Newcastle University

School of Engineering

Advanced adaptive modelling approaches in the evolution of vector/ cell manufacturing processes

by Joseph T. Emerson

A thesis submitted for the degree of Doctor of Philosophy

November 2020

Declaration of originality

I hereby declare that the contents and results presented in this thesis have been composed entirely by myself. The information sources in the review sections are indicated in the text and are appropriately referenced within this thesis.

Joseph T. Emerson

Newcastle University
November 2020

Abstract

The field of cell gene therapy has seen significant progress in recent years. The last decade has seen the licensing of the first Cell Gene Therapy (CGT) treatments in Europe and clinical trials have demonstrated safety and efficacy in the treatment of numerous severe inherited diseases of the blood, immune and nervous systems. Specifically, autologous viral vector-based CGT treatments have been the most successful to date. However, the manufacturing processes for these CGT treatments are at an early stage of development, and high levels of complexity, process variability and a lack of advanced process and product understanding in vector/cell manufacturing are hindering the development of new processes and treatments.

Here, Multivariate Data Analysis (MVDA) and Machine Learning (ML) techniques, which have not yet been widely exploited for the development of CGT processes, were leveraged to address some of the main hurdles in the development and optimisation of CGT processes. Principal component analysis (PCA) was primarily used for feature extraction to understand the main correlations and sources of variability within the process data, and to evaluate the similarities and differences between batches. Additionally, a sparse PCA algorithm was developed to ease the interpretation of the principal components with a large number of variables present in the dataset.

Predictive modelling techniques were utilized to model the relationships between process variables and critical quality attributes (CQAs) of the viral vector and cell drug products. The infectious titres of lentiviral vector (LV) products from both adherent cell cultures and suspension cell cultures were modelled and predicted successfully and critical process variables were identified with statistically significant correlations to this CQA. In cell drug product manufacturing, the LV copy number in the patient’s transduced cells was also modelled and process parameters in LV manufacturing and cell drug product manufacturing were linked to this CQA.

Overall, the modelling process recovered valuable information from historical process data from the early stages of process development. This data frequently remains unexploited, due to its commonly truncated and unstructured nature; however, this work showed that MVDA/ML techniques can yield beneficial insights despite less than ideal data structure and features.

Dedicated to my mother

Acknowledgments

Firstly, I would like to thank my main supervisor Prof. Jarka Glassey for her excellent support and guidance throughout the PhD journey. Jarka's support enabled me to develop as a researcher and professional, for which I am very grateful. I would also like to thank my second supervisor Dr Mark Willis for support during the PhD and during my undergraduate program. Mark's lectures on process control and research project on optimisation taught me valuable skills and raised my interests in Chemical Engineering, providing me with motivation to pursue the PhD. I want to thank Dr Moritz von Stosch for sharing this PhD opportunity with me and for his insightful and engaging lectures on multivariate data analysis, which again provided me with highly valuable skills and enabled me to take up this PhD project.

To my former industrial supervisor Dr Bo Kara, I owe thanks for setting up this PhD project at GSK and for his guidance and the opportunities that he provided me with during the first two years of the project. At GSK I would also like to thank everyone who engaged in the insightful technical discussions on thesis contents. Especially, I would like to thank Fritz Fiesser, Valerie Schuhl and Dr Cindy Jung for their support throughout the program.

I would also like to thank my family and friends for their support and encouragement while writing up the thesis. In particular, I would like to thank my loving partner Maria and my Dad for providing moral support throughout the PhD.

To my mother, to whom I owe everything, I would like to say thank you for encouraging me to learn from an early age and supporting me in my education throughout.

The project received funding from GlaxoSmithKline and the Engineering and Physical Sciences Research Council *via* the following studentship: EP/N509528/1.

Table of Contents

List of figures

List of tables

List of acronyms and abbreviations

ADA-SCID – adenosine deaminase severe combined immunodeficiency

ARSA – arylsulfatase A

ATP – adenosine triphosphate

AVE – adjusted variance explained

Avg – average

BM – bone marrow

Ca – calcium

$CaCl_2$ – calcium chloride

CDP – cell drug product

CEV – cumulative explained variance

CF – cell factory

CGT – cell gene therapy

CI – confidence interval

Cons. – consumption

CO_2 – carbon dioxide

CQA – critical quality attribute

CV – cross-validation

DNA – deoxyribonucleic acid

eGFP – enhanced green fluorescent protein

F-stat – F-statistic

GMP – good manufacturing practice

HBS – Hank’s balanced salt

HEK 293 – human embryonic kidney 293 cells

HIV-1 – human immunodeficiency virus-1

IS – index of sparsity

LASSO – least absolute shrinkage selection operator

LC – latent component

LP – linear programming

LV – lentiviral vector

MAE – mean absolute error

Max – maximum

MDOF – model degrees of freedom

Mg – magnesium

Min – minimum

MILP – mixed integer linear programming

MINLP – mixed integer nonlinear programming

ML – machine learning

MNC – mononuclear cells

MSE – mean square error

MVDA – multivariate data analysis

NaOH – sodium hydroxide

N_2 – nitrogen

NLP – nonlinear programming

NZL – number of zero loadings

O_2 – oxygen

PBS – phosphate-buffered saline

PC – principal component

PCA – principal component analysis

PEI – polyethylenimine

pH – potential of hydrogen

PLS – partial least squares

PO2 – partial oxygen pressure

PP – process parameter

SELO – seamless L0-norm

SLP – sequential linear programming

TNZL – total number of zero loadings

CQA – critical quality attribute

QC – quality control

R^2 – coefficient of determination

R^2_{Yperm} – coefficient of determination when Y is randomly permutated

RDOF – residual degrees of freedom

svPCA – sparse variable PCA

TU – transduction units

VIP – variable importance of projection

VSV-G – vesicular stomatitis virus- glycoprotein

h – hour

L – litre

mOsm – milliosmole

Chapter 1

Introduction

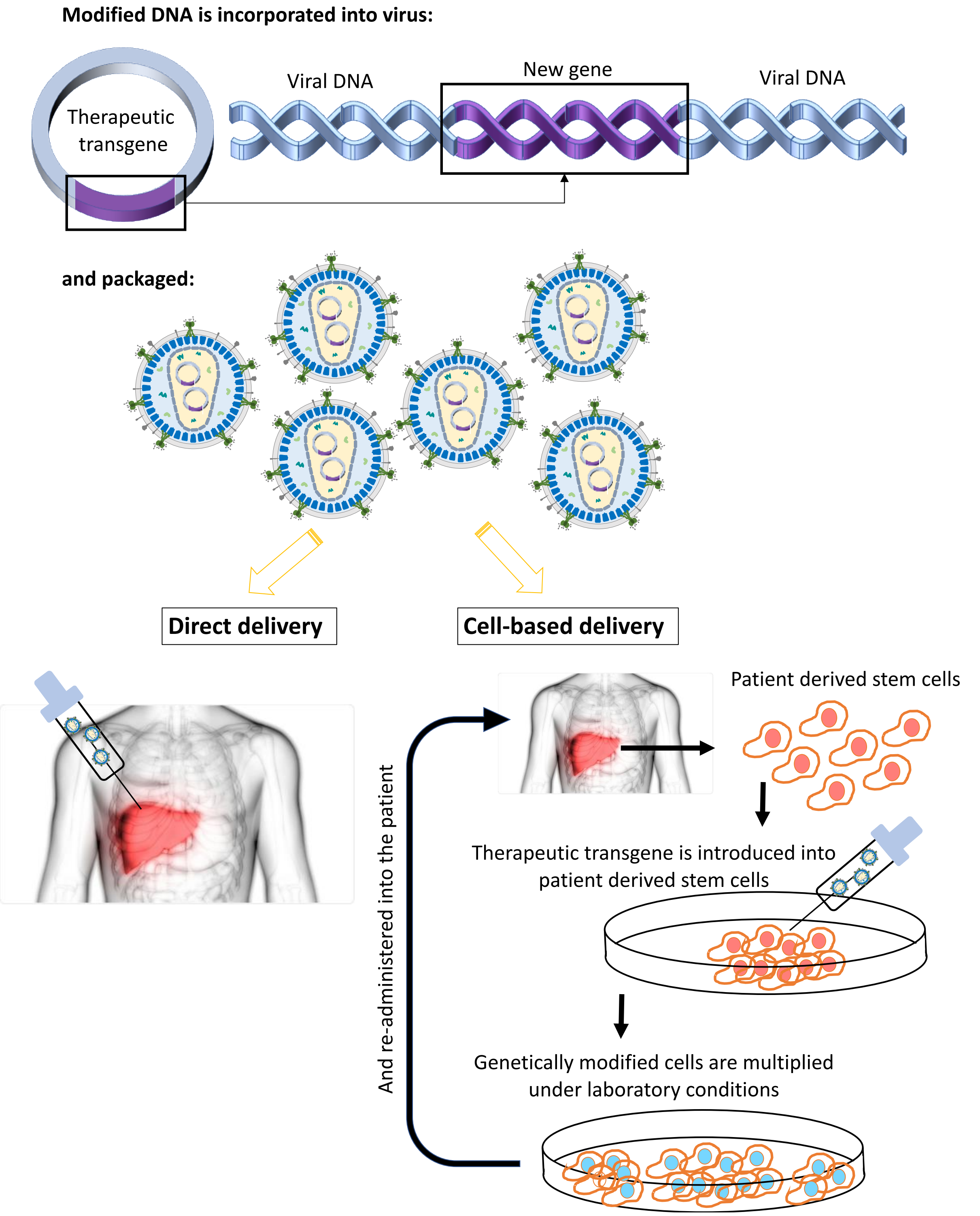
1.1 Research background

The search for cures to human disease is arguably one of the greatest challenges in modern society. Technological progress has expanded the discovery, production and delivery of therapeutic drugs, significantly improving the quality of life and life expectancy of people who have access to treatments. Even though many previously incurable diseases are now considered curable, there are thousands of diseases for which no treatment is currently available. Many of these diseases are genetic diseases, in which a change in the deoxyribonucleic acid (DNA) is observed, that differs from the healthy DNA sequence. Genetic diseases are of particular interest in the scientific community and society as they affect a large portion of the population and have a great socioeconomic impact (Khera et al., 2018). According to the National Human Genome Research Institute, around 280 million people on earth are affected by a genetic disease today ("National Human Genome Research Institute," 2020).

Genetic diseases, or disorders, are caused by genome abnormalities, such as a mutation in a single gene referred to as monogenic, mutations in multiple genes referred to as polygenic, or by a chromosomal abnormality. In a chromosomal abnormality, a portion of the chromosomal DNA is either missing, is in irregular form, or extra chromosomal DNA is present. The most common disease associated with chromosomal abnormality is Down syndrome (Rodwell and Aymé, 2014). Monogenic diseases are rarer than polygenic diseases but are believed to affect ~6% of the population at some point throughout their lives (Rodwell and Aymé, 2014). It is common for monogenic diseases to manifest at a young age as a result of inherited genes, even leading to premature death. Efforts to understand and treat monogenic diseases have resulted in the establishment of various rare disease initiatives, but the progress so far is largely limited due to the complexity of the available solutions (Prakash et al., 2016). Cystic fibrosis, Duchene muscular dystrophy, β-thalassemia and hemophilia are some of the most common monogenic diseases that still lack a well-established cure (Prakash et al., 2016). In contrast to monogenic diseases, polygenic diseases have been linked to environmental conditions and lifestyle. Polygenic diseases include both early and late-onset diseases such as asthma, height and body mass index characteristics, which fall under the early-onset diseases, and cardiovascular disease, diabetes, Altzheimer's, senile dementia, osteoarthritis and cancer which fall under the late-onset diseases (Oliynyk, 2019).

The most important tool in identifying genetic diseases was developed in 2003, as a result of the Human Genome Project (HGP). The HGP generated a physical and genetic map of the human genome. Sequencing the human genome provided the information necessary to generate a map for each human chromosome. Comparing the genome from one person to another, allowed scientists to determine the exact location or area that genes differ from a healthy to a person suffering from a genetic disease (Harold et al., 2009). Knowing the location of the abnormality in the genome later allowed for targeted methods aiming to provide treatment to the genetic disease. Even though the identification of the mutated genes and chromosomal abnormalities became possible, treatment strategies that alter the genes have not been widely applied yet. Treatment has heavily relied on either surgical options *i.e.* cardiovascular disease, blood transfusions *i.e.* β-thalassemia, and addressing the mutant or absent protein *i.e.* diabetes (Rosenberg and Rosenberg, 2012).

Gene-editing techniques for cell gene therapy treatments were initially applied on humans in the 1990s, giving the ability to directly treat a genetic disease at its genetic roots. Cell gene therapy (CGT) utilises the modification of human DNA in order to treat a genetic disease. With the use of genes as drugs, the malfunctioning gene can be replaced or counteracted within the cells affected by the genetic disease. CGT approaches can be characterised as *ex vivo*, *in situ*, or *in vivo*. In *ex vivo* gene therapy, the patient's cells are isolated and the target cell types are cultured *in vitro*, facilitating the selection, expansion and/or their differentiation process before they are genetically modified and introduced into the targeted tissue (Naldini, 2011). The *ex vivo* CGT method allows the modification and extensive characterisation of cells outside the patient's body. It is safe from an immunologic perspective and it allows good control over the process (Herrero et al., 2012; Suhonen et al., 2006). With *ex vivo* CGT, it is also possible to identify the cells that contain and produce the therapeutic gene in sufficient quantity and to control the production rate and level of the therapeutic gene expression (Bethesda, MD, National Institutes of Health, 2016). In *in situ* gene therapy genetic material is introduced into the target tissue directly, resulting in low transduction efficiency, often leading to re-appearance of the genetic disease (Davis and Cooper, 2007; Hu et al., 2007). Finally, in the *in vivo* method, the transfer of genetic material occurs through an appropriate vector into the target tissue. The *in vivo* method is a promising technique, once vector development issues and vector tissue targeting issues are improved (Nayerossadat et al., 2012). Figure 1-1 illustrates *in vivo* and *ex vivo* CGT approaches.



Modified DNA is incorporated into virus:
Therapeutic transgene
Viral DNA
New gene
Viral DNA
and packaged:
Direct delivery
Cell-based delivery
Patient derived stem cells
Therapeutic transgene is introduced into patient derived stem cells
Genetically modified cells are multiplied under laboratory conditions
And re-administered into the patient

Figure 1-1: Cell gene therapy approaches for the treatment of genetic diseases in humans. The genetically modified vector can be either introduced directly into the patient's cells (*in vivo* CGT), or they can be introduced into the patient stem cells in the lab and then be re-introduced into the patient (*ex vivo* CGT) (Worgall and Crystal, 2014).

1.2 Cell gene therapy

The following section provides a summary of the history of CGT before reviewing the current state of the art, including the progress that has been made with viral and non-viral approaches to CGT.

1.2.1 History of cell gene therapy

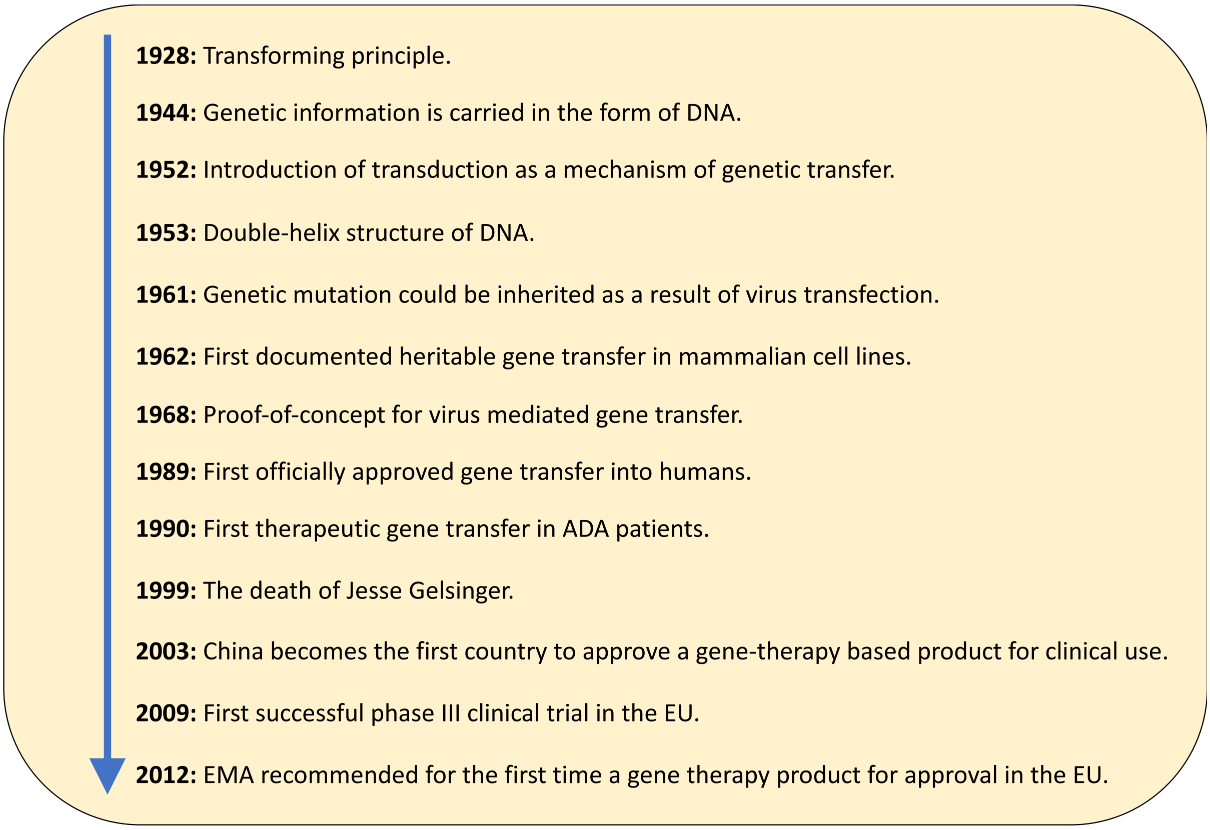
The idea for gene therapy was initially conceived in the early 1970s, almost 30 years before the completion of the HGP project. The main scientific discoveries that set the foundation for its development date back to the beginning of the 20th century and can be seen in **Error! Unknown switch argument.**. In 1928, Frederick Griffith described the transforming principle. Based on his findings, a non-virulent bacterium cell, Type I pneumococcus, could be transformed into a virulent type bacterium cell when it came in contact with an intracellular material from a virulent bacterium cell (Griffith, 1928). Up to this point, scientists believed that genes were composed of proteins and not DNA. In later years (1944), Avery, McLeod and McCarty showed that the transformation of a non-virulent bacterium cell to a virulent type bacterium cell was due to DNA transfer and not proteinic transfer. This was proven by precipitating the protein using chloroform, and repeating experiments based on the transforming principle by Griffiths. With this discovery, the research into understanding the molecular structure of DNA intensified, revolutionising the understanding of molecular genetics.

Another important discovery for the transfer of genetic material between bacteria was presented by Norgon D. Zinder and Joshua Lederberg in 1952 (ZINDER and LEDERBERG, 1952). Zinder and Lederberg observed that genetic recombination occurred between a drug resistant mutant of salmonella and the wild type, when the two colonies were separated by a fine glass filter. The genetic material only passed through the fine glass filter when a certain active substance transferred it through. After purification, they concluded that the substance responsible for carrying the DNA was a bacteriophage. Zinder and Lederberg named this phenomenon ‘genetic transduction’ and their discovery extended research into phages and eukaryotic viruses, as the material for gene transfer (Wirth et al., 2013). The capability of viruses to deliver genes into cells of interest became apparent in the 1960s. However, it was not possible at the time to strip the viruses of their pathology causing genes and to replace them with therapeutic genes that could be transferred to the patient’s genome. The technology to combine DNA from two different species was initially developed in the ‘70s and is still used today. With the use of enzymes, DNA is cut, synthesised and bound on specific genes that are then transported with vectors into the host organisms. The recombinant DNA technology led to the production of the first drugs for medical biotechnology, such as human insulin (Khan et al., 2016).

Even though recombinant DNA technology has been available since the ‘70s, the first human gene therapy trial took place in the early ‘90s. The slow progress from concept to clinical application is due to the complexity of the cellular and tissue barrier that needs to be overcome for gene transfer to be successful, whilst leaving the essential regulatory mechanisms of the cells uninterrupted (Giacca, 2010). Moreover, the amount of the gene-correcting cells must be sufficient in order to alter the genetic mutation, they must not trigger the immune system of the host, and they must survive long-term. The first human cell therapy treatments were conducted in 1989. Rosenberg and his collaborators used *ex vivo* modified tumour-infiltrating lymphocytes (TILs) to treat five patients with advanced melanoma (Rosenberg, 1992). The patients showed no signs of infections and the study showed that the tumour did not grow at the injection site. In two of the patients, there was no evidence of viable tumour cells after the surgical removal of the tumour three weeks after the gene-therapy treatment was initiated.

Following Rosenberg, Michael R. Blaese conducted the first gene-therapy trial with a therapeutic gene in 1990. The subject of this trial were two children with an adenosine deaminase severe combined immunodeficiency (ADA-SCID); a genetic disease that causes a severed or absent immune system due to a deficiency in adenosine deaminase (ADA). The two children were treated with white blood cells modified *ex vivo* to express the gene for ADA production in the body. Only one of the children exhibited a temporary response to the treatment, but the cause of the therapeutic effect could not be successfully identified, as the patient had also simultaneously received enzyme replacement therapy. In 1999, gene therapy encountered a major setback with the tragic death of Jesse Gelsinger. Jesse suffered from a milder form of ornithine transcarbamylase deficiency, a genetic disease of the liver which is usually fatal. However, he showed a genetic mutation of the ornithine transcarbamylase gene in part of his cells. Jesse was part of a clinical trial where he was injected with an adenoviral vector, carrying a corrected gene. After the injection, Jesse showed a severe immune response linked to the adenoviral vector and was declared brain dead after suffering from multiple organ failure (Wirth et al., 2013).

Over the next decade, substantial research went into improving the safety of viral vectors used in CGT treatments. In 2003, China became the first country to approve a CGT drug for the treatment of head and neck squamous cell carcinoma after five years of clinical trials with the only side-effect observed being self-limited fever (Pearson et al., 2004). Following China's success, the first successful phase three clinical trial was achieved in Europe in 2009. The trial targeted patients with malignant brain tumours, introducing an adenoviral vector with the therapeutic gene after the tumour had been surgically resected (Wirth et al., 2009). Since then numerous gene therapy trials have shown promising results, including trials for ADA-SCID (Aiuti et al., 2009), beta thalassemia (Cavazzana-Calvo et al., 2010), Wiskott-Aldrich syndrome (Boztug et al., 2010), X-linked severe combined immunodeficiency (Hacein-Bey-Abina et al., 2010) and haemophilia B (Jessup et al., 2011).



1928: Transforming principle.

1944: Genetic information is carried in the form of DNA.

1952: Introduction of transduction as a mechanism of genetic transfer.

1953: Double-helix structure of DNA.

1961: Genetic mutation could be inherited as a result of virus transfection.

1962: First documented heritable gene transfer in mammalian cell lines.

1968: Proof-of-concept for virus mediated gene transfer.

1989: First officially approved gene transfer into humans.

1990: First therapeutic gene transfer in ADA patients.

1999: The death of Jesse Gelsinger.

2003: China becomes the first country to approve a gene-therapy based product for clinical use.

2009: First successful phase III clinical trial in the EU.

2012: EMA recommended for the first time a gene therapy product for approval in the EU.

Figure 1-2: Key advances in the development of cell gene therapy. Image adapted from the work of Wirth *et al.* (Wirth et al., 2013).

1.2.2 Cell gene therapy today

Today, CGT has re-emerged as a promising treatment offering extraordinary therapeutic benefits, while showcasing a remarkable safety record in numerous phase I and II clinical trials (Naldini, 2015). The success of CGT today, is attributed to the engineering of improved viral and non-viral vectors (explained in detail in section 1.2.3) used to transduce the patient's cells, that have in turn improved the quality, safety and efficacy of the conducted clinical trials (Keeler et al., 2017; Kotterman et al., 2015; Naldini, 2015). These developments led large biopharmaceutical companies to increase research and manufacturing investments in CGT treatments between 2010 and 2016.

Successful clinical trials carried out so far have mainly focused on the treatment of severe genetic diseases of the blood, nervous and immune system (Naldini, 2015). For example, for the treatment of ADA-SCID a CGT product manufactured by Orchard Therapeutics Ltd., called Strimvelis, is currently available in Europe. The drug was used in a clinical trial with 12 children participating, and achieved a 100% survival rate outcome, at three years post-treatment. The intervention-free survival rate for this drug is 92% and shows significant reduction in severe infections in cases of children with ADA-SCID. Children diagnosed with ADA-SCID within the first six months of their lives, who remain untreated, suffer persistent infections that lead to the end of their lives before age one (Gaspar et al., 2006). There is also an available treatment for B-cell acute lymphoblastic leukaemia called Kymriah®, which is a lentivirus transduced autologous T-cell, and a treatment for inherited retinal disease called Luxturna®, utilising a recombinant adeno-associated virus (rAAV) (Golchin and Farahany, 2019). Following the progress so far and based on the current pipeline, the Food and Drug Administration (FDA) predictions are optimistic, expecting to approve between 10 and 20 new cell and gene therapy products *per* year by 2025 (Gottlieb and Marks, 2019).

