1}_n \hat{\boldsymbol{b}}^t||}$ is the standardised statistic and is constant on the set of all X and S mapping to the same D. Therefore $t(X)$ can be seen as a function of the scaled version of D alone i.e. $f(\frac{D}{||D||})$. Our interest parameter is Ψ. McCullagh (2009) shows that the distribution of an arbitrary $S \in \mathbb{S}(D)$ can be analytically derived as a singular Wishart distribution with a rank-deficient covariance matrix.

We first explain the linear transformation and its kernel applied to S necessary to formulate the marginal likelihood and then proceed with the derivation of the marginal likelihood in D. The proof is derived using McCullagh (2009).

LINEAR TRANSFORMATION AND KERNEL. Given a transformation matrix $\mathbb{L}$ with kernel $\mathcal{K}$, i.e. $\mathbb{L}\mathcal{K} = \mathbf{0}$ and a generalised Gaussian random variable in $\mathbb{R}^n$, $X \sim \mathcal{N}(\mathcal{K}, \boldsymbol{\mu}, \Sigma)$, then the linearly transformed vector $\mathbb{L}X$ is distributed as $\mathcal{N}(\mathbb{L}\boldsymbol{\mu}, \mathbb{L}\Sigma\mathbb{L}^t)$. Under $\mathcal{K} = \mathbf{1}_n$, two parameter values ($\boldsymbol{\mu}_1$, Σ_1) and ($\boldsymbol{\mu}_2$, Σ_2) are equivalent when $\mathbb{L}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \mathbf{0}$ and $\mathbb{L}(\Sigma_1 - \Sigma_2)\mathbb{L}^t = \mathbf{0}$ i.e. when $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \in \mathbf{1}_n$ and $(\Sigma_1 - \Sigma_2) \in \{\mathbf{1}_n\boldsymbol{v}^t + \boldsymbol{v}\mathbf{1}_n^t; \boldsymbol{v} \in \mathbb{R}^n\}$, the space denoted by $sym^2(\mathbf{1}_n \otimes \mathbb{R}^n)$. Equivalent parameter values denote the same distribution. Corresponding to the generalised distribution of X with kernel $\mathcal{K} = \mathbf{1}_n$, the similarity matrix $S = \frac{1}{d}XX^t$ is now distributed as $S \sim \mathcal{W}_d(\mathbf{1}_n, \Sigma)$. D exhibits the negative definiteness property i.e. $\mathbf{x}^t D\mathbf{x} = -2\mathbf{x}^t S\mathbf{x} \leq 0$ for any $\mathbf{x} : \mathbf{x}^t\mathbf{1}_n = 0$. The same property holds when $\mathbf{x}$ is replaced by a symmetric positive semi-definite matrix Q i.e. $QDQ = -2QSQ \leq 0$ for any $Q : Q\mathbf{1}_n = \mathbf{0}$.

Now we consider the case of having a generalised Gaussian random matrix for kernel $\mathcal{K}$: $X_{n\times d} \sim \mathcal{MN}(\mathcal{K}, M, \Omega)$ with mean matrix $M := \mathbf{1}_n\boldsymbol{b}^t$ where $\boldsymbol{b}_i$ is the i^{th}-column bias of X and covariance tensor $\Omega := \Sigma_{n\times n} \otimes I_d$. For the mean-shifted X, the exponent term in the matrix normal distribution of X will be:

$$(X - \mathbf{1}_n\hat{\boldsymbol{b}^t})^t\Sigma^{-1}(X - \mathbf{1}_n\hat{\boldsymbol{b}^t}). \tag{5.7}$$

The corresponding exponent term in the distribution of the transformed X, $\mathbb{L}X$, is now:

$$(X - \mathbf{1}_n\hat{\boldsymbol{b}^t})^t\mathbb{L}^t\left(\mathbb{L}\Sigma\mathbb{L}^t\right)^{-1}\mathbb{L}(X - \mathbf{1}_n\hat{\boldsymbol{b}^t}). \tag{5.8}$$

We define $Q = \Sigma\mathbb{L}^t\left(\mathbb{L}\Sigma\mathbb{L}^t\right)^{-1}\mathbb{L}$ or $\Psi Q = \mathbb{L}^t\left(\mathbb{L}\Sigma\mathbb{L}^t\right)^{-1}\mathbb{L}$ (where $\Psi = \Sigma^{-1}$) as a unique orthogonal projection with $\mathcal{K} = \mathbf{1}_n$. Q can be written as $(\mathbf{I} - \mathbf{1}_n(\mathbf{1}_n^t\Psi\mathbf{1}_n)^{-1}\mathbf{1}_n^t\Psi)$ which is the orthogonal projection onto the orthogonal complement of the space spanned by symmetric positive semi-definite Σ matrices constructed by $\Sigma + \mathbf{1}_n\hat{\boldsymbol{v}^t} + \hat{\boldsymbol{v}}\mathbf{1}_n^t; \boldsymbol{v} \in \mathbb{R}^n$. Note that Q is rank deficient with $rank = n-1$.

Based on $\mathbb{L}X$, the corresponding S follows a generalised Wishart distribution in d degrees of freedom $S \sim \mathcal{W}_d(\mathbf{1}, \Sigma_{n\times n})$. McCullagh (2009) shows that $D_{ij} = S_{ii} + S_{jj} - 2S_{ij}$ is a linear transformation on symmetric matrices with transformation kernel $\mathcal{K} = sym^2(\mathbf{1}_n \otimes \mathbb{R}^n)$, implying that D follows a generalised Wishart distribution $-D \sim \mathcal{W}_d(\mathbf{1}, 2\Sigma)$ defined with respect to a transformation kernel $\mathcal{K} = \mathbf{1} \subset \mathbb{R}^n$. The generalised distribution is different from the standard Wishart distribution in that Ψ is replaced by $\widetilde{\Psi} = \Psi Q = \Psi(\mathbf{I} - \mathbf{1}_n(\mathbf{1}_n^t\Psi\mathbf{1}_n)^{-1}\mathbf{1}_n^t\Psi)$ and the $|\cdot|$ symbol for determinant is replaced by the generalised $\det(\cdot)$ which is the product of non-zero eigenvalues of its argument. $\widetilde{\Psi}$ is rank deficient with $rank = n - 1$.

SHIFT- AND SCALE-INVARIANT MARGINAL LIKELIHOOD IN D. Using the above formulation of linear transformation and kernel on symmetric positive semi-definite S matrices, McCullagh (2009) derives the marginal likelihood in D based on the standardised statistic $t(X) = \frac{(X - \mathbf{1}_n\hat{\boldsymbol{b}^t})}{||X - \mathbf{1}_n\hat{\boldsymbol{b}^t}||}$ and the interest parameter $\alpha = \Psi$ (Equation 5.3). The nuisance parameters θ are bias estimates $\hat{\boldsymbol{b}}$ and scale parameter τ.

Given $X^o_{n\times d}$, the corresponding $S^o = \frac{1}{d}X^o(X^o)^t$ follows a central Wishart distribution [5] and its

[5] The central standard Wishart distribution is defined for $S^o = X^o(X^o)^t$. Throughout the chapter, we use $S^o = \frac{1}{d}X^o(X^o)^t$ so that d appears in the central Wishart distribution and can be later used as an annealing parameter in the inference procedure.

likelihood as a function of the inverse covariance Ψ is:

$$\mathcal{L}(\Psi; S^o) = |\Psi|^{\frac{d}{2}} \cdot \exp\left[-\tfrac{d}{2}\operatorname{tr}(\Psi S^o)\right]. \tag{5.9}$$

We consider the statistic for mean-shifted X as $(X - \mathbf{1}_n\hat{\boldsymbol{b}})$. In terms of this statistic, $S = \frac{1}{d}(X - \mathbf{1}_n\hat{\boldsymbol{b}}^t)(X - \mathbf{1}_n\hat{\boldsymbol{b}}^t)^t$ and Equation 5.9 becomes:

$$\mathcal{L}(\hat{\boldsymbol{b}}, \Psi; S) = |\Psi|^{\frac{d}{2}} \cdot \exp\left[-\tfrac{d}{2}\operatorname{tr}(\Psi S)\right]. \tag{5.10}$$

In Equation 5.10, we apply an arbitrary but fixed transformation $\mathbb{L}$ with kernel $\mathcal{K} = \mathbf{1}_n$ leading to $\Psi Q = \mathbb{L}^t\left(\mathbb{L}\Sigma\mathbb{L}^t\right)^{-1}\mathbb{L}$ and replace the determinant $|\cdot|$ symbol by the generalised $\det(\cdot)$ which is the product of non-zero eigenvalues of its argument (since Q is rank deficient) and obtain:

$$\mathcal{L}(\Psi; S) \propto \det(\Psi Q)^{\frac{d}{2}} \cdot \exp\left[-\tfrac{d}{2}\operatorname{tr}(\Psi Q S)\right]. \tag{5.11}$$

We substitute $\widetilde{\Psi} = \Psi Q = \Psi(\mathbf{I} - \mathbf{1}_n(\mathbf{1}_n^t\Psi\mathbf{1}_n)^{-1}\mathbf{1}_n^t\Psi)$ to arrive at the shift-invariant form for marginal likelihood in S:

$$\begin{aligned}\mathcal{L}(\Psi; S) &\propto \det(\widetilde{\Psi})^{\frac{d}{2}} \cdot \exp\left[-\tfrac{d}{2}\operatorname{tr}(\widetilde{\Psi}S)\right] \\ &\propto \det\left(\widetilde{\Psi}\right)^{\frac{d}{2}} \cdot \exp\left[-\tfrac{d}{2}\operatorname{tr}(\widetilde{\Psi}S)\right].\end{aligned} \tag{5.12}$$

The likelihood in Equation 5.12 is constant for all choices of $S \in \mathbb{S}(D)$ and hence it depends only on D. Using the negative definiteness property of D i.e. $\widetilde{\Psi}S = (-\frac{1}{2})\widetilde{\Psi}D$, Equation 5.12 can be written in terms of D as:

$$\mathcal{L}(\Psi; D) \propto \det(\widetilde{\Psi})^{\frac{d}{2}} \cdot \exp\left[\tfrac{d}{4}\operatorname{tr}(\widetilde{\Psi}D)\right]. \tag{5.13}$$

Equation 5.13 is the shift-invariant marginal likelihood in D based on the statistic $(X - \mathbf{1}_n\hat{\boldsymbol{b}})$ and the rank-deficient inverse covariance $\widetilde{\Psi}$.

To remove the scalar terms, we base the marginal likelihood on the standardised statistic $t(X) = \frac{(X - \mathbf{1}_n\hat{\boldsymbol{b}^t})}{||X - \mathbf{1}_n\hat{\boldsymbol{b}^t}||}$. Consider the scale parameter $\tau = \frac{1}{||X - \mathbf{1}_n\hat{\boldsymbol{b}}^t||}$. Equation 5.10 now becomes:

$$\mathcal{L}(\hat{\boldsymbol{b}}, \tau, \Psi; S) = |\tfrac{\Psi}{\tau^2}|^{\frac{d}{2}} \cdot \exp\left[-\tfrac{d}{2\tau^2}\operatorname{tr}(\Psi S)\right]. \tag{5.14}$$

Applying the same procedure as before i.e. using $\mathcal{K} = \mathbf{1}_n$ leading to ΨQ, replacing $|\cdot|$ with $\det(\cdot)$ symbol and substituting for $\widetilde{\Psi}$, we get:

$$\begin{aligned}\mathcal{L}(\tau, \Psi; S) &\propto \det\left(\frac{\widetilde{\Psi}}{\tau^2}\right)^{\frac{d}{2}} \cdot \exp\left[-\tfrac{d}{2\tau^2}\operatorname{tr}(\widetilde{\Psi}S)\right] \\ &\propto \tau^{-2\frac{(n-1)d}{2}} \cdot \det\left(\widetilde{\Psi}\right)^{\frac{d}{2}} \cdot \exp\left[-\tfrac{d}{2\tau^2}\operatorname{tr}(\widetilde{\Psi}S)\right]\end{aligned} \tag{5.15}$$

since $rank(\widetilde{\Psi}) = (n-1)$ and $\det(cA)^h = c^{h \cdot rank(A)}\det(A)^h$ for any constants c and h and a nonsingular matrix A. Notice here that the dependency on biases $\hat{\boldsymbol{b}}$ is removed.

Next, we differentiate Equation 5.15 and set the derivative to zero.

$$
\begin{aligned}
0 &= \frac{d(\mathcal{L}(\tau, \Psi; S))}{d\tau} \\
&= -2\tau^{-2\frac{(n-1)d}{2}} \cdot \exp\big(-\tfrac{d}{2\tau^2}\mathrm{tr}(\widetilde{\Psi}S)\big) \cdot (-\tfrac{d}{2})\mathrm{tr}(\widetilde{\Psi}S) \cdot \tau^{-3} + \\
&\exp\big(-\tfrac{d}{2\tau^2}\mathrm{tr}(\widetilde{\Psi}S)\big) \cdot \tau^{-2\frac{(n-1)d}{2}-1} \cdot (-2)^{\frac{(n-1)d}{2}}
\end{aligned} \tag{5.16}
$$

$$
\begin{aligned}
&2\tau^{-2\frac{(n-1)d}{2}} \cdot \exp\big(-\tfrac{d}{2\tau^2}\mathrm{tr}(\widetilde{\Psi}S)\big) \cdot (-\tfrac{d}{2})\mathrm{tr}(\widetilde{\Psi}S) \cdot \tau^{-3} = \\
&\exp\big(-\tfrac{d}{2\tau^2}\mathrm{tr}(\widetilde{\Psi}S)\big) \cdot \tau^{-2\frac{(n-1)d}{2}-1} \cdot (-2)^{\frac{(n-1)d}{2}}
\end{aligned} \tag{5.17}
$$

By cancelling terms and rearranging Equation 5.17, we obtain:

$$
\tau^2 = \frac{\mathrm{tr}(\widetilde{\Psi}S)}{n-1} \tag{5.18}
$$

and then substitute the expression for τ^2 back in Equation 5.15:

$$
\mathcal{L}(\Psi; S) \propto (\frac{\mathrm{tr}(\widetilde{\Psi}S)}{n-1})^{-\frac{(n-1)d}{2}} \cdot \det\big(\widetilde{\Psi}\big)^{\frac{d}{2}} \cdot \exp\left[-\frac{d}{2(\frac{\mathrm{tr}(\widetilde{\Psi}S)}{n-1})}\,\mathrm{tr}(\widetilde{\Psi}S)\right] \tag{5.19}
$$

where the dependency on τ vanishes.

Ignoring constant terms, we obtain the shift- and scale-invariant likelihood in S (McCullagh, 2009, Tunnicliffe-Wilson, 1989):

$$
\mathcal{L}(\Psi; S) \propto \det\big(\widetilde{\Psi}\big)^{\frac{d}{2}}\,\mathrm{tr}(\widetilde{\Psi}S)^{-\frac{(n-1)d}{2}} \tag{5.20}
$$

which is constant for all $S \in \mathbb{S}(D)$. Thus the likelihood depends only on (the scaled version of) D and by the negative definiteness property of D, we finally arrive at the shift- and scale-invariant marginal likelihood in D:

$$
\mathcal{L}(\Psi; \tfrac{D}{||D||}) \propto \det\big(\widetilde{\Psi}\big)^{\frac{d}{2}}\,\mathrm{tr}(-\tfrac{1}{2}\widetilde{\Psi}D)^{-\frac{(n-1)d}{2}} \tag{5.21}
$$

□

6

Automatic Archetype Analysis

6.1 Introduction

ARCHETYPES are defined as an original model, type or observation based on which similar things are patterned. Given observations of a multivariate dataset, archetype analysis aims at finding a small number of archetypes or *pure* data samples that optimally summarise the variation in the dataset. This summarisation is based on the precept that the data observations can be well represented as noisy convex mixtures of these archetypes and that the archetypes themselves are restricted to being convex combinations of the observations.

Archetype analysis as developed in Cutler and Breiman (1994) shows that archetypes are those *explaining* extremal points lying close to the convex hull [1] of the data. The traditional definition of an archetype is that it is in itself an *existing* observation. Since in practical applications, observations are never devoid of noise, this definition has been relaxed. Rather than assuming them to be existing observations, the archetypes are allowed to be a convex combination of the observations but still reside close to the convex hull. From the statistical viewpoint and for computational feasibility, this relaxed definition grants the archetypes more flexibility to interpret the observations. Therefore, archetype analysis can be seen as a technique where pure observations are used to minimise a set of archetypes given noise.

Archetype analysis and PCA. Like PCA, archetype analysis can also be seen as as a dimensionality-reduction technique. Whereas PCA is a technique based on maximum-variance projection,

[1] A convex hull is the smallest convex polygon containing all the observations of the multivariate dataset (Boyd and Vandenberghe, 2004). It can be visualised as a rubber band drawn taut around the data demarcating the data periphery.

archetype analysis is a projection along with a geometric constraint that the archetypes need to lie along the convex hull of the observations.

Applications. Examples of archetype analysis include analysis of compositional data in sedimentology to identify samples having *pure* geochemical compositions of sediments (Palmer and Douglas (2008)), in galaxy spectra studies (Chan et al. (2003)) to analyse new or evolving stellar constellations, in image analysis (Bauckhage and Thurau (2009)) for finding potential vision categories, in the analysis of the human genotope (Huggins et al. (2007)) for identifying informative *allele* or *single-nucleotide polymorphism* (SNP) locations and for text mining (Morup and Hansen (2012)) to group texts into *core* categories. Archetypes have also been used to study the evolution of species using phenotypic data (Shoval et al., 2012). In multispectral imaging, archetypes are known as *end members* and are used to extract *original* signals (Keshava (2003), Labitzke et al. (2012)). Archetype analysis has also made its foray into market segmentation where consumers or products are grouped into various heterogenous groups to provide marketing and advertising insights (Li et al., 2003). Archetypes are also widely used in studies dealing with petrology and palaeoecology for identifying archetypal rock patterns between continental tectonic plates (Hacker and Gans, 2005).

Focus of the current work. Conventional archetype analysis methods (Cutler and Breiman (1994), Bauckhage and Thurau (2009)) rely on RSS (residual sum of squares) for model selection. Although in low-noise datasets, the RSS curves are reliable for model selection due to their prominent knee regions, in high-noise settings the curves tend to be become uninformative as they flatten out. This has been verified in our Experiments section. Further, these models are sensitive to the initialisation of archetypes and therefore if the dataset possesses any structure (as shown in Figure 6.2.1 (centre and right)), archetype analysis becomes difficult, if the archetypes are not properly initialised.

The current work aims in addressing these two drawbacks jointly. For reliable model selection, we employ the Bayesian information criterion (BIC) (Schwarz, 1978). Even though BIC can be applied to conventional methods, in high-noise settings our experiments have shown that BIC curves tend to become unreliable (refer Figure 6.6.1 (d)). Since in these models the *degrees of freedom* $df = p$ (where p is the number of archetypes), assigning higher values of p to df only leads to overpenalising the complexity term in BIC, thereby making BIC favour models with lower p. In our current method, we have better access to efficiently compute the effective df for BIC. To overcome the dependency relating to a good initialisation of the archetypes, the archetypes are initialised to all the n observations since the model considers at most n archetypes sufficient to approximate the set of observations. Given larger datasets, this also calls for efficient methods. Thus, we base our work on the idea of enforcing grouped sparsity using the Group-Lasso formulation (Yuan and Lin (2006)). Since there are efficient methods that allow sampling of the solution paths of the Group-Lasso, stepwise model selection using BIC can be performed thereby effecting automatic archetype detection.

Outline of the chapter. Section 6.2 describes the assumed underlying process used to simulate archetypal data. In Section 6.3, the conventional archetype analysis is described and the existing

problems are elucidated. In Section 6.4, the automatic archetype analysis model is introduced, elaborating the algorithmic modifications brought about to that of the conventional methods. Section 6.5 details the model selection procedure using BIC scores in the automatic model. BIC score computation based on two forms of the degrees of freedom – *approximate* and *exact* – are described. Simulations and real-world experiments are discussed in Section 6.6. Section 6.7 summarises this work.

6.2 Data generative model and model learning

Definitions. A *convex* set X is a set of observations $\{\boldsymbol{x}_i\}$ for $i = 1, \cdots, n$ such that $\sum_{i=1}^{n} \lambda_i \boldsymbol{x}_i$ for $\lambda_i \geq 0, \sum_i \lambda_i = 1$. A *convex* hull C of the convex set X is the smallest convex set containing all the points of X i.e. $C = \{\sum_{i=1}^{n} \lambda_i \boldsymbol{x}_i : \lambda_i \geq 0, \sum_i \lambda_i = 1\}$. In other words, C is the set of convex combinations of any finite collection of observations contained in X and can also be seen as the *simplex* of the convex set.

6.2.1 Generative model

Assume there exists an underlying data generating density function that has a convex support. p points in $\mathbb{R}^d$ are sampled from this density function and serve as *archetypes*. The support of the density function is defined by the convex hull or the simplex generated by these p archetypes. Data observations are created as convex combinations of the p archetypes and additionally noise is added. Futher it is assumed that the set of data observations are in general position. For example, to generate a convex set X of n noisy observations from p archetypes, first sample p d-dimensional archetypes from a Gaussian distribution. Each of the remaining $(n - p)$ observations are convex combinations of the p archetypes where the weights are drawn from a p-dimensional Dirichlet distribution whose support is a $(p - 1)$-dimensional simplex. Gaussian noise is added to these $(n - p)$ observations.

Based on this generative model, in this current work, three different settings are outlined below.

- Setting 1: Here, the Gaussian samples are weighted with a symmetric Dirichlet distribution where the Dirichlet parameter, α, takes on the same value ($\alpha = 1$). This is equivalent to a uniform distribution over the simplex or is uniform over all observations in its support. Noisy data observations weighted using such a symmetric distribution are shown in Figure 6.2.1 (left) for $p = 3, n = 1000$ and $d = 10$.

- Setting 2: We consider weights from another symmetric Dirichlet distribution where $0 < \alpha < 1$. The Dirichlet density gets concentrated towards the edges of the simplex and this corresponds to a non-uniform distribution over the simplex. Gaussian samples weighted by such a Dirichlet distribution are shown in Figure 6.2.1 (centre) for $p = 4, n = 1000$ and $d = 100$.