1.2.3 Viral and non-viral vectors

The methods for *in vivo* and *ex vivo* gene delivery can be divided into two main categories, viral and non-viral, based on the type of gene transfer method that is used. In viral vector-based CGT treatments, the natural ability of viruses to transfer DNA to foreign cells is leveraged by modifying the virus's genome. The pathogenic parts of the virus's genome are stripped away, and a therapeutic gene is incorporated. The part of the virus' genome responsible for replication is also deleted, resulting in no immune system response from the patient. However, the system can still suffer from the immunogenicity of the virus, which can trigger an inflammatory response and lead to the degeneration of the transduced tissue. Toxic production can also occur, along with mutagenesis due to the insertion. Moreover, the ability of viral vector-based systems to target specific tissue cells is limited due to the tropism-specific nature of the virus, *i.e.* the distribution of cell surface receptors (Arbuthnot, 2015; Gardlík et al., 2005;

Nayerossadat et al., 2012). To solve the targeting limitations of viral vectors, designing viruses with specific receptors has allowed the retargeting of the viral vector to cells that were not receptive to the original virus (Wickham, 2003). Some viral vectors that have been previously used in CGT treatment are retroviral vectors, adenoviral vectors, adeno-associated vectors, lentiviruses, the herpes simplex virus (HSV), poxvirus vectors and the Epstein-Barr virus.

On the other hand, non-viral CGT treatments utilise chemical and physical methods to transfer DNA to targeted cells. Physical methods utilise electrical, ultrasonic, mechanical, hydrodynamic or laser-based energy so that the DNA can penetrate the targeted cells in a transient manner. Naked DNA, DNA particle bombardant by gene gun, ultrasound, magnetofection, hydrodynamic and electroporation have all been used as physical methods to transfer DNA into targeted cells (Nayerossadat et al., 2012). Chemical methods, which are more common than physical methods, are based on modifying the properties of nucleic acids to so that they have reduced hydrophilicity and neutralized charge, in order to increase cellular uptake. Chemical methods generally utilise nanomeric complexes, which consist of negatively charged nucleic acids compacted by polycationic nanomeric particles. The nanomeric complexes are stable enough to produce their bound nucleic acid upon degradation and usually enter the cells by endocytosis (Nayerossadat et al., 2012). In order for cationic particles to enter the targeted cells, they are attached to a lipid anchor or a DNA-binding cationic polymer (Liu et al., 2003), such as proteins (Boeckle and Wagner, 2006), small chemical compounds (Xu et al., 1999), antibodies (Wolschek et al., 2002), vitamins, carbohydrates (Chiu et al., 2004), and peptide ligands (Y. Zhang et al., 2004). Cationic systems offer benefits including low toxicity and antigenicity (capacity to bind with other receptors) and long-term expressions, but low efficiency compared to viral systems.

To date viral vector-based treatments have been more successful than non-viral vectors due to significantly higher transduction efficiency, *i.e.* the ability to express gene of interest in the target cells of the patient. Methods using non-viral methods have so far failed to achieve the transduction efficiency of viral vectors, but are characterised by higher availability, lower immunologic response, no limitations in the size of transgenic DNA that can be used and cost-efficiency in their production process (Hirai et al., 1997; Nayerossadat et al., 2012; Robertson et al., 1996). However, both viral and non-viral vector treatments need to overcome their drawbacks in order to be applied in clinical applications. In both treatments the extracellular and intracellular targeting and delivery needs to be improved, as well as the long-time expression of the transgene in the patient's cell. Finally, the toxicity and side-effects of viral and non-viral vector treatments on the human body needs to be minimised or suppressed completely (Nayerossadat et al., 2012).

1.3 Research aims

The main aim of this work is to utilise advanced modelling approaches to address some of the key manufacturing challenges in viral vector production and cell drug product manufacture. Specific challenges to be addressed include characterising the high levels of variability in viral vector production and cell drug product manufacture, assessing process comparability across different production scales, and deriving valuable process knowledge where there is currently a lack of advanced expertise. The objective is to deliver valuable insights from the modelling activities to the project sponsor, GlaxoSmithKline, to aid them with the development of their cell gene therapy manufacturing processes. Much of this work is in-line with the objectives of the Quality by Design initiative, which encourages detailed understanding of the process design and control space and relationships between process parameters and attributes of the product (Mishra et al., 2018). This work should thereby provide insights into appropriate modelling techniques, their benefits and the challenges involved in their application to CGT manufacturing processes, which should serve to support future work leveraging advanced multivariate data analytics in CGT manufacturing. To achieve the main aim, this work focuses on:

a) exploiting unsupervised learning methods, such as PCA, to assess process variability and comparability in viral vector production and cell drug product manufacture,
b) utilising supervised learning techniques, such as partial least squares (PLS) regression, to model the relationships between process parameters and critical quality attributes of the viral vector and cell drug products in viral vector production and cell drug product manufacture, respectively,
c) the development of several alternative programming approaches to sparse principal component analysis (PCA), in order to ease the interpretation of the PCA model and provide insights into process variability and comparability when applied to cell gene therapy manufacturing processes,
d) a review of the data-driven modelling approaches and the challenges faced in their application to cell gene therapy manufacturing processes.

These are the broad aims of this research. More detailed and specific objectives are provided further on, after reviewing the literature on cell gene therapy manufacturing processes, CGT manufacturing challenges and applications of multivariate data analysis to closely related bioprocesses.

Chapter 2

Literature review

2.1 Cell gene therapy manufacturing processes

2.1.1 Production processes for cell gene therapy treatments

As mentioned in section 1.1, there are two types of cell gene therapy that show promising results: *in vivo* and *ex vivo*. In viral *in vivo* gene therapy, the manufacturing process is essentially the manufacturing of the viral vector. For the production of lentiviral and adenoviral vectors (viruses commonly used in viral-based cell gene therapy treatments), the most well-established production methods rely on the transient transfection of plasmid DNA into host cells. Host cells for gene expression that are typically used are HEK293; derived from human embryotic kidney cells, HEK293T adherent cells; an adaption to the HEK293 cell line, PER.C6; an industrially relevant cell line for adenovirus manufacture and sf9 insect cells; which is a clonal isolate of Spodoptera frugipedra Sf21 cells, commonly used for the expression of recombinant proteins from baculovirus (Kotin, 2011; Manceur et al., 2017; Sharon and Kamen, 2018; van der Loo and Wright, 2015).

Typically, the upstream viral vector manufacturing process starts with cell thaw of frozen cells and seed train, followed by further cell expansion in the bioreactor, and finally transfection of plasmid DNA into host cells, followed by harvesting (Kotin, 2011; Merten et al., 2010; Schweizer and Merten, 2010). Seed train involves the generation of an adequate number of cells using various cultivation systems, *e.g.* T-flasks, shake flasks or roller bottles, in order to provide a sufficient number of cells for the inoculation of the bioreactor (Schweizer and Merten, 2010). Factors such as scale, type of viral vector and upstream processing methods impact the choice of downstream processing operations. In a typical downstream process, the first step is clarification, aiming to remove cells and cell debris. The clarification process is followed by the concentration of the feed stream, treatment with endonuclease, such as Benzonase® to digest host and plasmid DNA, size exclusion chromatography or anion exchange chromatography for removal of protein contaminants, diafiltration and sterile filtration (Kotin, 2011; Merten et al., 2010; Schweizer and Merten, 2010). In *ex vivo* gene therapy, the patient's cells are extracted, isolated and cultivated *in vitro* into a culture of the target cell types. This facilitates cell selection, expansion, and/or differentiation before or after genetic modification (Naldini, 2011). The focus of this thesis is on the manufacture of *ex vivo* cell gene therapy treatments, which are mainly applied to hematopoietic stem cells (HSCs), relevant to immunological and blood disorders (Marintcheva, 2018).

An overview of the production process for an *ex vivo* CGT treatment is presented in Figure 2-1, showing complex raw materials used and the product streams. There are a number of complex raw materials used for viral vector production and cell processing, including recombinant viral vector plasmids, modified producer cells, reagents, buffers, cytokines and culture media.

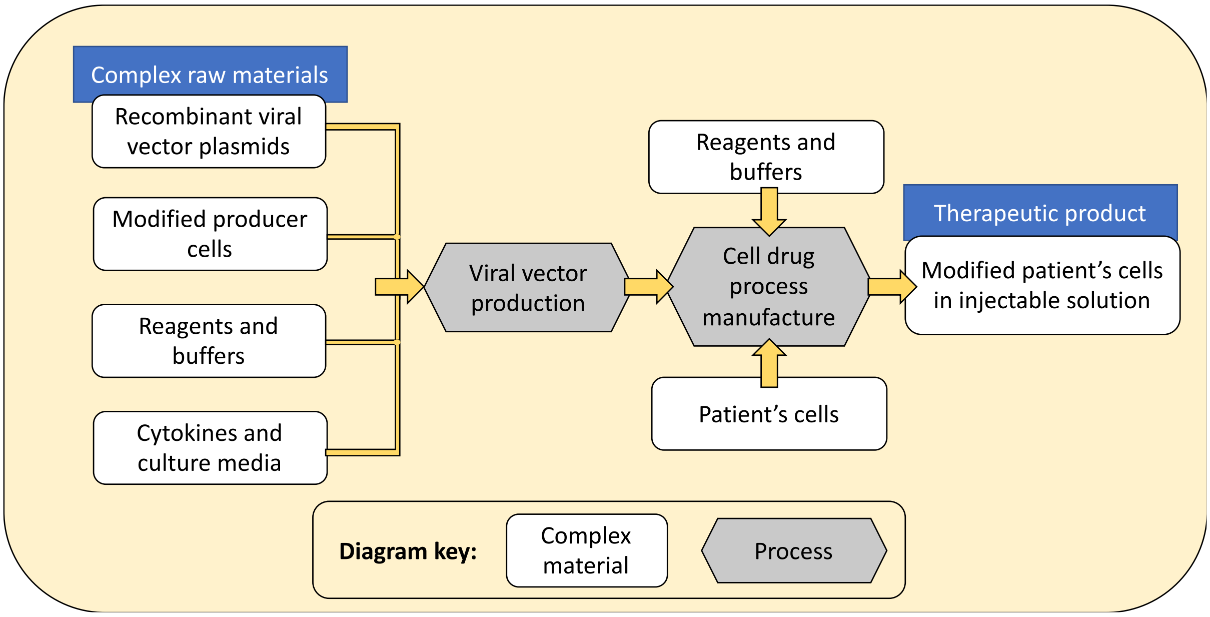


Figure 2-1: Schematic representation of the production process in ex vivo CGT, showing the two main production lines in the manufacturing process: viral vector production, followed by processing of the patient's cells ex vivo during CDP manufacture (Emerson et al., 2020; Kotin, 2011; Merten et al., 2010; Naldini, 2011; Schweizer and Merten, 2010).

The following section describes the manufacturing processes involved in lentiviral vector production and cell drug product manufacture.

2.1.2 3rd generation lentiviral vector packaging system and plasmid production

A viral vector system is derived from its parent genome by a series of genetic modifications, which are designed to remove the viral genes that are potentially pathogenic and to maintain and promote genes that are required for the replication of the virus in the intended virus-producing cells (Giacca, 2010). The third-generation lentivirus was originally developed in the Naldini and Trono laboratories (Gándara et al., 2018) and is a system that is widely used in R&D and clinical applications (Lundstrom, 2018; Merten et al., 2016; Milone and O'Doherty, 2018). The system is well-characterized, capable of delivering genetic material into both dividing and nondividing cells, stably integrates into the host cell genome, providing long-term expression in the target cells, and carries a large amount of genetic material compared to other viral vector systems (Gándara et al., 2018). Compared to previous generations, the third-generation lentiviral vector (LV) system offers enhanced safety due to a number of viral genes that were deleted from the system to create replication-incompetent and self-inactivating vectors (Gándara et al., 2018).

All generations of the LV system are based on the human immunodeficiency virus-1 (HIV-1). The third-generation system utilises four plasmids, including a transfer plasmid containing the transgene or therapeutic gene of interest, in addition to three other plasmids containing packaging genes, which are co-transfected into the producer cells with the transfer plasmid (Gándara et al., 2018). One of the plasmids codes for the *Rev* protein, which is a trans-activating protein responsible for HIV-1 protein expression, another encodes *Gag*, which encodes viral structural proteins and *Pol*, which encodes the retrovirus-specific enzyme reverse polymerase, and the final plasmid codes for the Vesicular stomatitis virus glycoprotein, which forms the virus envelope (Gándara et al., 2018).

Plasmid DNA for vaccine and cell therapy applications are commonly produced in bacterial fermentations, followed by downstream processing to concentrate and purify the plasmid DNA. The manufacture of plasmids with low variability and high quality (good manufacturing practices (GMP) standard) is an important aspect of CGT production (Lopes and Calado, 2018). More details on these processes can be found in the literature (Prather et al., 2003; Urthaler et al., 2012).

2.1.3 Adherent cell culture process for lentiviral vector production

This section provides a description of an adherent cell culture process for the production of LVs, which is similar to the process under study in chapter 5 of this thesis. Due to confidentially reasons, the exact process which is explored in chapter 5 is not described. However, Merten *et al.* (2010) and Ausubel *et al.* (2012) described adherent cell culture processes for the production of LVs by transient transfection, using a third-generation LV system and 293T producer cells, which are sufficiently similar to be consistent with the data that is analysed in chapter 5.

2.1.3.1 Producer cells

293T cells are commonly used for the production of LVs in adherent cell culture systems. 293T cells are derived from human embryonic kidney (HEK) 293 cells by transfection with a plasmid encoding for the simian virus 40 (SV40) large T antigen

(LTA) and a sub-clone was selected for its high yield of lentiviral vectors in transient transfection systems (Merten et al., 2010).

2.1.3.2 Seed train

The process begins with the thawing of cells from a working or master cell bank. The process described by Ausubel *et al.* (2012) begins with a working cell bank of 293T cells. Cells were grown in a medium supplemented with fetal bovine serum and incubated at 37ºC with 5% CO_2. Every three to four days cells were trypsinized and counted before being resuspended in fresh media. This process involved removing media from the cell culture flasks, washing the cells with phosphate-buffered saline (PBS) and removal of the waste. A mixture of Trypsin and PBS was then added to the cell culture flasks and the cells were incubated at 37ºC for 3-15 minutes. The purpose of the Trypsin addition was to process dislodge the cells and reduce aggregation. Next the Trypsin was deactivated by adding medium to the cell culture and cells were recovered using centrifugation. Following this, cells were reseeded into appropriately sized vessels and expanded further.

2.1.3.3 Expansion in cell factory stacks

In large-scale adherent culture processes, such as the process described by Merten *et al.* (2010), the 293T cells were then transferred to cell factory stacks for further expansion until the desired number of cells for transfection is obtained. The cell factory stacks consist of flat trays stacked vertically. These offer a larger volume than the flasks used in the seed train and feature a large surface area for the adherent 239T cells to attach to, with a relatively small footprint (Rout-Pitt et al., 2018). Cells are expanded until a sufficient number is obtained for the intended quantity of viral vectors to be produced.

2.1.3.4 Transfection

Continuing with the process described by Merten *et al.* (2010), after 3 days in the cell factory stacks, the medium was changed 2 hours prior to transfection. Cells were then transfected with a transfer plasmid and three packaging plasmids, using the Calcium Phosphate transfection method first described by Graham *et al.* (1977)*.* The medium was changed the following day and the viral supernatant was harvested twice, at 24 hours and 48 hours after transfection.

2.1.3.5 Downstream processing

Merten *et al.* (2010) provided a description for the downstream processing of GMP-grade LVs. The objective of downstream processing is the concentration and purification of the LVs, including the elimination of process and cell-derived contaminants, and the formulation of the final LV product.

The first step was the clarification of the harvested LVs using low retention membrane filters of decreasing pore size. Next, the filtered stock was treated with an endonuclease (Benzonase) overnight at 4ºC to remove plasmid DNA and residual DNA contaminants. After the filtration and Benzonase treatment, approximately two-thirds of the viral vector particles were recovered, and the infectivity dropped by less than two-fold. DNA content was reduced by around 85%. Anion-exchange chromatography was then used to capture the virus particles. The viral stock was pumped over a DEAE ion exchange chromatography column, and the column was washed with PBS and step-eluted with NaCl, leading to the removal of more than 99% total proteins. The eluate was then concentrated between 20 to 30-fold using tangential flow filtration. After this, the stock was passed through a hollow-fibre cartridge for further reduction of protein and DNA contents. The viral stock was then equilibrated into the formulation medium by size-exclusion chromatography. The collected eluate was then sterile filtered generating the final LV product (Merten et al., 2010).

2.1.4 Bioreactor-based suspension culture process for lentiviral vector production

The bioreactor-based production of LVs begins with the seed train, which is similar to that of the adherent cell culture process, with minor differences due to the use of different cell types and media (Thomas et al., 2013). HEK293-F and HEK 293-H are two industrially relevant cell lines that have been adapted for growth in suspension cultures using serum free media (Malm et al., 2020). A bioreactor-based suspension culture process for the production of LVs was described by Thomas *et al.* (2013). Initially in process development, 293-F cells were grown in shaker flasks using serum free media. The cells were transfected with helper and vector plasmids using 25kD linear polyethyleneimine (PEI) and the process was optimised based on the cell density at the time of transfection, the DNA to PEI ratio and the plasmid volume ratios. Following the optimization, the process was upscaled to a single-use bioreactor system. The cell growth rate in the shaker flasks was replicated in the bioreactor system by controlling the dissolved oxygen concentration and the pH of the culture. The cell culture was expanded from 2L to 10L in the bioreactor and then transfected with the optimized PEI transfection procedure. The LVs were then harvested and concentrated using ion exchange chromatography followed by size exclusion chromatography.

In this work, the downstream processing of the large-scale suspension culture productions was not investigated.

2.1.5 Cell drug product manufacture

In cell drug product (CDP) manufacture, the patient's cells are extracted, processed and transduced with the viral vectors*,* before being prepared for transplantation back to the patient. The essential steps of *ex vivo* gene therapy are isolation of the patient's cells and *in vitro* culture of the target cell types, to enable their isolation, expansion and/or differentiation before or after a genetic modification takes place (Naldini, 2011). Hematopoietic stem cell (HSC) gene therapies are some of the most successful and well-developed gene therapies to date (Morgan et al., 2017), partly due to the long-standing clinical experience in HSC transplantation (Naldini, 2011). HSC gene therapies have been used to treat

Merten *et al.* (2010) described the transduction of CD34+ cells (HSCs and progenitor cells) from umbilical cord blood using LVs containing the WAS gene. CD34+ cells were obtained from the umbilical cord blood by immunomagnetic selection and activated overnight by incubating the cells in medium supplemented with antibiotics and cytokines. Following selection of CD34+ cells, the preactivated cells were transduced with the LVs for 6 hours in the presence of hexadimethrine bromide. The next day cells were washed and tested. The final part of the process that remains is preparation of the cell drug product, which was not described.

2.1.6 Viral vector and cell drug product quality attributes

Critical quality attributes of the LV product described by Merten *et al.* (2010) included the infectious titre and the physical titre. The infectious titre is a measure of the concentration of virus particles capable of infecting cells to cause cytopathic effect following *in vitro* infection. Merten *et al.* (2010) determined the infectious titre by quantitative polymerase chain reaction (qPCR) through infection of HCT 116 cells with serial dilutions of viral vector (Palmer and Ng, 2004). The units of the infectious titre may be expressed as infectious genomes per millilitre or transducing units per millilitre. The physical titre is a measure of the concentration of total viral vector particles in vector preparation (Palmer and Ng, 2004). Merten *et al.* (2010) determined the physical titre using a so-called ELISA apparatus.

In the quality control of the GMP-grade LV product described by Merten *et al.* (2010), several important tests were carried out. Firstly, the LV product was tested to ensure that there were no replication-competent LV particles present. The transfer of plasmid

DNA, adenoviral and SV40 genomic sequences to target cells was tested by qPCR. Total protein content and total DNA content were measured by spectroscopy and spectrofluorimetry. Finally, the presence of cellular DNA was tested by qPCR. For more specific details on the testing methods, such as the qPCR approaches, the reader is referred to the paper (Merten et al., 2010).

The CQAs of the final cell drug product described by Merten *et al.* (2010) included the total cell count, cell viability and the number of vector copies per cell. Cells were counted by inverse microscopy and viability was measured by trypan blue dye exclusion. The number of vector copies integrated per cell was determined by qPCR as described by Charrier *et al.* (2007).

2.2 Cell gene therapy manufacturing challenges

2.2.1 Scale up and scale out

With the field of gene therapy experiencing significant progress in recent years, investment in clinical trials has increased and with it, the demand for viral vectors has grown. For clinical applications, it is also important that the viral vectors are high purity and high concentration and produced in accordance with current good manufacturing practice (CGMP) regulations. Popular and successful viral vectors, such as LVs and rAAVs, have traditionally been produced in laboratory scale systems using adherent cell lines. Such systems are severely limited in their practical scalability and cost effectiveness, which means that they are not suitable for meeting current and future viral vector demands (Manceur et al., 2017; Vlachakis, 2019). Adherent cell cultures are typically grown in flasks and transferred to cell factory stacks, which are multilayer flat plastic trays. These systems require a significant degree of manual handling during the cell expansion, transfection and harvesting phases of viral vector production (Merten et al., 2010; van der Loo and Wright, 2015). Henceforth, scaling cell factory stacks to larger sizes is neither practical nor efficient (McCarron et al., 2016).

Transient transfection techniques are also a challenge to scale up because it is difficult to efficiently transfect large numbers of cells and the process is susceptible to variation (van der Loo and Wright, 2015). The calcium-phosphate (CaP) coprecipitation method is challenging to implement at large-scale because it requires large volumes of plasmid DNA, serum or albumin (potential contaminants) are required in the culture to reduce the toxic effect of CaP on cells, and it is highly sensitive to variations in pH (McCarron et al., 2016). Lipid based reagents that are effective at small scale are toxic expensive making them less suitable for large scale transfection (van der Loo and Wright, 2015).

The transient transfection method that is reported to be the most suitable for scale up in the literature involves the use of polyethyleneimine (PEI). PEI is relatively inexpensive and less toxic to cells than CaP, and PEI mediated transfection is less sensitive to changes in conditions and can be conducted on adherent or suspension cultures, with or without serum (McCarron et al., 2016).

Transient transfection processes have been the ‘gold-standard’ production systems for GMP-grade viral vectors up to now (McCarron et al., 2016; Merten et al., 2010). This is largely due to the fact that transient transfection systems were fast to develop, in compassion to stable cell lines, which require extensive development to establish clonal production cell lines (Manceur et al., 2017; McCarron et al., 2016; Schweizer and Merten, 2010; van der Loo and Wright, 2015). However, transient transfection systems are not well suited to large scale production due to the high levels of variability associated with the transfection procedure and a requirement for large volumes of expensive plasmid DNA, with the potential to end up as a contaminant of the final cell drug product (McCarron et al., 2016).

Researchers and manufacturers of CGT treatments are developing scalable systems, such as bioreactor-based suspension cultures, including stirred tank, rocker, hollow-fibre and fixed bed bioreactors (McCarron et al., 2016; Merten et al., 2016; van der Loo and Wright, 2015), to overcome the challenges associated with traditional adherent cell culture systems. Additionally, stable cell lines are being developed due to their advantages for large-scale manufacturing, including the removal of need for expensive plasmid DNA, ability to produce virus over extended periods, and lower costs and complexity in downstream processing due to the fewer DNA impurities (McCarron et al., 2016). It is expected that stable cell lines will be established and become widely adopted in the future, as they offer the reduced process variability and manufacturing costs and increased safety compared to transient manufacturing processes (Merten et al., 2016).

Up until this point, the discussion has been related to the upstream manufacturing process for viral vectors. The downstream process is heavily dependent on the upstream process, as factors such as scale, choice of reagents and upstream methods impact the selection of downstream unit operations. Adapting downstream processes to deal with larger quantities of viral vector supernatant and changes to its composition is another area where manufacturers must focus on development. This will require manufacturers to transfer technology and scale up unit operations, utilising knowledge and cooperation from a wide range of scientific and engineering disciplines. A large amount of capital and time is required to be invested up front to optimise and validate the manufacturing processes in accordance with GMP guidelines to ensure the safety and efficacy of the product (van der Loo and Wright, 2015).

2.2.2 Process complexity and variability

High levels of process variability is a major challenge in CGT manufacturing. Process variability comes from two main sources: the manufacturing methods and the input materials (Cai et al., 2009; McCarron et al., 2016; Merten et al., 2010). In viral vector production, the producer cells, plasmid DNA, transfection reagents and culture media are all complex raw materials with the potential to introduce variability into the process (Emerson et al., 2020). Lot-to-lot variability in reagents, such as the HEPES-buffer saline that is used in transfection, has been reported to contribute significantly to process variability (van der Loo and Wright, 2015). Producer cells are sensitive to storage and process conditions and can undergo a limited number of passages before mutations start to cause adverse effects (Merten et al., 2010). Variability in storage and process conditions early in the seed train can cause differences in cell condition and metabolic state that amplify variability throughout the rest of the process (Streefland et al., 2013).

The use of complex mammalian cell lines, such as HEK 293 or 293T cells, heightens this issue because they are susceptible to variation and they introduce additional critical process parameters (CPPs) in cultivation compared to yeast or bacterial cultivations (Streefland et al., 2013). Identification of CPPs and implementation of tight control schemes is necessary to reduce batch-to-batch variability. The transfection process is an example of a procedure that requires tightly controlled conditions and is otherwise a key source of process variability (McCarron et al., 2016). As mentioned previously, adherent cell culture processes involve a high degree of manual handling, which can be a source of variability introduced by human operators. Manual tasks include passaging and counting cells, adding reagents and transferring materials between vessels (Kotin, 2011; Merten et al., 2010; Schweizer and Merten, 2010). The process operators are highly trained and skilled; however, manual process inevitably introduce more variability than machine automated processes. Furthermore, the systems are open systems to allow operators to interact with them and this creates a larger risk of contamination (Kotin, 2011; McCarron et al., 2016; Rout-Pitt et al., 2018;

van der Loo and Wright, 2015). The shift toward bioreactor production will reduce some of these issues as bioreactors are closed systems with a high degree of automation.

In cell drug product manufacture, the patient's cells are a key source of variability, as the cell characteristics can vary greatly from patient to patient. This can be partly attributed to genetic heterogeneities, epigenetic differences, or transcription regulation diversities (Stroncek et al., 2010). Additionally, cells used in autologous cell therapies can differ due disease type, treatment history or stage of the disease. Materials such as growth factors and cytokines used in cell drug product manufacture exhibit genetic polymorphisms and it is likely that these impact cell health and behaviour in vitro (Stroncek et al., 2010). The cell collection process can even contribute to variation in the quantity and viability of cells obtained (Stroncek et al., 2010).

Another important consideration is the high degree of complexity involved in CGT therapy manufacturing, which is inherent to the complex biological systems involved. Viral vector production is a particularly complex phase of the process, where many CPPs are involved and the mechanisms by which the process conditions impact viral vector production are not well understood. The viral vectors are complex biological nano particles, which are sensitive to their environment, meaning that carefully controlled conditions are required to prevent their degradation during downstream processing, formulation and storage (Emerson et al., 2020; Merten et al., 2016). Production of most viral vectors is biphasic, meaning that different conditions are required in the cell expansion phase versus the virus production phase, due to shifts in cell state and metabolic activity (Gálvez et al., 2012; Petiot et al., 2015). There is in general a lack of advanced process knowledge due to processes and products being at a relatively early stage of development (Kaemmerer, 2018; Vlachakis, 2019).

2.2.3 Material characterization and process measurements

Characterization of materials is a highly important aspect of CGT manufacturing. The complex raw materials, such as the culture media and reagents, should be well characterised to ensure quality and to understand the composition of materials going into the process. However, it is difficult to characterize all the materials going into the process as there is a lack of capable technology. Difficult to characterize materials include animal derived components in culture media, producer cells and the patient's cells (Li et al., 2010; Stroncek et al., 2010). Characterization of the viral vector and cell drug products is also essential so that the quality of the product can be controlled and so that the product attributes can be related to the manufacturing process variables to guide process development and optimisation. It is also critical to characterize the final cell drug product to ensure its safety, potency and efficacy (Merten et al., 2010). This represents a significant challenge in CGT manufacturing because assays are difficult and costly to develop and often there is a significant delay between collecting the sample and receiving the results. Moreover, the error in some of the key assays is high, for example the infectious titre assays for viral vectors have errors as high as 36% (Roldão et al., 2009).

The lack of advanced process understanding in CGT manufacturing is exacerbated by a lack of online process measurements for key molecular compounds and process parameters. Currently, the slow turnaround time associated with offline or at-line measurements for important process parameters such as the cell concentration, cell viability and properties of the culture, mean that it is difficult to develop online process monitoring and/or control schemes (Ansorge et al., 2011). This is a major limitation for the optimisation of viral vector manufacturing processes, since operators are unable to track batch progress and correct trajectories in real time (Emerson et al., 2020). Moreover, it hinders the development of process knowledge as it is not possible to learn more about the process mechanisms in the absence of good data on the chemical, physical and biological elements involved.

2.3 Multivariate data analysis in the chemical and biochemical industries

Multivariate data analysis covers an array of data-driven modelling approaches, which involve multiple variables in a single relationship or set of relationships (Hair, 2014). The advantage of these techniques over traditional univariate or bivariate techniques is that they capture the relationships between numerous variables simultaneously, often facilitating a reduction of dimensionality and a simplified interpretation of the overall dataset (Hair, 2014; Kirdar et al., 2007; A S Rathore et al., 2014). These techniques are highly relevant in the analysis of bioprocess manufacturing data, since these datasets are often complex with high dimensionality. Previously, practitioners have recognised that the use of univariate or bivariate techniques under these circumstances is likely to produce misleading results (A S Rathore et al., 2014). Moreover, theoretical/mechanistic process models are frequently not available for bioprocesses due to the large number of variable interactions and inherent complexity in biological systems. Reduced order more models, such as multivariate data analysis

(MVDA) models, are significantly easier to realise in many cases (O'Malley et al., 2012).

MVDA and machine learning (ML) techniques have a diverse range of uses, across different industries, and even within manufacturing. Some well-established uses of MVDA in chemical and biochemical processing are outlined in Figure 2-2. One simple distinction to make between different MVDA applications is between online and offline applications. Online applications include the use of MVDA models to make inferences from process data in order to monitor and/or control processes in real time or near-real time. Offline applications include retrospective investigations of historical process data to gain insights into process behaviour. Other applications, such as the characterization of materials, may be conducted online or offline depending on the objective. Here, some of the relevant applications of MVDA to mammalian and microbial process are reviewed to provide examples of the potential benefits that MVDA can bring to CGT processes. Both bacterial and mammalian cell cultures are of relevance since they are used in the manufacture of CGT treatments. Specifically, bacterial cell cultures are used for production of plasmids and mammalian cell cultures are used in viral vector production and processing of the patient's cells.

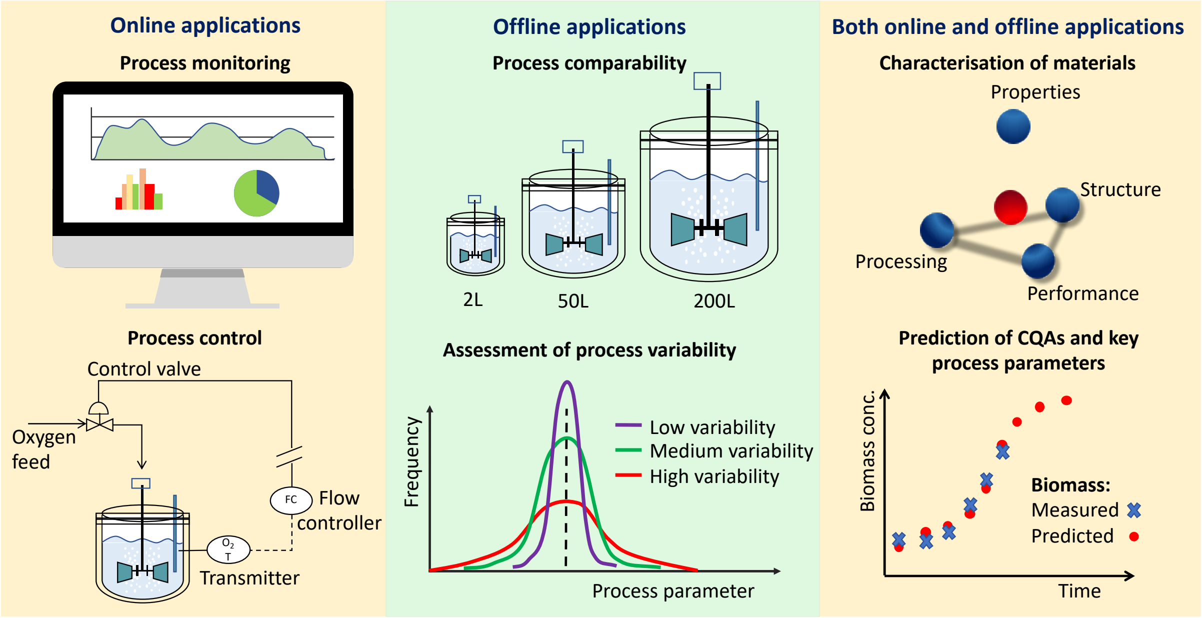


Figure 2-2: Applications of MVDA in bioprocessing (Emerson et al., 2020). Online applications include process monitoring and control. Offline applications include process comparability during scale-up and technology transfer and assessment of process variability. Other applications are both online and offline, such as characterisation of material and prediction of key process parameters.

2.3.1 Retrospective investigations of process data

Retrospective investigations of process data may be carried out using data that originates from a design of experiments (DoE) or using historical process data from production or development periods where no DoE was conducted. The approach to the analysis and the main challenges depend on the structure of the data and its provenance. Data originating from a DoE has key features which are uncorrelated by design, to allow for relationships between independent variables and the dependent variables, e.g. product quality attributes, to be observed under controlled conditions. While a DoE may produce more ideal data, there is an abundance of data that is produced during the development of processes and during production campaigns. This data is often influenced by external factors, and features multicollinearity, and at times too much or too little variation to observe the variable relationships. Nevertheless, this data has been shown to provide valuable insights and can offer a perspective on process behaviour that is not observable under DoE conditions. For example, insights into process drift or variability during production campaigns may be an important area of interest. MVDA has frequently been used for general investigations into historical process data, where the broad objective is to derive beneficial insights into process behaviour. There are also more specific research objectives which have been explored within these investigations, for example, these include evaluation of process variability, comparability across production scales, quality and robustness of process control or relationships between process parameters and product critical quality attributes (CQAs).

Le *et al.* (2012) analysed 243 batches from the manufacturing of a recombinant IgG molecule using Chinese Hamster Ovary (CHO) cells. The seed train was operated in the same manner for all batches (80L, 400L to 2000L) and the full production scale was operated in fed-batch mode. Inspection of the process data revealed that during the production period, there was significant variation in the pre-harvest antibody concentration, lactate concentration profiles and the maximum cell density. To gain a deeper understanding of the process behaviour, the authors decided to use supervised learning techniques, PLS regression and support vector regression (SVR), with process parameters as inputs to predict the final antibody titre and lactate concentration. It was found that the inclusion of data from early in the production greatly increased model predictive performance and that the history of the culture significantly influenced the process outputs. It was also observed that the lactate concentration profile provided a good reference for the state of the cell culture and the final antibody yield. Henceforth, it was suggested that the lactate concentration should be monitored to allow operators to intervene early and steer the metabolism of cells in the culture towards higher lactate consumption, and ultimately higher product titres.

Mercier *et al.* (2014) studied a process with high relevance to CGT processes due to the use of human cells (PER.C6 cell line) in a process at the early stages of development. The dataset contained 17 2L and 10 10L bioreactor-based cultivations operated in perfusion mode. Multiway PCA was used initially to assess the key features of variance and correlations within the data. MPCA facilitated the identification of the root causes for batch deviations and uncovered differences in process conditions between the 2L and 10L scales, which were previously considered to be comparable. The authors also developed multiway PLS regression models to predict the viability and doubling time of the cell culture; however, this was unsuccessful due to information missing from the early development dataset. The authors concluded that the value of early development process runs could be greatly improved by taking a more strategic and structured approach to experimentation from the very beginning of process development. Nevertheless, the authors stated that the utilisation of MVDA on early development datasets is a worthwhile endeavour, as they able to generate insights useful for process development and scale-up.

2.3.2 Process comparability

Evaluation of process comparability is an important usage of MVDA, which has been widely reported in the literature. In bioprocessing, process comparability is a highly relevant topic due to the fact that many processes are operated in batch or fed-batch mode, henceforth there is a need to evaluate batch-to-batch variability. Moreover, there is a requirement to assess process comparability throughout the phases of process development, scale-up and optimisation. PCA is a frequently exploited technique in this regard, since it has the ability to capture information from the whole set of process parameters in a reduced set of latent components, which highlight key features of variance in the data. This allows the researcher to more easily assess differences in process conditions between batches and to later identify the root cause (Kirdar et al., 2008).

Lopes and Calado (2018) analysed online and offline process data from 11 batches of *E. Coli* DH5-α in the production of plasmids (pVAX-LacZ). PCA was initially used to assess process variability and comparability and multivariate analysis of variance was used to evaluate the statistical significance of the differences identified. Following the feature extraction exercise, linear discriminant analysis (LDA) was used to predict the cell growth phase at numerous timepoints throughout the process, based on the rates of change in metabolite concentrations. The PCA and LDA models both provided a comparison of the performance of batches and the LDA model enabled prediction of the metabolic state of the culture that could be useful in process monitoring.

Rathore *et al.* (2014) investigated an unspecified cell culture process, with a dataset comprised of small-scale (2L), pilot-scale (2000L), and production-scale (15,000L) runs. PCA was used for feature extraction, where the two most important latent components identified two clusters of batches. One of the groups contained production-scale and pilot-scale runs that used specific raw material lots and was associated with atypical product attributes, elevated culture osmolality and elevated lactate and sodium concentrations. The PCA model provided a comparison of the process characteristics between the three scales and shed light on key correlated features within the data. Subsequently, PLS regression was used to link the process conditions and cell culture growth dynamics to the product quality attributes.

2.3.3 Characterization of materials

Characterisation of materials is sometimes a requirement, for example, in the quality control of a biopharmaceutical product it is necessary to know the concentrations of certain potential contaminants, and other times it is a beneficial, for example online characterisation of a bioreactor-based suspension culture can offer improved process monitoring and control. Raw materials are a key source of variability, particularly complex biological raw materials, which can have a significant influence on process performance. It is therefore desirable to characterize the raw materials before they're introduced to the process in order to minimize process variability and optimize product yield and quality attributes.

Lopes *et al.* (2004) analysed a fermentation process producing an active pharmaceutical ingredient (API), where the main nitrogen source for the culture was soyabean flour. From past experience, it was known that a simple change in the soyabean flour material lot could impact the yield of the API. To investigate this further, the authors took 25 soyabean flour samples from 25 fermentations and characterised them using FT-NIR reflectance spectroscopy. PCA was applied to the NIR spectra and the resulting latent components were passed to a Kohonen network with four output nodes. This enabled classification of the API concentration into four categories ranging from poor to very good. The accuracy of the model was around 70% and misclassifications were always within one neighbourhood of the true value; henceforth, the authors concluded that the model was able to successfully predict the process outcome using the FT-NIR characterisation of the soyabean flour feed material.

Li *et al.* (2010) explored the characterisation and quality assessment of media components used in a CHO cell culture for the production of recombinant proteins. The authors noted that in good biopharmaceutical manufacturing practice there are two key requirements, the first is the correct identification of raw material components, and the second is the quality assessment of raw materials prior to their use. With this in mind, the authors decided to characterize the material with different spectroscopic techniques and found that Raman spectroscopy was most suitable. After pre-processing the spectra, PCA and soft independent modelling of class analogy were successfully applied to the spectra to identify five chemically defined commercial media components. Analysis of the variance in the spectra also provided insights into the consistency of the media samples. The authors suggested that the strategy could be used in a number of applications, including for in-house sample handling, tracking and quality control.

2.3.4 Process monitoring and control

The online monitoring and control of bioprocesses is currently hindered by a lack of online sensors for the measurement of key process variables (Melcher et al., 2017; Rathore et al., 2010; Streefland et al., 2013). Without sensors to measure variables such as cell density or product concentration online, there is no basis upon which to build an online control scheme for optimising these critical process parameters. The ultimate objective is to set up online feedback control loops to optimise the product yield and quality (Melcher et al., 2017). Even online monitoring of these parameters could provide great benefits, as it would allow process operators to observe the batch trajectory and react to process deviations in real time (Chen et al., 2011). Due to the lack of direct online measurements, researchers have focused efforts on predicting key process variables, using known process variables and process analytical technology, such as spectroscopic methods, which are capable of providing online readings (Emerson et al., 2020).

Zheng and Pan (2016) studied a batch glutamate fermentation process, using *Corynebacterium glutamicum.* Such processes are difficult to control due to high levels of variability associated with the biological materials and due to a lack of online process measurements (Zheng and Pan, 2016). The authors therefore decided to develop a predictive model to enable the tracking of process performance. In their work, they developed a Gaussian process regression (GPR) model to predict the glutamate concentration online, using other online available process measurements as the model inputs. Model validation was carried out in 10 5L fermentation runs and it was found that the GPR model could provide effective guidance for online process control and optimisation of the glutamate yield.

Melcher *et al.* (2017) developed predictive models for key process parameters in a fed-batch *E.Coli* fermentation, including cell dry mass, optical density and protein concentration. Structured additive regression (STAR) models were used in combination with boosting to select predictor variables. The model inputs included online available process variables and 2D florescence spectroscopic data. The use of STAR models allowed incorporation on curvilinear and interaction effects and boosting enabled the most important spectroscopic frequencies and process variables to be determined and utilised in the prediction. The results showed that the STAR model could predict the cell dry mass, optical density and soluble protein concentration with relative errors of $\pm 3\%$, $\pm 6\%$ and $\pm 16\%$, respectively. The authors concluded that this would allow effective online monitoring.

Clavaud *et al.* (2013) conducted an MVDA study on process data from 10 production scale bioreactor (12,500L) runs, during the manufacture of monoclonal antibodies using CHO cells, cultivated in fed-batch mode. The process was monitored using a Fourier transform near infrared multiplex analyser. Initially, the resulting NIR spectra were pre-processed and analysed using PCA, which showed that a significant portion of the variance could be explained by deviations in the process trajectory. From. This information, it was clear that the spectra could be used to evaluate process variability and detect abnormal process behaviour. The authors followed up by using PLS regression to predict key process variables, including metabolite concentrations, viable cell density and product titre. The modelling results demonstrated accurate predictive performance for all of the media components that were modelled, leading the authors to conclude that the models could be effective in process monitoring and/or control.

2.4 Multivariate data analysis in cell gene therapy manufacturing

Multivariate data analysis has not yet been widely exploited in CGT manufacturing (Emerson et al., 2020). This may be attributed to the fact that CGT products and their manufacturing processes are still at a relatively early stage of development, with many changes taking place, such as development of stable cell lines and transition to scalable cell culture systems (McCarron et al., 2016; Merten et al., 2016; Ramirez, 2018). In the literature, MVDA has been leveraged to contribute towards solutions for several key manufacturing challenges that are faced in bioprocessing. This includes mammalian and bacterial cell bioprocesses for production of proteins, monoclonal antibodies and other active pharmaceutical ingredients. Many of the challenges that MVDA has already been used to tackle are relevant to CGT manufacturing, including characterization of materials, evaluation and targeted reduction of process variability, assessment of process comparability throughout process development, scale-up and technology transfer and providing insights into process behaviour. This thesis is mainly focused on using MVDA to generate beneficial insights into process variability, comparability and the relationships between manufacturing process variables and critical quality attributes of the viral vector and cell drug products.

2.5 Objectives of this work

The specific objectives of this work include the following:

- Provide a review of appropriate modelling techniques for bioprocess analytics and describe the particular challenges of applying these techniques to data from cell gene therapy manufacturing processes, detailing the steps taken to overcome challenges encountered.
- Development of linear and nonlinear programming approaches to solve and identify sparse PCA models, which ease the interpretability of the PCA model, enabling greater clarity of insights derived from the modelling activity.
- Analyse data from the adherent viral vector manufacturing process, the cell drug product manufacturing process and the bioreactor-based viral vector manufacturing process, in order to derive useful insights into process behaviour.

More specifically, for the adherent viral vector manufacturing process data, the objectives are to:

  - utilise unsupervised learning techniques to assess and evaluate within process variability and batch-to-batch clustering in both the upstream and downstream processes
  - explore supervised learning techniques and develop predictive models to identify critical process parameters and capture the relationship between process parameters and critical quality attributes of the viral vector product.

For the cell drug product manufacturing data, the objectives are to:

  - utilise unsupervised learning techniques to assess and evaluate within process variability and batch-to-batch clustering
  - develop predictive models to identify critical process parameters in CDP manufacture and capture the relationship between process parameters and critical quality attributes of the cell drug product.

For the cell drug product manufacturing data and adherent viral vector manufacturing data combined, the objectives are to:

  - develop predictive models to identify critical process parameters and capture the relationship between process parameters and critical quality attributes of the cell drug product. This time investigating whether the effect of process parameters from viral vector manufacturing are influential on critical quality attributes of the final cell drug product.

For the bioreactor-based viral vector manufacturing data, the objectives are to:

- utilise unsupervised learning techniques to assess within process variability and batch-to-batch clustering and to evaluate process comparability across different bioreactor production scales.
- develop predictive models to identify critical process parameters and capture the relationship between process parameters and critical quality attributes of viral vector product. Evaluate the performance of models used to make predictions across a range of bioreactor volumes.

Chapter 3

Model development and evaluation

In this chapter, a range of MVDA techniques are summarised, the theory is described with references to more detailed literature and examples of practical applications. Further to the presentation of multivariate models, key model development practices such as variable selection, data partitioning and cross validation are discussed. The material in this chapter is a foundation for the methodology implemented throughout this thesis and it also serves as a functional collection of data analysis techniques and practices, which are beneficial in a wide variety of disciplines.

3.1 Exploratory data analysis

Exploratory data analysis (EDA) is often carried out as the first phase of a MVDA investigation. The purpose of EDA is to assess the data structure and to highlight the key features of variance, correlations and patterns, which may be present. It also provides an overview of the dataset, such that similarities and differences between samples, systematic trends and outliers can be identified (Behrens, 1997; Biancolillo and Marini, 2018). EDA can provide many insights into the system under study and can help to identify specific areas that the practitioner should focus on in subsequent analysis or investigations. It provides information that is highly relevant to predictive modelling, therefore, EDA is usually carried out prior to the development of predictive models in order to guide selection of samples, variables and models. In this thesis, principal component analysis was the primary tool used for EDA.

3.1.1 Principal component analysis

Principal component analysis (PCA) is frequently used in MVDA as a feature extraction or pattern recognition technique (Hair, 2014). It is also used as a data compression technique, as it summarizes large datasets using a greatly reduced number of latent variables, which capture the maximum amount of information possible. PCA has proved to be a useful technique across many fields of science and engineering, including geophysical research (Li et al., 2013), bioinformatics (Zheng et al., 2012), signal processing (Harmouche et al., 2014) and chemometrics (Bro and Smilde, 2014), to name a few. There are numerous claims to the first use of PCA in the literature, the most famous early work was published by Karl Pearson in 1901. Hotelling (Hotelling, 1933) notably redeveloped the technique in the 1930s, when PCA took on the format that is most commonly used today.

3.1.1.1 Theory

Given a data matrix $\boldsymbol{X}$, comprised of I observations on J observed variables, the aim of PCA is to reduce this data to a set of k new variables, where k is small relative to J. The new variables are a weighted-linear combination of the original variables, known as variates, latent variables or principal components (PCs), which may be written as $\boldsymbol{t} = \boldsymbol{x}_1 \times \boldsymbol{w}_1 + \cdots + \boldsymbol{x}_J \times \boldsymbol{w}_J$. In matrix notation this becomes $\boldsymbol{t} = \boldsymbol{X}\boldsymbol{w}$, where $\boldsymbol{w}$ is the vector of loadings with elements $\boldsymbol{w}_j (j = 1, \dots, J)$ and $\boldsymbol{t}$ is a vector of scores. The weightings are determined by maximising the variance of each component in order to maximise the amount of information explained by the set of latent variables. The principal components are restricted to a set of principal axes where all components are mutually orthogonal so that covariance is eliminated. This design means PC one captures the most variation, and the variation captured in subsequent PCs decreases monotonically. Since multiplying $\boldsymbol{w}$ by an arbitrarily large number will make the variance of t also arbitrarily large, it is necessary to normalize the weighs, implemented by requiring that their norm, i.e. sum of squares, is equal to one (Bro and Smilde, 2014). Consider first a single PC. The formal problem is given by (3.1):

$$\underset{\|w\|=1}{\operatorname{argmax}}\, var(\boldsymbol{t}) \tag{3.1}$$

The variance of t is given by $\boldsymbol{w}^T\boldsymbol{Q}\boldsymbol{w}$ where $\boldsymbol{Q}$ is the covariance matrix

$$\boldsymbol{Q} = \frac{\boldsymbol{X}^T\boldsymbol{X}}{n-1} \tag{3.2}$$

It is assumed that the data matrix ***X*** is mean centred, so that all latent variables are also mean centred. The problem rewritten below is a standard problem in linear algebra.

$$\underset{\|w\|=1}{\operatorname{argmax}}(\boldsymbol{w}^T\boldsymbol{Q}\boldsymbol{w}) \tag{3.3}$$

If the variance explained by one PC is insufficient, a PCA model can be determined with multiple PCs, which will still represent a significant dimension reduction if the number of PCs is small compared to the original number of variables (Bro and Smilde, 2014). For models with more than one PC, the principal components subsequent to PC one are subject to an orthogonality constraint (3.4), to ensure that the variance explained by each of the PCs is independent.

$$for\ k \geq 2,\ \boldsymbol{w}_k^T\boldsymbol{w}_{k-i} = 0,\ for\ i = 1,..,k-1 \tag{3.4}$$

The scores and loadings vectors may be written into a standard regression equation, in this way PCA may be viewed as a modelling activity, and standard regression tools may be used to assess the quality of the model. For a PCA with multiple components, the model representation is given by equation (3.5):

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E} \tag{3.5}$$

where ***P*** is the loadings matrix and ***T*** is the scores matrix, which represents the observations in the latent variable space. The product of the scores and loadings matrices give an estimation of the original data.

$$\boldsymbol{T}\boldsymbol{P}^T = \widehat{\boldsymbol{X}} \tag{3.6}$$

To judge the quality of the scores as a summarizer of the original data, the explained variation of $\boldsymbol{t}$ may be calculated using equation (3.7).

$$explained\ variation = \frac{\|\boldsymbol{X}\|^2 - \|\boldsymbol{E}\|^2}{\|\boldsymbol{X}\|^2}100 \tag{3.7}$$

In this work, standard PCA was implemented using the inbuilt PCA function in MathWorks® software MATLAB.