- Setting 3: Clusters of compact convex sets are generated from Gaussians with different means and weighted using a symmetric Dirichlet distribution ($\alpha = 1$). Figure 6.2.1 (right) shows a dataset consisting of 3 such compact clusters with each cluster having 1000 observations generated from 4 10-d archetypes.

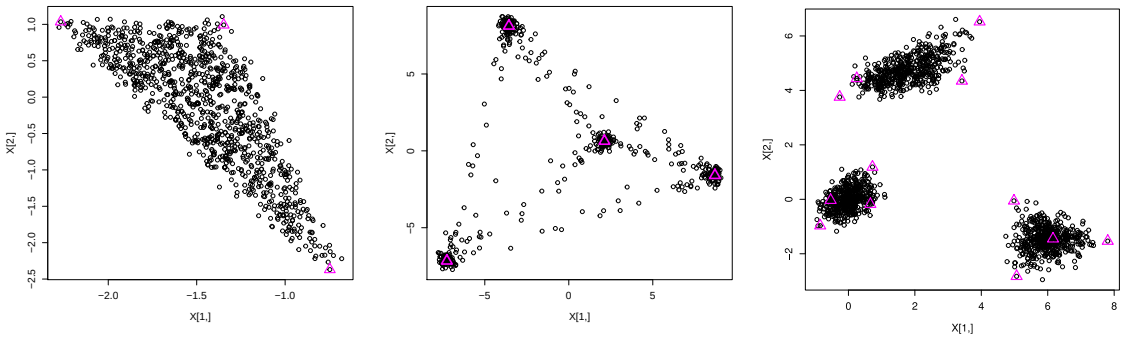


Figure 6.2.1: Left: Noisy convex set of Gaussian observations $X_{1000\times 10}$ weighted using a symmetric Dirichlet distribution ($\alpha = 1$). Observations (o) are generated using $Z_{3\times 10}$ (Δ). **Centre**: Noisy Gaussian observations $X_{1000\times 100}$ weighted using a symmetric Dirichlet distribution ($0 < \alpha < 1$). Samples tend to concentrate near the simplex edges. Observations are generated using $Z_{4\times 100}$. **Right**: 3 clusters of convex sets weighted using a symmetric Dirichlet distribution ($\alpha = 1$) where each convex set contains Gaussian observations $X_{1000\times 10}$ generated by a noisy convex combination of $Z_{4\times 10}$. All plots are the 2-d PCA projections of the data.

6.2.2 Model Learning

Given the set of noisy observations, $X_{n\times d}$, archetype analysis aims in fitting a noisy convex hull to the data. This is equivalent to finding the optimal set of archetypes $Z_{p\times d}$, that best describes X and that resides close to the convex hull of X with $p \ll n$. Figure 6.2.2 gives the graphical abstract with the top panel depicting the setup for $p = 3$ archetypes. The convex hull of X is the red outline.

6.3 Conventional Archetype Analysis – Model Description

We start with a description of the conventional archetype analysis. Given a data matrix $X_{n\times d}$ which is a noisy convex set of n observations $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)'$, where $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{id}) \in \mathbb{R}^d$ for $i = 1, \cdots, n$, the goal of archetype analysis is to find a sparse set of archetypes or *pure* samples $Z_{p\times d}$ with p archetypes $(\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_p)'$ where $p \ll n$ and $\boldsymbol{z}_j \in \mathbb{R}^d$ for $j = 1, \cdots, p$. The archetypes are such that the observations $\boldsymbol{x}_i$ are noisy convex combinations of these archetypes:

$$\boldsymbol{x}_i = Z'\boldsymbol{a}_i + \boldsymbol{\epsilon}_i, \qquad \text{for} \quad i = 1 \ldots n\,, \tag{6.1}$$

where $\boldsymbol{\epsilon}_i \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I})$ represents the stochastic nature of $\boldsymbol{x}_i$ and $\boldsymbol{a}_i$ is a composition vector such that $a_{ij} \geq 0$ and $\sum_{j=1}^{p} a_{ij} = 1$. The archetypes themselves are defined to be convex mixtures of the observations and reside close to the convex hull of the data. Hence archetypes are chosen to be a small number of points (typically smaller than n) residing in close proximity to this convex hull. The optimisation procedure in archetype analysis involves finding a small set of archetypes such that the error in approximation of the observations as convex mixtures of the archetypes is minimised.

Related work. Based on the above formulation, an iterative optimisation algorithm was introduced in Cutler and Breiman (1994). This version is still not computationally feasible for large datasets. A more scalable version was introduced in Bauckhage and Thurau (2009). The scalability

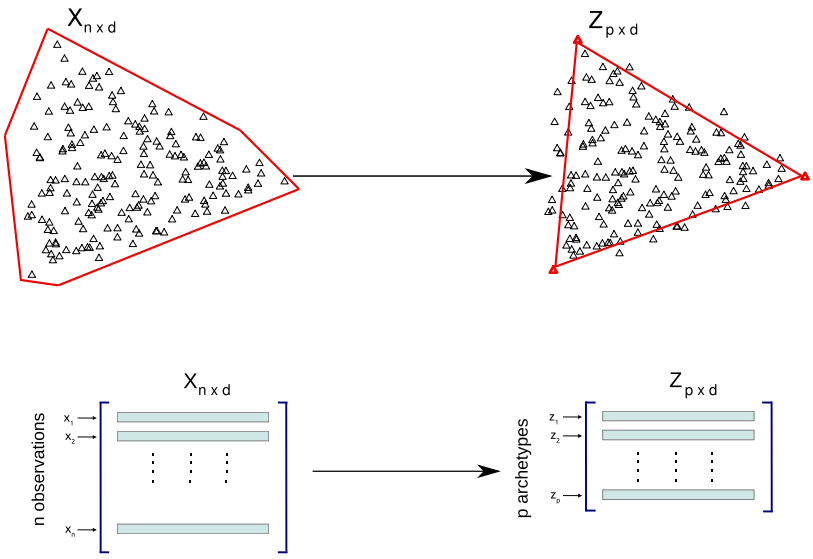


Figure 6.2.2: Graphical abstract. **Top**: Given $X_{n\times d}$, archetype analysis finds the optimal $Z_{p\times d}$ that best describes X and lie close to the convex hull of X. The convex hull of X is shown by the red outline around the observations (Δ). Plots show the 2-d PCA projections of the data. Bottom: In terms of matrix dimensions, we need to find the best p where $p \ll n$.

relied on choosing archetype candidates as points lying close to the convex hull. Another approach was introduced in Morup and Hansen (2012) that was based on kernelising the data prior to extracting the archetypes. An iterative weighted optimisation method was dealt with in Eugster and Leisch (2011) where robust archetypes were estimated in the presence of outliers since outliers normally plague the analysis of archetypes.

6.3.1 Conventional Archetype Analysis algorithm

The identification of archetypes is split into multiple steps as in Bauckhage and Thurau (2009). Given a set of archetypes Z, the compositions for all the data points can be estimated based on the optimisation problem:

$$\hat{A} = \text{argmin}_A \sum_{i=1}^{n} \|\boldsymbol{x}_i - \sum_{j=1}^{p} a_{ij}\boldsymbol{z}_j\|_2^2\,, \tag{6.2}$$

where $\|\cdot\|_2$ denotes the ℓ_2 norm of a vector. A is a $n \times p$ composition vector matrix which comprises of composition vectors $(\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_n)'$, where each $\boldsymbol{a}_i \in \mathbb{R}^p$ is represented as a row of the matrix A. Since $\boldsymbol{a}_i$ are composition vectors, additional constraints are imposed to ensure that each observation is a meaningful combination of the archetypes and that the observations are represented as mixtures of the archetypes:

$$a_{ij} \geq 0 \quad \text{and} \quad \sum_{j=1}^{p} a_{ij} = 1. \qquad \text{for} \quad i = 1 \ldots n\,. \tag{6.3}$$

Further, we assume that the archetypes lie within close proximity to the convex hull of the given

observations. Hence, each $\boldsymbol{z}_i$ can be expressed as a convex combination of the observations:

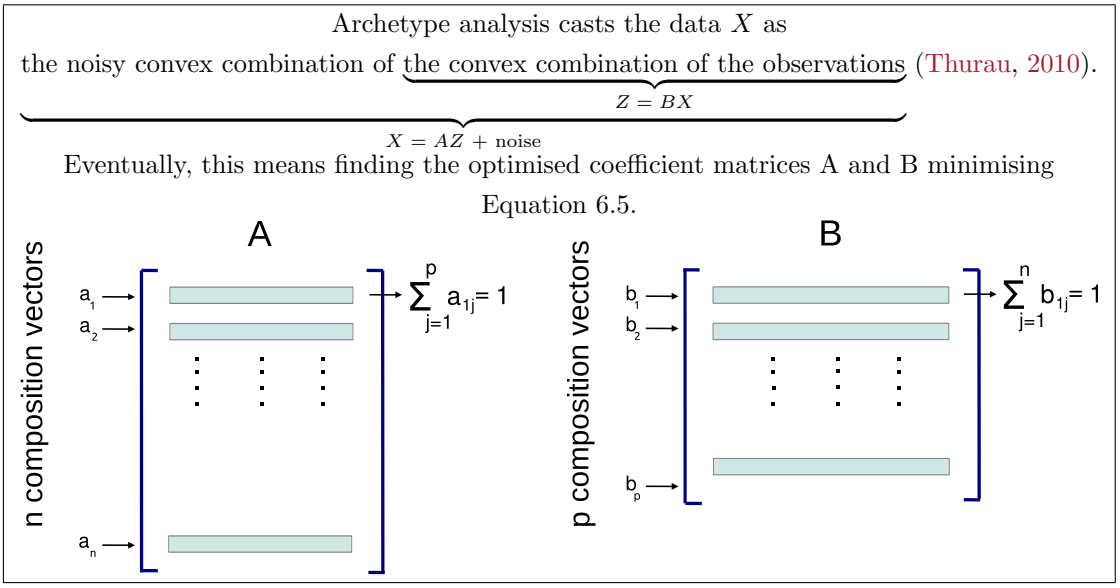
$$\boldsymbol{z}_i = \sum_{j=1}^{n} b_{ij} \boldsymbol{x}_j \,, \tag{6.4}$$

with the coefficients $b_{ij} \geq 0$ and $\sum_{j=1}^{n} b_{ij} = 1$. The coefficient vectors $\boldsymbol{b}_i$'s are represented as a $p \times n$ matrix $B = (\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_p)'$, with the rows comprising of composition vectors $\boldsymbol{b}_i \in \mathbb{R}^n$. This choice of $\boldsymbol{b}_i$'s assures that the archetypes resemble the data and that they are a convex mixture of the data. A and B are row-stochastic matrices meaning that every row sums to 1.

In matrix form, with $X \in \mathbb{R}^{n \times d}$, $Z \in \mathbb{R}^{p \times d}$, $A \in \mathbb{R}^{n \times p}$ and $B \in \mathbb{R}^{p \times n}$, a suitable choice of archetypes will minimise the residual sum of squares (RSS) problem:

$$\begin{aligned} (\hat{A}, \hat{B}) &= \text{argmin}_{A,B} \|X - AZ\|_2^2 = \text{argmin}_{A,B} \|X - ABX\|_2^2 \\ \text{s.t.} \quad & b_{ij} \geq 0, \;\; \sum_{j=1}^{n} b_{ij} = 1, \quad a_{ij} \geq 0, \;\; \sum_{j=1}^{p} a_{ij} = 1 \,. \end{aligned} \tag{6.5}$$

Thus the overall problem of finding a fixed predetermined number of archetypes for a given set of observations translates into a constrained optimisation problem involving two sets of coefficients $\{a_{ij}\}$ and $\{b_{ij}\}$.



The optimisation-based formulation given in Equation 6.5 is solved using an alternating least squares procedure described in Algorithm 2 (Bauckhage and Thurau (2009)).

Complexity analysis for conventional methods. The complexity of one iteration of the algorithm in Cutler and Breiman (1994) is given by $\mathcal{O}(n^2 p)$. To reduce this complexity further, the approach in Bauckhage and Thurau (2009) was to resort to preselecting the archetypal candidates by subsampling a set of n_* points ($n_* \ll n$) that reside on the convex hulls obtained from random

Algorithm 2 Conventional Archetype Analysis Algorithm

A : **Initialise** $Z_{p\times d}$

repeat

B : **determine** coefficients a_{ij} by minimising $||x_i - Z'a_i||_2^2$ s.t. $a_{ij} \geq 0$ and $\sum_j^p a_{ij} = 1$ for $i = (1, ..., n)$.

C : **solve** OLS equation $\widehat{Z} = (A'A)^{-1}A'X$ to compute the intermediate archetypes $\widehat{Z}$ using updated a_{ij}.

D : **determine** coefficients b_{ij} by minimising $||\hat{z}_j - X'b_j||_2^2$ s.t. $b_{ij} \geq 0$ and $\sum_i^n b_{ij} = 1$ for $j = (1, ..., p)$.

E : **Update** the archetypes by setting $Z = BX$.

until *end criterion based on the RSS approximation error.*

2D projections. This, thereby, reduced the complexity to $\mathcal{O}(n_*^2 p)$.

6.3.2 Problems with the conventional methods

Model Selection mechanism. Model selection in conventional methods is generally done using RSS decay curves. The RSS problem for archetype analysis is given in Equation 6.5. The decay curves are obtained by plotting different RSS values against corresponding p values. Model selection in RSS decay curves is facilitated by observing a prominent knee region in the curve, where the p corresponding to the knee is the optimal model choice.

In the simulations we conducted on high-noise datasets, it is observed that there are no clear knee regions since the curves tend to flatten out (as shown in Figure 6.6.1 (b)). A larger p will lead to better approximation of the dataset but not without undesirable fitting to the noise in the dataset. Therefore, RSS decay curves for model selection can be unreliable in high-noise settings. Model selection in high-noise settings using BIC is also problematic in these models. The classic BIC computation is: $BIC(df) = n \cdot \log(\frac{RSS(df)}{n}) + df \log(n)$ where n is the number of observations and df is the number of free parameters to be estimated. The first term $n \cdot \log(\frac{RSS(df)}{n})$ represents the goodness of fit and the second term $df \log(n)$ measures complexity. For conventional models, $df = p$. For higher values of p, the complexity term is overpenalised. Thus, BIC favours models with lower p leading to model underfitting. This is clearly exhibited in Figure 6.6.1 (d).

Sensitivity to initialisation of archetypes. Conventional methods are not only dependant on the input p but also on the initialisation of $Z_{p\times d}$. For example, in datasets consisting of observations drawn from a non-uniform distribution (Figure 6.2.1 (centre)) or having a structure like clusters of compact convex sets (Figure 6.2.1 (right)), conventional methods will find it hard to approximate the noisy observations since it requires a sound *a priori* knowledge of not only p but also their initialisation (Step **A**, Algorithm 2). Our experimental results (Section 6.6.1) confirm that a right choice for p as well as a good initialisation of archetypes at the onset of the algorithm is necessary for the proper functioning of conventional methods.

6.4 Automatic Detection of the Number of Archetypes

In this work, the aforementioned drawbacks of the conventional models – unreliable RSS/BIC curves for model selection in high noise cases and the sensitivity to p and initialisation of $Z_{p\times d}$ – are addressed jointly. To facilitate automatic archetype extraction, the archetype analysis problem is stated using a Group-Lasso formulation (Yuan and Lin (2006)) together with a well-defined criterion for model selection. A refresher to Group-Lasso is presented in Appendix 8.2. The fact that we can also efficiently sample the solution path of the Group-Lasso offers a well-defined model selection procedure.

6.4.1 Sparse Archetype Selection using the Group-Lasso

We start with a $n \times d$ data matrix X as before. Since n is the maximum number of archetypes that would be needed to represent data, we consider a $n \times d$ matrix Z (instead of $p \times d$ as referred in Section 6.3) which assumes that at most n archetypes are required to represent data. Our goal is to formulate an optimisation problem for identifying a sparse set of archetypes which translates to obtaining a sparse matrix Z where most of the rows are zero. Hence the non-zero rows of Z will culminate as the selected archetypes of the data.

This type of a sparsity attainment can be related to the Group-Lasso formulation (as defined in Yuan and Lin (2006)) which involves solving a linear regression problem with the goal of achieving grouped sparsity in the regression coefficients. The solution path of the Group-Lasso is efficiently computed using a fast active-set algorithm defined in Roth and Fischer (2008). As in the Group-Lasso, we use similar constraints on the matrix Z to impose grouped sparsity where the groups are the rows of the matrix and the aim is to obtain sparsity at a group level. This is achieved by imposing a $\ell_{1,2}$ norm constraint on the rows of the matrix Z. The modified optimisation problem with $X \in \mathbb{R}^{n\times d}$, $Z \in \mathbb{R}^{n\times d}$, $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times n}$, is now:

$$\begin{aligned}(\hat{A}, \hat{B}) &= \text{argmin}_{A,B}\|X - AZ\|_2^2 = \text{argmin}_{A,B}\|X - ABX\|_2^2 \\ &\text{s.t.} \quad b_{ij} \geq 0, \;\; \sum_{j=1}^{n} b_{ij} \leq 1, \;\; a_{ij} \geq 0, \;\; \sum_{j=1}^{n} a_{ij} \leq 1 \\ &\text{s.t.} \quad \sum_{j=1}^{n} \|z_j\|_2 \leq \kappa \qquad (\ell_{1,2} \text{ norm on rows of } Z)\end{aligned} \tag{6.6}$$

where κ is the tuning parameter.

To solve this optimisation problem we use the same alternating least squares idea used in previous methods, however with several algorithmic changes namely:

1. To compute the constrained optimised set of coefficients $\{a_{ij}\}$ and $\{b_{ij}\}$ using Equation 6.6 (and steps B and D respectively in Algorithm 2), we implement the *Monotone Incremental Forward Stage-wise Regression* (MIFSR) (Hastie et al. (2007)). More details are given in Section 6.4.2.

2. Instead of solving the intermediate archetypes $\hat{Z} = \text{argmin}_Z\|X - AZ\|_2^2$ using ordinary least squares *(OLS)* (step C in Alg. 2), we now introduce the Group-Lasso optimisation step. This

is elaborated in Section 6.4.3. Since the solution path of Group-Lasso can be sampled at steps of κ using the fast active-set algorithm (Roth and Fischer (2008)), BIC can be computed at these κ steps allowing a well-defined model selection procedure. Model selection using BIC is described in Section 6.5.

To further accelerate our Group-Lasso based archetype analysis model, we utilise two preprocessing steps.

1. Since archetype analysis is sensitive to outliers, a preprocessing step to remove them using the Outlier-Pursuit technique (Xu et al., 2010) is implemented.

2. The archetypes are located in close vicinity of the convex hull of the dataset. Making use of this fact, archetypal *candidates* can be preselected to be amongst those points close to the convex hull. This approach was used in Bauckhage and Thurau (2009).

The new algorithm incorporating these changes is given in Algorithm 3.

Algorithm 3 Group-Lasso extension for archetype analysis

Preprocessing X:
- Removal of outliers using the Outlier Pursuit method (Xu et al., 2010).
- Preselecting archetypal candidates (Bauckhage and Thurau, 2009).

A : **Initialise** $Z_{n\times d}$

repeat

B : **Determine** coefficients a_{ij} by minimising $\|\boldsymbol{x}_i - Z'\boldsymbol{a}_i\|_2^2$ s.t. $a_{ij} \geq 0$ and $\sum_{j=1}^{n} a_{ij} \leq 1$ for i^{th} row, $i = (1, \ldots, n)$ using MIFSR (Hastie et al., 2007). Refer Algorithm 4.

C : **Solve** Equation (6.8) for $\boldsymbol{z}^{GL}$ to obtain $\hat{\boldsymbol{z}}^{GL}$ using the Active-set algorithm in Roth and Fischer (2008). Refer Algorithm 5.

D : **Determine** coefficients b_{ij} by minimising $\|\hat{\boldsymbol{z}}_j - X'\boldsymbol{b}_j\|_2^2$ s.t. $b_{ij} \geq 0$ and $\sum_{j=1}^{n} b_{ij} \leq 1$ for i^{th} row, $i = (1, \ldots, n)$ using MIFSR (Hastie et al., 2007). Refer Algorithm 4.

E : **Update** the archetypes by setting $Z = BX$.

until *end criterion*

Next, the algorithmic changes are discussed in detail.

6.4.2 Monotone Incremental Forward Stage-wise Regression (MIFSR)

To compute the constrained optimised set of coefficients $\{a_{ij}\}$ and $\{b_{ij}\}$ using Equation 6.6 (and steps B and D respectively in Algorithm 2), we implement the Monotone Incremental Forward Stage-wise Regression (MIFSR) as introduced in Hastie et al. (2007). This heuristic is used for closely approximating Equation 6.6 to reduce the computational complexity further rather than directly solving the quadratic program.

For instance, the respective optimisation problem in terms of the MIFSR for step B can be written

as:

$$\min_{\boldsymbol{a}_i} \|\boldsymbol{x}_i - Z'\boldsymbol{a}_i\|_2^2 \quad \text{s.t } a_{ij} \geq 0 \quad \text{and} \quad \sum_{j=1}^{n} a_{ij} \leq 1 \quad \text{for } i^{th} \text{ row } (i = 1 \dots n\,). \tag{6.7}$$

Algorithm 4 depicts MIFSR for step B that involves the optimisation of $\boldsymbol{a}_i$. Step D involves the similar optimisation for $\boldsymbol{b}_j$.

Algorithm 4 MIFSR algorithm for step B

1: Start with $\boldsymbol{r} = \boldsymbol{x}_i - mean(\boldsymbol{x}_i)$, $a_{ij} = 0$.
2: Find predictor $\hat{\boldsymbol{z}}_j$ most positively correlated with $\boldsymbol{r}$.
3: Update $a_{ij} \leftarrow a_{ij} + \epsilon$. ($\epsilon$ is a predefined stepsize parameter)
4: Update $\boldsymbol{r} \leftarrow \boldsymbol{r} - \epsilon\hat{\boldsymbol{z}}_j$.
5. Repeat steps 2 and 3 until no predictor has any correlation with $\boldsymbol{r}$.

Complexity Analysis for MIFSR. By construction, Algorithm 4 terminates after κ/ϵ steps meaning that there is a fixed number of iterations that neither depend on n or p but only on κ and ϵ which are constant in this setting. The only cost involved then would be that of the correlation (Step 2 of Algorithm 4) that involves a matrix multiplication between a matrix of $\mathbb{R}^{n\times d}$ and a vector $\in \mathbb{R}^d$, thereby having a worst-case complexity of $\mathcal{O}(nd)$.

6.4.3 Group-Lasso optimisation step

The next modification to Algorithm 2 is the computation of intermediate archetypes. Instead of solving $\hat{Z} = \text{argmin}_Z \|X - AZ\|_2^2$ using ordinary least squares *(OLS)* (Step B of Algorithm 2), we now introduce the Group-Lasso optimisation step:

$$\hat{\boldsymbol{z}}^{GL} = \text{argmin}_{\boldsymbol{z}^{GL}} \|\boldsymbol{x}^{GL} - \mathcal{A}\boldsymbol{z}^{GL}\|_2^2 \quad \text{s.t.} \quad \sum_{j=1}^{n} \|\boldsymbol{z}_j^{GL}\|_2 \leq \kappa\,, \tag{6.8}$$

where in terms of the standard Group-Lasso formulation, $\mathcal{A} \in \mathbb{R}^{nd\times nd}$, $\boldsymbol{x}^{GL} \in \mathbb{R}^{nd}$ and $\boldsymbol{z}^{GL} \in \mathbb{R}^{nd}$ i.e.

$$\mathcal{A} = \begin{pmatrix} \boldsymbol{a}_1 & \mathbf{0}_n & \cdots & \mathbf{0}_n & & \boldsymbol{a}_n & \mathbf{0}_n & \cdots & \mathbf{0}_n \\ \mathbf{0}_n & \boldsymbol{a}_1 & \mathbf{0}_n & \cdots & & \mathbf{0}_n & \boldsymbol{a}_n & \mathbf{0}_n & \cdots \\ & \ddots & & & \cdots & & \ddots & & \\ \mathbf{0}_n & \mathbf{0}_n & \cdots & \boldsymbol{a}_1 & & \mathbf{0}_n & \mathbf{0}_n & \cdots & \boldsymbol{a}_n \end{pmatrix}, \quad \boldsymbol{x}^{GL} = \begin{pmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_n \end{pmatrix}, \quad \boldsymbol{z}^{GL} = \begin{pmatrix} \boldsymbol{z}_1 \\ \vdots \\ \boldsymbol{z}_n \end{pmatrix}.$$

Solving equation 6.8 gives the entire ensemble of solutions for Z that traces the different models corresponding to $p = (1, \cdots, n)$. This is solved using the fast active-set algorithm described in Roth and Fischer (2008) that uses a projected gradient method (Kim et al. (2006)) and that allows to sample the solution path at various steps of κ. Algorithms 5 and 6 present the details for the active-set and projected-gradient methods respectively.

Algorithm 5 Active-set Algorithm

A : **Initialise** : set active set $AS = \{j_0\}$, $\boldsymbol{z}_{j_0}^{GL}$ arbitrary with $\left\|\boldsymbol{z}_{j_0}^{GL}\right\|_2 = \kappa, active_\kappa = 0, step_\kappa$, $\kappa_{set} = \{\}$.

B : **Iterate** :

$active_\kappa = active_\kappa + step_\kappa$; Add $active_\kappa$ to κ_{set}.

Optimise over the current AS using the projected gradient method (Kim et al., 2006) with $active_\kappa$. Refer Algorithm 6.

Define set $AS_* = \left\{j \in AS : \left\|\boldsymbol{z}_j^{GL}\right\|_2 > 0\right\}$.

Adjust the active set $AS = AS_*$.

C : **Model Selection** :

Compute BIC scores for κ steps in κ_{set} using Equation 6.9.

Find κ_{min}, the κ value that minimises the BIC curve.

For $j \in AS$ at $\kappa_{min} \rightarrow$ construct $\hat{\boldsymbol{z}}^{GL}$.

D : **Return**: $\hat{\boldsymbol{z}}^{GL}$

Algorithm 6 Projection onto the $\ell_{1,2}$-norm ball

B1 : **Gradient** : At time $t-1$, set $\boldsymbol{d} = \left(\boldsymbol{z}^{GL^{t-1}} - s\nabla_{\boldsymbol{z}^{GL}}(\|\boldsymbol{x}^{GL} - \mathcal{A}\boldsymbol{z}^{GL}\|_2^2)\right)$ and $AS_{temp} = AS$, where s is the step-size parameter.

B2 : **Projection** :

$\mathcal{M}_j = \|\boldsymbol{d}_j\|_2 + \frac{\kappa - \sum_j \|\boldsymbol{d}_j\|_2}{|AS|} \quad \forall j \in AS_{temp}$.

If $\mathcal{M}_j \geq 0 \ \forall j \in AS_{temp}$

Go to **B3**

Else

Update $AS_{temp} = \{j : \mathcal{M}_j > 0\}$

Repeat **B2**.

B3 : **New solution**:

$\forall j \in AS_{temp}$, set $\boldsymbol{z}_j^{GL^t} = \boldsymbol{d}_j \frac{\mathcal{M}_j}{\|\boldsymbol{d}_j\|_2}$.

For the rest $j \in AS$, $j \notin AS_{temp}$, set $\boldsymbol{z}_j^{GL^t} = 0$.

B4 : **Return**: $\boldsymbol{z}^{GL^t}$

6.4.4 Further Acceleration of our Algorithm

To further accelerate our archetype analysis model, we utilise two preprocessing steps as described below.

Dimensionality reduction with robust PCA. The first aspect of preprocessing involves dimensionality reduction which aims at reducing d. Real-world datasets are usually of high dimensions which call for the use of dimensionality reduction techniques such as PCA that project the data to a low-dimensional manifold. It becomes relevant to use PCA in finding such low-rank projections in the context of archetype analysis. This is due to the fact that a set of convex mixtures of p archetypes cannot lie on a subspace greater than p and since convex sets are linear manifolds, the search for these linear manifolds is justified using PCA-based projections. However, PCA is highly susceptible to outliers and thus it becomes essential to filter out the outliers before performing PCA. We resort to a robust version of PCA as given in Xu et al. (2010) that deals with Outlier Pursuit.

This method involves decomposing the data matrix X as $X = X_L + X_C$ where X_L is the low-rank matrix comprising the true subspace of the non-outlier points and X_C the column-sparse matrix denoting presence of outliers. Through robust PCA, we estimate X_L that represents the uncorrupted data. Details of the method are given in Xu et al. (2010).

Preselecting the archetype candidates. After obtaining the outlier-free data matrix with reduced dimensionality, we focus our attention on reducing the number of possible archetypes from n to a lower number for computational gains. Archetypal candidates can be chosen from amongst those observations located close to the convex hull of the data.

For preselecting the archetype candidates we use the approach as in Bauckhage and Thurau (2009). Here they consider that the convex hull can be seen as a polytope in $\mathbb{R}^d$. Given a transformation matrix $\mathcal{M} \in \mathbb{R}^{d^* \times d}$ and a real vector $t \in \mathbb{R}^{d^*}$, the main theorem in polytope theory states that every image of a polytope under an affine transformation $\pi : \boldsymbol{x} \mapsto \mathcal{M}\boldsymbol{x} + t$ is also a polytope (Henk et al. (1997), Bauckhage and Thurau (2009)). This means that for every point on the convex hull of X, there exists a linear map under whose image the point also appears on the convex hull of the image. Using this, the technique is then to take the union of as many such points forming the convex hull of different 2D projections of X with the view to recover the true convex hull of X. Calculating the convex hull in 2D is easier since the worst-case combinatorial complexity of calculating the convex hull for n observations in d dimensions increases exponentially with d as $\mathcal{O}(n^{\lfloor d/2 \rfloor + 1})$ (Skiena, 1997).

6.5 Model Selection

Selecting a sparse set of archetypes according to Equation 6.8, however, involves tuning the parameter κ which controls the level of sparsity in the solution. Since different κ return different parsimonious models, model selection is required to select one amongst these models. Although cross-validation is generally used for model selection, it can tend to be computationally expensive.

We use the BIC scoring mechanism for Group-Lasso as detailed in Yuan and Lin (2006) for model selection. BIC, proposed by Schwarz (1978), is an information criterion based on the regression models goodness of fit and complexity. An increase in the model's goodness of fit increases BIC

whereas an increase in the model complexity, penalises BIC. Since a lower BIC indicates a favourable model, BIC penalises larger models more heavily and would tend to prefer smaller models which in our problem setting would mean preferring a smaller p.

The BIC score for Group-Lasso is given as:

$$\text{BIC}(\hat{\mu} \equiv \hat{\mathcal{A}}\hat{\boldsymbol{z}}^{GL}) = \frac{\|\boldsymbol{x}^{GL} - \hat{\mu}\|_2^2}{\tau\sigma^2} + \frac{\log(\tau)}{\tau} \cdot \widehat{df}(\hat{\mu})\,, \tag{6.9}$$

where $\tau = nd$, $\hat{\mathcal{A}}$ and $\hat{\boldsymbol{z}}^{GL}$ are the estimated values of $\mathcal{A}$ and $\boldsymbol{z}^{GL}$ based on a particular κ value. The degrees of freedom of Group-Lasso, $\widehat{df}(\hat{\mu})$, can be computed using both the *approximate* (Yuan and Lin (2006)) and *exact* (Vaiter et al. (2012)) forms which are discussed in turn below.

6.5.1 'Approximate' degrees of freedom for Group-Lasso

The approximate degrees of freedom for Group-Lasso (Yuan and Lin, 2006) is given as:

$$\widehat{df}(\hat{\mu}) = \underbrace{\sum_{j=1}^{n} \mathcal{I}(\|\hat{\boldsymbol{z}}_j^{GL}\|_2 > 0)}_{|AS|} + \sum_{j=1}^{n} \frac{\|\hat{\boldsymbol{z}}_j^{GL}\|_2}{\|\hat{\boldsymbol{z}}_j^{GL_{LS}}\|_2}(d-1)\,, \tag{6.10}$$

where $\mathcal{I}(\cdot)$ is the indicator function, $\hat{\boldsymbol{z}}^{GL}$ is the estimated value of $\boldsymbol{z}^{GL}$, $|AS|$ is the number of active groups where the set of active groups is $AS := \left\{j : \hat{\boldsymbol{z}}_j^{GL} \neq 0\right\}$, $\|\hat{\boldsymbol{z}}_j^{GL}\|_2$ is the ℓ_2 norm of $\hat{\boldsymbol{z}}_j^{GL}$ and $\|\hat{\boldsymbol{z}}_j^{GL_{LS}}\|_2$ is the ℓ_2 norm of the least-square estimate of $\hat{\boldsymbol{z}}_j^{GL}$.

6.5.2 'Exact' degrees of freedom for Group-Lasso

The exact degrees of freedom for Group-Lasso (Vaiter et al., 2012) is given as:

$$\begin{aligned} \widehat{df}(\hat{\mu}) &= \text{tr}\left(\mathcal{A}_{AS}\left(\mathcal{A}'_{AS}\mathcal{A}_{AS} + \tau\,\mathbb{Z} \circ \mathbb{P}\right)^{-1}\mathcal{A}'_{AS}\right) \\ &= \text{tr}\left(\left(\mathcal{A}'_{AS}\mathcal{A}_{AS} + \tau\,\mathbb{Z} \circ \mathbb{P}\right)^{-1}\mathcal{A}'_{AS}\mathcal{A}_{AS}\right) \end{aligned} \tag{6.11}$$

where $\tau > 0$ and is the regularisation parameter, $\circ$ is the Hadamard (element-wise) multiplication of two matrices having same dimensions and $\mathcal{A}_{AS}$ is the $\mathcal{A}$ matrix whose columns correspond to the set of active groups $AS := \left\{j : \hat{\boldsymbol{z}}_j^{GL} \neq 0\right\}$.

We look in detail at the terms within the inversion of Equation 6.11.

$$\mathbb{Z} = \begin{pmatrix} \mathfrak{z} & \mathbf{0}_{|AS|\times|AS|} & \cdots & \mathbf{0}_{|AS|\times|AS|} \\ \mathbf{0}_{|AS|\times|AS|} & \mathfrak{z} & \mathbf{0}_{|AS|\times|AS|} & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{|AS|\times|AS|} & \mathbf{0}_{|AS|\times|AS|} & \cdots & \mathfrak{z} \end{pmatrix} \in \mathbb{R}^{|AS|d\times|AS|d}$$

where $|AS|$ is the number of active groups.

$$\mathfrak{z} = \begin{pmatrix} \frac{1}{\|\hat{\boldsymbol{z}}^{GL}_{AS[1]}\|} & 0 & \cdots & 0 \\ 0 & \frac{1}{\|\hat{\boldsymbol{z}}^{GL}_{AS[2]}\|} & 0 & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\|\hat{\boldsymbol{z}}^{GL}_{AS[|AS|]}\|} \end{pmatrix} \in \mathbb{R}^{|AS|\times|AS|}.$$

The projection matrix $\mathbb{P}$ is given as:

$$\mathbb{P} = \begin{pmatrix} \mathfrak{p} & \mathbf{0}_{|AS|\times|AS|} & \cdots & \mathbf{0}_{|AS|\times|AS|} \\ \mathbf{0}_{|AS|\times|AS|} & \mathfrak{p} & \mathbf{0}_{|AS|\times|AS|} & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{|AS|\times|AS|} & \mathbf{0}_{|AS|\times|AS|} & \cdots & \mathfrak{p} \end{pmatrix} \in \mathbb{R}^{|AS|d\times|AS|d}$$

and

$$\mathfrak{p} = I_{|AS|} - \begin{pmatrix} \frac{\hat{\boldsymbol{z}}^{GL}_{AS[1]}(\hat{\boldsymbol{z}}^{GL}_{AS[1]})'}{\|\hat{\boldsymbol{z}}^{GL}_{AS[1]}\|^2} & 0 & \cdots & 0 \\ 0 & \frac{\hat{\boldsymbol{z}}^{GL}_{AS[2]}(\hat{\boldsymbol{z}}^{GL}_{AS[2]})'}{\|\hat{\boldsymbol{z}}^{GL}_{AS[2]}\|^2} & 0 & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{\hat{\boldsymbol{z}}^{GL}_{AS[|AS|]}(\hat{\boldsymbol{z}}^{GL}_{AS[|AS|]})'}{\|\hat{\boldsymbol{z}}^{GL}_{AS[|AS|]}\|^2} \end{pmatrix} \in \mathbb{R}^{|AS|\times|AS|}$$

where $\mathfrak{p}$ is the projector orthogonal to $\hat{\boldsymbol{z}}^{GL}_{AS}$.

$\mathbb{Z}$ and $\mathbb{P}$ are shown to be block-diagonal matrices. Since the expensive operation in Equation 6.11 is that of the matrix inversion of $(\mathcal{A}'_{AS}\mathcal{A}_{AS}+\tau\,\mathbb{Z}\circ\mathbb{P})^{-1}$, we exploit a block-diagonal structure of the terms present in the inversion. For this, we permute the columns in $\mathcal{A}_{AS}$ to obtain the block-diagonal structure as shown below.

$$\mathcal{A}_{AS} = \begin{pmatrix} \boldsymbol{a}_{AS[1]} & \boldsymbol{a}_{AS[2]} & \cdots & \boldsymbol{a}_{AS[|AS|]} & \ldots & \mathbf{0}_n & \mathbf{0}_n & \cdots & \mathbf{0}_n \\ \mathbf{0}_n & \mathbf{0}_n & \mathbf{0}_n & \cdots & \ldots & \mathbf{0}_n & \mathbf{0}_n & \mathbf{0}_n & \cdots \\ & \ddots & & & \ldots & & \ddots & & \\ \mathbf{0}_n & \mathbf{0}_n & \cdots & \mathbf{0}_n & \ldots & \boldsymbol{a}_{AS[1]} & \boldsymbol{a}_{AS[2]} & \cdots & \boldsymbol{a}_{AS[|AS|]} \end{pmatrix} \in \mathbb{R}^{nd\times|AS|d}.$$

Therefore $\mathcal{A}'_{AS}\mathcal{A}_{AS}$ takes the following block-diagonal structure:

$$\mathcal{A}'_{AS}\mathcal{A}_{AS} = \begin{pmatrix} \mathtt{A} & \mathbf{0}_{|AS|\times|AS|} & \cdots & \mathbf{0}_{|AS|\times|AS|} \\ \mathbf{0}_{|AS|\times|AS|} & \mathtt{A} & \mathbf{0}_{|AS|\times|AS|} & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{|AS|\times|AS|} & \mathbf{0}_{|AS|\times|AS|} & \cdots & \mathtt{A} \end{pmatrix} \in \mathbb{R}^{|AS|d\times|AS|d}$$

where $\mathtt{A}$ is a symmetric matrix given as:

$$\mathtt{A} = \begin{pmatrix} \boldsymbol{a}'_{AS[1]}\boldsymbol{a}_{AS[1]} & \boldsymbol{a}'_{AS[1]}\boldsymbol{a}_{AS[2]} & \cdots & \boldsymbol{a}'_{AS[1]}\boldsymbol{a}_{AS[|AS|]} \\ \boldsymbol{a}'_{AS[2]}\boldsymbol{a}_{AS[1]} & \boldsymbol{a}'_{AS[2]}\boldsymbol{a}_{AS[2]} & \cdots & \boldsymbol{a}'_{AS[2]}\boldsymbol{a}_{AS[|AS|]} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{a}'_{AS[|AS|]}\boldsymbol{a}_{AS[1]} & \boldsymbol{a}'_{AS[|AS|]}\boldsymbol{a}_{AS[2]} & \cdots & \boldsymbol{a}'_{AS[|AS|]}\boldsymbol{a}_{AS[|AS|]} \end{pmatrix} \in \mathbb{R}^{|AS|\times|AS|}.$$

This indicates that to construct the block-diagonal $\mathcal{A}'_{AS}\mathcal{A}_{AS}$, it is sufficient to compute $\mathtt{A}$ only once and diagonalise it using eigenvalue decomposition i.e. $\mathtt{A} = \mathtt{V}\Lambda\mathtt{V}'$ where $\mathtt{V} \in \mathbb{R}^{|AS|\times|AS|}$ consists of the eigenvectors of $\mathtt{A}$ and $diag(\Lambda) = (\lambda_1, \cdots, \lambda_{|AS|})$ is the diagonal matrix with the corresponding eigenvalues, λ_i. Using $\mathtt{A} = \mathtt{V}\Lambda\mathtt{V}'$, the per-block operation for the inversion terms in Equation 6.11 can be written as:

$$(\mathtt{A} + \tau_{\mathfrak{z}} \circ \mathfrak{p}\mathtt{I})^{-1} = ((\mathtt{V}\Lambda\mathtt{V}') + \tau_{\mathfrak{z}} \circ \mathfrak{p}\mathtt{I})^{-1}. \tag{6.12}$$

Equation 6.11 now reads as:

$$\widehat{df}(\hat{\mu}) = \mathrm{tr}(\mathtt{MN})$$

where

$$\mathtt{M} = \begin{pmatrix} ((\mathtt{V}\Lambda\mathtt{V}') + \tau_{\mathfrak{z}} \circ \mathfrak{p}\mathtt{I})^{-1} & \mathbf{0}_{|AS|\times|AS|} & \mathbf{0}_{|AS|\times|AS|} \\ \mathbf{0}_{|AS|\times|AS|} & ((\mathtt{V}\Lambda\mathtt{V}') + \tau_{\mathfrak{z}} \circ \mathfrak{p}\mathtt{I})^{-1} & \cdots \\ \vdots & \vdots & \ddots \\ \mathbf{0}_{|AS|\times|AS|} & \mathbf{0}_{|AS|\times|AS|} & ((\mathtt{V}\Lambda\mathtt{V}') + \tau_{\mathfrak{z}} \circ \mathfrak{p}\mathtt{I})^{-1} \end{pmatrix}$$

and

$$\mathtt{N} = \begin{pmatrix} \mathtt{V}\Lambda\mathtt{V}' & \mathbf{0}_{|AS|\times|AS|} & \mathbf{0}_{|AS|\times|AS|} \\ \mathbf{0}_{|AS|\times|AS|} & \mathtt{V}\Lambda\mathtt{V}' & \cdots \\ \vdots & \vdots & \ddots \\ \mathbf{0}_{|AS|\times|AS|} & \mathbf{0}_{|AS|\times|AS|} & \mathtt{V}\Lambda\mathtt{V}' \end{pmatrix}.$$

Therefore,

$$
\begin{aligned}
\widehat{df}(\hat{\mu}) &= \operatorname{tr}\begin{pmatrix} ((\mathtt{V}\Lambda\mathtt{V}') + \tau_{\mathfrak{Z}} \circ \mathfrak{p}\mathtt{I})^{-1}\mathtt{V}\Lambda\mathtt{V}' & \mathbf{0}_{|AS|\times|AS|} & \mathbf{0}_{|AS|\times|AS|} \\ \mathbf{0}_{|AS|\times|AS|} & ((\mathtt{V}\Lambda\mathtt{V}') + \tau_{\mathfrak{Z}} \circ \mathfrak{p}\mathtt{I})^{-1}\mathtt{V}\Lambda\mathtt{V}' & \cdots \\ \vdots & \vdots & \ddots \\ \mathbf{0}_{|AS|\times|AS|} & \mathbf{0}_{|AS|\times|AS|} & ((\mathtt{V}\Lambda\mathtt{V}') + \tau_{\mathfrak{Z}} \circ \mathfrak{p}\mathtt{I})^{-1}\mathtt{V}\Lambda\mathtt{V}' \end{pmatrix} \\
&= \sum_d \operatorname{tr}\Big(((\mathtt{V}\Lambda\mathtt{V}') + \tau_{\mathfrak{Z}} \circ \mathfrak{p}\mathtt{I})^{-1}\mathtt{V}\Lambda\mathtt{V}'\Big) \quad \text{(sum over traces of the block matrices on the diagonal)} \\
&= \sum_d \operatorname{tr}\Big((\mathtt{V}(\Lambda + \tau_{\mathfrak{Z}} \circ \mathfrak{p}\mathtt{I})\mathtt{V}')^{-1}\mathtt{V}\Lambda\mathtt{V}'\Big) \\
&= \sum_d \operatorname{tr}\Big(\mathtt{V}^{'-1}(\Lambda + \tau_{\mathfrak{Z}} \circ \mathfrak{p}\mathtt{I})^{-1}\mathtt{V}^{-1}\mathtt{V}\Lambda\mathtt{V}'\Big) \\
&= \sum_d \operatorname{tr}\Big(\mathtt{V}(\Lambda + \tau_{\mathfrak{Z}} \circ \mathfrak{p}\mathtt{I})^{-1}\underbrace{\mathtt{V}^{-1}\mathtt{V}}_{\mathtt{I}}\Lambda\mathtt{V}'\Big) \quad (\text{since } \mathtt{V}^{'-1} = (\mathtt{V}^{-1})' = (\mathtt{V}')' = \mathtt{V}) \\
&= \sum_d \operatorname{tr}\Big((\Lambda + \tau_{\mathfrak{Z}} \circ \mathfrak{p}\mathtt{I})^{-1}\Lambda\Big) \\
&= \underbrace{\sum_d \underbrace{\sum_{i=1}^{|AS|} \frac{\lambda_i}{\lambda_i + \frac{\tau}{\|\hat{\boldsymbol{z}}_i^{GL}\|}\Big(1 - \frac{\hat{\boldsymbol{z}}_i^{GL}(\hat{\boldsymbol{z}}_i^{GL})'}{\|\hat{\boldsymbol{z}}_i^{GL}\|^2}\Big)}}_{\text{per-block summation } |AS| \text{ times}}}_{\text{overall summation over } d \text{ blocks}}
\end{aligned}
\tag{6.13}
$$

Throughout the rest of the paper and in the experiments, we resort to using the *exact* degrees of freedom given in Equation 6.13 for BIC score computation to evaluate models obtained with different κ values.