3.1.1.2 Determining the number of principal components

As mentioned previously, the initial components explain the most variation while later components explain a smaller percentage. Usually, it is not desirable to explain all of the variance in the original dataset because it is known to contain a degree of noise. Numerous methods have been described for selecting the optimal number of principal components, where the aim is to capture all of the important features of variance while rejecting principal components containing a large percentage of noise. One popular method is known as the scree test, where the eigenvalues or percentage of variation explained is plotted for each corresponding principal component. The number of principal components is then determined based on the point at which the decrease in variance explained becomes linear. This point indicates that the model is starting to capture a larger percentage of noise and therefore represents a good stopping point (Bro and Smilde, 2014). Figure 3-1 shows an example of the scree test applied to a PCA model trained on a classic dataset. The scree test method was the method chosen in this work.

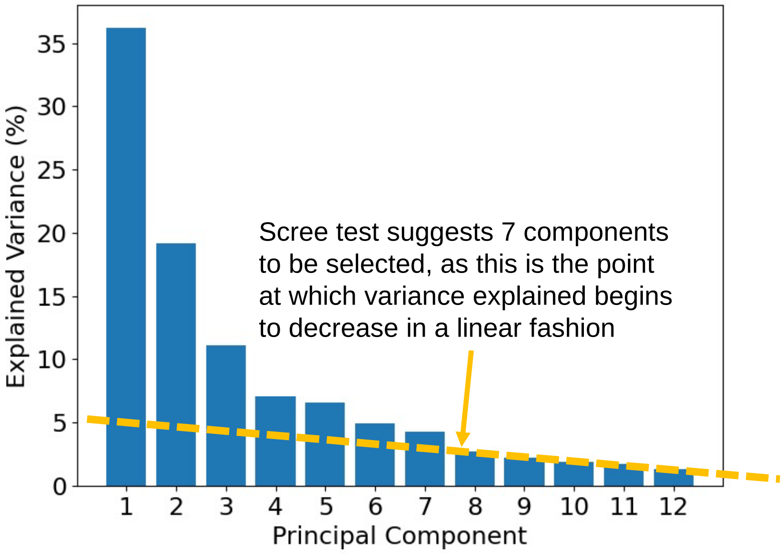


Figure 3-1 Explained variance versus number of components in a PCA model for a classic wine dataset. The dashed line shows the point at which the loadings begin decreasing in a linear fashion.

3.1.2 Parallel coordinates plots

Parallel coordinates plots map each observation in a dataset as a line on a graph, which tracks its values across multiple variables. An example is shown in Figure 3-2.

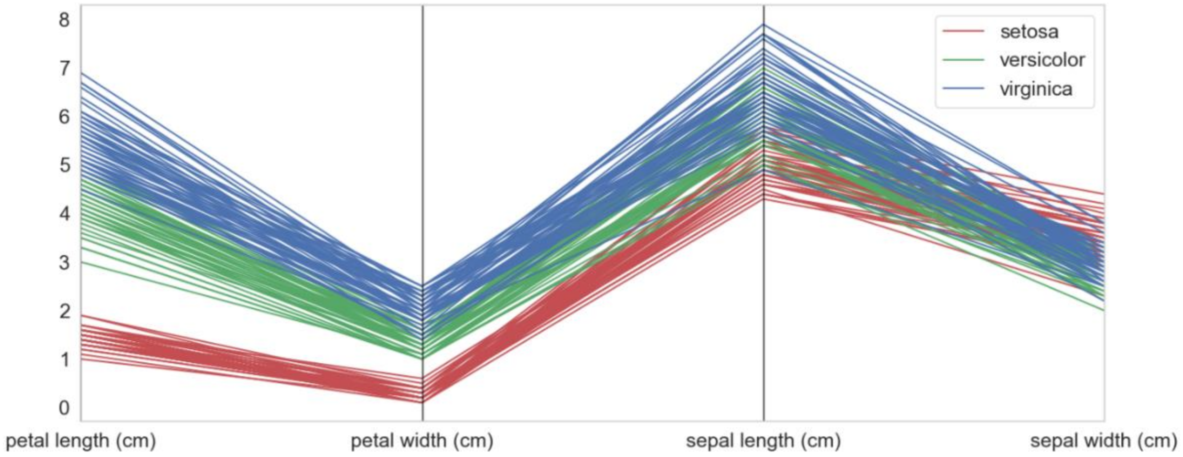


Figure 3-2: Example of a parallel coordinates plot for Fisher's Iris dataset. The plot shows four variables recorded for 3 different plant species, which are represented by the different colours on the plot.

Parallel coordinates plots allow the observer to identify trends and correlations between multiple variables while utilising colour to show an additional dimension, which can be a categoric or continuous response variable for example. Figure 3-2 for example shows four variables: petal length, petal width, sepal length and sepal width and their relationship to a categoric variable, the plant species to which the samples belong. Each sample is given by one line on the plot. In this plot, it is easy to see that petal length and petal width are positively correlated with one another because the majority of the lines on the plot do not cross over one another, i.e. they remain in the same order on the y-axis. Additionally, the fact that these two variables provide separation of the colours (plant species) indicates that both are predictive of the plant species. Conversely, sepal length and sepal width are less powerful predictors for classifying the plant species. In this thesis, parallel coordinates plots were used to further evaluate features of process variability that were identified with PCA.

3.1.3 Other EDA tools

Additional EDA tools that were used in this thesis include histograms, bivariate scatter plots and the correlation matrix. All three of these techniques were used for data exploration to understand the characteristics of the data and to provide initial insights into the trends and correlations present.

3.1.3.1 Histograms

Histograms are used to visualise the distribution of variables, which is an important characteristic to understand when trying to make inferences from data. A histogram plots the distribution of a numeric variable on a bar graph. Each bar typically represents a range of values called a bin and the height of each bar represents the frequency of datapoints that fall within the corresponding bin. Figure 3-3 shows an example of a histogram for Fisher's Iris dataset. The histogram shows the distribution of sepal width for the Virginica plant species, which closely follows a normal distribution.

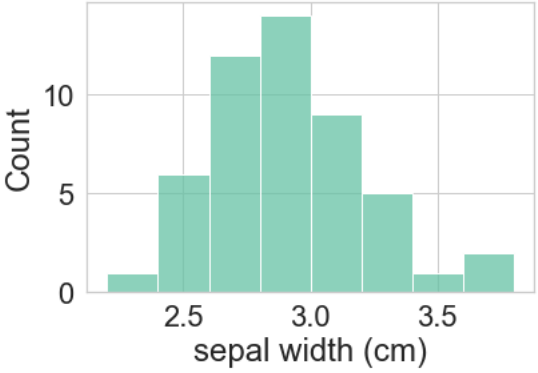


Figure 3-3 Histogram example for the sepal width of the Virginica species found in the Fisher's Iris dataset.

Statistical techniques often make assumptions about the distribution of data and histograms can be used to check which techniques are applicable and whether the assumptions are valid. The plots can also help to identify outliers and gaps in the data that may need to be addressed. Typical distributions that can be identified using histograms include symmetric unimodal (normal distribution), left or right skewed, uniform, bimodal or multimodal distributions. More information on histograms and distributions can be found in (Hair, 2014).

3.1.3.2 Scatter plots

Bivariate scatter plots use markers to show the datapoints for one variable plotted against another. The plots can be implemented for one independent variable versus another independent variable and for an independent variable versus a dependent variable. They allow for visualisation of any correlation between the two variables, whether it be linear or nonlinear, hence they quickly provide useful insights into the variable relationships in a dataset. Additionally, if there are classes or categories present in the data, colouring the datapoints by the class highlights any differences between the classes that are present in the two variables. This can be a useful exercise for determining whether the variables may be good predictors in a classification task. Figure 3-4 shows a matrix of scatterplots for the four variables in the Fisher's Iris dataset in the off-diagonal elements, while the diagonal elements display a histogram for each variable.

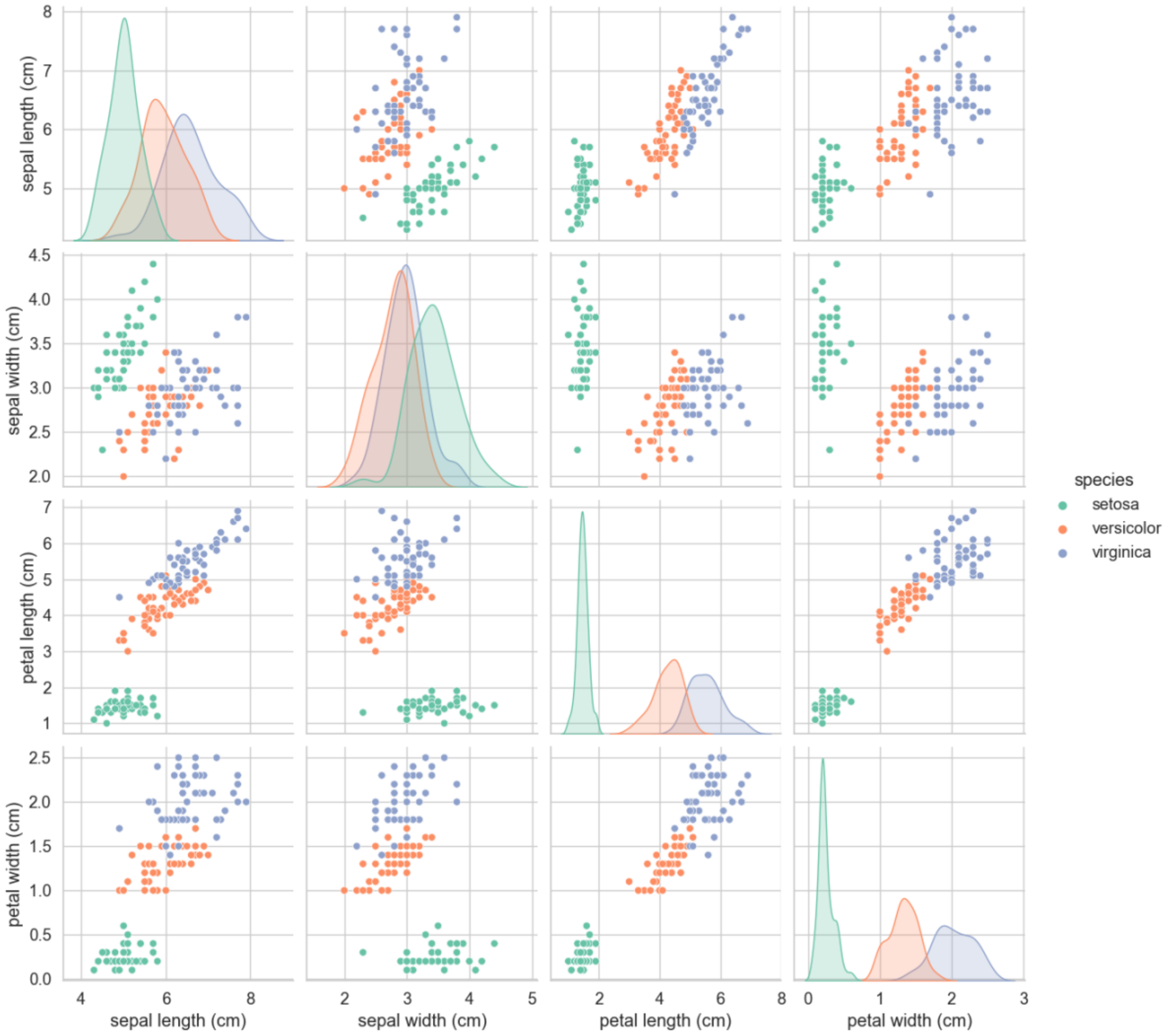


Figure 3-4 A matrix of scatter plots for the Fisher's Iris dataset on the off-diagonal elements, while the diagonal shows histograms for each variable

Figure 3-4 shows that there is a strong positive linear correlation between petal length and petal width. It is also evident that these two variables are good predictors of plant species. The other scatter plots show some weaker correlations for certain species, for example petal length and sepal length correlate positively for the versicolor and virginica species. Some of these observations were also made with the parallel coordinates plot (Figure 3-2) of the Fisher's Iris dataset, demonstrating that EDA techniques can serve purposes that overlap, although they generally excel at different things. The scatter plots can highlight linear and nonlinear corelations with more clarity than parallel coordinates (Hair, 2014).

3.1.3.3 Correlation matrix and heatmaps

The correlation matrix is a useful tool for quickly evaluating the Pearson's linear correlation coefficient between each pairing of variables in a multivariate dataset. It can be visualised by plotting the coefficient matrix as a heatmap, where positive correlations are one colour and negative correlations are another and the depth of the colour can be made to scale with the magnitude of the coefficient. An example of a heatmap for Fisher's Iris data is shown in Figure 3-5.

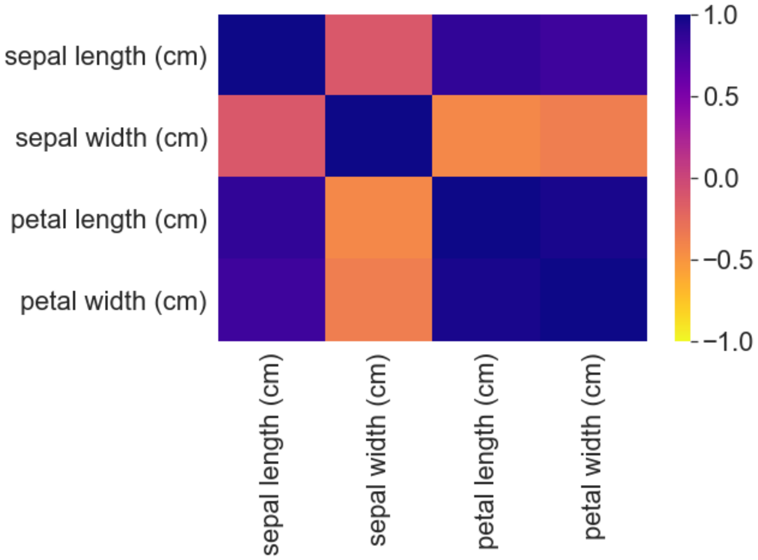


Figure 3-5 Heatmap of a correlation matrix for the four independent variables in the Fisher's Iris dataset.

The heatmap in Figure 3-5 shows that there are no significant negative correlations, but it highlights the large positive correlation between petal length and petal width. It also highlights the moderate positive correlation between sepal length and petal width/length. Heatmaps are good for getting an overview of the linear correlations in a dataset and can be used to display a relatively large number of variables at once.

3.2 Predictive modelling

Predictive modelling techniques model the relationship between a set of independent variables (model inputs) and one or more response variable. If the response variable is categorical, the prediction task is described as classification; alternatively, if the response variable is a continuous metric variable, the prediction task is described as regression. Once the prediction model has been trained to model the relationship between the predictors and the response, it can be used to predict the outcome of future events (Kuhn and Johnson, 2013). This means the model can be used to explore hypothetical scenarios to guide decision making and the model structure and coefficients can provide insights into the system from which the data originates.

Important applications of predictive models in bioprocessing where previously explored in chapter 2, section 2.3. In this thesis, multiple linear regression (MLR) and partial least squares (PLS) regression were used for prediction of critical quality attributes of the viral vector and cell drug products.

3.2.1 Multiple linear regression

MLR is a classic statistical technique with origins dating back to the late nineteenth century (Stanton, 2001). It has been widely used by scientists and statisticians to understand the relationships between variables in systems of interest. It is a fundamental technique in statistics, which is still frequently used today. MLR models the relationship between two or more regressors or predictor variables and a response variable of interest. The goal of MLR is to use the known values for a set of predictor variables to predict the unknown value of the response variable (Hair, 2014).

3.2.1.1 Theory

Given a set of predictor variables, $\boldsymbol{X}$, and a response variable, $\boldsymbol{y}$, the goal of MLR is to identify a set of coefficients, $\boldsymbol{\beta}$**,** that minimises the sum of squares difference between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$, where $\hat{\boldsymbol{y}}$ is the estimate of $\boldsymbol{y}$ given by the MLR model, as in equation (3.8).

$$\hat{\boldsymbol{y}} = \boldsymbol{\beta X} + c \tag{3.8}$$

Here, the set of coefficients, $\boldsymbol{\beta}$, is a vector with *m* elements corresponding to each of the *m* variables in $\boldsymbol{X}$ and c is a scalar intercept term. The problem of fitting the MLR model to training data to identify the coefficients is solved using a least squares optimisation algorithm (Hair, 2014), where the objective is to minimise the error, ε, given by equation (3.9)

$$\varepsilon = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3.9}$$

where *n* is the number of observations. The quality of the model fit to the training data and estimation of the model’s predictive capability are both important aspects to evaluate when developing a regression model. Error and goodness-of-fit metrics are discussed further in section 3.3.

3.2.2 Partial least squares regression

Partial least squares (PLS) regression is a popular regression model that is widely used in science and engineering domains. For example, PLS has been used in chemometrics to relate high-dimensional spectral data to chemical composition (Leardi, 2000) and in geophysical research to relate sea temperatures to hurricane activity (Smoliak et al., 2010). PLS regression was first introduced by Wold (1966) who demonstrated its application in the field of econometrics.

Like PCA, PLS is a latent variable method meaning that the original variables are summarised on a set of latent variables, which are weighted linear combinations of the original variables. However, unlike PCA, PLS is a supervised learning technique meaning that it models the relationship between a set of independent variables, ***X***, and a set of response variables, ***Y***. When the number of predictor variables is large compared to the number of observations, ***X*** is likely to be singular and the multivariate linear regression approach is no longer feasible (Abdi, 2010), due to the multicollinearity. PLS regression solves this problem by decomposing ***X*** to latent variables, which overcomes problems with multicollinearity. This property makes PLS very attractive for modelling linear relationships between ***X*** and ***Y*** when the data is high dimensional and/or when J is large and the data contains a high degree of multicollinearity.

3.2.2.1 Theory

In PLS regression both **X** and **Y** are decomposed to a set of latent vectors. The decomposition of ***X*** and ***Y*** is carried out simultaneously with a constraint that the resulting latent vectors must maximise the covariance between ***X*** and ***Y*** (Abdi, 2010)**.** The decomposition of ***X*** is given by (3.10):

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{\boldsymbol{T}} + \boldsymbol{E} \tag{3.10}$$

Here, ***T*** is the score matrix, ***P*** is the loading matrix and ***E*** is an error matrix, corresponding to ***X***. In some algorithms ***T*** is normalised using (3.11):

$$\boldsymbol{T}^{T}\boldsymbol{T} = \boldsymbol{I} \tag{3.11}$$

where ***I*** is the identity matrix. ***Y*** is estimated as described by (3.12):

$$\widehat{\boldsymbol{Y}} = \boldsymbol{TBQ}^{T} \tag{3.12}$$

where ***B*** is a diagonal matrix with the regression weights in the diagonal elements and **Q** (D x R) is the loading matrix corresponding to ***Y***. The columns of ***T*** are referred to as latent vectors. In order to determine ***T,*** such that covariance between ***X*** and ***Y*** is maximised, additional conditions are required (Abdi, 2010). Specifically, two sets of weights ***w*** and ***q*** are created.

$$\boldsymbol{t} = \boldsymbol{Xw} \;\; and \;\; \boldsymbol{u} = \boldsymbol{Yq} \tag{3.13}$$

with the constraints that $\boldsymbol{w}^T\boldsymbol{w} = 1$, $\boldsymbol{t}^T\boldsymbol{t} = 1$ and $\boldsymbol{t}^T\boldsymbol{u}$ is maximal. After the first latent vector is obtained, it is subtracted from both ***X*** and ***Y*** and the procedure is reiterated until ***X*** becomes a null matrix (Abdi, 2010).

The determination of the regression weights and the loadings is an iterative process. An explanation of the NIPALS algorithm follows. The first step is to create two matrices: ***E*** = ***X*** and ***F*** = ***Y***. These matrices are then mean centred and scaled to a standard deviation of one. A vector ***u*** is also initialized with random values. The following steps are then carried out (the symbol ∝ is used because the result of the operation is normalized):

Step 1 – estimate weights

$$\boldsymbol{w} \propto \boldsymbol{E}^{T}\boldsymbol{u}$$

Step 2 – estimate ***X*** factor scores

$$\boldsymbol{t} \propto \boldsymbol{E}\boldsymbol{w}$$

Step 3 – estimate ***Y*** weights

$$\boldsymbol{q} \propto \boldsymbol{F}^{T}\boldsymbol{t}$$

Step 4 – estimate ***Y*** scores

$$\boldsymbol{u} = \boldsymbol{F}\boldsymbol{q}$$

Step 5 – if ***t*** has not converged, then go back to step 1, if ***t*** has converged

Step 6 – compute the value of ***b*** which is used to predict ***Y*** from ***t***

$$\boldsymbol{b} = \boldsymbol{t}^{T}\boldsymbol{u}$$

Step 7 – compute the factor loadings for **X**

$$\boldsymbol{p} = \boldsymbol{E}^{T}\boldsymbol{t}$$

Step 8 – deflate the ***E*** and ***F*** matrices

$$\boldsymbol{E} = \boldsymbol{E} - \boldsymbol{t}\boldsymbol{p}^{T}$$

$$\boldsymbol{F} = \boldsymbol{F} - b\boldsymbol{t}\boldsymbol{q}^{T}$$

Step 9 – store the vectors ***t, u, w*** and ***q*** in their corresponding matrices, and store the scalar *b* as a diagonal element of ***B***.

Step 10 – if ***E*** is a null matrix, then the whole set of latent vectors has been obtained, otherwise the procedure can be re-iterated from *Step 1* to determine additional latent vectors.

3.2.3 Variable selection methods

Variable selection is a key part of predictive modelling, as models containing redundant variables are unnecessarily complex and often have reduced predictive capability compared to models with redundant variables removed (Arlot and Celisse, 2010; Hastie et al., 2001). Furthermore, it is often of interest to interpret the model coefficients to gain understanding of the relationships between predictor variables and the response variables. In this case, redundant variables increase the risk of misinterpretations (Mehmood et al., 2012).

In this thesis, some pre-selection of variables was carried out to remove variables that could be confidently ruled out with low/no relevance to the analysis. Variables that were not thought to be relevant but could not confidently be ruled out were kept in the analysis. This is because one of the objectives of the analysis was to increase understanding of variable inter-relationships, where there was a recognised lack of existing knowledge. Henceforth, it was important not to bias the models and limit the learning opportunities by discarding variables. After pre-selection, two rule-based approaches to variable selection were utilised in order to select the best predictor variables for the models based on trends captured by the data. The two variable selection methods employed were the variable importance for PLS projection (VIP) method and a forward variable selection method.

3.2.3.1 VIP selection method

The VIP method is a variable selection procedure for PLS models (Andersen and Bro, 2010; Chong and Jun, 2005; Mehmood et al., 2012). The VIP score is a measure of the importance of each variable in the model, based on its PLS weight for each latent variable and the percentage of variance in ***y*** that each latent variable explains. In the VIP selection method, a PLS model is trained using all of the variables available and the VIP score for each variable is calculated. Subsequently, variables with a VIP score lower than a predetermined threshold are removed from the model. The VIP score for a variable *j* is given by equation (3.14).

$$VIP_j = \sqrt{\frac{p}{\sum_{m=1}^{M} SS(b_m.t_m)} \sum_{m=1}^{M} \boldsymbol{w}_{mj}^2 \, SS(b_m.t_m)} \tag{3.14}$$

where *p* is the number of variables, $SS(b_m.t_m)$ is the percentage of variance in ***y*** that is explained by latent variable *m* and $\boldsymbol{w}_{mj}$ is the PLS weight for the *j*th variable on latent variable *m.* A threshold of one has commonly been used as a variable selection criterion because the average VIP score is equal to one (Mehmood et al., 2012). However, 0.83 and 1.21 have been reported to yield more relevant variables depending on the features of the dataset. Chong and Jun (2005) found that for datasets with low proportion, high multicollinearity, or an equal coefficients structure, a threshold of greater than one is more appropriate.

One drawback to the VIP selection method is that it is dependent upon the initial model capturing the important relationships between ***X*** and ***Y*** (Andersen and Bro, 2010). In cases where there are too many redundant variables, the noise in the model is can cause the VIP selection method to fail. An alternative method is forward variable selection.

3.2.3.2 Forward variable selection method

Forward variable selection is a classic statistical approach to variable selection, based on selecting the variables that offer the best performance one-by-one (Andersen and Bro, 2010). A procedure for forward variable selection is described by the following steps:

1. Select the first variable based on the variable with the highest Pearson's correlation coefficient with the response.
2. Add and then remove each of the remaining variables to the model to calculate model performance metrics. Select the variable that explains the most variation in ***y***, when added to the model*.*
3. Keep adding variables to the model according to step 2 until the adjusted R^2 increases by less than 0.01.

In chapter 6, forward variable selection was carried out with 1000 randomly selected subsets to derive a range of alternative models. Each subset included 2/3rds of the observations randomly selected from the training data. The decision was made to use two thirds of the data in each subset because this was found to work well in practice, i.e. it provided diverse models for evaluation and it used enough of the data to ensure that relevant variables were selected. The advantage of forward variable selection compared to VIP selection method is that it does not require a good PLS model to be obtained prior to variable selection. Forward variable selection was used in chapter 6 because it was difficult to identify a set of predictor variables that offered good PLS model performance prior to variable selection; henceforth, the VIP was not suitable.

3.2.3.3 Stepwise variable selection

Stepwise variable selection methods utilise a set of rules to select variables for a regression model and they allow variables to be both added and removed as the steps are executed. These methods offer increased flexibility by adopting rules from both forward and backward selection methods, which often arrive at different results. In the work conducted in Chapters 5 and 6, a stepwise selection method was explored and was found to identify good predictive models, however, these models did not offer improved performance over the alternative methods that were explored, and the variables selected were in agreement with the alternative methods. Hence, the stepwise selected models are not presented. The stepwise selection method that was employed is outlined below:

1. Starting with no variables in the model. The first variable is selected based on the highest correlation coefficient with the output variable
2. Add the next variable with the highest t-value and acceptable multicollinearity (tolerance value less than 0.6)
3. Check the t-value of variables in the model, remove any that are insignificant ($t< 1.9$, $\alpha=0.05$, x DoF) and remove any variables with tolerance less than 0.6
4. Repeat steps 2 and 3 until there are no more variables with a significant t-value

3.2.4 Representing nonlinear effects

PCA and PLS are multivariate linear models because the model coefficients are linear. Nonlinear effects can be represented by these models by applying nonlinear transformations to the variables prior to passing them into the PLS or PCA model. The practitioner may want to apply data transformations or add nonlinear effects based on theoretical reasons related to the nature of the variables, or for data derived reasons, where data observations indicate that a transformation may be necessary or beneficial (Hair, 2014). Two common nonlinear transformations were utilised in this thesis: curvilinear effects and moderator effects.

3.2.4.1 Curvilinear effects

In cases where the relationship between a predictor variable and the response is known to be nonlinear, this can be taken into account by applying power transformations to the independent variable and adding the power terms to the model. This can be used to implement polynomial relationships with any degree, although its best to use the simplest model that offers good performance, due to the potential for overfitting the data (Hair, 2014). In this thesis, quadratic terms were tested and used to improve model fit when nonlinearities were present in the regression errors.

3.2.4.2 Moderator effects

Sometimes in regression, there are cases where the relationship between an independent variable and the response is affected by another independent variable. This type of effect is referred to as a ‘moderator’ or ‘interaction’ effect. One common moderator effect is the bilinear moderator, which can be represented by creating a new variable that is the cross-product of the two independent variables. A bilinear moderator effect for variables *i* and *j* can be formed by adding the term $(\boldsymbol{X}_i\boldsymbol{X}_j)$ to the model. The interpretation of a regression coefficient changes slightly when moderator effects are used. Consider, the following regression equation with two variables and one moderator effect:

$$\boldsymbol{Y} = b_0 + b_1\boldsymbol{X}_1 + b_2\boldsymbol{X}_2 + +b_3\boldsymbol{X}_1\boldsymbol{X}_2 \tag{3.15}$$

Here, the regression coefficient b_3 indicates the unit change in the effect of $\boldsymbol{X}_1$ as $\boldsymbol{X}_2$ changes. In the unmoderated relationship, the coefficient b_1 represents the effect of $\boldsymbol{X}_1$ across all levels of $\boldsymbol{X}_2$. The same is true for b_2. Henceforth, in unmoderated regression coefficients b_1 and b_2 are averaged across levels of the other independent variables, whereas in a moderated relationship they are separate from the other independent variables (Hair, 2014). The bilinear moderator effect was found to have a statistically significant effect on models used in this work. Other transformations that were explored but were not found to be significant included log transforms and inverse transforms.

3.3 Model validation

Model validation techniques play a critical role in the development of predictive models, since they test a model's significance, robustness and predictive ability (Rücker et al., 2007). Model validation methods allow practitioners to compare the quality of different models to select the best model and to evaluate its performance.

3.3.1 Regression metrics

The coefficient of determination (R^2) provides a measure of the amount of variance in the response, ***y***, that is explained by the regression model. R^2 values close to 1 indicate that the regression model performs well and explains a large percentage of the variance in ***y***, while a measure close to zero indicates poor performance. An R^2 value of 0.6 can be said to explain approximately 60% of the variance in the response (Hair, 2014). Negative R^2 values indicate that the model provides predictions of ***y*** that are worse than if all response had been predicted as the mean of ***y***.

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}} = \frac{\sum_{i=1}^{N}(\boldsymbol{y}_i - \widehat{\boldsymbol{y}}_i)^2}{\sum_{i=1}^{N}(\boldsymbol{y}_i - \bar{\bar{\boldsymbol{y}}}_i)^2} \tag{3.16}$$

The adjusted coefficient of determination (R^2_{adj}) is adjusted for the number of variables in the regression model. There is a slight penalty applied to the R^2_{adj} score, as variables are added to the model, which is a basic method to stop redundant variables being added to a regression model.

$$R_{adj}^2 = \frac{(1-R^2)(N-1)}{N-p-1} \tag{3.17}$$

The mean absolute error and mean square error provide estimates of the generalisation error of a predictive model. The generalisation error is discussed in the next section.

$$MAE(x) = \frac{1}{N-1}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{3.18}$$

$$MSE(x) = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{3.19}$$

In chapters 5, 6 and 7, primarily the MAE and R^2 metrics were used to evaluate regression models, with a focus on the out-of-sample testing through either cross-validation or hold-out testing. The MAE was preferred over the MSE because it is less sensitive to outliers and the R^2 metric provides an intuitive estimation of the variance explained by the model. These metrics were evaluated together with a visual assessment of the model fit, which is an important sensibility check.

3.3.2 Bias, variance and model complexity

The generalisation performance of a statistical or machine learning model describes its prediction capability on independent test data that has not been used to train the model (Hastie et al., 2001). Evaluation of the generalisation performance is of critical importance in practice because it guides model selection and provides a measure of quality of the chosen model. The concepts of bias and variance, which are influenced by model complexity, play an important role in determining the generalisation ability of a data-driven model. Consider data that is partitioned into two parts: a training set and a test set. A model given by (3.20) is fitted to the training data and used to predict the response for the test data.

$$\boldsymbol{y} = f(x) + e \tag{3.20}$$

The model error in both training and testing can be described by (3.21).

$$Error = E\left[\left(\boldsymbol{y} - \hat{f}(x)\right)^2\right] \tag{3.21}$$

The error can be further decomposed as:

$$Error(x) = \left(E\left[\hat{f}(x) - f(x)\right]\right) + E\left[(\hat{f}(x) - E\left[\hat{f}(x)\right])^2\right] + \sigma_e^2 \tag{3.22}$$

$$Error(x) = Bias^2 + Variance + Irreducible\ Error \tag{3.23}$$

The irreducible error is the error that cannot be reduced by developing a good model, for example, noise in data contributes to error that cannot be reduced no matter how good the model (Hastie et al., 2001; Rodríguez et al., 2010). The variance term represents the magnitude of variation in the predicted responses, which can be attributed to models that capture noise in the training data. Such models are described as 'over-fitted' to the training sample. The bias term describes the accuracy of the model, where a high bias model fails to capture the relationship between ***X*** and **y**. Models with high bias are described as 'under-fitted' to the training data (Hastie et al., 2001).

Both bias and variance depend upon the complexity of the model. A model that is too simple, e.g. containing too few predictors, will be under-fitted to the training data with high bias. The model will explain insufficient variation in the response because there are relationships that the model fails to capture. Conversely, a complex model, e.g. with too many predictors, will often be over-fitted to the data with high variance. This is because the complex model has many degrees of freedom to capture the variance in the training data, including the noise component, and it therefore becomes highly specified to the training data. Figure 3-6 illustrates the concepts of bias, variance and model complexity.

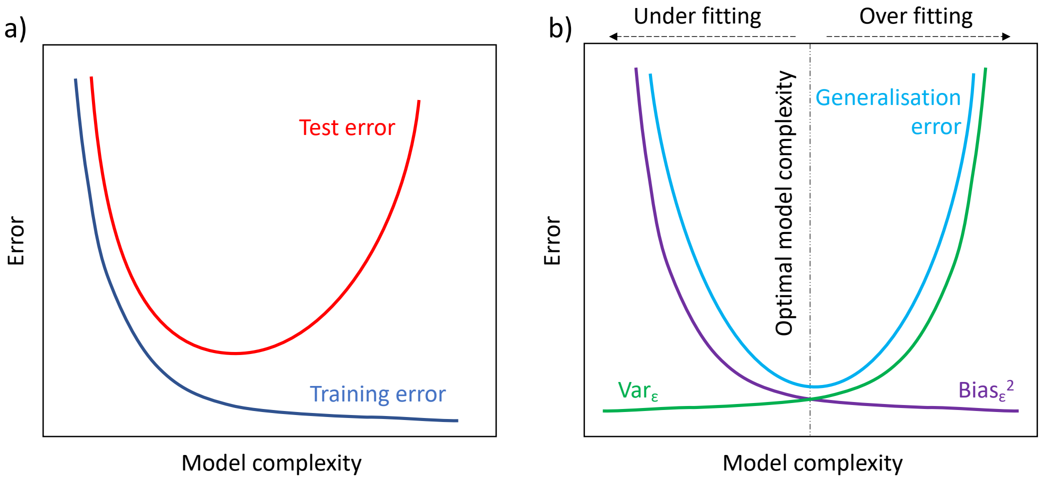


Figure 3-6: a) Training error and test error versus model complexity, b) bias-variance trade-off versus model complexity. The optimal model with minimal generalisation error is the model with minimal test error. Figure adapted from (Hastie et al., 2001).

A model that is under-fitted fails to explain variation in the response variable for the training data, because it doesn't capture the relationships between the predictors and the response. It will also fail to explain sufficient variation in the response variables for independent test samples. The underfitted model therefore has poor goodness of fit and performance metrics for both the training data and the test data. A model that is over-fitted will explain a significant portion of the variation in the response variable for the training data. However, due to the noise captured by the model, it will not predict the response accurately for independent test samples. The training metrics and the training goodness of fit can be misleading due to the fact that increasing model complexity will always lead to improved training performance metrics and goodness of fit. In other words, training metrics provide very little insight into model over-fitting and maximising training performance alone will result in models with high variance. Testing models on unseen data is therefore very important for controlling model complexity and avoiding overfitting the model.

Models that are under-fitted and over-fitted both have poor generalisation ability to unseen data. The ideal model is therefore the model which minimises both variance and bias, although to some extent there is a trade-off between the two competing objectives (Hastie et al., 2001). Selection of predictor variables and determination of the appropriate degree of model complexity therefore have a direct impact on model predictive performance.

3.3.3 Cross validation

The concept of generalisation ability and data partitioning can be traced back to Larson (1931). Larson showed that when a model is trained through re-substitution, where all samples in a dataset are used for both training and validation, the resulting model was heavily biased due to the model fitting the noise component in the data. Such models gave poor predictions on unseen data that was not used in model training. Consequently, Larson proposed that data could be split randomly into a training set and a validation set, which could be used to evaluate - and avoid - model over-fitting. The idea of splitting the data in two parts: a training set and test set is now commonly referred to as the hold-out method, which is the simplest form of cross-validation (CV) (Arlot and Celisse, 2010).

In addition to fitting a model, there are two main objectives of CV (Arlot and Celisse, 2010; Hastie et al., 2001). These are:

1. Model selection – estimating the performance of a number of alternative models, often with varied complexity. For example, with PLS regression the practitioner must determine which variables to use and how many latent variables to use. CV may be used to evaluate the predictive performance of models with different predictors and different numbers of latent variables.
2. Model assessment – having selected the best performing model, estimating its generalisation error on unseen data.

Since Larson's early work on data partitioning, CV techniques have been expanded upon to provide more robust estimates of model predictive performance (generalisation ability). A simple extension, when a sufficiently large volume of data is available, is the partitioning of the data into three parts: a training set, a validation set and a test set. The training set is used to train the models, the validation set is used to estimate the prediction error to evaluate and compare models, and the test set is used to assess the prediction error of the selected model (Hastie et al., 2001). The number of observations to assign to each part is a decision that depends upon the signal-to-noise ratio and the number of observations available. Hastie *et al.* (2001) suggested that a typical split might be 50% for training, and 25% each for validation and testing, as shown in Figure 3-7.

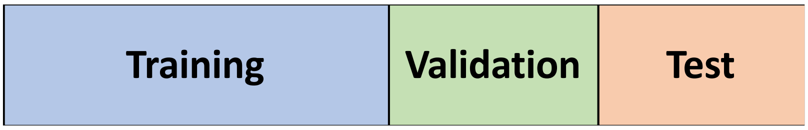


Figure 3-7 Partitioning of data into three parts: training (50%), validation(25%) and testing (25%), adapted from Hastie *et al.* (2001).

While the three-way data partitioning can be effective, the results can vary greatly depending upon on the partitioning, i.e. which observations are included in the training, validation and test sets (Arlot and Celisse, 2010). Ideally, all three partitions should contain samples that are representative of the population, without bias towards a specific set of conditions (Efron and Gong, 1983). Indeed, there are algorithms for assigning observations to the training, validation and testing sets to partition the data in the least biased way. One such algorithm is the Kennard-Stone algorithm, which partitions the data into a training and test set based the Euclidean distance between observations over the predictor variable space (Kennard and Stone, 1969). The objective of this algorithm is to select observations for the test set with uniform distribution over the predictor variable space.

Other forms of cross validation are also available, which reduce the dependence of the prediction error on how the data is divided by splitting the data into more subsets (Stone, 1974). These include:

- Leave-one-out (LOO) CV
- K-fold CV
- Repeated K-fold CV

3.3.3.1 K-fold cross-validation

5-fold CV is an example of K-fold CV where the data is split into five subsets, as illustrated in Figure 3-8. In 5-fold cross validation, the data is split into five folds (equally sized portions with the exception of remainders) and the model is trained (Tr) five times on 4/5ths of the data, each time leaving out a different portion (1/5th) of the data for validation (Va). The validation error (MSE or MAE) is then averaged across the five folds to estimate the model’s prediction error. In K-fold CV, each data point is included in the training set exactly K-1 times. As K is increased, the variance of the prediction error estimate is reduced (Rodríguez et al., 2010). However, the computational demand is increased as the model training algorithm must be executed K times. 5-fold and 10-fold CV are popular due to their lower computational and demand, low bias and sufficiently stable error estimates (Rodríguez et al., 2010).

3.3.3.2 Leave-one-out cross-validation

LOO CV takes this logic to its extreme by leaving out exactly one data point for validation each time, which can be extremely computationally expensive for large datasets (Rodríguez et al., 2010; Stone, 1974).

3.3.3.3 Repeated K-fold cross-validation

Repeated K-fold CV involves repeating K-fold CV *x* times, each time permuting the order of the dataset so that different observations are grouped together in the K folds. This is another way to reduce the variance of the prediction error estimate. The larger the number *x,* the lower the variance. In the literature, anywhere from 5 to 10,000 repeats have been used (Fushiki, 2011). The main idea is to use a sufficient number of repeats for the error estimate to become stable, such that it varies little each time repeated K-fold CV error is calculated.

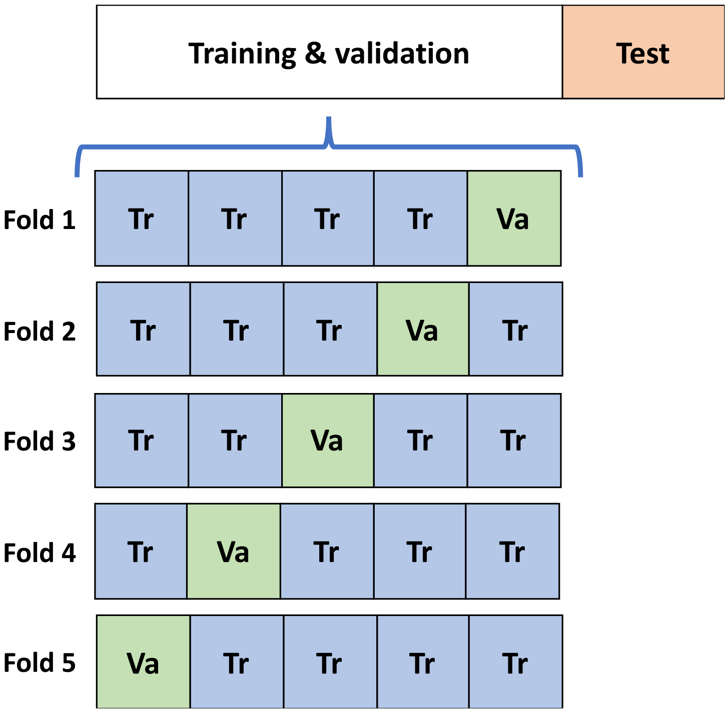


Figure 3-8 Illustration of data partitioning with a hold-out test set and a training/validation set used for model training and 5-fold cross validation. In 5-fold cross validation, the data is split into five folds (equally sized portions) and the model is trained (Tr) five times, each time leaving out a different portion of the data for validation (Va). The validation error (MSE or MAE) is averaged across the five folds to estimate the model's prediction error.

In Figure 3-8, there are initially two sets of data: a training/validation set and a hold-out test set. The training/validation set may be used to evaluate different models using repeated K-fold CV. For a PLS model, the number of latent variables can be varied, and the optimal model can be selected as the model with the minimum prediction error estimate obtained from K-fold CV. The model can then be trained using all of the observations in the training/validation set and the prediction error can then be estimated one more time, using the test set that was held out of model training.

3.3.4 Bootstrap sampling

Bootstrapping is a method of testing that uses random sampling with replacement. By repeatedly fitting a model of interest to bootstrap samples, it is possible to evaluate variation in key model performance statistics and model coefficients. The bootstrap is based on the law of large numbers, which indicates that, with sufficient data the empirical distribution will be a good estimate of the true distribution for the population (Efron and Gong, 1983). In this work, the parametric bootstrap was used to determine confidence intervals for the regression coefficients of PLS models, as the parametric bootstrap makes no assumptions about the underlying distribution (Efron and Tibshirani, 1986). The confidence intervals were determined by recording the model coefficients obtained from fitting the model to 2000 bootstrap samples. The 5% and

95% confidence intervals were then determined by obtaining the 5th and 95th quantiles from the vector of estimates for each model coefficient.

The confidence intervals for the regression coefficients provide insights into the reliability of the regression model and the significance of each predictor variable. For example, regression coefficients with confidence intervals that cross zero are considered insignificant and can be removed from the model. Wide confidence intervals are a sign of instability in the model, indicating that the model structure and variable selection may be wrong or that the data are of poor quality.

3.3.5 Testing chance correlations with Y-randomization

Y-randomization tests are carried out to detect and quantify chance correlations between the predictor variables and the response, which may be present in a prediction model (Kiralj and Ferreira, 2009; Rücker et al., 2007). Prediction models are designed to capture a physical or cause-and-effect relationship between the predictors and the response variables that are part of a mechanism. However, it is known that adding variables to a regression model will always result in an increase in the quality of the model fit to the training data, irrespective of whether the predictors contain relevant information. This is because the model fits the noise in the data and learns any chance correlations that may be present. ***Y***-randomization involves fitting the prediction model multiple times, each time keeping the original matrix of predictors, ***X***, and randomly permuting the values in ***Y***. The performance metrics are then averaged across multiple randomizations. The resulting models should have poor performance metrics because the mechanistic connection between ***X*** and ***Y*** has been broken. The better the performance metrics when the model is fitted to ***Y***-randomized data, the greater the presence of chance correlations in the dataset. ***Y***-randomization has been widely utilised in the development of quantitative structure-activity relationship models, where large numbers of predictors are used, making the models susceptible to chance correlations (Kiralj and Ferreira, 2009). In this work, the average coefficient of determination, for the model fit to the ***Y***-randomized data, was used to evaluate the chance correlation. Kiralj and Ferreira (2009) classified the chance correlation based on the coefficient of determination, as described by (3.24) to (3.27).

$$R^2_{yrand} < 0.2 \rightarrow no\ chance\ correlation \tag{3.24}$$

$$0.2 < R^2_{yrand} < 0.3 \rightarrow negligible\ chance\ correlation \tag{3.25}$$

$$0.3 < R^2_{yrand} < 0.4 \rightarrow tolerable\ chance\ correlation \tag{3.26}$$

$$0.4 < R^2_{yrand} \rightarrow recognized\ chance\ correlation \tag{3.27}$$

3.3.6 Data partitioning

As discussed in the cross-validation section (section 3.3.3), partitioning data is an important part of model validation. K-fold CV, LOO CV and repeated K-fold CV are based on splitting the data randomly. This works well with these techniques because the data is split into numerous folds; henceforth, the resulting performance metrics are averaged across numerous folds and are not heavily dependent on one single partition. In contrast, when a hold-out set is utilised (based on one single partition), it is important to ensure that the hold-out and training sets are both representative of the full range of conditions observed in the data. In this work, the Kennard-Stone algorithm was used to identify the most suitable hold-out test set.

3.3.6.1 Kennard-Stone algorithm

The Kennard-Stone algorithm carries out the following process to assign observations to training and test sets:

1. Select a pair of observation that are furthest apart, based on Euclidean distance over the independent variable space. These observations are assigned to the training set and removed from the list of unassigned observations.
2. Compute the Euclidean distance between each of the unassigned observations and the observations in the training set. Add the observation with the largest Euclidean distance from the observations in the training set.
3. Stop when adding to the training set when the desired test-training ratio has been achieved

3.4 Data pre-processing

Data pre-processing involves essential transformations and processing of the data to get it into the correct and most useful format, prior to carrying out modelling or MVDA.

3.4.1 Dummy variable coding

Dummy variables are used to convert categorical variables into a numerical format, which can be interpreted by statistical/ML models. Models such as principal component analysis and partial least squares regression, which are used in this thesis, are designed to work with metric variables only, i.e. quantitative data where the quantity describes the degree to which the subject may be characterised by the attribute (Hair, 2014). Categoric data cannot be simply represented by different numbers because the varying quantity does not reflect a physical attribute.

3.4.1.1 Indicator coding

One way to code dummy variables is to generate *L*-1 metric variables to represent a categorical variable with *L* levels. Taking a coin toss as an example, a coin toss has two levels and one dummy variable can model these two levels by setting the variable to 1 for heads, and 0 for tails, or vice versa. The omitted category is termed the reference category. With this type of coding, the model coefficients assigned to dummy variables represent the category differences from the reference category (Hair, 2014).

3.4.1.2 Effects coding

An alternative to the indicator coding system is called effects coding, where the reference category is assigned a value of minus one across the set of dummy variables. With this type of coding, the model coefficients assigned to the dummy variables represent deviations from the mean of all categories (Hair, 2014).

3.4.2 Unfolding of 3D data

Most statistical and ML models, including PCA and PLS, deal with data in a 2D array. Traditionally the two dimensions represent variables and observations. However, there are many data sources that feature a third dimension. Temporal variation is a common example. Batch or fed-batch processes are common sources of 3D data in the chemical and biochemical industries, since variables are often recorded online and offline, where they are traced over the duration of the batch (Kourti, 2003).

Unfolding is a method by which 3D data is transformed into a 2D array, by combining two of the original dimensions along one axis of the 2D array. The interpretation of the unfolded data and the resulting models is dependent upon which two dimensions are combined and which one is preserved. For temporal data, and for batch process data, the most common unfolding method is to preserve the batch or observation dimension on the first axis (rows) and to combine the variable and time dimensions along the second axis (columns) (Ündey et al., 2003). This is illustrated in Figure 3-9.

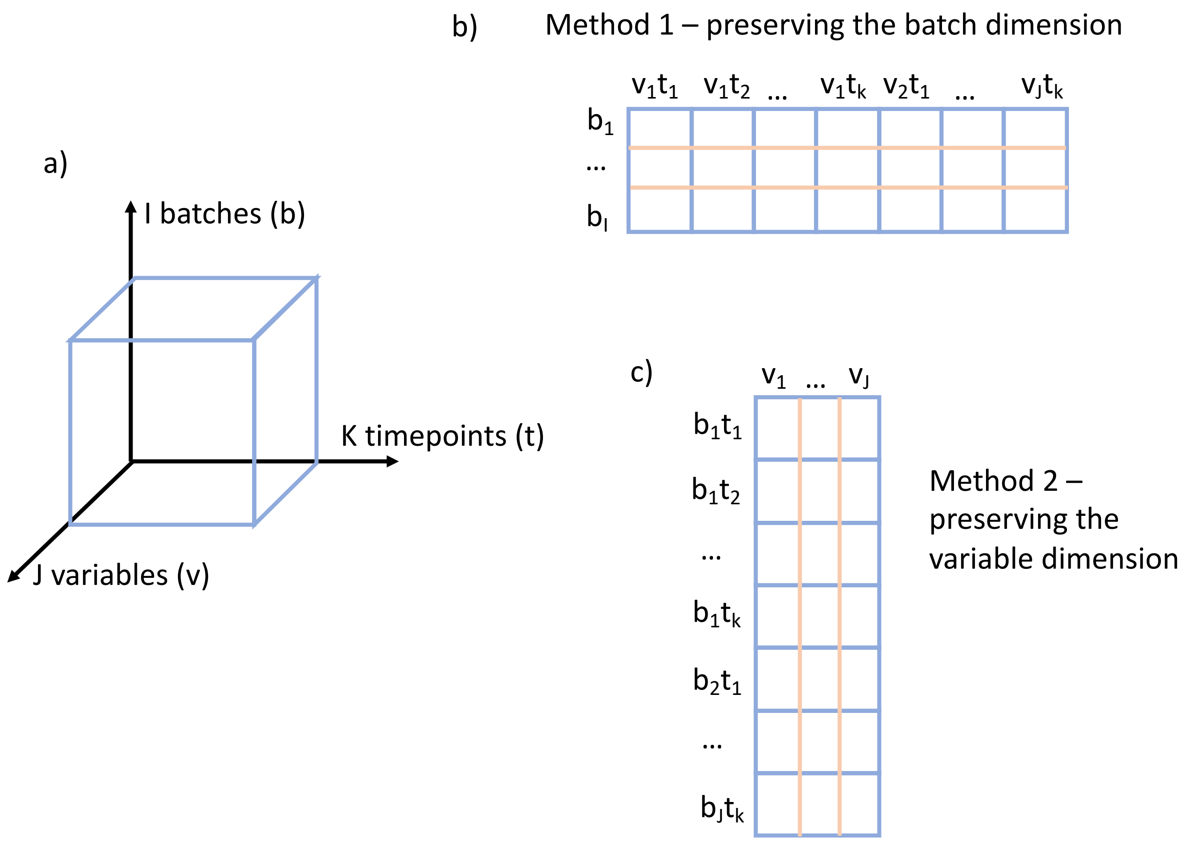

b)
Method 1 – preserving the batch dimension
v_1t_1 v_1t_2 … v_1t_k v_2t_1 … v_Jt_k
b_1
…
b_I
a)
I batches (b)
K timepoints (t)
J variables (v)
c)
v_1 … v_J
b_1t_1
b_1t_2
…
b_1t_k
b_2t_1
…
b_Jt_k
Method 2 – preserving the variable dimension

Figure 3-9 Illustration of 3D data unfolded into a 2D array. This figure shows a) 3D batch process data, b) unfolding of the data preserving the batch dimension and c) unfolding of the data preserving the variable dimension.