Complexity analysis for Model Selection using BIC. Equation 6.9 computes BIC for a particular κ. Since the active-set algorithm permits sampling of the solution path at discrete sets of κ, corresponding BIC scores can be computed stepwise. Since no additional costs are involved in computing the BIC scores over the entire solution path, it renders our method to be computationally efficient. Choosing the best model from amongst parsimonious models boils down to merely observing the minimum attained in the BIC curves.

6.6 Experiments

6.6.1 Simulations

Simulation example I. We generate two datasets, one with low Gaussian noise and another with high noise using a Student-t distribution (known to confuse traditional PCA but not the robust PCA). Each dataset consists of $n = 1000$ observations in $\mathbb{R}^{10}$ generated from $p = 3$ archetypes using Setting 1 (see Section 6.2.1). The datasets are subject to preprocessing as described in Section 6.4.4:

first for dimensionality reduction using robust PCA followed by reducing the number of archetype candidates to those points residing near the convex hull of the data.

For comparison, we run our algorithm versus the conventional algorithm (Bauckhage and Thurau (2009)) for archetype detection. For Bauckhage and Thurau (2009), we compute the percentage decay in the RSS error against p. The percentage decay is given as: $100 \cdot \frac{RSS(p)}{RSS(1)}$. Refer Figure 6.6.1 (top row). In the low-noise setting, the knee in the RSS decay curve is at $p = 3$ signalling the right model whereas in the high-noise case, the curve hardly exhibits a knee region thereby making the curve uninformative for model selection.

Concomitantly, we plot the BIC scores for both noise settings for the conventional methods (Figure 6.6.1 (middle row)). The classic BIC computation is: $BIC(df) = n \cdot \log(\frac{RSS(df)}{n}) + df \log(n)$ where n is the number of observations and df is set to p. For the low-noise case, BIC has a minimum at the correct model $p = 3$ whereas for the high-noise setting, we have observed that there hardly occurs a curve with a minimum but that the BIC scores always increase with increasing p. Thus even the BIC scores fail to signal the right p in high-noise cases.

For our Group-Lasso based method, we have a more direct access to efficiently compute the effective df needed for BIC (Equation 6.13). We plot the BIC curves computed at discrete steps of κ (using Equation 6.9) versus the κ values (Figure 6.6.1 (bottom row)). In both noise settings, the BIC curves show a clear minimum that serves to determine the right p and is aptly positioned at $p = 3$, enabling automatic model selection.

Thus, we have experimentally shown that for conventional methods, although RSS/BIC curves can aid model selection in low-noise cases, in high-noise settings they tend to be unreliable. The model selection using BIC for our Group-Lasso based method performs better even in high-noise settings.

Simulation example II: Noisy convex sets generated from a non-uniform density. Figure 6.2.1 (centre) shows a noisy convex dataset $X_{1000\times100}$ generated using 4 archetypes from a non-uniform density using Setting 2 (Section 6.2.1). Figure 6.6.2 depicts the performance of conventional methods. These methods find it hard to extract archetypes as they heavily rely on p and the archetype initialisation. Plots (a)–(d) show the optimised archetypes (□) for an increasing $p = 3, 4, 6$ and 8 where these values are known *a priori*. A smaller number of archetypes (for example 3) is not sufficient to represent the observations well whereas a larger number of archetypes may or may not approximate the noisy observations depending on the initialisation of Z at the onset. For example, plots (b) and (c) show 4 and 6 archetypes respectively that have been extracted but wrongly approximate the data and in plots (e) and (f), 4 and 6 archetypes are correctly retrieved but after a good random initialisation of Z.

In Figure 6.6.3, we show the performance of our Group-Lasso based method on the same noisy convex set. Our method clearly extracts all 4 archetypes since it does not depend on any *a priori* known value for p. Moreover, Z is initialised to all the $n = 1000$ observations of X since the model considers at most n archetypes sufficient to approximate X, thereby desensitising our method to initialisations of Z.

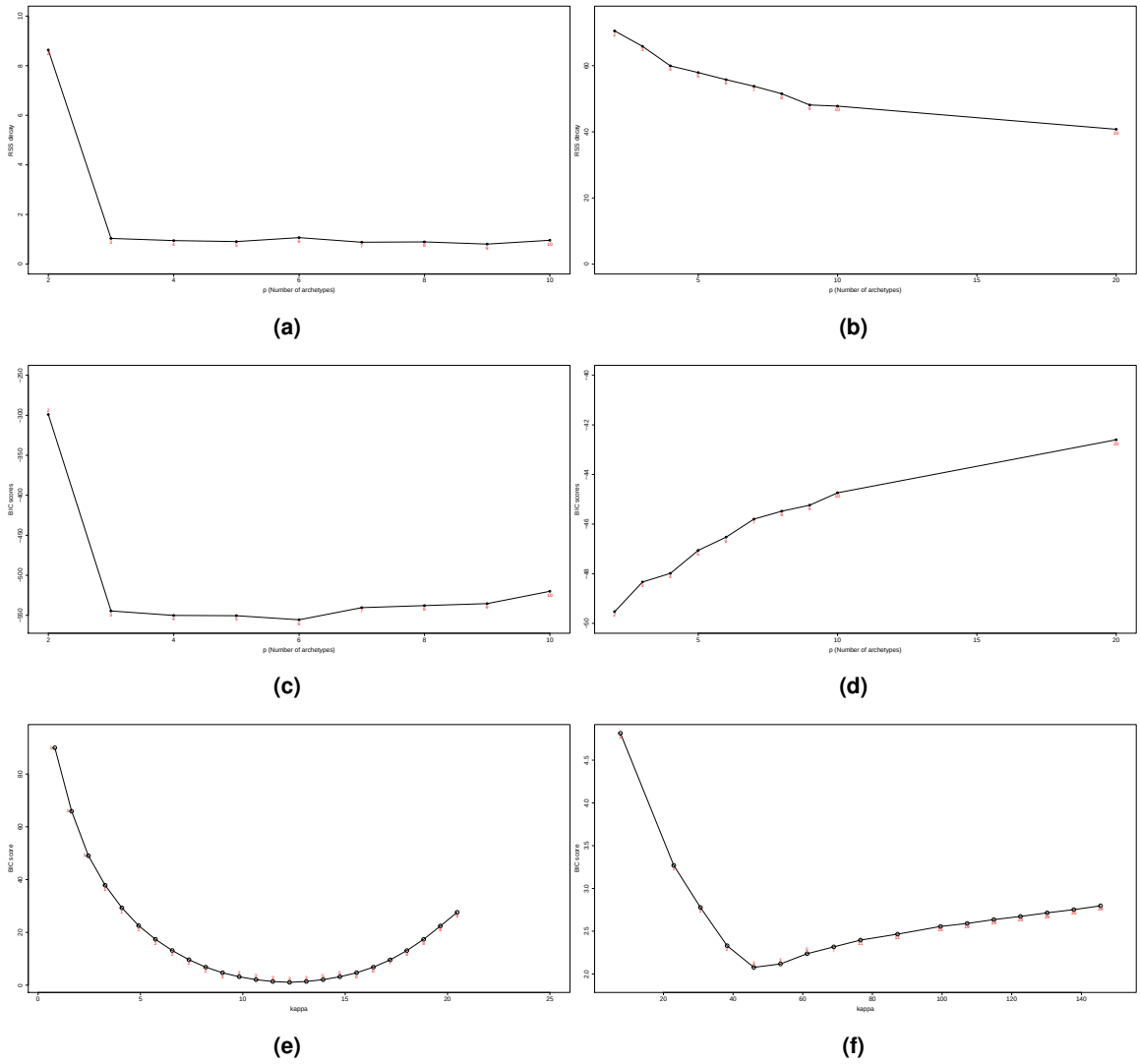
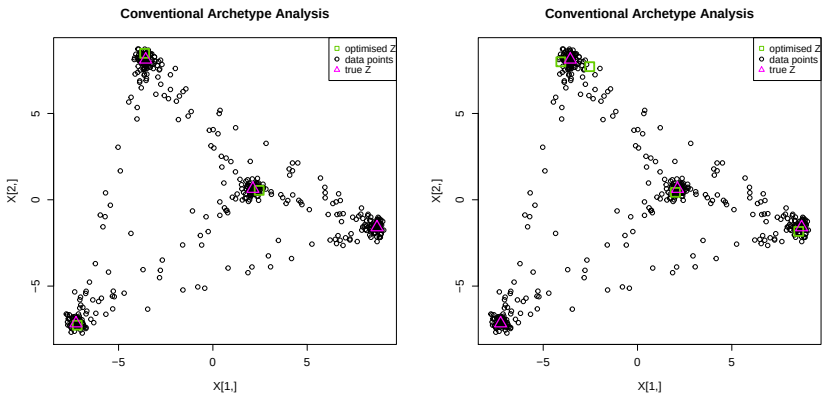
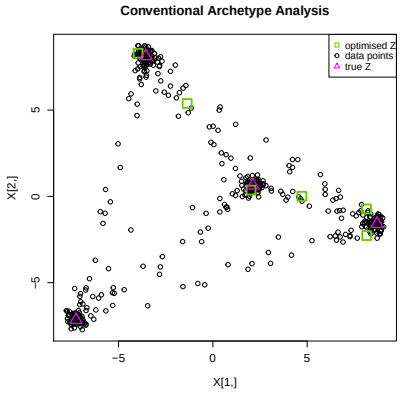


Figure 6.6.1: Comparison of the Group-Lasso based method with that of Bauckhage and Thurau (2009). Number of archetypes are indicated along the curves. $\mathbf{Left}$ column: low-Gaussian noise. $\mathbf{Right}$ column: high Student-t noise.
For Bauckhage and Thurau (2009):
(a) RSS decay curves plotted against p archetypes for a low-Gaussian noise dataset.
(b) RSS decay curves plotted against p archetypes for a high Student-t noise dataset.
(c) BIC scores plotted against p archetypes for a low-Gaussian noise dataset.
(d) BIC scores plotted against p archetypes for a high Student-t noise dataset.
For $\mathbf{Group\text{-}Lasso\ based}$:
(e) BIC scores versus κ for the low-Gaussian noise dataset.
(f) BIC scores versus κ for the high Student-t noise dataset.
Both datasets were generated using $p = 3, n = 1000$ and $d = 10$. For conventional methods, RSS/BIC curves perform well in low-noise settings whereas in high-noise cases, model selection will be cumbersome due to no clear *knee region* in the RSS curves or incorrect minimum in the BIC scores. On the other hand, our Group-Lasso based model automatically identifies 3 archetypes as shown clearly by the prominent minimum in the BIC curves, in either noise setting.

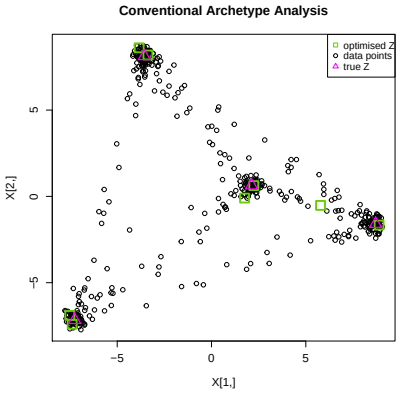


(a) Using 3 archetypes (□).

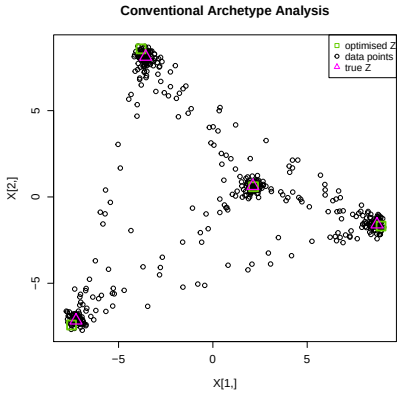
(b) Using 4 archetypes (□).



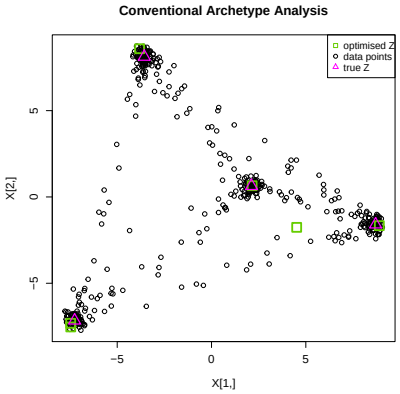
(c) Using 6 archetypes (□).



(d) Using 8 archetypes (□).



(e) Using 4 archetypes (□), after a good random initialisation of Z.



(f) Using 6 archetypes (□), after a good random initialisation of Z.

Figure 6.6.2: Performance of conventional methods on a convex set generated from a non-uniform density as shown in Figure 6.2.1 (centre): Here $X_{1000\times100}$ and $Z_{4\times100}$. Conventional methods find it hard to extract archetypes for such convex sets as they heavily rely on p and the archetype initialisation. (a)–(d): Plots showing the optimised archetypes (□) for an $a\ priori$-known increasing number of archetypes viz. $p = 3, 4, 6$ and 8. A smaller number of archetypes (for example 3) is not sufficient to represent the observations well whereas a larger number of archetypes may or may not approximate the noisy observations depending on the initialisation of Z. For example, (b) and (c) show 4 and 6 archetypes extracted but that wrongly approximate the data. (e) and (f) show 4 and 6 archetypes correctly retrieved but only after a proper random initialisation of Z. All plots show the 2-d PCA projections of the data.

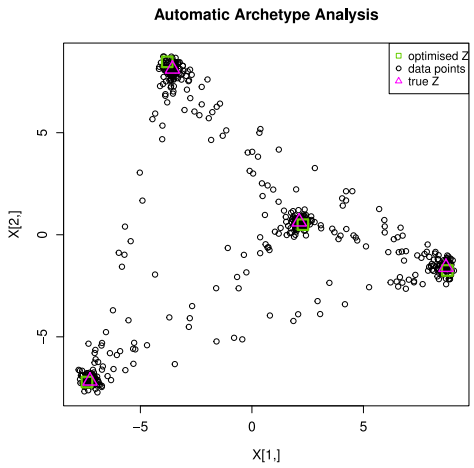


Figure 6.6.3: Group-Lasso based method on a convex set generated from a non-uniform density as given in Figure 6.2.1 (centre): Given a convex set $X_{1000\times100}$ generated from 4 archetypes (△), the Group-Lasso based method is able to automatically identify all 4 archetypes (□). Plot shows the 2-d PCA projection of the data.

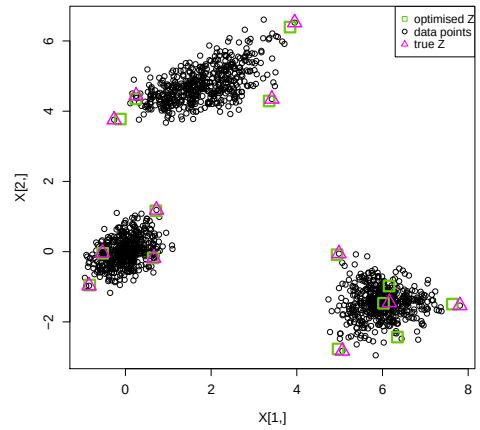
Simulation example III: Dataset containing clusters of compact convex sets. For this simulation, we compare the different methods on a dataset that contains clusters of noisy convex sets generated according to Setting 3 (Section 6.2.1). Refer Figure 6.2.1 (right). In this experiment, we generate 3 convex sets with each convex set $X_{1000\times10}$ being noisy convex combinations of $Z_{4\times10}$. Figure 6.6.4 (a) shows that the Group-Lasso based method optimises the data using 14 archetypes (□) of which all 12 original archetypes are found. Plots (b)–(d) show the optimised archetypes for conventional methods using $4, 8$ and 12 archetypes (□) respectively. The plots clearly show the inability of conventional methods to approximate the right p which again heavily depends on the initialisation of Z.

6.6.2 Real-world experiments

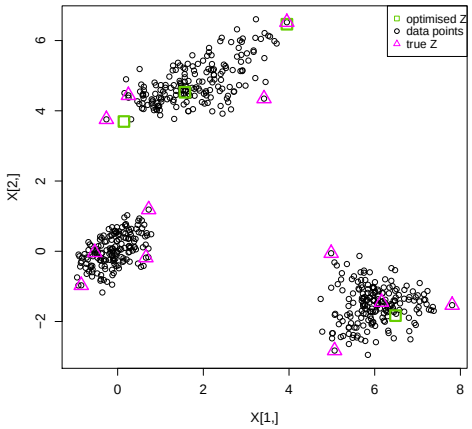
Text categorisation using Reuters Corpus Volume 1. As a real-world example, we apply our archetype analysis model for categorising texts where the focus is to obtain automatic annotations of the text corpus leading to potential new categories as opposed to manually-provided categories. Since the term frequency (TF) of a document can be described as an ideal convex mixture of words, in archetype analysis the pursuit would be to find those archetypal documents that can be meaningfully interpreted as a convex combination of legitimate words comprising one of the main categories.

We use the Reuters Corpus Volume 1 (*RCV1*), an archive of news documents manually categorised and made available through Lewis et al. (2004). The four categories reflecting the content of the corpus are *CCAT*: Corporate/Industrial, *ECAT*: Economics, *GCAT*: Government/Social and *MCAT*: Markets. The dataset we use consists of 23,149 TF-IDF normalised documents with their corresponding labels and 57,180 words. We compute the Gram matrix of this dataset and apply kernel-PCA that results in the dataset having 23,149 documents and a reduced dimensionality of 200 words.

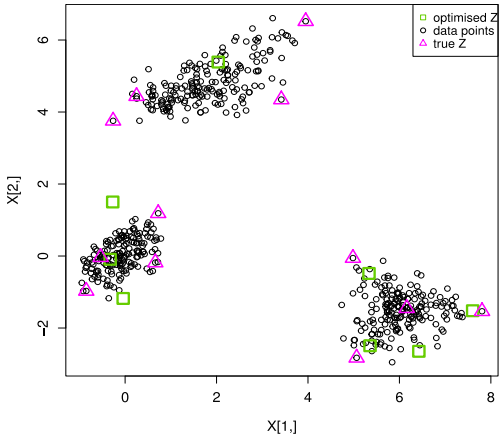
We apply our Group-Lasso based method on this corpus and retrieve 89 archetype documents



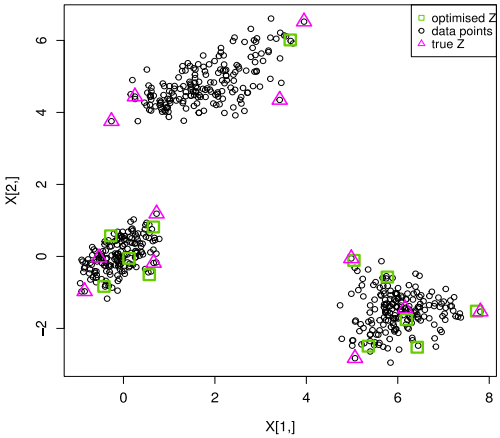
(a) Group-Lasso based method: 12 archetypes (□) are correctly extracted.



(b) Conventional methods with 4 archetypes (□).



(c) Conventional methods with 8 archetypes (□).



(d) Conventional methods with 12 archetypes (□).

Figure 6.6.4: Performance of Group-Lasso based method and conventional methods on a dataset having clusters of compact convex sets: Given a dataset that consists of 3 compact convex sets where each convex set consists of 1000 10-d observations generated from 4 10-d archetypes (△). (a) The Group-Lasso based method optimises the dataset using 14 archetypes (□) of which all 12 original archetypes are found. (b)–(d) Plots with the optimised archetypes for conventional methods using $4, 8$ and 12 archetypes (□) clearly showing the inability to approximate the right p which again heavily depends on the initialisation of Z. All plots show the 2-d PCA projections of the data.

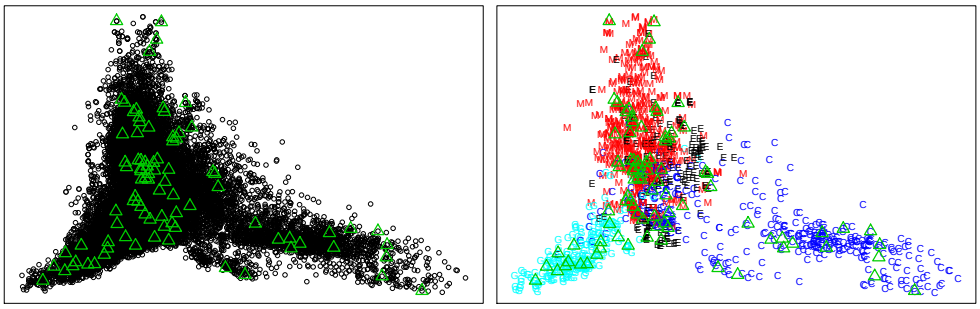


Figure 6.6.5: Archetypes (green triangles) plotted against the entire 23K documents (left) and against documents, categorically annotated of the *RCV1* corpus (right). The landscape of 89 archetypal documents is clearly spread across all four categories.

as shown in Figure 6.6.5. Analysing the archetypes identified, it is obvious that all 89 archetypes are spread out across all the four core categories. Another interesting result is that the archetypes also capture all the high-frequency terms denoting rare words present in the corpus (see Figure 6.6.6 (a and b)). Rare words are akin to the most informative words in a document and are given by their *inverse document frequency* (IDF). Thus the identified archetypes can be seen as those apex documents in the corpus meaningfully representing the core categories. Next, we plot the BIC scores using Equation 6.9 for the different parsimonious models obtained for various values of κ (see Figure 6.6.6 (c)). The BIC scores are computed by following the solution path of the Group-Lasso in successive steps of κ. The scores of each model along with the corresponding number of archetypes supporting that model are shown. We also plot the RSS curves obtained at these stepwise intervals of κ. From the plot, it is clear that the RSS curve cannot be used for reliable model selection since there is no prominent *knee* in the curve. On the other hand, the BIC curve clearly depicts a *minimum* emphasising that automatic model selection made possible by the active-set algorithm of Group-Lasso works well in reclaiming the unknown number of archetypal documents.

A word cloud [2] comprising the most informative words from amongst the 89 archetypal documents is plotted in Figure 6.6.6 (d). The higher the IDF of rare words, the bigger the font it is depicted in the word cloud. Next, the per category IDF is plotted in Figure 6.6.7 along with the corresponding word clouds. Thus, one can use the archetype analysis method to further finegrain existing major categories into their archetypal documents and also delineate news highlights or crucial topics per category.

Archetypal compounds from amongst active chemical compounds. As a second real-world example we use our model to identify archetypal compounds from a list of 456 active chemical compounds which also includes all the 25 currently available anti-HIV drugs. The entire list of active compounds is found at `http://www.dtp.nci.nih.gov/docs/aids/aids_data.html`.

We begin by first accessing the SMILES strings (Weininger, 1988) that encode the chemical structure of the compounds. Using these strings, the *chemical hashed fingerprints* of compounds

[2] Word clouds are plotted using the R package *wordcloud* (Fellows, 2012).

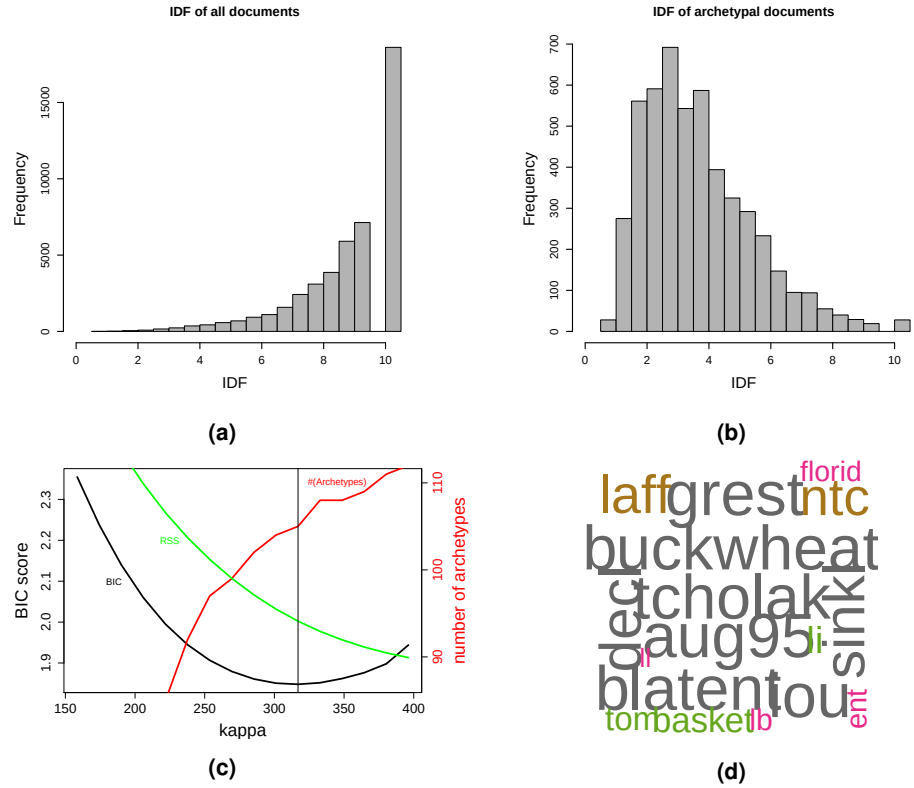


Figure 6.6.6: Inverse Document Frequency (IDF) of (a) all documents and (b) archetype documents. The archetype documents successfully capture the high-frequency terms present in the *RCV1* corpus. (c) BIC scores and number of archetypes plotted against different models obtained for various κ values. (d) Word cloud comprising informative words (tokens) in the 89 archetypal documents. The tokens comprise of stock market ticker symbols or commodities procured.

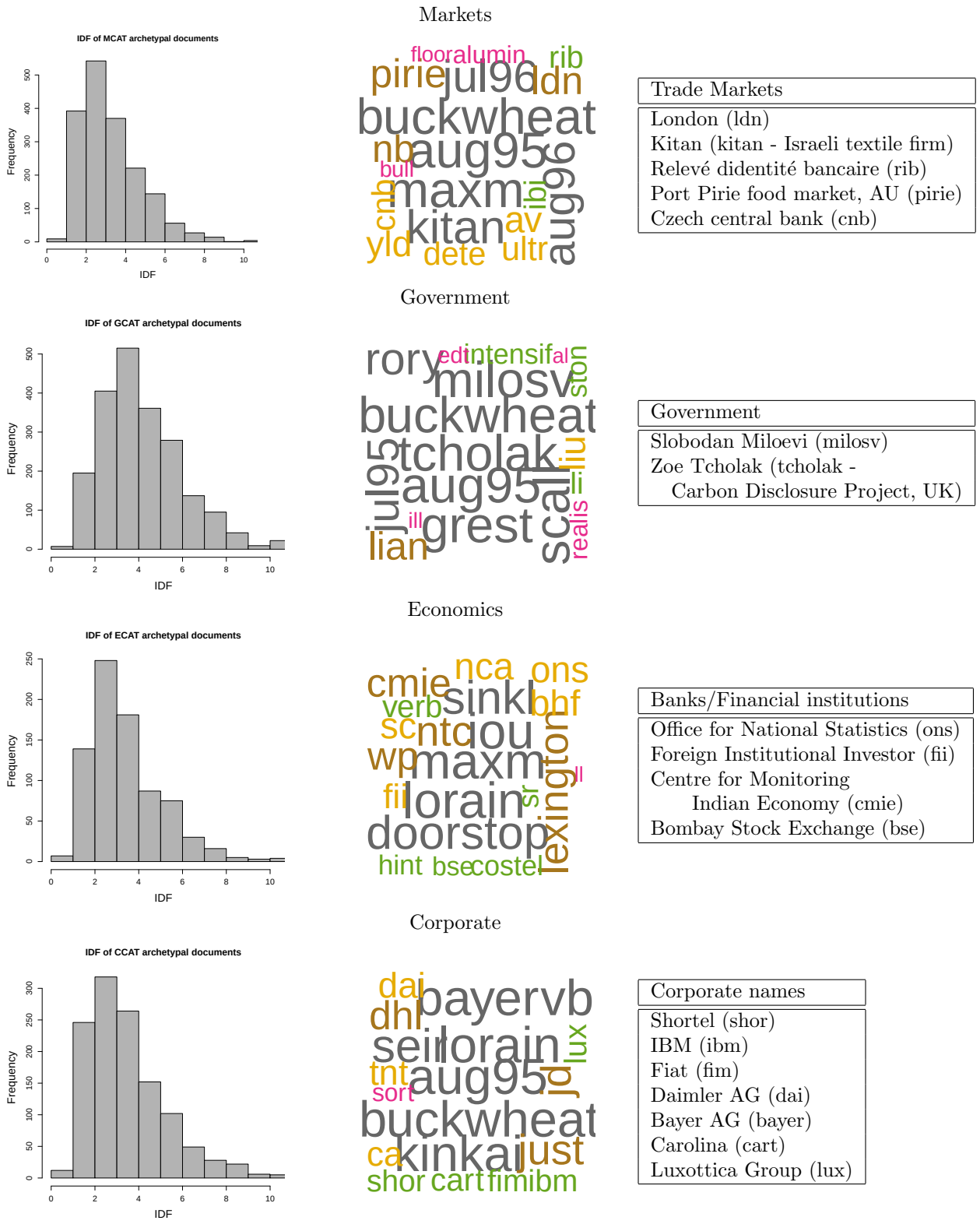


Figure 6.6.7: IDF (left column), Word cloud (central column) and trending topics (right column) of archetypal documents per category: Markets, Government, Economics and Corporate. The word clouds denote the most important words (tokens) constituting archetypal documents and are given by the rarity of the words obtained from the IDF. The rarer the words are, the more information they contain and the bigger the font they are depicted in. Here, the words having an IDF ≥ 4 are plotted.

Such dominating words can possibly be used to carve out trending topics or get an insight into words that contribute highly to a particular category. The tokens are stock market ticker symbols (*NASDAQ* or *NYSE*), commodities procured or names of government officials.

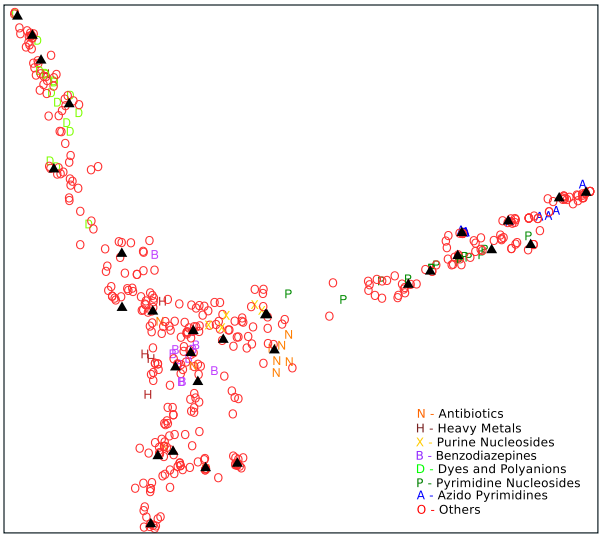
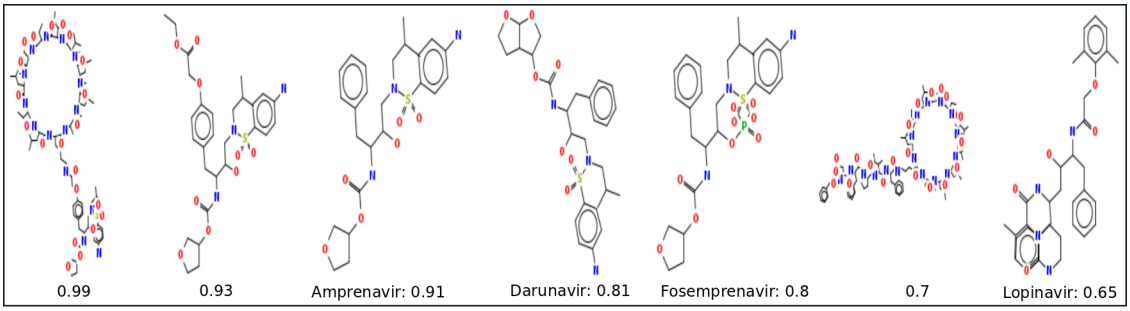


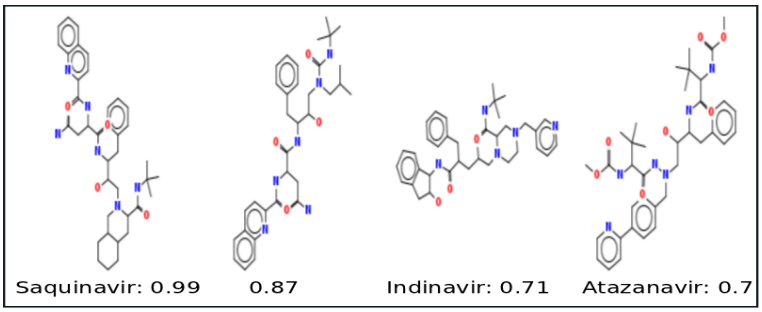
Figure 6.6.8: The 456 active chemical compounds are annotated by the chemical classes they represent and the 29 archetypal compounds found by our archetype analysis algorithm are denoted by ▲. It is evident that the archetypal compounds neatly spread across all 456 compounds.

are computed. A pairwise similarity matrix S, the *Tanimoto* kernel, is constructed by the pairwise Tanimoto association scores (Rogers and Tanimoto, 1960) between the compounds' fingerprints. An eigenvalue decomposition of S is performed with a projection onto a reduced-dimensional space to obtain a reduced dataset with 456 drugs and 16 dimensions. Our Group-Lasso method is applied on this dataset and extracts 29 archetypal compounds from amongst the 456 active compounds. The resulting landscape of all 456 active compounds annotated by the chemical classes and the archetypes are shown in Figure 6.6.8.

To analyse the archetypal compounds further, we look at how they explain any of the 25 currently-available anti-HIV drugs. Depending on their mode of inhibiting HIV-1 viral infection and progression within the host, all currently active anti-HIV drugs are designated to any one of the functional groups: "Nucleoside reverse transcriptase inhibitors (NRTI)", "Non-nucleoside reverse transcriptase inhibitors (NNRTI)", "Protease inhibitors", "Integrase inhibitors", or "Entry inhibitors". See also Section 1.2.2. Drugs within a specific group are known to show almost the same resistance profiles (Johnson et al., 2010). For any anti-HIV drug amongst the available 25, we plot the *most influential* archetypes based on similar, higher-valued archetypal weights. Figure 6.6.9 (a and b) illustrate two different convex sets of currently-available anti-HIV drugs as archetypes and their *explaining* archetypal compounds. Interestingly, from the figures, we see that some of the archetypal compounds themselves represent anti-HIV drugs. For example in Figure 6.6.9 (a), the majority of archetypal compounds are also anti-HIV drugs, all constituting the *Protease Integrase* family and thereby sharing the same resistance profiles. This sharing of resistance profiles is very well reflected by their grouping as archetypal compounds, that is seen in the similarity of archetypal weights, and is also strickingly similar in their underlying chemical structures. The chemical structures of the compounds are plotted using the R package *rcdk* (Guha (2007)). Medically, this grouping can be very useful to predict HIV-1 *cross-resistance* depending on the location of the compound within this



(a)



(b)

Figure 6.6.9: Sets of most influential archetypal compounds shown with their chemical structures and corresponding archetypal weights. Amongst these archetypal compounds, anti-HIV drugs have been identified. (a) One set of explaining archetypes of which four are anti-HIV drugs namely, *Amprenavir*, *Darunavir* and *Fosemprenavir* and *Lopinavir*, that constitute the *Protease inhibitors* group. (b) Another set of explaining archetypes of which three are anti-HIV drugs namely, *Saquinavir*, *Indinavir* and *Atazanavir*, also from the *Protease inhibitors* group. The archetypal drugs identified in each set are known to exhibit similar resistance profiles and this grouping is well captured by our method that makes use of chemical structural similarity in identifying archetypes. Another use of such a grouping can be that chemical compounds similar to archetypal drugs can be looked into further for antiretroviral drug discovery studies.

approximate convex set. Such information could be used to benefit new drug discovery studies and therapeutic protocols.

6.7 Conclusion

Archetype analysis involves the identification of representative objects from amongst a set of multivariate data such that the observations can be expressed as a noisy convex combination of these representative objects. Conventional archetype analyses rely on RSS curves for model selection. In high-noise settings, the curves tend to become uninformative due to no sudden change in the decay. Another drawback is that these methods are also sensitive to the initialisation of archetypes at the onset of the algorithm. If the dataset consists of some structure, then these methods have difficulties in extracting the right archetypes.

In the current work, we address these problems through a Group-Lasso formulation together with a well-defined criterion, BIC, for model selection. Further, the archetypes are initialised to all the observations desensitising our method to archetype initialisation. With the usage of larger datasets,

this would require efficient methods, and we therefore use the Group-Lasso to enforce grouped sparsity. Since the Group-Lasso solution ensemble can be sampled at discrete steps using a fast active-set method, BIC can be computed stepwise for model selection, thereby effecting automatic archetype identification. Model selection is experimentally shown to perform well even in high-noise cases and also with structured data. Both the simulations and real-world experiments bring out the proficiency of our Group-Lasso based archetype analysis method over conventional methods.

7

Conclusion and Future directions

In this concluding chapter, we summarise the 3 models presented in the thesis along with possible extensions.

Facet I

HIV Haplotype Inference using a propagating Dirichlet Process Mixture Model. The first facet of the thesis analyses the genetically-diverse HIV populations present in an infected patient's blood samples using Next-generation Sequencing (NGS) data. The data are shorter viral strains called *reads*. Understanding genetic diversity is crucial for further insights into the evolution of drug-resistant viral lineage within an infected host, disease progression and for *personalised medication* where drugs are prescribed to a patient based on his/her viral lineage. The puzzle is in matching every read to its parent strain or *haplotype*. Given error-prone reads with limited lengths, the main modelling challenge is that non-overlapping reads do not have any suitable *a priori* pairwise similarity measure; this leads to a *non-standard* clustering problem. None of the previous approaches have provided a convincing strategy to solve this issue. In this work we successfully overcome this problem by introducing a propagating Dirichlet Process Mixture Model that adaptively chooses the number of haplotypes. A Gibbs sampler is used for inferring the unknown haplotypes from the error-prone reads. The model is computationally efficient and requires only a few input parameters. From our results based on simulated reads we can see that the model's performance is stable under simulations conducted with varying diversities and read lengths. Experiments with real data also confirmed the model's performance.

Future outlook:

1. Looking ahead, specific regions of the HIV genome can be analysed such as fully-conserved regions, mutation *hotspots* or ART-drug binding target sites. This would provide deeper insights into these regions' mechanisms towards evolution of drug-resistant mutants and aid proactive drug design.

2. In terms of alignment, aligners could take into account frameshifts that in turn would reduce the number of false haplotypes. Frameshifts are caused by insertions or deletions of 1 or 2 bases that shift the reading frame and disturb the amino-acid triplet encoding. Currently, frameshifts are ignored while aligning reads to a reference genome. Frameshift detectors for metagenomes (such as MetaGeneTack (Tang et al. (2013))) can be used to provide prior information to aligners.

3. There are challenges involved in the handling of huge amounts of data churned out by NGS platforms and these are paving the way for wide-scale *cloud computing* (Schadt et al. (2010)) and *Big Data* initiatives (Har (2012), Golden et al. (2013), Lewis et al. (2013)).

Facet II

The second facet of the thesis takes the initial steps to identify similarity patterns between anti-HIV drugs and active chemical compounds. At present there are only 25 commercialised anti-HIV drugs spread over 5 functional groups based on the drug's mode of viral attack. When a viral lineage becomes resistant to a particular drug, it tends to show resistance to other drugs in the same drug group, a property called *cross-resistance*, thus limiting drastically the number of available drugs for prescription. This situation demands proactive drug development and thus, an indepth understanding of similarities between the current drugs and active chemical compounds is necessary. This is done by examining a landscape of active chemical compounds that also contains the drugs. With respect to this, we have developed two models in the thesis:

TiWnet – network inference. We develop a fully probabilistic approach, *Translation-invariant Wishart network* (TiWnet), to infer networks from pairwise Euclidean distances obtained from kernel matrices of n objects. Traditional models (Lasso-type methods), are based on the central Wishart likelihood parametrised by the inverse covariance and sparsity of the latter is usually enforced by some penalty term. Assuming a central Wishart, however, is equivalent to assuming that the origin of the coordinate system is known. If these methods use on input only kernel matrices, they would rely on an assumed origin for any such kernel rather than the relevant pairwise distance information of the kernel. This might lead to biased networks. The method we developed is specifically designed to work with pairwise distances since the likelihood depends only on these distances. Combining this likelihood with a prior suited for sparse network recovery, we are able to extract sparse networks using only pairwise distances. Network inference can now be carried out on any such distance matrix induced by a Mercer kernel on graphs, probability distributions or more complex structures. We also present an efficient MCMC sampler for TiWnet making it applicable to medium-size instances. For further higher-node cardinalities, we show the possibility

of inferring module networks using kernels defined on probability distributions over groups of nodes. Comparisons with competing methods demonstrate the high quality of networks obtained from TiWnet. Given a set of chemical compounds which also includes anti-HIV drugs, we construct kernels using the *SMILES* string encodings of the compounds. The network extracted using the kernels can be used to read out cross resistance properties shared amongst compounds from different chemical classes and drugs' functional groups.

Future outlook:

1. We can consider deploying pairwise distance data into the regression-based neighbourhood-selection method of Meinhausen and Bühlmann (2006) for network recovery. For this, there would be n independent regression models, one each to find the neighbourhood for every node. The translation-invariant Wishart likelihood is used and regularised with the ℓ_1 penalty.

2. Conditional covariance selection has been presented in Kolar et al. (2010a) where the neighbourhoods of nodes are conditioned on a random variable that holds information about the associations between nodes. The problem was cast as a logistic regression model with a $\ell_{1,2}$ penalty. We can extend TiWnet for estimating such a conditional covariance matrix. For this, we look at the weighted RSS problem which is $(X'\boldsymbol{\beta} - y)^t \text{diag}(K)(X'\boldsymbol{\beta} - y)$ where K has the weights on its diagonal. These weights can be regarded as the values the conditioning random variable takes. The similarity kernel is given by $S_{weighted} = \text{sqrt}(\text{diag}(K))XX^t\text{sqrt}(\text{diag}(K))$ which can be used to compute pairwise distances needed for TiWnet.

3. In TiWnet, we use a prior similar to the spike and slab prior of Mitchell and Beauchamp (1988). A new class of priors over covariance matrices that parametrises the partial autocorrelations is studied in Daniels and Pourahmadi (2009). Such a prior can be devised for TiWnet that induces a marginal uniform prior allowing values between $(-1, 1)$ on the entries of a covariance matrix.