Unfolding the data in this way, the columns become variable-time instances, thus retaining a sensible physical meaning and capturing information on the dynamics of the process. The resulting model coefficients represent effects associated with the variable at a given point in time. Comparison of model coefficients, for a given variable at different time points, shows how the variable relationships change over time. With high sampling frequency, there will be a high degree of multicollinearity due to the autocorrelation in the profile of each variable. The use of latent variable methods is able to overcome problems associated with multicollinearity (Hair, 2014).

The second most common approach to unfolding batch process data is to preserve the variable dimension. Using this unfolding approach with PCA results in scores that show the development of a batch over time, i.e. its trajectory, and the loadings will reflect the correlation of the variables from an 'overall' perspective without taking into account time dependency and correlation (Ramos et al., 2021). In this work, method 1 (preserving batch dimension) is used in order to capture the dynamics of the variables and their correlation.

3.4.3 Scaling and centring

Scaling and centring data are standard practices in MVDA and machine learning, which serve several purposes, depending on the particular model being used. For example, PCA requires the data to be mean centred because without mean centring, the first principal component can correspond with the mean of the data instead of the direction of maximum variance. Henceforth, mean-centring data for PCA is critical to avoid misinterpretations of the PCA model (Bro and Smilde, 2014). Conversely, multivariate regression and PLS regression can be implemented directly without mean centring the data, although the interpretation of model coefficients changes when the data is centred versus uncentred (Seasholtz and Kowalski, 1992).

Centring removes the offset from the data and allows the practitioner to focus on differences between observations rather than similarities. In regression, the model coefficients indicate the effect of unit increases in the independent variable on the response variable. When the data is centred, the model coefficients refer to deviations from the mean henceforth negative coefficients. When the data is centred, the model coefficients reflect the impact of unit changes in that variable (van den Berg et al., 2006). Henceforth, centring can be used to improve the interpretability of regression models.

Scaling is carried out in order to remove the impact of different units and scales, so that variables of large magnitude do not dominate the latent projections. Additionally, the standardised regression coefficients obtained from models using scaled data, allow direct comparison of the importance of each variable (Hair, 2014).

Chapter 4

Sparse principal component analysis

In this chapter, sparse PCA is introduced as an alternative method to standard PCA that derives simplified principal components, which are easier to evaluate and interpret. Henceforth, sparse PCA is beneficial in applications where interpretation of the PCA model is important, including applications explored in this thesis. In the introduction, works focusing on the interpretability of PCA are reviewed, including sparse PCA methods, which have been widely researched in recent years. Four different optimisation programs for sparse PCA are then developed as an alternative to other sparse PCA methods in the literature. In the results section, the performance of the four optimisation programs is evaluated and compared to one another, as well as to other popular sparse PCA approaches in the literature. The comparison is carried out using standard datasets from the literature and synthetically constructed datasets. The best performing technique, out of the four optimisation programs developed in this chapter, was then utilised in Chapters 5 and 6 to perform sparse PCA on the data from adherent viral vector production and cell drug product manufacture.

4.1 Introduction

Sparse PCA is a variation of standard PCA, which seeks to obtain sparse loadings; consisting of a combination of zero and nonzero elements. Loadings of the standard PCA model which are close to zero typically don’t contribute significantly to the variation in scores; however, they must still be taken into account when analysing the contribution of the original variables. In sparse PCA, the idea is to force those small magnitude loadings to zero, to end up with sparse principal components where the variables with nonzero values contribute significantly to the scores and the variables with zero loadings may be ignored entirely. This significantly simplifies the interpretation of the PCA model and provides insights into the system under study with greater clarity.

The difficulty of interpreting PCA models is an issue that has long been recognised. Rotation techniques were initially developed for the closely related factor analysis in 1931 (Thurstone, 1931) and later for PCA (Jolliffe, 1989), these involve rotation of the principal component axes in order to obtain a ‘simple structure’. There are however several drawbacks to rotation techniques, including the fact that they involve complicated post-processing of the PCA solution with many options for the rotation method. These can lead to different rotated solutions, which impact model interpretation (Jolliffe, 1995). Informal thresholding techniques have frequently been used in practice, whereby the PCA solution is modified and loadings smaller than a given magnitude are artificially set to zero. This simple *ad hoc* approach may be effective in some cases, however, it relies on an arbitrary choice for the threshold and can be misleading in several ways (Jolliffe, 1995).

More recently, other informal thresholding techniques, heuristics and algorithms have been developed. Farcomeni (2009) proposed a branch and bound algorithm to find the best sparse dimension reduction of a matrix and suggested methods to choose the number of nonzero loadings for each principal component. This algorithm allowed the user to specify the number of nonzero loadings directly, or to specify the degree of sparsity based on an objective function, which maximises variance and penalises nonzero loadings. Ma (2013) developed an iterative thresholding technique to obtain sparse eigenvectors, based on the orthogonal iteration method in matrix computation with an additional thresholding step to determine sparse basis vectors for the subspace. Gajjar *et al.* (2017) utilised a genetic algorithm with exhaustive and non-exhaustive search approaches to evaluate different possible sparse solutions. Several other "greedy" methods have been proposed for the solution of all possible sparse combinations; however, these are computationally expensive (Jolliffe and Cadima, 2016).

The same interpretation problem arises in multiple regression, where variable selection techniques are used to reduce the number of variables in the model and ease interpretation. The Least Absolute Selection and Shrinkage Operator (LASSO) is a variable selection technique introduced by Tibshirani (1996), which produces accurate and sparse models through the use of a cost function, which simultaneously minimizes model error and penalises model complexity. The first computational approach to sparse PCA was developed by Jolliffe *et al.* (2003), where they applied the LASSO constraint to the formulations of PCA, in a technique they named SCoTLASS. Consider the first principal component and the variance maximization definition of PCA, outlined in chapter 3, section 3.1.1. The SCoTLASS constrained sparse principal component analysis problem is defined by (4.1) and (4.2):

$$\max_{W^TW=1} Tr(\boldsymbol{W}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{W}) \tag{4.1}$$

$$subject\ to,\ \|\boldsymbol{W}\|_1 \leq t \tag{4.2}$$

Here, *t* is a tuning parameter to be selected to induce sparsity in the loadings, ***W***. When $t \geq \sqrt{\boldsymbol{W}}$, the standard PCA solution is obtained. As *t* is reduced from $\sqrt{\boldsymbol{W}}$, the loadings progressively shrink with some eventually shrinking to zero. For $t = 1$, there must be exactly one nonzero loading for each principal component. Given an appropriate value for *t*, SCoTLASS succeeds in producing sparse principal components with zero and nonzero coefficients; however, a drawback to SCoTLASS is that it is computationally expensive to solve (Zou and Hastie, 2005).

Zou and Hastie (2005) developed the elastic net, a generalised version of the LASSO, which is a convex combination of the ridge and LASSO penalties, based on the L2-norm and L1-norm, respectively. The elastic net overcame a limitation of the LASSO whereby in cases where the number of variables, p, exceeds the number of samples, n, at most n variables can be selected. Zou *et al.* (2006) later used the elastic net to produce sparse principal components, in a technique which they called sparse PCA. Their sparse PCA approach was based on formulating PCA as a regression type optimisation problem with the elastic net penalty integrated into the regression criterion, leading to principal components with sparse loadings. They also suggested an algorithm with improved computational efficiency compared to SCoTLASS, making their sparse PCA approach viable for larger problems.

Since the introduction of sparse PCA by Zou *et al.* (2006)*,* numerous other approaches have been developed and ‘sparse PCA’ has become the general term used to describe techniques which produce sparse principal components. Several works have utilised L1-regularisation to produce solutions for sparse PCA. d’Aspremont *et al.* (2007) developed a semi-definite programming approach, called direct sparse PCA (DSPCA), which utilised a convex relaxation to L1-constrained sparse PCA. Convex relaxation is a technique in operational research to ease the computational burden of difficult nonconvex problems. For large problems, d’Aspremont *et al.* suggested a Nesterov’s smooth minimization technique to solve DSPCA efficiently. d’Aspremont *et al.* (2008) further developed this work with a greedy algorithm to speed up computation. Journee *et al.* (2010) developed a generalized power method to solve the Lagrangian form of the SCoTLASS problem (4.1) and (4.2), as given by (4.3).

$$\max_{\boldsymbol{W}^T\boldsymbol{W}=1} Tr(\boldsymbol{W}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{W}) - \lambda\|\boldsymbol{W}\|_1 \tag{4.3}$$

The generalized power method made use of a further reformulation of (4.3) that provides great computational savings when $p \gg n$, as well as an objective function which is differentiable and convex (Journée et al., 2010). Other works utilised a convex combination of L1 and L2 norm penalties, with approaches similar to that of Zou *et al.* (2006). Qi *et al.* (2013) developed a new penalty function, which uses a tuning parameter to determine how much the L1-norm and L2-norm penalties contribute to the overall penalty. They called this approach sparse PCA by choice of norm. They also proposed an efficient iterative algorithm to solve the optimization problem. Xiaoshuang *et al.* (2013) developed a sparse PCA approach using a joint L2,1-norm penalty term, which was based on a modification of SPCA algorithm developed by Zou *et al.* (2006), where the elastic net penalty was replaced with the joint L2,1 norm. They also relaxed the orthogonality constraint to provide more flexibility in the features that are selected, allowing variables with less important features to be ignored, with zero loadings across all components.

A key drawback to the L1 and L2-norm regularisation approaches is that non-sparse model parameters are known to shrink as the regularisation weight is increased, which can lead to models with incorrect structure and biased model coefficients (Willis and von Stosch, 2017; Zou et al., 2006). A more natural choice for the regularisation penalty function is the L0-norm (4.4), which directly penalises the total number of nonzero model parameters.

$$P = \sum_{j=1}^{m} \|\boldsymbol{W}\|_0 \tag{4.4}$$

L0-regularisation does not induce shrinkage of nonzero model parameters, allowing the correct structure to be identified as well as unbiased values for the nonzero model parameters (Ulfarsson and Solo, 2011; Willis and von Stosch, 2017). This provides clear benefits over L1 and L2-regularisation; however, the solution is more difficult to obtain since the L0-norm is non-convex and discontinuous. Regularisation problems utilising L0-norm penalties are known to be NP-hard. Consequently, L0-regularisation does not scale well to problems with larger numbers of variables, although various transformations and L0-norm approximations have been used to circumvent this issue.

A few works in the literature have developed approaches to sparse PCA based on L0-norm penalties. Ulfarson and Solo (2011) developed a technique called sparse variable noisy PCA (svnPCA), based on L0-regularisation of a noisy PCA model, which is an alternative representation of PCA. svnPCA completely removes some variables from the PCA model by simultaneously zeroing the loadings of some variables across all of the components, acting as a variable selection method for PCA. In addition to their L1-regularised sparse PCA method, Journee *et al*. (2010) implemented L0-regularisation with the generalised power method that was described previously. Their reformulation of the objective function and efficient algorithm allows L0-regularisation to be implemented with large datasets, including cases where the number of variables exceeds the number of observations in the data.

In this work, linear and nonlinear programs are developed for sparse PCA by L0-regularisation and approximate L0-regularisation. L0-regularisation was explored for the benefits that were outlined previously and L0-approximations were used to ease the computational demand to solve sparse PCA with larger datasets. Linear and nonlinear programming solvers are readily available in software, such as Excel, Matlab and Python; henceforth, these approaches were explored as accessible alternatives to the L0-regularised sparse PCA solutions in the literature.

4.2 Method

In this work, four different approaches to sparse PCA were developed and tested utilising different optimisation programs:

1. Mixed integer nonlinear programming (MINLP)
2. Nonlinear programming (NLP)
3. Mixed integer linear programming (MILP)
4. Linear programming (LP)

All four of the approaches were based on configuring sparse PCA as an optimisation problem, where the sparse principal components were determined sequentially using an appropriate cost function. The sparse PCA optimisation problem is inherently nonlinear as the objective function and constraints feature nonlinear functions of the decision variables. However, it is possible to solve nonlinear problems using linear solvers through sequential linear programming (SLP). SLP obtains the solution to a nonlinear problem by finding the solutions to a sequence of linear subproblems approximating it. In this work, linear programming approaches utilising SLP were explored as well as nonlinear programming approaches to evaluate and compare their performance. Additionally, both the L0-norm penalty and relaxed approximations to the L0-norm penalty were explored. The integer programming approaches, MINLP and MILP, were utilised to implement the full L0-norm penalty. The non-integer variants, NLP and LP, were used to apply approximate L0-norm penalties, which ease the computational demand.

There are three main features required to define the optimisation problem, which are implemented in all four approaches. These are:

1. The objective function
2. Normalisation of the loadings
3. Orthogonality constraints

The objective function is used to maximise the variance of the principal components and at the same time induce sparsity. The basic form of the cost function is given by equation (4.5):

$$J = -\boldsymbol{p}^{\boldsymbol{T}}\boldsymbol{Q}\boldsymbol{p} + \lambda Pen \tag{4.5}$$

Here, $\boldsymbol{p}$ is the loadings vector, $\boldsymbol{Q}$ is the covariance matrix, λ is the regularisation/ penalty parameter and Pen is the penalty function. This equation represents a trade-off between the variance captured by the principal component and the sparsity of the loadings, which can be controlled by adjusting the penalty parameter λ. The variance is given by the function $\boldsymbol{p^T Q p}$, since this term is nonlinear with respect to the loadings, it can only be implemented directly with nonlinear programming solvers. The MINLP approach allows direct implementation of the nonlinear functions which are required to set up the sparse PCA optimisation problem, as well as implementation of the full L0-norm penalty. For this reason, the MINLP model will be explained first in full. The other three approaches will be explained by detailing the aspects which deviate from the MINLP model.

4.2.1 Mixed integer nonlinear programming (MINLP)

The objective function for the NLP approach is given by equation 4.6.

$$J(\lambda) = -\boldsymbol{p}^T \boldsymbol{Q} \boldsymbol{p} + \lambda \|\boldsymbol{p}\|_0 \tag{4.6}$$

The implementation of the L0-norm penalty requires the use of a vector of binary variables, $\boldsymbol{\delta}$, which is the same length as $\boldsymbol{p}$, so that each element of the loadings vector has a corresponding binary parameter. Using constraints given in equation 4.8 and 4.9, the binary variable, $\boldsymbol{\delta}_j$, is set to one when the corresponding loading is nonzero and set to zero when the corresponding loading is zero. The NLP cost function given by equation 4.7 is then equivalent to equation 4.6.

$$J(\lambda) = -\boldsymbol{p^T Q p} + \lambda \sum_{j=1}^{N} \boldsymbol{\delta}_j \tag{4.7}$$

$$L_j \boldsymbol{\delta}_j \leq p_j \leq U_j \boldsymbol{\delta}_j, \qquad (j = 1, \dots, N) \tag{4.8}$$

$$\boldsymbol{\delta}_j \in \{0,1\}, \qquad (j = 1, \dots, N) \tag{4.9}$$

$$-1 \leq \boldsymbol{p}_j \leq 1, \qquad (j = 1, \dots, N) \tag{4.10}$$

$$\boldsymbol{p}^T \boldsymbol{p} \leq 1 \tag{4.11}$$

$$for\ k \geq 2, \qquad \boldsymbol{P}_k{}^T \boldsymbol{P}_{k-i} = 0, \quad (i = 1, \dots, k-1) \tag{4.12}$$

The decision variables for the NLP model are the loadings, $\boldsymbol{p}$ $(j = 1, \dots, N)$, and the binary variables, $\boldsymbol{\delta}_j$ $(j = 1, \dots, N)$. The upper and lower bounds, U_j and L_j, are set to 1 and -1 respectively. The loadings are also bound between -1 and 1. The constraint given in equation 4.11 is in place to normalise the loadings vector to ensure that the optimisation problem is bounded. Note that the constraint given by (4.11) is a nonlinear function of the decision variables and therefore requires approximation in the linear programming approaches. The constraint given by (4.12) is the orthogonality constraint, which is applied to all principal components that are subsequent to principal component 1, whenever there is more than one principal component to be determined.

4.2.2 Nonlinear programming (NLP)

The NLP approach is the same as the MINLP approach with the exception of the penalty term in the cost function. The NLP approach is a relaxed version of the MINLP approach, which uses an approximation to the L0-norm penalty. In the literature, several alternative approximate L0-norm penalty functions have been proposed, including the exponential (Bradley and Mangasarian, 1998), log (Weston et al., 2003) and seamless-L0 (SELO) (Dicker et al., 2013) penalty functions. The formulas for these are listed in Table *4-1*.

Table 4-1: Popular L0-norm approximations from the literature. In these formulae, *p* represents the loadings vector, and ϵ_1, ϵ_2 and ϵ_3 are tuning parameters that affect the shape of the penalty.

Penalty function	Formula
Log	$\sum_{j=1}^{N} log(\lvert \boldsymbol{p}_j \rvert + \epsilon_1)$
Exponential	$1 - e^{-\epsilon_2 \lvert \boldsymbol{p}_j \rvert}$
SELO	$\frac{1}{\log{(2)}} \log\left(\frac{\lvert \boldsymbol{p}_j \rvert}{\lvert \boldsymbol{p}_j \rvert + \epsilon_3} + 1\right)$

For this work, the SELO penalty function was selected as the literature indicates that the SELO penalty is a smooth function that very closely approximates the L0-norm (Dicker et al., 2013; Shi et al., 2018). The cost function for the NLP approach can be obtained by substituting the L0-norm approximations in Table *4-1* into equation (4.13). The only additional modification to the constraints is the relaxation of the binary variables, $\boldsymbol{\delta}_j$, so that they may take any value in the range of 0 to 1. Constraint (4.8) ensures that $\boldsymbol{\delta}_j$ is assigned to the absolute value of its corresponding loading parameter. Consider the SELO approximation as an example, the final cost function is given by equation (4.14).

$$J(\lambda) = -\boldsymbol{p}^{\boldsymbol{T}}\boldsymbol{Q}\boldsymbol{p} + \lambda Pen \tag{4.13}$$

$$J = -\boldsymbol{p}^{\boldsymbol{T}}\boldsymbol{Q}\boldsymbol{p} + \lambda \sum_{j=1}^{N} \frac{1}{\log(2)} \log\left(\frac{\boldsymbol{\delta}_j}{\boldsymbol{\delta}_j + \epsilon_3} + 1\right) \tag{4.14}$$

4.2.3 Mixed integer linear programming (MILP)

As mentioned in the introduction to this section, the LP approaches utilise SLP to solve a sequence of linear subproblems which approximate the nonlinear sparse PCA optimization problem. The nonlinear functions which are to be approximated include the variance term in the cost function and the normalisation of the loadings. The first order Taylor series approximations to the variance term, $p^T\boldsymbol{Q}p$, and the normalisation of the loadings described by constraint (4.11) are given by (4.15) and (4.16), respectively.

$$2\boldsymbol{p}_k^T\boldsymbol{Q}\boldsymbol{p}_{k-1} - \boldsymbol{p}_{k-1}^T\boldsymbol{Q}\boldsymbol{p}_{k-1} \tag{4.15}$$

$$2\boldsymbol{p}_k^T\boldsymbol{p}_{k-1} - \boldsymbol{p}_{k-1}^T\boldsymbol{p}_{k-1} \leq 1 \tag{4.16}$$

To ensure convergence of the solution, the cutting plane method (Kelley, 1960) is used to reduce the search space with successive iterations. The application of this technique involves storing constraint described by equation 4.16 from all previous iterations and applying them in the constraints for the current iteration. This reduces the search space with each iteration until the solution meets the prescribed convergence criterion given by (4.17).

$$|(2\boldsymbol{p}_k^T\boldsymbol{Q}\boldsymbol{p}_{k-1} - \boldsymbol{p}_{k-1}^T\boldsymbol{Q}\boldsymbol{p}_{k-1}) - \boldsymbol{p}^{\boldsymbol{T}}\boldsymbol{Q}\boldsymbol{p}| \leq \epsilon_4 \tag{4.17}$$

Here, ϵ_4 is a sufficiently small number, for example, 1x10^{-8}. This convergence criterion requires that the absolute difference between the linearized variance term and the nonlinear variance term is smaller than ϵ_4. The final cost function is given by equation 4.18. The MILP cost function utilises the full L0-norm penalty, with the same implementation as the MINLP approach where $\boldsymbol{\delta}_j$ is the binary variable indicating whether the corresponding loading is zero or nonzero. Notice that the term $\boldsymbol{p}_{k-1}^T\boldsymbol{Q}\boldsymbol{p}_{k-1}$ was dropped from the cost function as it is a constant.

$$J = -2\boldsymbol{p}_k^T\boldsymbol{Q}\boldsymbol{p}_{k-1} + \lambda\sum_{j=1}^{N}\boldsymbol{\delta}_j \tag{4.18}$$

4.2.4 Linear programming (LP)

The LP model is the relaxed version of the MILP approach, where the only differences arise in the cost function. The LP model makes use of the same L0-norm approximations as the NLP model; however, since the L0-norm approximations are nonlinear, they were linearized using a first order Taylor series approximation in combination with SLP. A first order Taylor series approximation of the penalty function at the k+1 iteration is given by equation (4.19):

$$F(\boldsymbol{p}_{k+1}) \approx F(\boldsymbol{p}_{\boldsymbol{k}}) + \left.\frac{dF(\boldsymbol{p})}{d\boldsymbol{\delta}}\right|_k (\boldsymbol{\delta}_{k+1} - \boldsymbol{\delta}_k) \tag{4.19}$$

Since $F(\boldsymbol{p}_k)$ and $\left.\frac{dF(\boldsymbol{p})}{d\boldsymbol{\delta}}\right|_k \boldsymbol{\delta}_k$ are constants, they will not affect the optimal solution, leaving the cost function as described by equation (4.20).

$$J_{k+1}(\lambda) = -2\boldsymbol{p}_{k+1}^T\boldsymbol{Q}\boldsymbol{p}_k + \textstyle\sum_{i=1}^{N}\left.\frac{dF(\boldsymbol{p})}{d\boldsymbol{\delta}}\right|_k \boldsymbol{\delta}_{k+1} \tag{4.20}$$

Table *4-2* below provides the $\left.\frac{dF(\boldsymbol{p})}{d\boldsymbol{\delta}}\right|_k$ term for the L0-norm approximations, which may be substituted into equation 4.20 to derive the cost function at the k+1 iteration in terms of the relaxed binary variables $\boldsymbol{\delta}_{j,k}$.

Table 4-2: Formulae for the linearised versions of the penalty functions listed in Table 4-1.

	$\boldsymbol{F(p)}$	$\frac{\boldsymbol{dF(p_{j,k})}}{\boldsymbol{d\delta_{j,k}}}$
Log	$\sum_{j=1}^{N} log(\lvert\boldsymbol{p}_j\rvert + \epsilon_1)$	$\frac{1}{\boldsymbol{\delta}_{j,k} + \epsilon_1}$
Exponential	$1 - e^{-\epsilon_2\lvert\boldsymbol{p}_j\rvert}$	$\epsilon_2 e^{-\epsilon_2\lvert\boldsymbol{\delta}_{j,k}\rvert}$
SELO	$\frac{1}{\log{(2)}}\log\left(\frac{\lvert\boldsymbol{p}_j\rvert}{\lvert\boldsymbol{p}_j\rvert + \epsilon_3} + 1\right)$	$\frac{1}{\log{(2)}}\left(\frac{\epsilon_3}{(2\boldsymbol{\delta}_{j,k} + \epsilon_3)(\boldsymbol{\delta}_{j,k} + \epsilon_3)}\right)$

4.2.5 Summary of the four programming approaches to sparse PCA

The LP and MILP approaches to sparse PCA utilised SLP to linearise the nonlinear objective function and constrains, with the cutting plane technique used to converge on a solution. The NLP and MINLP iterative solutions as the nonlinear objective function and constraints were implemented directly, and with the NLP approach, a multi-start procedure was employed. Table ***4-3*** summarises the differences in the objective function and the normalisation of the loadings for the four approaches.

Table 4-3: Summary table showing the objective function and constraint for normalisation of the loadings, that were used in the four sparse PCA methods.

	Objective function	**Normalisation of loadings**
MINLP	$J = -\boldsymbol{p}^{\boldsymbol{T}}\boldsymbol{Qp} + \lambda\lVert\boldsymbol{p}\rVert_0$	$\boldsymbol{p}^T\boldsymbol{p} \le 1$
NLP	$J = -\boldsymbol{p}^T\boldsymbol{Qp} + \sum_{j=1}^{N}\frac{1}{\log{(2)}}\log\left(\frac{\boldsymbol{\delta}_j}{\boldsymbol{\delta}_j + \epsilon_3} + 1\right)$	$\boldsymbol{p}^T\boldsymbol{p} \le 1$
MILP	$J = -2\boldsymbol{p}_k^T\boldsymbol{Qp}_{k-1} + \lambda\lVert\boldsymbol{p}\rVert_0$	$2\boldsymbol{p}_k^T\boldsymbol{p}_{k-1} - \boldsymbol{p}_{k-1}^T\boldsymbol{p}_{k-1} \le 1$
LP	$J = -2\boldsymbol{p}_k^T\boldsymbol{Qp}_{k-1}$ $+\sum_{j=1}^{N}\frac{1}{\log{(2)}}\left(\frac{\epsilon_3}{(2\boldsymbol{\delta}_{j,k} + \epsilon_3)(\boldsymbol{\delta}_{j,k} + \epsilon_3)}\right)$	$2\boldsymbol{p}_k^T\boldsymbol{p}_{k-1} - \boldsymbol{p}_{k-1}^T\boldsymbol{p}_{k-1} \le 1$

The MILP and LP approaches were solved using the *intlinprog* and *linprog* functions in Matlab version 2019a, respectively. NLP was solved using the *fmincon* function in Matlab version 2019a. Finally, MINLP was solved using the *bonmin* solver, available in the OPTI toolbox, using MATLAB version 2019a. Table *4-4* provides more information on the solver settings that were used throughout.

Table 4-4: Details of solvers and solver options used in the LP, MILP, NLP and MINLP approaches to sparse PCA

Method	**Solver and solver options**
LP	Solver: *linprog* Convergence tolerance = 1x10^{-7} Non-default options: None Loadings initial values: all set to a small number e.g. 0.01
MILP	Solver: *intlinprog* Convergence tolerance = 1x10^{-7} Non-default options: None Loadings initial values: all set to a small number e.g. 0.01
NLP	Solver: *fmincon* Non-default options: algorithm = 'Sequential Quadratic Programming' (default = 'Interior-Point Algorithm'), max iterations = 2x10^{5} (default = 1x10^{3}), max function evaluations = 6x10^{5} (default = 3x10^{3}) Loadings initial values: Multi-start procedure with initial values obtained from a Latin hypercube design
MINLP	Solver: *bonmin* (OPTI toolbox) Non-default options: max iterations = 9x10^{5} (default = 1x10^{3}), max function evaluations = 2x10^{6} (default = 3x10^{3}) Loadings initial values: all set to a small number e.g. 0.01

Note that for the LP and MILP solvers, non-default options were explored to try to improve the speed of convergence, however, none were found to offer significantly improved performance.

4.2.6 Tuning of the regularisation parameter and evaluation of model performance

A common requirement of regularisation methods is the determination of the optimal regularisation penalty or degree of sparsity. Typically, this involves optimising a selected criterion, such as the Bayesian information criterion, Akaike information criterion, or the index of sparsity (IS), which serve to optimise the trade-off between variance explained and degree of sparsity (Gajjar et al., 2017; Journée et al., 2010; Qi et al., 2013; Willis and von Stosch, 2017). In this work, the IS was the selected criterion used to optimise the variance-sparsity trade-off. The formula for the IS is given by (4.21).

$$IS = \frac{V_a V_S}{V_O^2} \frac{TNZL}{ml} \tag{4.21}$$

Here, V_a, V_s and V_o are the adjusted, unadjusted and ordinary total variance, respectively, TNZL is the total number of zero loadings, m is the number of principal components and l is the number of variables. In order to optimise the IS, a simple search method was carried out to evaluate solutions over a range of penalty weights, using sufficiently small intervals to ensure that a near optimal solution could be obtained. This heuristic procedure was automated within a script in Matlab. The IS was the primary metric used to evaluate and compare sparse PCA solutions. The adjusted variance explained and the TNZL were also calculated and recorded to compare the variance explained and the degree of sparsity in the solutions.

4.2.7 Example datasets and data pre-processing

Two datasets were utilised to compare the performance of the four programming approaches to sparse PCA with one another, and with popular sparse PCA approaches from the literature.

Pitprops data

The pitprops is a classic dataset used to evaluate PCA models that was first introduced by Jeffers (1967). The pitprops dataset consists of 13 variables and 180 observations. The data contains several underlying factors, where the variables are correlated to one another to varying degrees. For this reason, the underlying structure is considered to be relatively complex and it therefore provides a good benchmarking dataset to compare the performance of sparse PCA solutions, as several authors have done (Farcomeni, 2009; Gajjar et al., 2017; Journée et al., 2010; Shen and Huang, 2008; Ulfarsson and Solo, 2011; Zou et al., 2006).

Synthetic data

While the pitprops data is good for evaluating the quality of sparse PCA solutions, it is relatively small with only 13 variables. To test the scalability of the sparse PCA solutions a number of synthetic datasets were generated with different numbers of variables and underlying factors. In total, seven datasets were generated, with underlying factors and number of variables shown in Table *4-5*.

Table 4-5: Number of variables and number of factors in seven synthetic datasets.

Dataset	1	2	3	4	5	6	7
No. of variables	10	20	30	40	60	100	200
No. of factors	2	2	3	3	5	5	5

For example, dataset 2 with 20 variables was generated as follows. Two underlying factors V_1 and V_2 were assigned to the 20 variables by adding an independent gaussian noise component to each one. V_1 was assigned to the first eight variables and V_2 was assigned to variables 9 to 20.

$$\boldsymbol{V}_1 \sim N(0{,}290) \tag{4.22}$$

$$\boldsymbol{V}_2 \sim N(0{,}300) \tag{4.23}$$

$$\boldsymbol{X}_i = \boldsymbol{V}_1 + \epsilon_i^1,\ \ \epsilon_i^1 \sim N(0{,}1),\ \ i = 1{,}2{,}\ldots{,}8 \tag{4.24}$$

$$\boldsymbol{X}_i = \boldsymbol{V}_1 + \epsilon_i^2,\ \ \epsilon_i^2 \sim N(0{,}1),\ \ i = 8{,}9{,}\ldots{,}20 \tag{4.25}$$

Data pre-processing

Following the standard practice outlined in chapter 3, section 3.4.3, the data were autoscaled using equation (4.26).

$$\boldsymbol{X}_{scaled} = \frac{\boldsymbol{X}-\overline{\boldsymbol{X}}}{\sigma_X} \tag{4.26}$$

PCA was then conducted on the covariance matrix, as calculated by equation (4.27):

$$\boldsymbol{Q} = \frac{\boldsymbol{X}^T\boldsymbol{X}}{n-1} \tag{4.27}$$

where $\boldsymbol{X}$ is the scaled data and n is the number of observations.

4.3 Results

This section begins with a comparison of the four approaches to sparse PCA, in order to select the best approaches based on the quality of the solution and the ability to scale to larger problem sizes. The pitprops dataset was explored initially to test and compare the quality of the sparse PCA solutions to one another.

4.3.1 Comparison of the four programming approaches to sparse PCA

The four programming approaches to sparse PCA are compared here, using the pitprops dataset and synthetic datasets.

4.3.1.1 Pitprops data

The pitprops dataset was initially used to evaluate and compare the performance of the four programming approaches to sparse PCA. The pitprops dataset is particularly useful for comparing the quality of the solutions obtained, due to the numerous underlying factors present within the data. For all of the approaches, when the penalty parameter is sufficiently small, the loadings are all nonzero. Increasing the penalty leads to increased sparsity in the loadings and consequently a reduction in the variance explained by each principal component. To determine the optimal penalty parameter and corresponding sparse PCA solution, the penalty parameter was varied across an appropriate range of values and the index of sparsity was recorded. The solution which maximised the index of sparsity was then determined to be the optimal solution. An example of this tuning of the regularisation parameter is provided for the NLP approach later in section 4.3.2. Key model performance metrics for the optimal solution for each of the four approaches are displayed in Table ***4-6***, including the adjusted variance explained, the index of sparsity and the time taken to solve.

Table 4-6: Comparison of the four different approaches to sparse PCA: MILP, LP, MINLP and NLP, when fitting a six-component model to the pitprops dataset. These near optimal solutions were obtained after a trial and error search through different penalty terms to maximise the index of sparsity for each of the four approaches.

	Penalty applied (λ)	Adjusted variance explained (%)	Index of sparsity (IS)	Time to solve (s)
MILP	0.15	79.6	0.488	381.6
LP	1.6	75.4	0.425	51.5
MINLP	0.1	79.6	0.488	23.5
NLP	0.03	79.6	0.489	25.2

On the basis of the IS value, the MILP, MINLP and NLP approaches were evenly matched, producing IS values of 0.49. The solutions were nearly identical with respect to the values of the loadings and the elements selected as nonzero were consistent across all three approaches. With the LP approach, it was not possible to identify a single regularisation parameter that could be applied to all components to produce a competitive IS value. For example, a penalty value of 0.17 resulted in no nonzero elements being selected in component 6, yet component 1 has too many nonzero elements compared to the optimal solutions obtained with the other three approaches. As a result, the maximum IS value obtained with the LP approach was significantly lower than the other three approaches. To demonstrate this, Table ***4-7*** shows a comparison of the loadings produced from the LP and MILP approaches. It is possible to obtain a better solution using the LP approach by varying the regularisation parameter for each principal component individually; however, this is laborious and offers no advantage compared to the alternative approaches, which are able to obtain good solutions while only requiring the identification of a single penalty parameter.

To highlight the benefit of sparse PCA over standard PCA, Table ***4-8*** shows the loadings for the first 6 components of standard PCA. Comparing these to the sparse loadings produced from the LP and MILP approaches, it is clear that sparse loadings reduce the level of complexity involved in interpreting the principal components. Additionally, Figure 4-1 shows a biplot for standard PCA next to a biplot of the sparse PCA model produced by the MILP approach. Figure 4-1 shows that the scores of the first two principal components are impacted minimally by the choice of model, however, the loadings are easier to interpret because the majority of small loadings present in the PCA model are set to zero, leaving the large contributors only. Here, the sparse PCA model clearly shows that PC1 represents contributions from variables 1 and 2 and 6 to 10, which are all positively correlated, while principal component 2 represents variables 3 and 4, which are again positively correlated.

Table 4-7: Comparison of the MILP and LP sparse PCA solutions for the pitprops dataset. The rows Topdiam to Diaknot present the loadings for the 13 variables in the pitprops dataset. The percentage of explained variance (PEV) for each component and the number of zero loadings (NZL) are provided in the bottom two rows.

Variable	**MILP, λ = 0.15**						**LP, λ = 1.8**					
	SPC1	**SPC2**	**SPC3**	**SPC4**	**SPC5**	**SPC6**	**SPC1**	**SPC2**	**SPC3**	**SPC4**	**SPC5**	**SPC6**
Topdiam	0.42		0.29				0.44				-0.08	
Length	0.43		0.29				0.45		0.01			
Moist		0.71					0.03	-0.73				
Testg		0.71					0.08	-0.66				
Ovensg			-0.50			0.73				1.00		
Ringtop	0.27		-0.42				0.24					
Ringbut	0.40		-0.34				0.40					
Bowmax	0.31						0.29					
Bowdist	0.38						0.38					
Whorls	0.40						0.40	0.18				
Clear				-1.00								
Knots					1.00		-0.03				-1.00	
Diaknot			0.53			0.68	-0.01		1.00			
PEV (%)	31.3	14.2	14.2	7.4	6.6	5.9	31.8	14.5	7.5	7.2	6.4	0.0
NZL	6	11	7	12	12	11	2	10	11	12	11	13

Table 4-8 Loadings of the first 6 principal components of a standard PCA model of the pitprops data

Variable	Standard PCA					
	SPC1	SPC2	SPC3	SPC4	SPC5	SPC6
Topdiam	-0.30	-0.25	-0.18	-0.08	0.06	-0.22
Length	-0.31	-0.23	-0.20	-0.08	0.06	-0.25
Moist	0.03	-0.59	0.22	-0.09	-0.21	0.29
Testg	-0.08	-0.50	0.44	-0.05	-0.22	0.11
Ovensg	-0.18	0.29	0.60	0.16	-0.30	-0.45
Ringtop	-0.32	-0.05	0.31	0.05	0.38	-0.22
Ringbut	-0.34	0.00	0.06	0.03	0.20	-0.11
Bowmax	-0.33	-0.01	-0.24	0.11	-0.15	0.20
Bowdist	-0.33	-0.14	-0.21	0.02	0.06	-0.08
Whorls	-0.34	0.00	-0.13	-0.05	0.01	0.06
Clear	0.21	-0.31	-0.14	0.89	0.08	-0.19
Knots	0.31	-0.20	0.16	-0.22	0.69	-0.12
Diaknot	0.29	-0.21	-0.26	-0.31	-0.33	-0.66

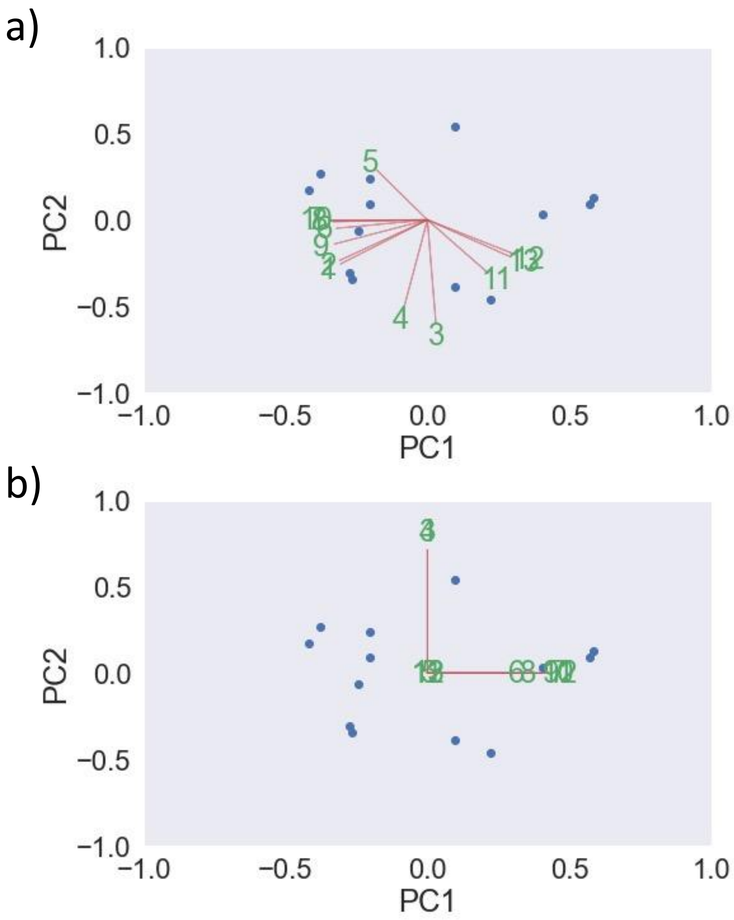


Figure 4-1 PCA biplot of components 1 and 2 for the pitprops data. a) standard PCA model, b) sparse PCA model obtained using the MILP approach.

4.3.1.2 Synthetic data

Synthetic datasets were generated, as described in section 4.2.7, to test the scalability of the four sparse PCA approaches. For this test, the synthetic data has been generated with very simple underlying factors which are easy to identify because the aim here is to test scalability rather than accuracy and precision. The smallest synthetic dataset (dataset 1) consists of ten variables and two underlying factors, while the largest (dataset 7) contains 200 variables and five underlying factors. Taking synthetic dataset 2 as an example, two underlying factors are present. Factor 1 is composed of correlated variables 1 to 8, and factor 2 includes correlated variables 9 to 20. The PCA solution was considered successful when the nonzero loadings correctly identified the underlying factors in the data. For example, Table *4-9* shows the sparse PCA solution for dataset 2 obtained using the NLP approach, which correctly identifies the two underlying factors.

Table 4-9: NLP sparse PCA solution for synthetic dataset 2, featuring two principal components (PC1, PC2).

Variable	PC1	PC2
1	0.0	0.4
2	0.0	0.4
3	0.0	0.4
4	0.0	0.4
5	0.0	0.4
6	0.0	0.4
7	0.0	0.4
8	0.0	0.4
9	-0.3	0.0
10	-0.3	0.0
11	-0.3	0.0
12	-0.3	0.0
13	-0.3	0.0
14	-0.3	0.0
15	-0.3	0.0
16	-0.3	0.0
17	-0.3	0.0
18	-0.3	0.0
19	-0.3	0.0
20	-0.3	0.0

Table ***4-10*** shows the time taken to obtain solutions for the seven synthetic datasets. Time recordings in Table ***4-10*** indicate that all of the underlying factors were successfully identified in the time stated (s), where no time is given, the respective SPCA approach did not produce a solution with the solver settings specified in Table

4-4. In most cases where the solver did not produce a solution, the solver began to converge extremely slowly, and used up excessive computer memory. The information in Table ***4-10*** is also displayed graphically in Figure 4-2, which helps to show the rate of increase in time to solve when the number variables and factors is increased.

Table 4-10: Time to solve (seconds) synthetic datasets 1 to 5, for each of the SPCA approaches. The number of variables (n) and the number of PCs that were determined for each dataset is stated in the table. The shaded boxes indicate that no solution was obtained.

	MILP	**LP**	**MINLP**	**NLP**
Dataset 1 (n=10, 2PCs)	29.6	2.1	27.5	8.2
Dataset 2 (n=20, 2PCs)	▨	7.9	579.3	10
Dataset 3 (n=30, 3PCs)	▨	15.3	2966.6	18.5
Dataset 4 (n=40, 3PCs)	▨	474.1	▨	43.2
Dataset 5 (n=60, 5PCs)	▨	1340.9	▨	129.5
Dataset 6 (n=100, 5PCs)	▨	▨	▨	657.4
Dataset 7 (n=200, 5PCs)	▨	▨	▨	3664.9

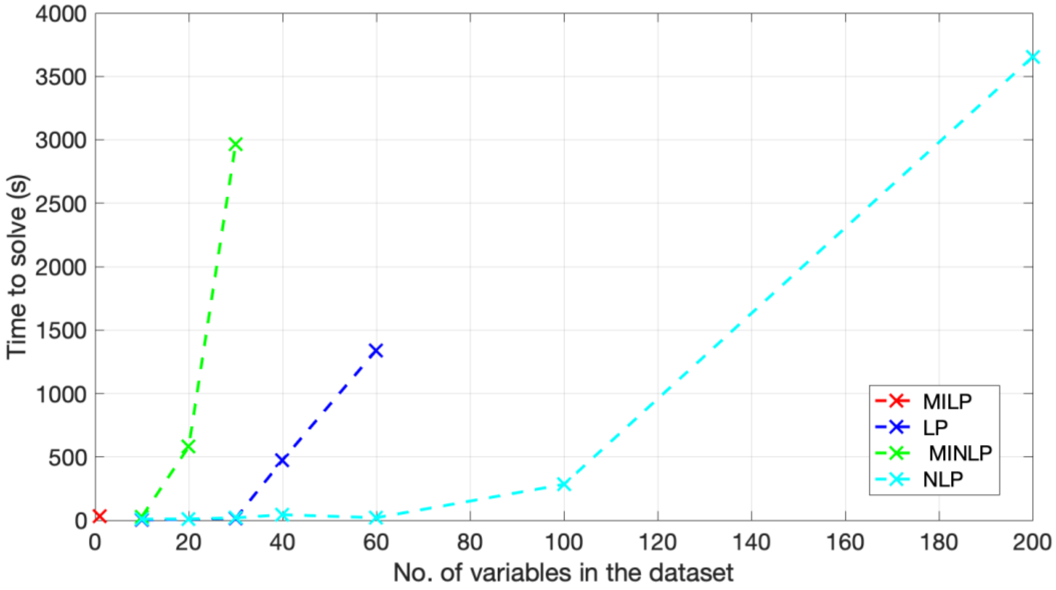


Figure 4-2: Time to solve versus the no. of variables in the dataset, for the seven synthetic datasets, for the four programming approaches to sparse PCA

Table *4-10* shows that the MILP approach lacks scalability compared to the other approaches, since it was only able to produce a solution for dataset 1 containing 10 variables. The MILP approach attempts to solve the full L0-constrained optimisation problem, where the objective function is nonconvex and discontinuous, and the computational complexity is known to be NP-hard. The results show that the scalability of the MILP approach to L0-regulairised sparse PCA is limited by the computational complexity of the problem. Henceforth in practice, the MILP approach is applicable to small problems with fewer than approximately 20 variables.

The LP approach, a relaxed version of the MILP approach, was able to produce solutions for all five synthetic datasets. The time to solve the LP problem increased exponentially with the number of variables; however, the rate of increase was lower compared to the MILP approach and the time to solve remains reasonable with up to 60 variables. This result shows that smooth approximations to the L0-norm penalty greatly reduce the computational complexity, leading to a solution within a reasonable number of iterations and decreasing the solution time. Using the LP approach with the synthetic datasets, it was possible to identify a single penalty parameter that obtained the correct loadings to represent the underlying factors in the data, whereas the LP solution for the pitprops dataset was substandard. This is likely due to the relatively simple structure of the synthetic data compared to the pitprops data.

The MINLP approach produced solutions for datasets 1 to 3, although for datasets 2 and 3, it was the slowest of the three successful approaches. As with the LP and MILP approaches, the solve time increased exponentially with the number of variables. Compared to the MILP approach, the MINLP solved the full L0-constrained optimisation problem with greater efficiency. This may be partly attributed to the need to carry out SLP with the MILP approach and to differences in the performance of the solver used in the MILP (Matlab, *intlinprog*) and MINLP (Matlab OPTI toolbox, *bonmin*) approaches.

Finally, the NLP approach produced solutions for all five datasets, and it was significantly faster than the LP approach for the two largest datasets. When compared to the MINLP approach, the use of the L0-norm approximation made the NLP approach significantly faster and scalable to larger problems, while retaining the ability to produce sparse loadings with the correct underlying factors identified. In both cases, MILP versus LP and MINLP versus NLP, the relaxed non-integer variants outperformed their respective integer programming variants with regards to scalability.

This outcome demonstrates that the high computational complexity associated with the L0-norm penalty limits its scalability, and that L0-norm approximations offer a more practical alternative. It is possible that different solvers may offer improved performance. Additionally, more powerful computer hardware would likely reduce the computation time and increase the scalability of the full L0-constrained optimisation problems. However, with the hardware utilised in this work and the solvers that were chosen, the full L0-constrained approaches were not suitable for problems with more than 30 variables.

4.3.1.3 Summary of the tests on the pitprops and synthetic datasets

The LP approach was shown to have limitations with the pitprops dataset due to the inability to identify a single regularisation parameter that could be applied to all principal components to produce a competitive solution. The MILP approach produced competitive solutions for the pitprops dataset; however, it lacked scalability as it was unable to produce a solution for the 20-variable synthetic dataset. The results indicate that the MILP approach is limited to somewhere between 10 and 20 variables. The MINLP approach was the more successful approach using the full L0-norm penalty, as it successfully scaled to the 30-variable dataset. There are problems within chemical engineering and other domains, which do not exceed 30 variables and the MINLP approach is a viable method to apply full L0-regularisation in these instances, with commonly available hardware and solvers. However, the NLP approach is scalable to problems with 200 variables and considering the quality of the solutions obtained with the pitprops dataset, the relaxed NLP solution was just as good as the MINLP solution, thus providing no reason to select the MINLP approach over the NLP approach. There may be some instances where the practitioner desires to use the full L0-norm penalty, in most cases however, it appears that the NLP approach is the best of the four approaches when considering the quality of the solution, the time to solve and scalability.

4.3.2 NLP multi-start

The NLP solver in Matlab, *intlinprog*, requires initial values for model parameters, which are the loadings in this case. The initial values can impact the rate of convergence and the final solution obtained; however, it is difficult to predict the effect of the initial values and to determine what they should be. The initial values should not violate the model constraints. For the sparse PCA model, options for the initial values include (i) loadings from standard PCA, (ii) all zeros, (iii) all *x*, where *x* denotes a small positive or negative number within the problem boundaries or (iv) randomly chosen small numbers of magnitude less than one within the problem boundaries.

Due to the fact that it's difficult to predict the impact of the initial values on the solution, it was decided that a multi-start procedure would be used to produce multiple solutions with different starting points. To generate a set of starting points with a wide coverage across the space of possible starting points, a Latin hypercube design was applied. In order to ensure that the starting points where within the problem boundaries, the starting points were scaled to a maximum absolute value of $\sqrt{1/n}$, where n is the number of variables. The number of starts is a parameter that needs to be specified. For a larger number of starts, the final solution obtained is likely to be closer to the global optimum, due to increased coverage of the space of possible starting points. The drawback is increased computational demand and time to solve. Henceforth, the number of starts must be selected with this trade-off in mind. Figure 4-3 shows the index of sparsity, variance explained and total number of zero loadings plotted against the regularisation parameter for models identified from 10, 50, 200 and 400 starts.