4. We are yet to provide the necessary and sufficient conditions for consistent sparse network recovery using TiWnet.

5. For an extremely large number of nodes, variational approximations methods can be explored for model speed-up (Jordan et al., 1999).

Automatic Archetype Analysis. Archetype analysis involves the identification of representative objects from amongst a set of multivariate data such that the observations can be expressed as a noisy convex combination of these representative objects. Conventional archetype analyses rely on RSS decay curves for model selection, which in high-noise settings tend to break down due to no prominent knee region. Another drawback is that these methods are sensitive to the initialisation of archetypes at the onset of the algorithm. This is crucial for a dataset having a structure like clusters of convex sets, where these methods have difficulties in extracting the right archetypes. In the current work, we address these problems through a Group-Lasso formulation together with a well-defined criterion, BIC, for model selection. Further, the archetypes are initialised to all the observations desensitising our method to archetype initialisation. Since the usage of larger datasets requires efficient methods, we use the Group-Lasso to enforce grouped sparsity. The Group-Lasso

solution ensemble can be sampled at discrete steps using a fast active-set method allowing BIC computation stepwise for model selection, thereby effecting automatic archetype identification. Model selection is experimentally shown to perform well even in high-noise cases and also with structured data. The method is applied to extract archetypes from a set of active chemical compounds including anti-HIV drugs. From the resulting set of archetypal compounds, one can draw deeper insights into the functional similarities that can be shared between archetypal drugs and their explaining set of compounds.

Future outlook:

1. One extension would be to amplify the domain scope of archetype analysis. We could apply our Group-Lasso based method on datasets used in, for example, phenotypic data analysis (Shoval et al. (2012)), galaxy spectra studies (Chan et al. (2003)) or consumer-behaviour studies (Li et al. (2003)) and verify existing results.

2. Since we have only empirically observed our model's performance in high noise settings, the next step is to provide theoretical justifications for the same.

8 Appendix

8.1 Appendix: Networks extracted using *graph lasso*

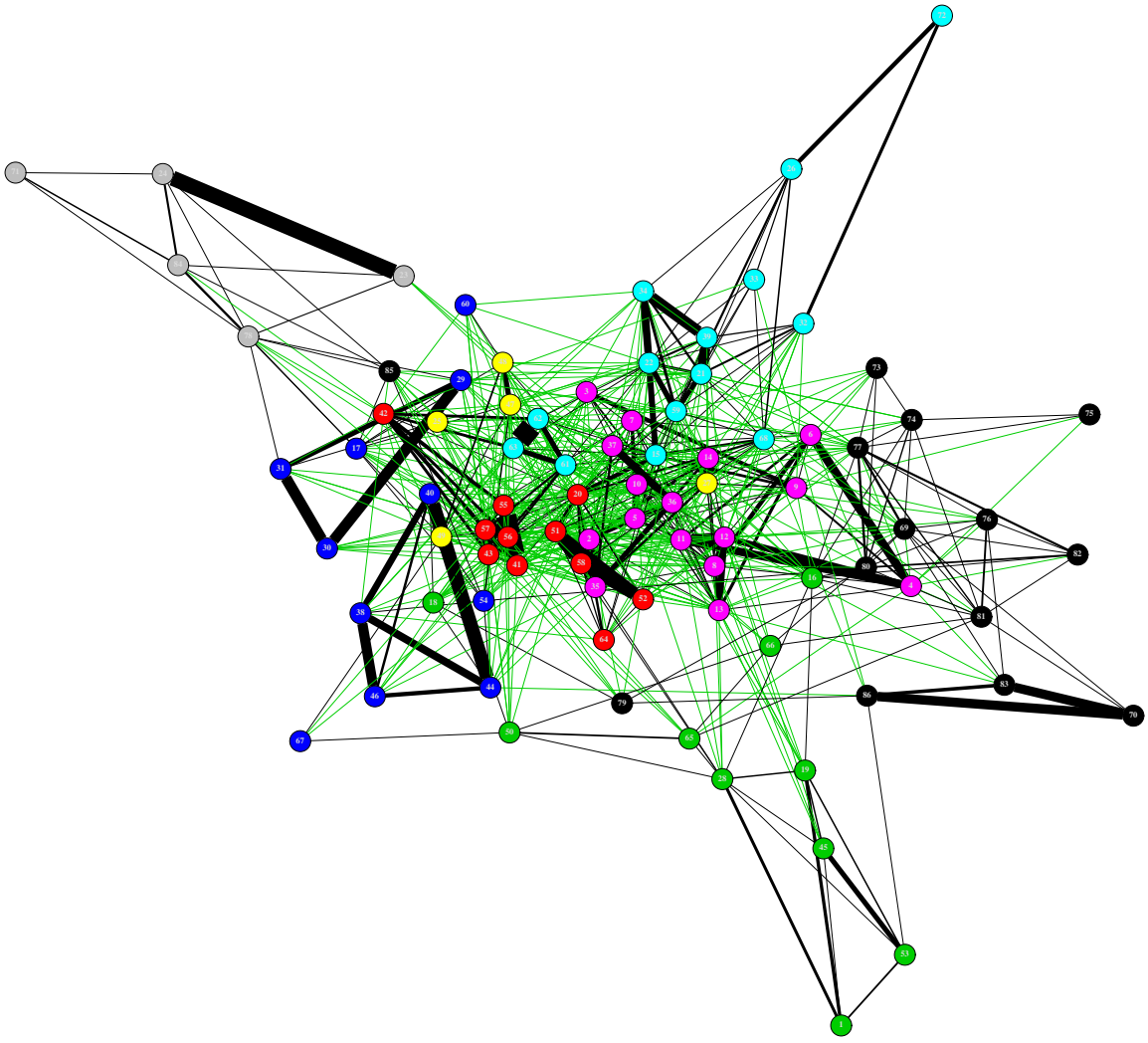


Figure 8.1.1: Network of chemical anti-HIV compounds inferred by *graph lasso* method (Friedman et al., 2007, 2009) with a small ℓ_1 penalty.

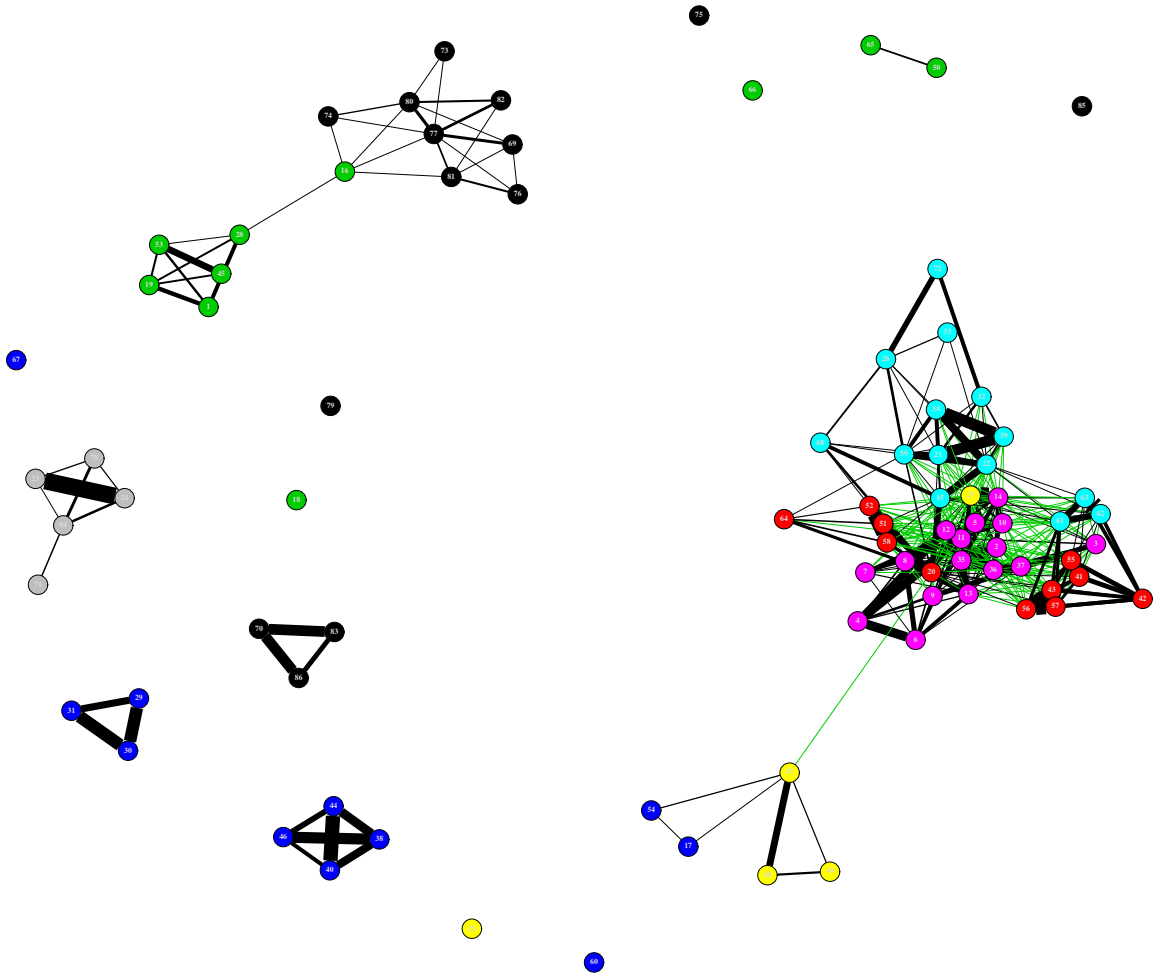


Figure 8.1.2: Network of chemical anti-HIV compounds inferred by *graph lasso* method (Friedman et al., 2007, 2009) with a medium-sized ℓ_1 penalty.

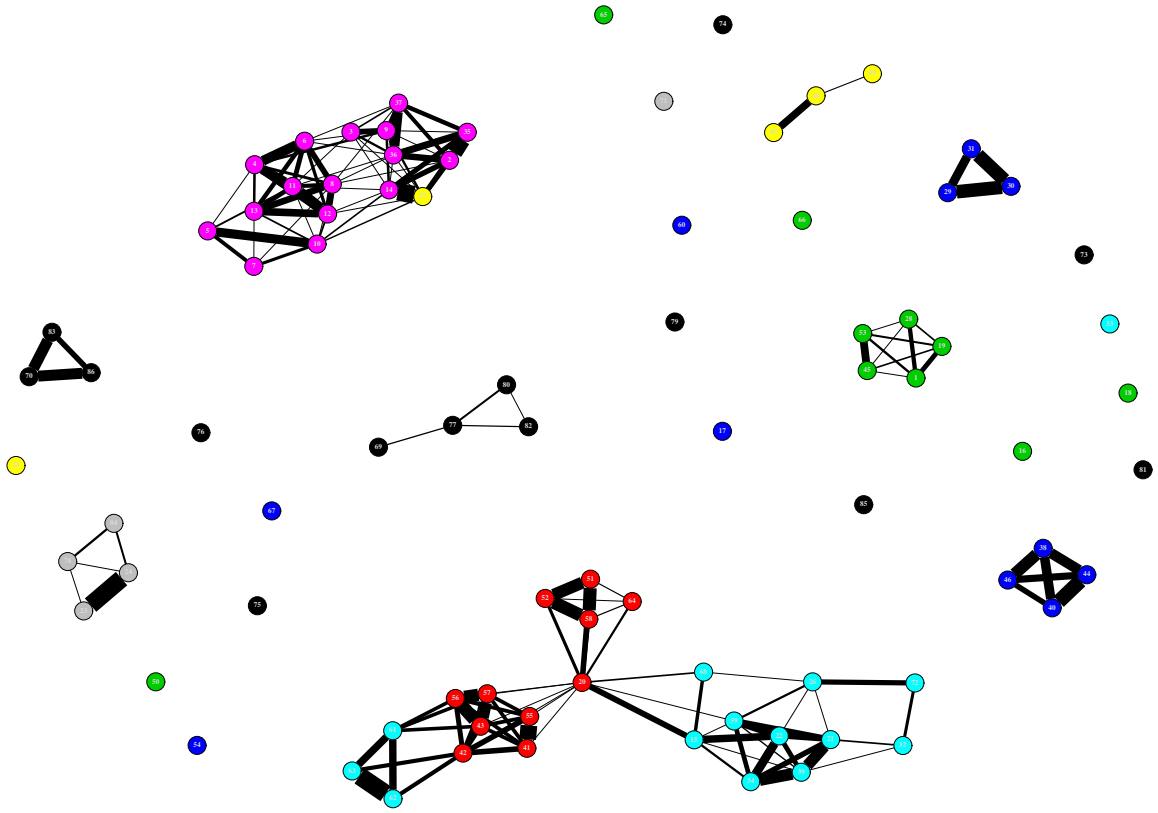


Figure 8.1.3: Network of chemical anti-HIV compounds inferred by *graph lasso* method (Friedman et al., 2007, 2009) with a large ℓ_1 penalty.

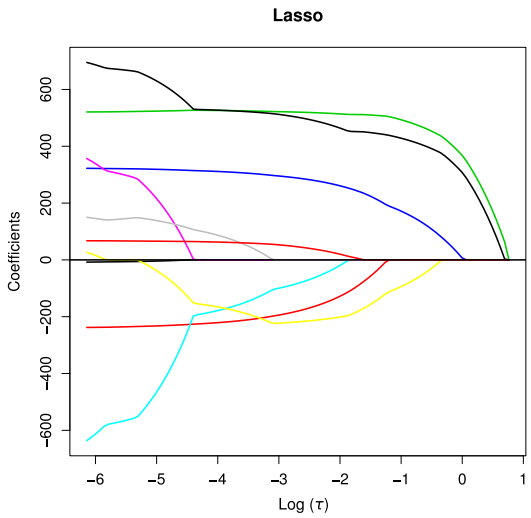


Figure 8.2.1: Lasso estimate profiles for the diabetes data $X_{442\times10}$. When $\tau = 1$ (*i.e.* $\log(\tau) = 0$), 3 features enter the model. When $\tau = 0$ (*i.e.* $\log(\tau) = -\infty$), all the coefficients $\hat{\boldsymbol{\beta}}$ have non-zero values and thus all the features are also active.

8.2 Appendix: Primer to Group Lasso

Consider a response vector $\boldsymbol{y} \in \mathbb{R}^d$ and a feature (predictor) matrix $X \in \mathbb{R}^{n\times d}$ where $X = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)'$ and each feature $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{id}) \in \mathbb{R}^d$ is placed as a row in X. We assume that $\boldsymbol{y}$ and X are centred i.e. mean $= 0$. The standard linear regression problem can be stated as:

$$min_{\boldsymbol{\beta}} \|(\boldsymbol{y} - X'\boldsymbol{\beta})\|_2^2 \tag{8.1}$$

where $\|\cdot\|_2$ denotes the ℓ_2 norm of a vector, $\boldsymbol{\beta} \in \mathbb{R}^n$ are the weights associated with each feature $\boldsymbol{x}_i$ i.e. $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_n)'$. When the number of features n is much greater than the dimension d, Equation 8.1 fails. As a workaround, the *Least absolute shrinkage and selection operator (Lasso)* was introduced in Tibshirani (1996) or *basis pursuit* in signal processing by Chen et al. (1998) which regularised Equation 8.1 using a ℓ_1 penalty norm as follows (Knight and Fu, 2000):

$$\hat{\boldsymbol{\beta}} = argmin\{\|\boldsymbol{y} - X'\boldsymbol{\beta}\|_2^2 + \underbrace{\tau}_{\geq 0} \underbrace{\|\boldsymbol{\beta}\|_1}_{\sum_{j=1}^n |\beta_j|} \} \tag{8.2}$$

In the Lasso estimator, the degree of sparsity is controlled indirectly via the penalty weight τ. For $\tau = 0$, the full linear regression model is employed, whereas increasingly many features are deleted from the model as $\tau \to \infty$. This can be seen in Figure 8.2.1 where Lasso estimate profiles ($\hat{\beta}$) are provided for the diabetes data set $X_{442\times10}$ (used in Efron et al. (2004)). When $\tau = 0$ (*i.e.* $\log(\tau) = -\infty$), all the βs are active and the model corresponds to the standard linear regression problem. As τ increases, the βs are gradually zeroed out one by one, thereby dropping out the corresponding features from the model. Thus, the ℓ_1 penalty norm encourages sparsity in the Lasso estimate $\hat{\beta}$ i.e. the solution returned consists of certain non-zero entries in β which allow only those corresponding

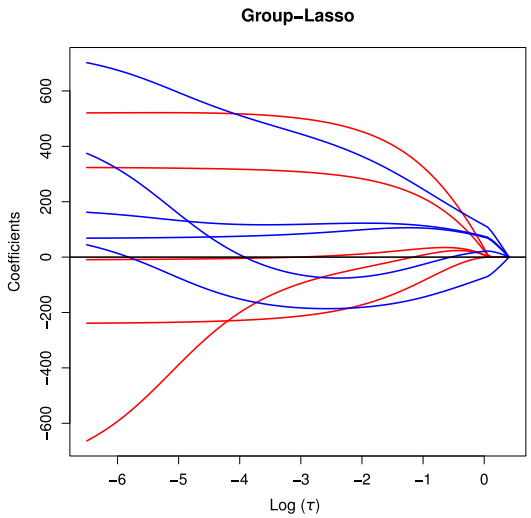


Figure 8.2.2: Group-Lasso estimate profiles for the diabetes data $X_{442\times 10}$, assuming 2 groups of features. When $\tau = 1$ (*i.e.* $\log(\tau) = 0$), the blue group of features is active whereas the red group is inactive. As τ decreases further, both the groups are active and likewise all the features are also active.

features (or rows) to be selected in X.

As seen, in Lasso the features are selected individually. If the features have a grouped structure, then it is practically meaningful to identify important groups (factors/subsets) than individual features (Yuan and Lin (2006)). For example, consider X to be divided into F factors (groups) with d_f, the number of elements in the f^{th} factor i.e. $X = [X_1', \cdots, X_F']'$ where $X_f \in \mathbb{R}^{d_f \times d}$ for $f = 1, \cdots, F$ and $\sum_{j=1}^{F} d_f = n$. X_f is orthonormalised[1] i.e. $X_f X_f' = I_{df}$ for $f = 1, \cdots, F$. When $d_f = 1$, there is no grouped structure in the feature space and the setup is exactly the same as for the Lasso case. To be able to select groups of features, Yuan and Lin (2006) proposed the *Group-Lasso*, by penalising the grouped coefficients in a manner similar to *Lasso*. Here, they use a $\ell_{1,2}$ penalty norm over the grouped coefficients. The convex optimisation problem being solved here to get the Group-Lasso estimates $\hat{\boldsymbol{\beta}}$ is as follows:

$$\hat{\boldsymbol{\beta}} = argmin\{||\boldsymbol{y} - \sum_{f=1}^{F} X_f'\beta_f||_2^2 + \underbrace{\tau}_{\geq 0} \underbrace{\sum_{f=1}^{F} \|\beta_f\|_2}_{\ell_{1,2} \text{ norm}}\} \tag{8.3}$$

The Group-Lasso penalty is a sum (ℓ_1-norm) over the ℓ_2-norms of the coefficient vectors β_f pertaining to groups of features and this induces sparsity at the group level. Here $\boldsymbol{\beta} = (\beta_1', \cdots, \beta_F')'$ and each $\beta_f \in \mathbb{R}^{d_f}$, $f = 1, \cdots, F$. Figure 8.2.2 gives the Group-Lasso estimate profiles for the same diabetes data $X_{442\times 10}$ as used in Lasso, this time grouping the 10 features into 2 groups. As $\log(\tau) = 1$, none of the groups are active. As $\log(\tau)$ decreases, *groups* of coefficients become active - first the blue group of 5 features and subsequently the red group with the remaining 5 features - and likewise the corresponding *grouped* features make their way into the regression model.

[1]The condition is further relaxed for non-orthonormal matrices: *sparse group lasso* (Simon et al. (2013)).

References

Interpreting DNA sequence. URL http://delliss.people.cofc.edu/virtuallabbook/LabReadings/Interpreting_DNA_SequenceREV.pdf.

What is a chromatogram. URL http://www.dnabaser.com/help/samples/what%20is%20a%20chromatogram.html.

Mosaik – The MarthLab. URL https://github.com/wanpinglee/MOSAIK/wiki/QuickStart.

WordNet Search – 3.1. URL http://wordnetweb.princeton.edu/perl/webwn?s=virulence.

454 glossary. 1996. URL http://www.454.com/glossary.

454 sequencing. 2007. URL http://www.454.com/downloads/news-events/how-genome-sequencing-is-done_FINAL.pdf.

SNP-vs-SNP. 2009. URL http://www.politigenomics.com/2009/07/snp-vs-snp.html.

NATIONAL HIV/AIDS STRATEGY FOR THE UNITED STATES. 2010. URL http://aids.gov/federal-resources/national-hiv-aids-strategy/nhas.pdf.

HIV Drug resistance fact sheet. 2011. URL http://www.who.int/hiv/facts/drug_resistance/en/.

AIDS map, 2012. URL http://www.aidsmap.com/v634738045080000000/file/1052204/AHD_2012_Web.pdf.

Harvard School of Public Health - The promise of Big Data. 2012. URL http://www.hsph.harvard.edu/news/magazine/spr12-big-data-tb-health-costs/.

HIV and AIDS. 2012. URL http://www.nhs.uk/conditions/hiv/pages/introduction.aspx.

Drug Information - HIV AIDS Drugs - HIV medication, 2013. URL http://www.aidsmeds.com/list.shtml.

Amplicon Wikipedia, 2013. URL http://en.wikipedia.org/wiki/Amplicon.

Centers for Disease Control and Prevention - Prevention Benefits of HIV Treatment, 2013. URL http://www.cdc.gov/hiv/prevention/research/tap/.

What is HIV Antiretroviral treatment? | Foundcare, 2013. URL http://www.foundcare.org/HIV-Antiretroviral-Treatment.

HIV/AIDS Antiretroviral Drugs Classes, 2013. URL http://www.niaid.nih.gov/topics/HIVAIDS/Understanding/Treatment/pages/arvdrugclasses.aspx.

Primer Wikipedia, 2013. URL http://en.wikipedia.org/wiki/Primer_%28molecular_biology%29.

Genevera I. Allen and Robert Tibshirani. Transposable regularized covariance models with an application to missing data imputation. *The Annals of Applied Statistics*, 4(2):764–790, 2010.

Andre Altmann, Peter Weber, Carina Quast, Monika Rex-Haffner, Elisabeth B. Binder, and Bertram Müller-Myhsok. vipR: variant identification in pooled DNA using R. *Bioinformatics (Oxford, England)*, 27(13):i77–i84, 2011.

Animashree Anandkumar, Vincent Tan, and Alan S. Willsky. High-Dimensional Graphical Model Selection: Tractable Graph Families and Necessary Conditions. *Advances in Neural Information Processing Systems 24*, pages 1863–1871, 2011.

Michael Ashburner, Catherine A. Ball, Judith. A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1):25–29, 2000.

Irina Astrovskaya, Bassam Tork, Serghei Mangul, Kelly Westbrooks, Ion I. Mandoiu, Peter Balfe, and Alex Zelikovsky. Inferring viral quasispecies spectra from 454 pyrosequencing reads. *BMC Bioinformatics*, 12(S-6):S1, 2011.

Jean-Marc Aury, Corinne Cruaud, Valérie Barbe, Odile Rogier, Sophie Mangenot, Gaelle Samson, Julie Poulain, Véronique Anthouard, Claude Scarpelli, François Artiguenave, and Patrick Wincker. High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies. *BMC Genomics*, 9(1):603+, 2008.

Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data. *Journal of Machine Learning Research*, 9:485–516, 2008.

Daniel Barthel, Jonathan D. Hirst, Jacek Blazewicz, Edmund K. Burke, and Natalio Krasnogor. ProCKSI: a decision support system for Protein (Structure) Comparison, Knowledge, Similarity and Information. *BMC Bioinformatics*, 8(416):3250–3264, 2007.

Robert G. Bartle and Donald R. Sherbert. *Introduction to real analysis*. John Wiley & Sons Canada, Limited, 2000.

Christian Bauckhage and Christian Thurau. Making Archetypal Analysis Practical. In *Proceedings of the 31st DAGM Symposium on Pattern Recognition*, pages 272–281, 2009.

Niko Beerenwinkel. HIV-1 whole-genome quasispecies analysis by ultra-deep sequencing and computational haplotype inference to determine the mechanisms of drug resistance development. *Swiss National Science Foundation proposal form*, 2009.

Niko Beerenwinkel and Osvaldo Zagordi. Ultra-deep sequencing for the analysis of viral populations. *Current Opinion in Virology*, 1(5):413–418, 2011.

Niko Beerenwinkel, Huldrych Günthard, Volker Roth, and Karin Metzner. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Frontiers in Microbiology*, 3:329+, 2012a.

Niko Beerenwinkel, Karin J. Metzner, and Volker Roth. Tutorial: Inferring Genetic Diversity from Next-generation Sequencing Data: Computational Methods and Biomedical Applications. *European Conference on Computational Biology*, 2012b. URL http://www.eccb12.org/t4.

Mohamed-Ali Belabbas and Patrick J. Wolfe. Fast low-rank approximation for covariance matrices. In *IEEE Workshop on Computational Advances in Multi-Sensor Processing*, pages 293–296, 2007.

James O. Berger, Brunero Liseo, and Robert L. Wolpert. Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science*, 14(1):1–28, 1999.

Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1995.

Peter J. Bickel and Elizaveta Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36:2577–2604, 2008.

David. Blackwell and James. B. Macqueen. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1:353–355, 1973.

David M. Blei and Michael I. Jordan. Variational Methods for the Dirichlet Process. In Carla E. Brodley, editor, *Proceedings of the International Conference on Machine learning (ICML)*, volume 69 of *ACM International Conference Proceeding Series*. ACM, 2004.

Sebastian Bonhoeffer and Martin A. Nowak. Pre-Existence and Emergence of Drug Resistance in HIV-1 Infection. *Proceedings: Biological Sciences*, 264(1382):631–637, 1997.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Leo Breiman. Better Subset Regression Using the Nonnegative Garrote. *Technometrics*, 37(4), 1995.

Wray Buntine and Marcus Hutter. A Bayesian Review of the Poisson-Dirichlet Process, 2010. URL http://arxiv.org/abs/1007.0296.

Frederic D. Bushman, Gary J. Nabel, and Ronald Swanstrom. *HIV: From Biology to Prevention and Treatment*. Cold Spring Harbor Perspectives in medicine. Cold Spring Harbor Laboratory Press, 2012. ISBN 9781936113408.

Carlos M. Carvalho and Mike West. Dynamic matrix-variate graphical models. *Bayesian Analysis*, 2(1):69–97, 2007.

George Casella and Edward I. George. Explaining the Gibbs Sampler. *The American Statistician*, 46(3):167–174, 1992.

Ben H. P. Chan, Daniel A. Mitchell, and Lawrence E. Cram. Archetypal analysis of galaxy spectra. *Monthly Notices of the Royal Astronomical Society*, 338(3):790–795, 2003. ISSN 1365-2966.

Scott Shaobing Chen, David L. Donoho, Michael, and A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20:33–61, 1998.

John M. Coffin and Harold E. Varmus. Etiologic Agents. *Retroviruses*, 1997.

Anamaria Crisan, Rodrigo Goya, Gavin Ha, Jiarui Ding, Leah M. Prentice, Arusha Oloumi, Janine Senz, Thomas Zeng, Kane Tse, Allen Delaney, Marco A. Marra, David G. Huntsman, Martin Hirst, Sam Aparicio, and Sohrab Shah. Mutation Discovery in Regions of Segmental Cancer Genome Amplifications with CoNAn-SNV: A Mixture Model for Next Generation Sequencing of Tumors. *PLoS ONE*, 7(8):e41551, 08 2012.

Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, UK, 2000.

Adele Cutler and Leo Breiman. Archetypal Analysis. *Technometrics*, pages 338–347, 1994.

Michael J. Daniels and Mohsen Pourahmadi. Modeling covariance matrices via partial autocorrelations. *Journal of Multivariate Analysis*, 100(10):2352–2363, 2009.

Alexandre d'Aspremont, Onureena Banerjee, and Laurent El Ghaoui. First-Order Methods for Sparse Covariance Selection. *SIAM Journal on Matrix Analysis and Applications.*, 30(1):56–66, 2008.

Easley David and Kleinberg Jon. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, New York, NY, USA, 2010.

Arthur P. Dempster. Covariance Selection. *Biometrika*, 28:157–175, 1972.

Aarti N. Desai and Abhay Jere. Next-generation sequencing: ready for the clinics? *Clinical Genetics*, 81(6):503–510, 2012.

José A. Díaz-García, Ramón Gutierrez Jáimez, and Kanti V Mardia. Wishart and Pseudo-Wishart distributions and some applications to shape theory. *Journal of Multivariate Analysis*, 63:73–87, 1997.

Esteban Domingo and John. J. Holland. RNA VIRUS MUTATIONS AND FITNESS FOR SURVIVAL. *Annual Review of Microbiology*, 51(1):151–178, 1997.

Esteban Domingo and C. Perales. From Quasispecies Theory to Viral Quasispecies: How Complexity has Permeated Virology. *Mathematical Modelling of Natural Phenomena*, 7:105–122, 1 2012. ISSN 1760-6101.

Esteban Domingo, Donna Sabo, Tadatsugu Taniguchi, and Charles Weissmann. Nucleotide sequence heterogeneity of an RNA phage population. *Cell*, 13(4):735–744, 1978.

Marcus Droege and Brendon Hill. The Genome Sequencer FLX System – Longer reads, more applications, straight forward bioinformatics and more complete data sets. *Journal of Biotechnology*, 136(12):3–10, 2008.

Richard Durbin, Sean Eddy, Anders Krogh, and Graeme Mitchison. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press. 1998.

Richard L. Dykstra. Establishing the Positive Definiteness of the Sample Covariance Matrix. *The Annals of Mathematical Statistics*, 41:2153–2154, 1970.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

Manfred Eigen. Self-organization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 58:465–523, 1971.

Manfred Eigen. New concepts for dealing with the evolution of nucleic acids. *Cold Spring Harbor Symposia on Quantitative Biology*, 52:307–319, 1987.

Manfred Eigen. The origin of genetic information: viruses as models. *Gene*, 135(1-2):37–47, 1993.

Manfred Eigen. On the nature of virus quasispecies. *Trends Microbiol*, 4(6):216–218, 1996.

Manfred Eigen and Peter Schuster. The Hypercycle. A Principle of Natural Self-Organisation. Part A: Emergence of the Hypercycle. *Naturwissenschaften*, 64(11):541–565, 1977.

Manfred Eigen and R. Winkler. *Steps towards life: a perspective on evolution*. Stufen zum Leben. Oxford University Press, Incorporated, 1992.

Manfred Eigen, John McCaskill, and Peter Schuster. Molecular quasi-species. *The Journal of Physical Chemistry*, 92(24):6881–6891, 1988.

Adam C. English, Stephen Richards, Yi Han, Min Wang, Vanesa Vee, Jiaxin Qu, Xiang Qin, Donna M. Muzny, Jeffrey G. Reid, Kim C. Worley, and Richard A. Gibbs. Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology - slides. *PLoS ONE*, 7(11):e47768, 11 2012. URL http://pacb.com/pdf/Poster_Upgrading_Reference_Genomes_PacBio_RS_Long_Read.pdf.

Paul Erds and Alfréd Rényi. On the Evolution of Random Graphs. In *PUBLICATION OF THE MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES*, pages 17–61, 1960.

Nicholas Eriksson, Lior Pachter, Yumi Mitsuya, Soo-Yon Rhee, Chunlin Wang, Baback Gharizadeh, Mostafa Ronaghi, Robert W. Shafer, and Niko Beerenwinkel. Viral population estimation using pyrosequencing. *PLoS Computational Biology*, 4(5):e1000074, 2008.

Michael D. Escobar and Mike West. Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90:577–588, 1994.

Manuel J. A. Eugster and Friedrich Leisch. Weighted and robust archetypal analysis. *Computational Statistics and Data Analysis*, 55(3):1215–1225, 2011.

Brian Sidney Everitt and David J. Hand. In *Finite Mixture Distributions*, Monographs on Applied Probability and Statistics, pages 1–143. Chapman and Hall, 1981.

Warren J. Ewens. The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, 3(1):87–112, 1972.

Brent Ewing and Phil Green. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Research. *Genome Research*, 8:175–185, 1998.

Ian Fellows. Wordcloud. *http://cran.r-project.org/web/packages/wordcloud*, 2012.

Thomas S. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209–230, 1973.

Thomas S. Ferguson. Prior Distributions on Spaces of Probability Measures. *The Annals of Statistics*, 2(4):615–629, 1974.

Francesca Finotello, Enrico Lavezzo, Paolo Fontana, Denis Peruzzo, Alessandro Albiero, Luisa Barzon, Marco Falda, Barbara Di Camillo, and Stefano Toppo. Comparative analysis of algorithms for whole-genome assembly of pyrosequencing data. *Briefings in Bioinformatics*, 13(3):269–280, 2012.

Alan F. Fleming. Opportunistic infections in AIDS in developed and developing countries. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 84(1):1–6, 1990.

David A. Freedman. On the Asymptotic Behavior of Bayes' Estimates in the Discrete Case. *The Annals of Mathematical Statistics*, 34(4):1386–1403, 1963.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9:432–441, 2007.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. glasso: Graphical lasso - estimation of Gaussian graphical models. R package version 1.4, 2009.

Bela A. Frigyik, Amol Kapila, and Maya R. Gupta. Introduction to the Dirichlet Distribution and Related Processes. Technical Report 206, 2010.

Alan E. Gelfand and Athanasios Kottas. A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 11:289–305, 2002.

Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

Moritz Gerstung, Christian Beisel, Markus Rechsteiner, Peter Wild, Peter Schraml, Holger Moch, and Niko Beerenwinkel. Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nature Communications*, page 811, 2012.

Subhashis Ghosal. The Dirichlet Process, Related Priors, and Posterior Asymptotics. In Nils Lid Hjort, Chris Holmes, Peter Müller, and Stephen G. Walker, editors, *Bayesian Nonparametrics*. Cambridge University Press, 2010.

Andre Gilles, Emese Meglecz, Nicolas Pech, Stephanie Ferreira, Thibaut Malausa, and Jean F. Martin. Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genomics*, 12(1):245+, 2011.

Aaron Golden, S. Djorgovski, and John Greally. Astrogenomics: big data, old problems, old solutions? *Genome Biology*, 14(8):129+, 2013.

Rajarshi Guha. Chemical Informatics Functionality in R. *Journal of Statistical Software*, 18(6), 2007.

Arjun K. Gupta and Daya K. Nagar. *Matrix Variate Distributions*. Chapman and Hall/CRC ISBN 978-1584880462, 1999.

Bradley R Hacker and Philip B Gans. Continental collisions and the creation of ultrahigh-pressure terranes: Petrology and thermochronology of nappes in the central Scandinavian Caledonides. *Geological Society of America Bulletin*, 117(1-2):117–134, 2005.

Nancy F. Hansen, Jared J. Gartner, Lan Mei, Yardena Samuels, and James C. Mullikin. Shimmer: detection of genetic alterations in tumors using next-generation sequence data. *Bioinformatics*, 29(12):1498–1503, 2013.

David A. Harville. Bayesian Inference for Variance Components Using Only Error Contrasts. *Biometrika*, 61(2):383–385, 1974.

Trevor Hastie, Jonathan Taylor, Robert Tibshirani, and Guenther Walther. Forward Stagewise Regression and the Monotone Lasso. *Electronic Journal of Statistics*, 1:1–29, 2007.

Charlotte Hedskog, Mattias Mild, Johanna Jernberg, Ellen Sherwood, Göran Bratt, Thomas Leitner, Joakim Lundeberg, Björn Andersson, and Jan Albert. Dynamics of HIV-1 Quasispecies during Antiviral Treatment Dissected Using Ultra-Deep Pyrosequencing. *PLoS ONE*, 5(7):e11345+, 2010.

Martin Henk, Jürgen Richter-Gebert, and Günter M Ziegler. Basic Properties Of Convex Polytopes. In *Handbook of discrete and computational geometry, Chapter 13*, pages 243–270. CRC Press, Boca, 1997.

Martin S. Hirsch, Huldrych F. Günthard, Jonathan M. Schapiro, Françoise Brun Vézinet, Bonaventura Clotet, Scott M. Hammer, Victoria A. Johnson, Daniel R. Kuritzkes, John W. Mellors, Deenan Pillay, Patrick G. Yeni, Donna M. Jacobsen, and Douglas D. Richman. Antiretroviral Drug Resistance Testing in Adult HIV-1 Infection: 2008 Recommendations of an International AIDS Society-USA Panel. *Clinical Infectious Diseases*, 47(2):266–285, 2008.

Steve Hoffmann, Christian Otto, Stefan Kurtz, Cynthia M. Sharma, Philipp Khaitovich, Jörg Vogel, Peter F. Stadler, and Jörg Hackermüller. Fast Mapping of Short Sequences with Mismatches, Insertions and Deletions Using Index Structures. *PLoS Computational Biology*, 5(9):e1000502, 2009.

Myles Hollander and Douglas A. Wolfe. *Nonparametric Statistical Methods, 2nd Edition*. Wiley-Interscience, 1999.

T. Déirdre Hollingsworth, Roy M. Anderson, and Christophe Fraser. HIV-1 Transmission, by Stage of Infection. *Journal of Infectious Diseases*, 198(5):687–693, 2008. URL http://jid.oxfordjournals.org/content/198/5/687.abstract.

Austin Huang, Rami Kantor, Allison DeLong, Leeann Schreier, and Sorin Istrail. QColors: An algorithm for conservative viral quasispecies reconstruction from short and non-contiguous next generation sequencing reads. In *BIBM Workshops*, pages 130–136, 2011.

Peter Huggins, Lior Pachter, and Bernd Sturmfels. Toward the Human Genotope. *Bulletin of Mathematical Biology*, 69:2723–2735, 2007. ISSN 0092-8240.

Hemant Ishwaran and Lancelot F. James. Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association*, 96:161–173, 2001.

Hemant Ishwaran and Mahmoud Zarepour. Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics*, 30(2):269–283, 2002.

Tony Jebara, Risi Kondor, and Andrew Howard. Probability Product Kernels. *Journal of Machine Learning Research*, 5:819–844, 2004.

Gareth M. Jenkins, Michael Worobey, Christopher H. Woelk, and Edward C. Holmes. Evidence for the Non-quasispecies Evolution of RNA Viruses. *Molecular Biology and Evolution*, 18(6):987–994, 2001.

Harry Joe. Families of m-variate distributions with given margins and $m(m-1)/2$ bivariate dependence parameters. In L. Rüschendorf, B. Schweizer, and M.D. Taylor, editors, *Distributions with Fixed Marginals and Related Topics*, volume 28 of *IMS lecture notes*, pages 120–141. AMS, 1996.

Jason K. Johnson, Dmitry M. Malioutov, and Alan S. Willsky. Walk-Summable Gaussian Networks and Walk-Sum Interpretation of Gaussian Belief Propagation. *Technical Report – 2650, LIDS, MIT*, 2005a.

Jason K. Johnson, Dmitry M. Malioutov, and Alan S. Willsky. Walk-Sum Interpretation and Analysis of Gaussian Belief Propagation. In *Advances in Neural Information Processing Systems 18*, pages 579–586, 2005b.

Victoria A. Johnson, Francoise Brun-Vezinet, and Bonaventura Clotet et al. Update of the drug resistance mutations in HIV-1: Dec 2010. *Topics in HIV medicine*, 18(5):156–163, 2010.

Vladimir Jojic, Tomer Hertz, and Nebojsa Jojic. POPULATION SEQUENCING USING SHORT READS: HIV AS A CASE STUDY. *Pacific Symposium of Biocomputing*, pages 114–125, 2008.

Michael I. Jordan, Zoubin Ghahramani, Tommi Jaakkola, and Lawrence K. Saul. An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37(2):183–233, 1999.

Ingrid M. Keseler, Julio Collado-Vides, and Alberto Santos-Zavaleta et al. Ecocyc: a comprehensive database of Escherichia coli biology. *Nucleic Acids Research*, 39:D583–D590, 2011.

Nirmal Keshava. A Survey of Spectral Unmixing Algorithms. *Lincoln Lab Journal*, 14(1):55–78, 2003.

Yuwon Kim, Jinseog Kim, and Yongdai Kim. Blockwise Sparse Regression. *Statistica Sinica*, 16: 375–390, 2006.

Martin Kircher and Janet Kelso. High-throughput DNA sequencing – concepts and limitations. *Bioessays*, 32(6):524–536, 2010.

Edward C Klatt. *Pathology of AIDS*, volume 24. 2013. URL http://library.med.utah.edu/WebPath/AIDS2013.PDF.

Keith Knight and Wenjiang Fu. Asymptotics for lasso-type estimators. *Annals of Statistics*, pages 1356–1378, 2000.

Daniel C. Koboldt, Qunyuan Zhang, David E. Larson, Dong Shen, Michael D. McLellan, Ling Lin, Christopher A. Miller, Elaine R. Mardis, Li Ding, and Richard K. Wilson. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, 22(3):568–576, 2012.

Scott Koenig, Anthony J Conley, Yambasu A Brewah, Gary M Jones, Simon Leath, Lynn J Boots, Victoria Davey, Guiseppi Pantaleo, James F Demarest, Charles Carter, et al. Transfer of HIV-1-specific cytotoxic T lymphocytes to an AIDS patient leads to selection for mutant HIV variants and subsequent disease progression. *Nature medicine*, 1(4):330–336, 1995.

Mladen Kolar, Ankur P. Parikh, and Eric P. Xing. On Sparse Nonparametric Conditional Covariance Selection. In *Proceedings of the 27th International Conference on Machine Learning*, pages 559–566, 2010a.

Mladen Kolar, Le Song, Amr Ahmed, and Eric P. Xing. Estimating Time-Varying Networks. *Annals of Applied Statistics*, 4(1):94–123, 2010b.

Natalio Krasnogor and David A. Pelta. Measuring the Similarity of Protein Structures by Means of the Universal Similarity Metric. *Bioinformatics*, 20(7):1015–1021, 2004.

Björn Labitzke, Serkan Bayraktar, and Andreas Kolb. Generic visual analysis for multi- and hyperspectral image data. *Data Mining and Knowledge Discovery*, pages 1–29, 2012. ISSN 1384-5810. doi: 10.1007/s10618-012-0283-9.

Peter R. Lamptey, Jami L. Johnson, and Marya Khan. The Global Challenge of HIV and AIDS. *Population Bulletin*, 61(1), 2006.

Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, 2012.

Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25–10, 2009.

Steffen L. Lauritzen. *Graphical Models*. Oxford University Press, 1996. ISBN 0-19-852219-3.

David D. Lewis, Yiming Yang, Tony G. Rose, Fan Li, G. Dietterich, and Fan Li. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5: 361–397, 2004.

Seth C. Lewis, Rodrigo Zamith, and Alfred Hermida. Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods. *Journal of Broadcasting and Electronic Media*, 57(1):34–52, 2013.

Heng Li. Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. *Bioinformatics*, 28(14):1838–1844, 2012.

Heng Li and Richard Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 26(5), 2010.

Ming Li, Xin Chen, Xin Li, Bin Ma, and Paul M.B. Vitanyi. The Similarity Metric. *IEEE Transactions on Information Theory*, 50(12):3250–3264, 2004.

Shan Li, Paul Wang, Jordan Louviere, and Richard Carson. Archetypal analysis: A new way to segment markets based on extreme individuals. *A Celebration of Ehrenberg and Bass: Marketing Knowledge, Discoveries and Contribution. Proceedings of the ANZMAC 2003 Conference*, pages 1674–1679, 2003.

James J. Lipsky. Antiretroviral drugs for AIDS. *The Lancet*, 348(9030):800–803, 1996.

Lawrence A. Loeb, John M. Essigmann, Farhad Kazazi, Jue Zhang, Karl D. Rose, and James I. Mullins. Lethal mutagenesis of HIV with mutagenic nucleoside analogs. *Proceedings of the National Academy of Sciences*, 96(4):1492–1497, 1999. doi: 10.1073/pnas.96.4.1492.

Chengwei Luo, Despina Tsementzi, Nikos Kyrpides, Timothy Read, and Konstantinos T. Konstantinidis. Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PLoS ONE*, 7(2):e30087+, 2012.

Abe Macher, David Thomas, and Sindy M Paul. Contraindicated antiretroviral drug combinations. *New Jersey medicine: the journal of the Medical Society of New Jersey*, 100(9 Suppl):41, 2003.

Daniel MacLean, Jonathan D. Jones, and David J. Studholme. Application of 'next-generation' sequencing technologies to microbial genetics. *Nature Reviews Microbiology*, 7(4):287–296, 2009.

Louis M Mansky. Retrovirus mutation rates and their role in genetic variation. *Journal of General Virology*, 79(6):1337–1345, 1998.

Louis M. Mansky and Howard M. Temin. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *Journal of Virology*, 69(8):5087–5094, 1995.

Elaine R. Mardis. The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, 24(3):133–141, 2008.

Michael Marmor, Ki Hertzmark, Su M. Thomas, Pi N. Halkitis, and Mi Vogler. Resistance to HIV infection. *Journal of Urban Health*, 83(1):5–17, 2006.

Guillaume Martin-Blondel, Karine Sauné, Vinh Vu Hai, Bruno Marchou, Pierre Delobel, Jacques Izopet, Lise Cuzin, and Patrice Massip. Factors associated with a strictly undetectable viral load in HIV-1-infected patients. *HIV Medicine*, 13(9):568–73, 2012.

André F. T. Martins, Mário A. T. Figueiredo, Pedro M. Q. Aguiar, Noah A. Smith, and Eric P. Xing. Nonextensive entropic kernels. In *Proceedings of the 25th International Conference on Machine Learning*, pages 640–647, 2008.

Peter McCullagh. MARGINAL LIKELIHOOD FOR DISTANCE MATRICES. *Statistica Sinica*, 19: 631–649, 2009.

Peter McCullagh and Jie Yang. How many clusters? *Bayesian Analysis*, 3:101–120, 2008.

Ahmed S. Mehanna. *Rationale of Design of Anti-HIV Drugs*. John Wiley and Sons, Inc., 2003.

Nicolai Meinhausen and Peter Bühlmann. High dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 38:1436–1462, 2006.

Michael L. Metzker. Sequencing technologies – the next generation. *Nature Reviews Genetics*, 11 (1):31–46, 2010.

Karin J. Metzner, Alexandra U. Scherrer, Benjamin Preiswerk, Beda Joos, Viktor von Wyl, Christine Leemann, Philip Rieder, Dominique Braun, Christina Grube, Herbert Kuster, Jürg Böni, Sabine Yerly, Thomas Klimkait, Vincent Aubert, Hansjakob Furrer, Manuel Battegay, Pietro L. Vernazza, Matthias Cavassini, Alexandra Calmy, Enos Bernasconi, Rainer Weber, Huldrych F. Günthard, and the Swiss HIV Cohort Study. Origin of Minority Drug-Resistant HIV-1 Variants in Primary HIV-1 Infection. *Journal of Infectious Diseases*, 2013.

Steve Meyer R Ph. How HIV drugs work. *HIV Treatment Series III: Part Two of Five*, 2004. URL http://www.thebody.com/content/art968.html(FromTestPositiveAwareNetwork).

Toby J. Mitchell and John J. Beauchamp. Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988.

Pejman Mohammadi, Sébastien Desfarges, István Bartha, Beda Joos, Nadine Zangger, Miguel Muñoz, Huldrych F. Günthard, Niko Beerenwinkel, Amalio Telenti, and Angela Ciuffi. 24 Hours in the Life of HIV-1 in a T Cell Line. *PLoS Pathogens*, 9(1):e1003161+, 2013.

Morten Morup and Lars Kai Hansen. Archetypal analysis for machine learning and data mining. *Neurocomputing*, 80(0):54–63, 2012. ISSN 0925-2312.

Robb J. Muirhead. *Aspects of Multivariate Statistical Theory*. Wiley New York, 1982.

Giuseppe Narzisi and Bud Mishra. Comparing De Novo Genome Assembly: The Long and Short of It. *PLoS ONE*, 6(4):e19175, 2011.

Radford M. Neal. Bayesian Mixture Modeling by Monte Carlo Simulation. Technical report, 1991.

Radford M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS*, 9(2):249–265, 2000.

Matteo Negroni and Henri Buc. Mechanisms of retroviral recombination. *Annual Review of Genetics*, 35:275–302, 2001.

Martin A Nowak. What is a quasispecies? *Trends in Ecology and Evolution*, 7(4):118–121, 1992.

David I. Ohlssen, Linda D. Sharples, and David J. Spiegelhalter. Flexible random-effects models using Bayesian semi-parametric models: applications to institutional comparisons. *Statistics in Medicine*, 26(9):2088–2112, 2007. ISSN 1097-0258.

Mark J. Palmer and Grant B. Douglas. A Bayesian statistical model for end member analysis of sediment geochemistry, incorporating spatial dependences. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 57(3):313–327, 2008. ISSN 1467-9876.

Desmond Patterson and Robin Thompson. Recovery of Inter-Block Information when Block Sizes are Unequal. *Biometrika*, 58(3):545–554, 1971.

Luc Perrin and Amalio Telenti. HIV Treatment Failure: Testing for HIV Resistance in Clinical Practice. *Science*, 280:1871–1873, 1998.

Jim Pitman. *Combinatorial stochastic processes*, volume 1875 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2006.

Jim Pitman and Marc Yor. The Two-Parameter Poisson-Dirichlet Distribution Derived from a Stable Subordinator. *The Annals of Probability*, 25(2):855–900, 1997.

Sandhya Prabhakaran, Karin J Metzner, Alexander Boehm, and Volker Roth. Recovering Networks from Distance Data. *Journal of Machine Learning Research Workshop and Conference Proceedings*, 25:349–364, 2012.

Sandhya Prabhakaran, David Adametz, Karin J. Metzner, Alexander Boehm, and Volker Roth. Recovering Networks from Distance Data. *Machine Learning*, 92(2-3):251–283, 2013a.

Sandhya Prabhakaran, Melanie Rey, Osvaldo Zagordi, Niko Beerenwinkel, and Volker Roth. HIV Haplotype Inference Using a Propagating Dirichlet Process Mixture Model. *IEEE/ACM Transactions on Computational Biology and Bioinformatics [Epub ahead of print]*, 2013b.

Bradley D. Preston. Reverse Transcriptase Fidelity and HIV-1 Variation. *Science*, 10:228–229, author reply 230–231, 1997.

Bradley D. Preston and Joseph P. Dougherty. Mechanisms of retroviral mutation. *Trends in Microbiology*, 4(1):16–21, 1996.

Mattia Prosperi, Luciano Prosperi, Alessandro Bruselles, Isabella Abbate, Gabriella Rozera, Donatella Vincenti, Maria Solmone, Maria Capobianchi, and Giovanni Ulivi. Combinatorial analysis and algorithms for quasispecies reconstruction using next-generation sequencing. *BMC Bioinformatics*, 12(1):5+, 2011.

Mattia C. F. Prosperi and Marco Salemi. QuRe: software for viral quasispecies reconstruction from next-generation sequencing data. *Bioinformatics*, 28(1):132–133, 2012.

Christopher Quince, Anders Lanzen, T Curtis, R Davenport, N Hall, I Head, L Read, and W Sloan. Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature Methods*, 6: 639–641, 2009.

Christopher Quince, Anders Lanzen, Russell Davenport, and Peter Turnbaugh. Removing Noise From Pyrosequenced Amplicons. *BMC Bioinformatics*, 12(1):38+, 2011.

Muthannan A. Ramakrishnan, Zheng Jin J. Tu, Sushmita Singh, Ashok K. Chockalingam, Marie R. Gramer, Ping Wang, Sagar M. Goyal, My Yang, David A. Halvorson, and Srinand Sreevatsan. The feasibility of using high resolution genome sequencing of influenza A viruses to detect mixed infections and quasispecies. *PLoS ONE*, 4(9), 2009.

Andrew Rambaut, David Posada, Keith A. Crandall, and Edward C. Holmes. The causes and consequences of HIV evolution. *Nature Reviews Genetics*, 5(1):52–61, 2004.

Carl Edward Rasmussen. The Infinite Gaussian Mixture Model. In *In Advances in Neural Information Processing Systems 12*, pages 554–560. MIT Press, 2000.

Daniel C. Richter, Felix Ott, Alexander F. Auch, Ramona Schmid, and Daniel H. Huson. MetaSim – A Sequencing Simulator for Genomics and Metagenomics. *PLoS ONE*, 3(10):e3373+, 2008.

David J. Rogers and Taffee T. Tanimoto. A computer program for classifying plants. *Science*, 132: 1115–1118, 1960.

Volker Roth and Bernd Fischer. The Group-Lasso for generalized linear models: uniqueness of solutions and efficient algorithms. In *ICML '08*, pages 848–855. ACM, 2008.

Eric E. Schadt, Michael D. Linderman, Jon Sorenson, Lawrence Lee, and Garry P. Nolan. Computational solutions to large-scale data management and analysis. *Nature Reviews Genetics*, 11(9): 647–657, 2010.

Melanie Schirmer, William T. Sloan, and Christopher Quince. Benchmarking of viral haplotype reconstruction programmes: an overview of the capacities and limitations of currently available programmes. *Briefings in Bioinformatics*, 2012.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, 10:1299–1319, 1998.

Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, 1978.

Becky Schweighardt, Terri Wrin, Duncan A Meiklejohn, Gerald Spotts, Christos J Petropoulos, Douglas F Nixon, and Frederick M Hecht. Immune escape mutations detected within HIV-1 epitopes associated with viral control during treatment interruption. *Journal of acquired immune deficiency syndromes (1999)*, 53(1):36, 2010.

Jayaram Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

Robert W. Shafer and Jonathan M. Schapiro. HIV-1 drug resistance mutations: an updated framework for the second decade of HAART. *AIDS reviews*, 10(2):67–84, 2008.

Yuan K. Shen. Markov Properties. 2011. URL http://people.csail.mit.edu/yks/documents/classes/mlbook/pdf/chapter15.pdf.

Jay Shendure and Hanlee Ji. Next-generation DNA sequencing. *Nature Biotechnology*, 26(10): 1135–1145, 2008.

Oren Shoval, Hila Sheftel, Guy Shinar, Yuval Hart, Omer Ramote, Avi Mayo, Erez Dekel, Kathryn Kavanagh, and Uri Alon. Evolutionary Trade-Offs, Pareto Optimality, and the Geometry of Phenotype Space. *Science*, 336(6085):1157–1160, 2012.

Christian Sigg, Bernd Fischer, Björn Ommer, Volker Roth, and Joachim Buhmann. Nonnegative CCA for Audiovisual Source Separation. In *In IEEE Workshop on Machine Learning for Signal Processing*, pages 253–258, 2007.

Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

Steven S. Skiena. *The Algorithm Design Manual*. New York: Springer-Verlag, 1997.

Pavel Skums, Nicholas Mancuso, Alexander Artyomenko, Bassam Tork, Ion Mandoiu, Yury Khudyakov, and Alex Zelikovsky. Reconstruction of viral population structure from next-generation sequencing data using multicommodity flows. *BMC Bioinformatics*, 14:1471–2105, 2013.

Patrizia F. Stifanelli, Teresa M. Creanza, Roberto Anglani, Vania C. Liuzzi, Sayan Mukherjee, and Nicola Ancona. A comparative study of Gaussian Graphical Model approaches for genomic data. *arXiv preprint arXiv:1107.0261*, 2011.

Michael R. Stratton, Peter J.Campbell, and P. Andrew Futreal. The cancer genome. *Nature*, 458 (5):719–724, 2009.

Shiyuyun Tang, Ivan Antonov, and Mark Borodovsky. MetaGeneTack: ab initio detection of frameshifts in metagenomic sequences. *Bioinformatics*, 29(1):114–116, 2013.

Yee W. Teh. Dirichlet Processes. In *Encyclopedia of Machine Learning*. Springer, 2010.

Christian Thurau. Nearest Archetype Hull Methods for Large-Scale Data Classification. In *International Conference on Pattern Recognition*, pages 4040–4043. IEEE, 2010.

Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society (Series B)*, 58:267–288, 1996.

Armin Töpfer, Osvaldo Zagordi, Sandhya Prabhakaran, Volker Roth, Eran Halperin, and Niko Beerenwinkel. Probabilistic Inference of Viral Quasispecies Subject to Recombination. *Journal of Computational Biology*, 20(2):113–123, 2013.

Granville Tunnicliffe-Wilson. On the Use of Marginal Likelihood in Time Series Model Estimation. *Journal of the Royal Statistical Society, Series B*, 51:15–27, 1989.

Harald Uhlig. On singular Wishart and singular multivariate Beta distributions. *Annals of Statistics*, 22:395–405, 1994.

Samuel Vaiter, Charles Deledalle, Gabriel Peyré, Jalal Fadili, and Charles Dossal. The Degrees of Freedom of the Group Lasso. *http://arxiv.org/abs/1205.1481*, 2012.

Vladimir Vapnik. *Statistical Learning Theory*. John Wiley and Sons, New York, 1998.

Santosh S. Vempala. *The Random Projection Method*. Series in Discrete Mathematics and Theoretical Computer Science. AMS, 2004.

Emanuel N. Vergis and John W. Mellors. NATURAL HISTORY OF HIV-1 INFECTION. *Infectious disease clinics of North America*, 14:809–825, 12 2000.

Julia E. Vogt, Sandhya Prabhakaran, Thomas J. Fuchs, and Volker Roth. The Translation-invariant Wishart-Dirichlet Process for Clustering Distance Data. In *Proceedings of the 27th International Conference on Machine Learning*, pages 1111–1118, 2010.

Chunlin Wang, Yumi Mitsuya, Baback Gharizadeh, Mostafa Ronaghi, and Robert W. Shafer. Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome Research*, 17(8):1195–201+, 2007.

David Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988.

Kelly Westbrooks, Irina Astrovskaya, David Campo, Yury Khudyakov, Piotr Berman, and Alex Zelikovsky. HCV quasispecies assembly using network flows. In *Proceedings of the 4th International Conference on Bioinformatics Research and Applications*, ISBRA'08, pages 159–170, Berlin, Heidelberg, 2008. Springer-Verlag.

Joe Whittaker. *Graphical Models in Applied Multivariate Statistics*. Wiley, 1990.

Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust PCA via outlier pursuit. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 2496–2504. 2010.

Xiao Yang, Sriram P. Chockalingam, and Srinivas Aluru. A survey of error-correction methods for next-generation sequencing. *Briefings in Bioinformatics*, 14(1):56–66, 2012.

Izumi Yoshida, Wataru Sugiura, Junko Shibata, Fengrong Ren, Ziheng Yang, and Hiroshi Tanaka. Change of Positive Selection Pressure on HIV-1 Envelope Gene Inferred by Early and Recent Samples. *PLoS ONE*, 6(4):e18630, 04 2011.

Shipeng Yu. *Advanced probabilistic models for clustering and projection.* PhD thesis, Ludwig Maximilians University Munich, 2006a. URL http://edoc.ub.uni-muenchen.de/archive/00005884/.

Shipeng Yu. Advanced probabilistic models for clustering and projection – slides. 2006b. URL http://www.dbs.informatik.uni-muenchen.de/~spyu/.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2006.

Ming Yuan and Yi Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

Osvaldo Zagordi, Lukas Geyrhofer, Volker Roth, and Niko Beerenwinkel. Deep Sequencing of a Genetically Heterogeneous Sample: Local Haplotype Reconstruction and Read Error Correction. *Research in Computational Molecular Biology*, 5541:271–284, 2009.

Osvaldo Zagordi, Lukas Geyrhofer, Volker Roth, and Niko Beerenwinkel. Deep Sequencing of a Genetically Heterogeneous Sample: Local Haplotype Reconstruction and Read Error Correction. *Journal of Computational Biology*, 17(3):417–428, 2010a.

Osvaldo Zagordi, Rolf Klein, Martin Däumer, and Niko Beerenwinkel. Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. *Nucleic Acids Research*, 38(21):7400–7409, 2010b.

Osvaldo Zagordi, Arnab Bhattacharya, Nicholas Eriksson, and Niko Beerenwinkel. ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. *BMC bioinformatics*, 12(1):119+, 2011.

Osvaldo Zagordi, Martin Däumer, Christian Beisel, and Niko Beerenwinkel. Read length versus Depth of Coverage for Viral Quasispecies Reconstruction. *PLoS ONE*, 7(10):e47046, 10 2012a.

Osvaldo Zagordi, Armin Töpfer, Sandhya Prabhakaran, Volker Roth, Eran Halperin, and Niko Beerenwinkel. Probabilistic inference of viral quasispecies subject to recombination. In *Proceedings of the 16th Annual international conference on Research in Computational Molecular Biology*, RECOMB'12, pages 342–354, Berlin, Heidelberg, 2012b. Springer-Verlag.

Alon Zaslaver, Anat Bren, and Michal Ronen et al. A comprehensive library of fluorescent transcriptional reporters for Escherichia coli. *Nature Methods*, 3(8):623–628, 2006.

Jun Zhang, Rod Chiodini, Ahmed Badr, and Genfa Zhang. The impact of next-generation sequencing on genomics. *Journal of Genetics and Genomics*, 38(3):95–109, 2011.

Shuheng Zhou, John Lafferty, and Larry Wasserman. Time Varying Undirected Graphs. *Machine Learning*, 83:295–319, 2010.

Sandhya Prabhakaran

Department of Mathematics and Computer Science,
University of Basel,
Bernoulistrasse 16, Basel, CH-4056
Switzerland

Herbergsgasse 7,
Basel, CH-4051
Switzerland

CONTACT

sandhya.prabhakaran@unibas.ch, sandhyaprabhakaran@gmail.com
Phone: 0041 - 076 634 2412
Website: https://sites.google.com/site/sandhyaprabhakaran/

EDUCATION

Ph.D, Mathematics and Computer Science
Ph.D THESIS: Machine Learning Methods for HIV/AIDS Diagnostics and Therapy Planning. [Magna Cum Laude / High Distinction]
University of Basel, Switzerland February 2009 - January 2014

Master of Science (M.Sc), Artificial Intelligence (specialism: Intelligent Robotics)
Master THESIS: Multi-scale, Reactive Motion Planning with Deformable Linear Objects. [Distinction with First class]
University of Edinburgh, Scotland, UK Sep 2007-December 2008

Bachelor of Technology (B.Tech), Computer Engineering
Bachelor project: Remote monitoring of SCADA machines using Java/CORBA architecture. [Distinction with First class]
College of Engineering, Chengannur (CUSAT), India May 1997-August 2001

COMPUTER SKILLS

R, MatLab, C++, COBOL, JCL, IBM 370 Assembler

PUBLICATIONS

"HIV Haplotype Inference using a propagating Dirichlet Process Mixture Model", Sandhya Prabhakaran, Melanie Rey, Osvaldo Zagordi, Niko Beerenwinkel and Volker Roth. ***IEEE/ACM Transactions on Computational Biology and Bioinformatics***, [Epub ahead of print], 2013.

- An extended abstract was presented at the ***Machine Learning in Computational Biology (MLCB) Workshop***, **NIPS** 2010.

"Recovering networks from distance data", Sandhya Prabhakaran, David Adametz, Karin J. Metzner, Alexander Böhm and Volker Roth. ***Machine Learning Journal***, volume 92:2-3, pages 251–283, 2013.

- A conference proceeding has appeared here: ***Asian Conference of Machine Learning (ACML'12), Journal of Machine Learning Research Workshop and Conference Proceedings***, volume 25, pages 349–364, 2012.
  [Best Student Paper award]

"Probabilistic Inference of Viral Quasispecies Subject to Recombination", Armin Töpfer,

Osvaldo Zagordi, Sandhya Prabhakaran, Volker Roth, Eran Halperin and Niko Beerenwinkel. ***Journal of Computational Biology***, volume 20:2, pages 113–123, 2013.

- A conference proceeding has appeared here: ***The 16th Annual International Conference on Research in Computational Molecular Biology (RECOMB'12)***, pages 342–354, 2012.

"Automatic Model Selection in Archetype Analysis", Sandhya Prabhakaran, Sudhir Raman, Julia E. Vogt and Volker Roth. 34^{th} ***DAGM/OAGM Symposium***, volume 7476 of Lecture Notes in Computer Science, page 458–467, 2012.

"The Translation-invariant Wishart-Dirichlet Process for Clustering Distance Data", Julia E. Vogt, Sandhya Prabhakaran, Thomas J. Fuchs and Volker Roth. ***The 27th International Conference on Machine Learning (ICML'10)***, pages 1111–1118, 2010. [Best Paper award runner-up]

TEACHING EXPERIENCE

University of Basel — Autumn semesters 2010-2012

Life Science Seminar Project Assistant for MatLab and C++.

WORK EXPERIENCE

IBM Software Laboratories — November 2004 - August 2007

Bangalore, India — (System Software Engineer)

- IBM Mainframe Operating system (zOS) development. Involved specifically in APPC/MVS – Advanced Program to Program Communication on zOS, implemented through zOS' network stack.

U.S. Technology Global — November 2001 - November 2004

Kerala, India — (Software Engineer)

- IBM Mainframe Application development.

OTHERS

Recipient of the *Scottish International Scholarship Programme* (SISP) (2007-08) under the *Fresh Talent Initiative*. The programme offers fully-funded scholarships for 22 Commonwealth students each year that cover Master programmes in UK.

Won the 1^{st} place for Best Poster at the Informatics Jamboree (2008) held at University of Edinburgh:
Multi-scale, Reactive Motion Planning with Deformable Linear Objects

Co-authored an IP disclosure on Data Transfer from JCL to COBOL with Padmaraj Meethal, IBM Software Laboratories. Awarded Publish status on January 9th 2008. (Disclosure Number: IPCOM000167543D).

Article on IBM developerworks (January 2007):
Converting z/OS assembler code to COBOL

LANGUAGES

Fluent: English, Hindi, Malayalam
Conversational: German, Sanskrit, Tamil

Beginner: French, Japanese, Arabic

INTERESTS Yoga & meditation, Tae Bo, hiking, running and reading.