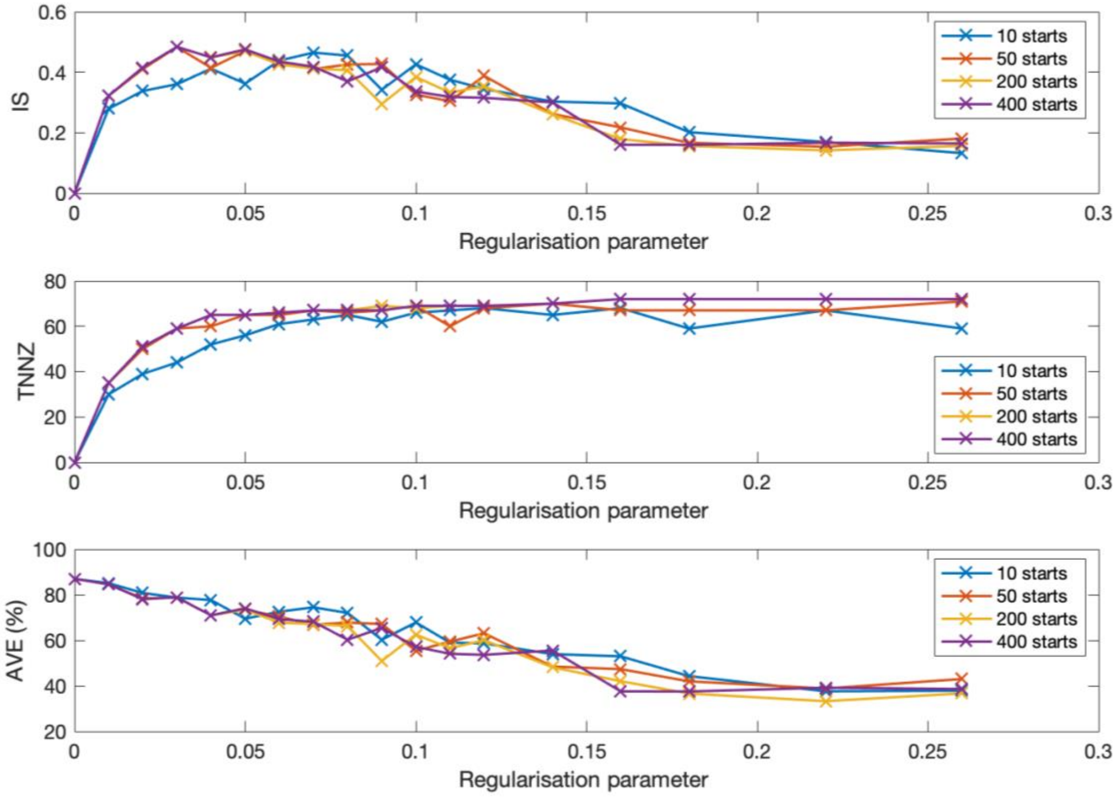


Figure 4-3: (a) Index of sparsity, (b) total number of zero loadings and (c) adjusted variance explained versus the regularisation parameter, for sparse PCA models identified after 10, 50, 200 and 400 starts.

The index of sparsity is zero when λ is zero and the variance explained by the model is maximal, due to complete nonzero loadings in the unpenalized solution. As λ increases, the TNZL increases, variance explained decreases and the IS increases because the first loadings to be reduced to zero are the ones which explain a small portion of the variance. The IS peaks when the trade-off between sparsity and variance explained reaches an optimal balance. The 400, 200 and 50 start models reach maximum IS at the same point (λ = 0.03), while the 10 start model peaks at $\lambda = 1$. The different solutions are obtained due to the presence of local optima. As λ increases further, the TNZL increases and the variance explained decreases. The IS decreases as the loss of variance explained by the model outweighs the gain in sparsity. For models obtained from 400 starts, the TNZL increases monotonically with the increasing magnitude of λ; whereas for the models produced from 10 starts, the TNZL fluctuates as λ increases. This contrast can be attributed the fact that there is more variation in the quality of the solutions (distance from global optimum) for the 10 start models, because the initial values have reduced coverage.

When it comes to determining the number of starts that should be used to solve a given problem, the number of starts is likely to increase based on the number of variables because there are more solutions. It is difficult to know prior to investigating the problem whether or not the number of starts will be sufficient. Here, 50 starts were sufficient to obtain the best solution (max IS), that was also identified using 200 and 400 starts.

4.3.3 Comparison of the NLP approach to alternatives in the literature

The following section first compares the NLP sparse PCA solution for the pitprops data with other popular solutions in the literature. Table ***4-11*** shows the NLP method sparse PCA and the solution developed by Gajjar *et al.* (2017), which used a greedy search algorithm to search through all possible sparse combinations. Table ***4-12*** shows the sparse PCA solutions by Zou *et al.* (2006), utilising the elastic net penalty function, and Farcomeni (2009), using a branch and bound algorithm to optimise the degree of sparsity. The total adjusted variance explained, the total number of zero loadings and the SI for these four solutions are displayed in Table ***4-13***. Additionally, the SI and total number of zero loadings for one of the solutions obtained by Journee *et al.* (2010), using the generalised power method is presented. Together these represent high performing sparse PCA solutions obtained by a variety of alternative methods.

Table 4-11: Sparse PCA solution for the pitprops data using the NLP approach, compared to the sparse PCA solution obtained by Gajjar et al. (2017).

Variable	**NLP**						**Gajjar *et al.* (2017)**					
	SPC1	**SPC2**	**SPC3**	**SPC4**	**SPC5**	**SPC6**	**SPC1**	**SPC2**	**SPC3**	**SPC4**	**SPC5**	**SPC6**
Topdiam	0.42		0.29				-0.5					
Length	0.43		0.29				-0.51					
Moist		0.71						0.77				
Testg		0.71						0.64				
Ovensg			-0.50			0.73	0.15		0.64			
Ringtop	0.27		-0.42						0.61			
Ringbut	0.40		-0.34				-0.23		0.46			
Bowmax	0.31						-0.36					
Bowdist	0.38						-0.44					
Whorls	0.40						-0.32			0.23		
Clear					1.00					-0.97		
Knots				1.00							-1	
Diaknot			0.53			0.68						1
NZL	6	11	7	12	12	11	6	11	10	11	12	12
PEV	31.3	14.2	14.2	7.4	6.6	5.9	29.3	14.2	13.4	8.0	6.7	6.0

Table 4-12: Sparse PCA solutions for the pitprops data, produced by Zou et al. (2006) and Farcomeni (2009).

Variable	Zou *et al.* (2006)						Farcomeni (2009)					
	SPC1	SPC2	SPC3	SPC4	SPC5	SPC6	SPC1	SPC2	SPC3	SPC4	SPC5	SPC6
Topdiam	-0.48						-0.423					
Length	-0.47						-0.43		-0.283			
Moist		0.79						0.707				
Testg		0.62						0.707				
Ovensg	0.18		0.66						0.6			0.704
Ringtop			0.59				-0.268		0.455			
Ringbut	-0.29		0.47				-0.403					
Bowmax	-0.34	-0.03	-0.05				-0.313					
Bowdist	-0.41						-0.379					
Whorls	-0.38						-0.4					
Clear				-1						1		
Knots					-1						1	
Diaknot						1			-0.594			0.71
NZL	6	10	9	12	12	12	6	11	9	12	12	11
PEV	29.1	14.4	13.4	7.4	6.7	6.0	31.3	14.2	11.7	7.5	6.6	5.9

Table 4-13: Comparison of the performance metrics for sparse PCA solutions for the pitprops data.

Method	NZL	Adjusted variance explained (%)	Index of sparsity
Gajjar *et al.* (2017)	62	77.6	0.501
Zou *et al.* (2006)	61	77.2	0.483
NLP	59	79.6	0.489
Farcomeni (2009)	61	77	0.47
Journee *et al.* (2010)	60	76.7	Not provided

The NLP approach produced a sparse PCA solution with the second highest IS value in Table ***4-13***. Zou *et al.* (2006) obtained a solution with more zero loadings; however, the NLP approach scored higher on the IS, indicating a better ratio of adjusted explained variance to nonzero loadings in the model. The greedy approach developed by Gajjar *et al.* (2017) achieved the highest IS. The branch and bound approach developed by Farcomeni (2009) performed similarly to the NLP method, with a slightly lower IS score. Journee *et al.* (2010) did not provide the loadings or the IS values for their pitprops solutions, so the IS was not obtained. However, from the variance explained and number of zero loadings, it can be seen that their solution offers similar performance, on the pitprops data, to the other methods in Table ***4-13***.

Overall, the results show that with small datasets, such as the pitprops dataset, the NLP approach to sparse PCA is capable of delivering high quality solutions, with similar performance to popular sparse PCA approaches in the literature. As far as scalability is concerned, the NLP approach was previously shown to solve problems with 100 variables in a reasonable time, allowing for heuristic procedures to be implemented to determine the optimal regularisation parameter. With a problem of 200 variables in size, the NLP approach was capable of identifying 5 sparse principal components in approximately 1 hour. While this is still feasible, it is close to the limit of what is practical when there is a need to optimise the regularisation parameter, given that there are significantly faster alternatives in the literature. Efficient approaches for large and/or high dimensional datasets include the generalised power method (Journée et al., 2010) for L0 and L1-regularised solutions, and the DSPCA method by d’Aspremont *et al.* (2007) for L1-regularisation.

4.4 Conclusions

In this work, LP, MILP, NLP and MINLP approaches were used to derive sparse PCA models with L0-norm and approximate L0-norm penalties applied to the PCA loadings. The LP approach produced substandard solutions due to difficulty achieving an appropriate degree of sparsity. The MILP approach performed well with the pitprops dataset (13 variables); however, it proved to be impractical for larger problems after failing to obtain a solution for the 20-variable synthetic data, with the solver settings that were specified. The MINLP approach also produced a good solution for the pitprops data and scaled to larger problems than the MILP approach (30-variable synthetic dataset was solved); although ability to scale to larger problems was still limited. The NLP method performed the best out of the four approaches, as it produced high quality solutions, comparable to high-performing sparse PCA algorithms in the literature, and it scaled to problems with up to 200 variables. This scalability is sufficient to make the NLP method a feasible choice for many PCA applications that are encountered. In chemical engineering for example, PCA has regularly been applied to chemical process data, which often features less than 100 variables. Conversely, in bioinformatics it is common to analyse high-dimensional genome data consisting of several thousand variables, where the NLP sparse PCA method would be infeasible or impractical compared to alternative sparse PCA approaches in the literature. The NLP approach to sparse PCA that was developed here will later be applied to the CGT process data in chapters 5 and 6. The CGT process data used in chapters 5 and 6 featured approximately 80 and 160 process variables for analysis with sparse PCA, respectively.

Chapter 5

Analytics for viral vector production: adherent cell process

This chapter focuses on an adherent cell process for viral vector manufacturing and the use of MVDA to investigate historical process data, with the overall aim to derive beneficial insights into process behaviour and to contribute towards solutions for critical manufacturing challenges. The MVDA approach was conducted in two phases: 1) feature extraction was carried out using PCA and the sparse PCA approach that was developed in chapter 4, 2) predictive modelling was carried using regression methods to model the relationship between process variables and critical quality attributes of the viral vector product.

5.1 Introduction

The most well-established method for GMP-grade lentiviral vector production is through the transient transfection of adherent cell cultures (Ausubel et al., 2012; Merten et al., 2016; Rout-Pitt et al., 2018). In this chapter, an adherent culture process for LV production, and the following downstream processing of the LVs, is investigated using MVDA to address some of the key manufacturing challenges described in chapter 2. These include high levels of variability in materials and production methods and a lack of advanced process knowledge, due to the inherent complexity in the system and process that are still in development.

The sparse PCA algorithm that was developed in chapter 4 was applied here to carry out feature extraction to provide insights into process variability and correlations between process parameters. This was followed by predictive modelling using PLS regression to model the relationships between process parameters from LV production and the infectious titre of the LV product. The infectious titre is a CQA of the LV product, which is a measure of the concentration of infectious particles produced from each batch. The aim was to identify critical process parameters influencing the infectious titre and to quantify the relationship to shed light on process behaviour.

5.1.1 Data

For this work, data was combined from the production of two different viral vector products to treat different unspecified diseases. The manufacturing process is the same for the two products, the only difference being the therapeutic plasmid that is used in the 3rd generation lentiviral vector system. 29 batches of LV manufacturing were available for the analysis, 13 of these were for treatment 1 and 16 batches were for treatment 2. The datasets were readily integrated as they contained all of the same variables. The LV manufacturing data consisted of 98 manufacturing process variables i.e. input parameters and 16 dependent variables recorded on the certificate of analysis for the final LV product. Of the 98 process variables, 17 were categorical and 81 were metric.

Table 5-1 Unit operations and a description of the types of variables available in the adherent viral vector production dataset

Unit operation	No of variables	Description of variable types
Cell expansion	36	**Metric** Number, volume, concentration and viability of cells, volumes of reagents and volume of growth media **Categorical** Material lots for growth media and reagents
Transfection	15	**Metric** Number, volume, concentration and viability of cells, volumes of reagents, volume of growth media and volume and concentration of plasmids, duration of process **Categorical** Material lots for growth media, reagents and plasmids
Endonuclease treatment	8	Temperature, oscillation speed, volumes of reagents **Categorical** Material lots for reagents
Ion exchange chromatography	12	**Metric** Number, volume, concentration and viability of cells, column settings, volumes before and after processing **Categorical** Column settings
Concentration step	4	Volumes before and after processing, concentration method **Categorical** Type of concentration method
Sterile filtration	5	Filter area **Categorical** Filter type

5.2 Methods

5.2.1 Pre-processing of process data

The pre-processing of the cross-sectional process data involved typical steps, such as the handling of missing data and representation of categorical variables as dummy variables. The following section details the pre-processing steps that were taken to obtain datasets for the PCA, sparse PCA and PLS analyses that were conducted.

5.2.1.1 Processing of cross-sectional process data

1. The variables in the dataset were evaluated for missing data and variables with significant portions of missing data (greater than 20% across more than 20% of the batches) were removed from the analysis.
2. After removing missing data, categoric variables were coded as dummy variables, meaning that *x* new variables were created to replace the original categoric variable. Here *x* denotes the number of categories minus one. See chapter 3, section 3.4.1 for details of dummy variable coding.
3. For PCA, sparse PCA and PLS, the data inputs were auto-scaled, i.e. each variable was transformed by subtracting the mean and dividing by the standard deviation. Mean centring and scaling were both carried out due to the improved interpretability of the results, which they provide, for more details the reader is referred to chapter 3, section 3.4.3.

5.2.1.2 Processing of product quality data

The product quality data modelled here was the infectious titre LV product as recorded on the certificate of analysis for each batch. The infectious titre was the response variable that was predicted with PLS regression modelling. The response variable required scaling. Autoscaling was applied to the response variable, consistent with the scaling applied to the input variables. The measurement error on a typical infectious titre assay is reported to be around 35-40% in the literature (Roldão et al., 2009), which is the approximate error of the infectious titre used in this work.

5.2.2 PCA and sparse PCA

The sparse PCA algorithm that was developed in chapter 4 was applied to the cross-sectional process data to carry out feature extraction. Sparse PCA was used to obtain simplified principal components for easier interpretation compared to standard PCA; however, standard PCA was also applied to the data in order to compare the results and check whether there were any important differences in the model outputs. For details on the sparse PCA algorithm, the reader is referred to the nonlinear programming approach to sparse PCA, which was described in chapter 4, section 4.2.2. For more details on standard PCA, see chapter 3, section 3.1.1.

5.2.3 Development of PLS regression models

In this chapter, PLS models were developed to predict the infectious titre of the LV product using the process parameters from cell expansion and downstream processing as predictor variables. The following list details the steps taken to develop the PLS models:

1. Initially predictor variables were pre-selected by ruling out duplicate variables and variables of no relevance to the analysis. The pre-selection was kept to a minimum to allow the data to reveal key correlations and the variable selection procedure to remove redundant variables of low predictive power.
2. The scaled predictor variables in the matrix $\boldsymbol{X}$, and the scaled response variable in the vector $\boldsymbol{y}$, from the training dataset, were then passed to the repeated K-fold cross-validation script in MATLAB (details on the repeated K-fold cross validation method are provided in chapter 3, section 3.3.3). This script carried out 5-fold cross validation, with 2000 repeats, for models with n latent components, where n was varied from 1 to 8.
3. The model with the smallest cross validation mean absolute error (MAE) score was selected and the number of latent components was determined.
4. Next the PLS model was fitted to the whole of the training data and the variable importance of projection (VIP) selection method (detailed in chapter 3 section 3.2.2) was used to select the most important predictor variables from the model. The threshold VIP cut-off was varied between 1 and 1.4, due to the high degree of multicollinearity in the data.
5. The reduced model was put through repeated 5-fold cross validation (2000 repeats) and the number of latent components was varied from 1 to 10.
6. The optimal model (minimum cross validation) was then selected and fitted to the whole of the training data to evaluate model fit and determine model parameters.
7. After identifying the reduced model, residual plots were checked and some nonlinearity in the errors was observed. Consequently, nonlinear model terms were investigated by generating new variables with squared terms and interaction terms. See chapter 3, section 3.2.3 for information on nonlinear transformations.
8. The VIP selection technique was reapplied to the model with additional nonlinear variables and those with low VIP scores (less than 0.8) were removed.
9. Using the bootstrap technique detailed in chapter 3, section 3.3.4, the stability and distribution of the model regression coefficients were evaluated with 2000 bootstrap samples. This provided confidence intervals for the standardised regression coefficients (beta coefficients).

5.2.3.1 PLS model performance evaluation

Model performance was evaluated through repeated K-fold cross validation (CV MAE and CV R^2), model fit to the whole training data (visual inspection, MAE and R^2) and bootstrapping (stability of regression coefficients).

5.3 Results and discussion

5.3.1 Sparse PCA results and comparison with PCA

Sparse PCA and standard PCA were applied to the process variables from adherent LV production to develop an understanding of the corelations between the process parameters and the main features of variance in the manufacturing data. Figure 5-1 below shows the percentage of variance explained by each of the first six principal components for sparse PCA and standard PCA. The information in Figure 5-1 may be used to decide how many principal components should be used in the model, as discussed in chapter 3, section 3.1.1. In this case, six principal components are sufficient to explain around 60% of the variation in the data and subsequent components explain less than 5% each. These initial components provide information about the correlation between the original variables and batchwise information on process conditions, where the clustering and separation of batches highlights similarities and differences in process conditions. The variation in later components is the result of smaller deviations in process conditions, where the signal to noise ratio is small and therefore the observations are of less significance (Bro and Smilde, 2014). For the purpose of this exploratory analysis, the first six components are sufficient to highlight the key features in the data, while looking into components seven and beyond would yield limited insights.

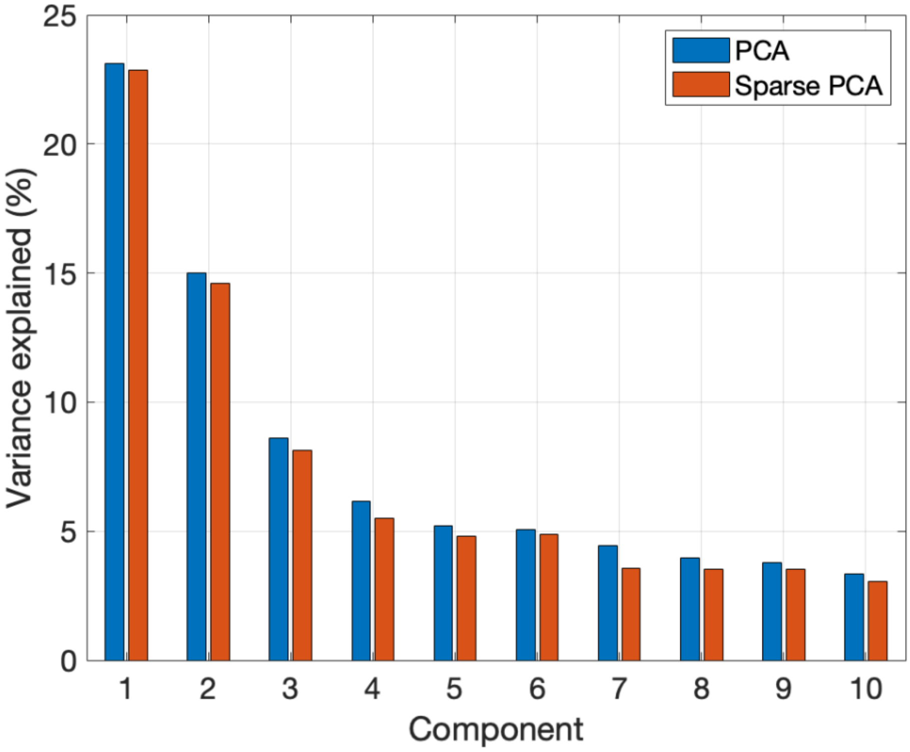


Figure 5-1: Variance explained by the first 10 principal components in the standard PCA and sparse PCA models. The PCA models were built on process variables from adherent cell culture viral vector production.

It can be seen in Figure 5-1 that the sparse PCA model explains slightly less variation in each principal component than standard PCA. This is the expected result due to the sparse loadings (see chapter 4, section 4.1); however, the reduction in explained variance is relatively small and the advantage gained in the ease of interpretation of the sparse principal components is significant. Table 5-2 demonstrates this with a comparison of the number of nonzero loadings on each principal component (PC) for sparse and standard PCA.

Table 5-2: Number of nonzero loadings in each principal component (PC).

		PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**
Sparse PCA	No. of nonzero loadings	39	27	28	17	14	13
	Cumulative variance explained (%)	22.8	37.5	45.6	51.1	55.9	60.8
PCA	No. of nonzero loadings	53	53	53	53	53	53
	Cumulative variance explained (%)	23.1	38.1	46.7	52.8	58.0	63.1

Table 5-2 shows that the first sparse principal component in sparse PCA has the most nonzero loadings and that there is a trend where the number of nonzero loadings decreases in subsequent components, making interpretation easier. Sparse principal components 2 and 3 feature around half of the original variables and components 4 to 6 feature close to one quarter of the original variables. The simplified principal components come at a relatively small cost to the explained variance with a drop of 2.3% compared to standard PCA across the 6-component model.

5.3.1.1 Components 1 and 2

Figure 5-2 shows the scores for the PCA (Figure 5-2a) and sparse PCA (Figure 5-2b) models for components 1 and 2.

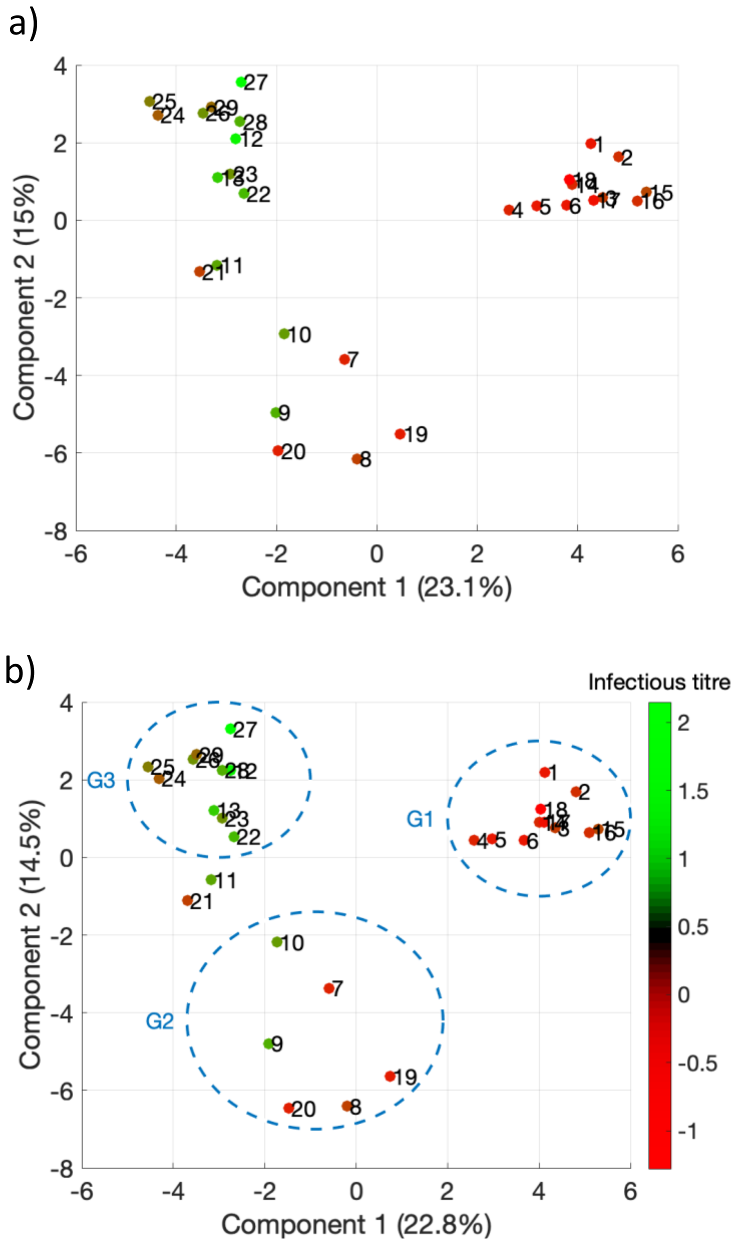


Figure 5-2: PCA (a) and sparse PCA (b) scores plots for principal components 1 and 2.

The scores on the first two components are very similar for both models and there are three main clusters present, circled in blue. The plot markers were coloured from red to green, representing low to high infectious titre, respectively. Interestingly, the three clusters present on components 1 and 2 show clear differences in the infectious titre, which indicates that the process conditions represented by the clusters are correlated to the infectious titre. In particular, the batches circled in group 1 all score high on component 1 and show a low infectious titre, while the opposite is true for group 3. Group 2 batches show mixed levels of infectious titre, scoring in between group 1 and 3 on component 1 and scoring low on component 2, which appears to be less strongly correlated to infectious titre. Observing correlations to the product CQAs is not the goal of this exercise, however, correlations that are identified are informative about process behaviour and provide useful information for the predictive modelling which follows.

To better understand the process conditions that are resulting in the clustering of the batches in Figure 5-2, it is necessary to investigate the loadings for components 1 and 2. The sparse PCA model will be investigated here because of the advantage it offers through the relative simplicity of the sparse loadings vector. To observe the main variables contributing to the scores, the variables with nonzero loadings were plotted on a parallel coordinates diagram and the batches were coloured by the group that they were assigned to in Figure 5-2. Figure 5-3 and Figure 5-4 below show parallel coordinates plots for component 1 variables with positive and negative loadings, respectively. The reader is referred to chapter 3, section 3.1.2 for an explanation of parallel coordinates plots.

In Figure 5-3, the main variables responsible for the separation of the groups on component 1 are material lots, the number of cells seeded during cell expansion, cell suspension volumes, rest time before transfection, parameters of the ion exchange chromatography column and the area of the filter used for sterile filtration.

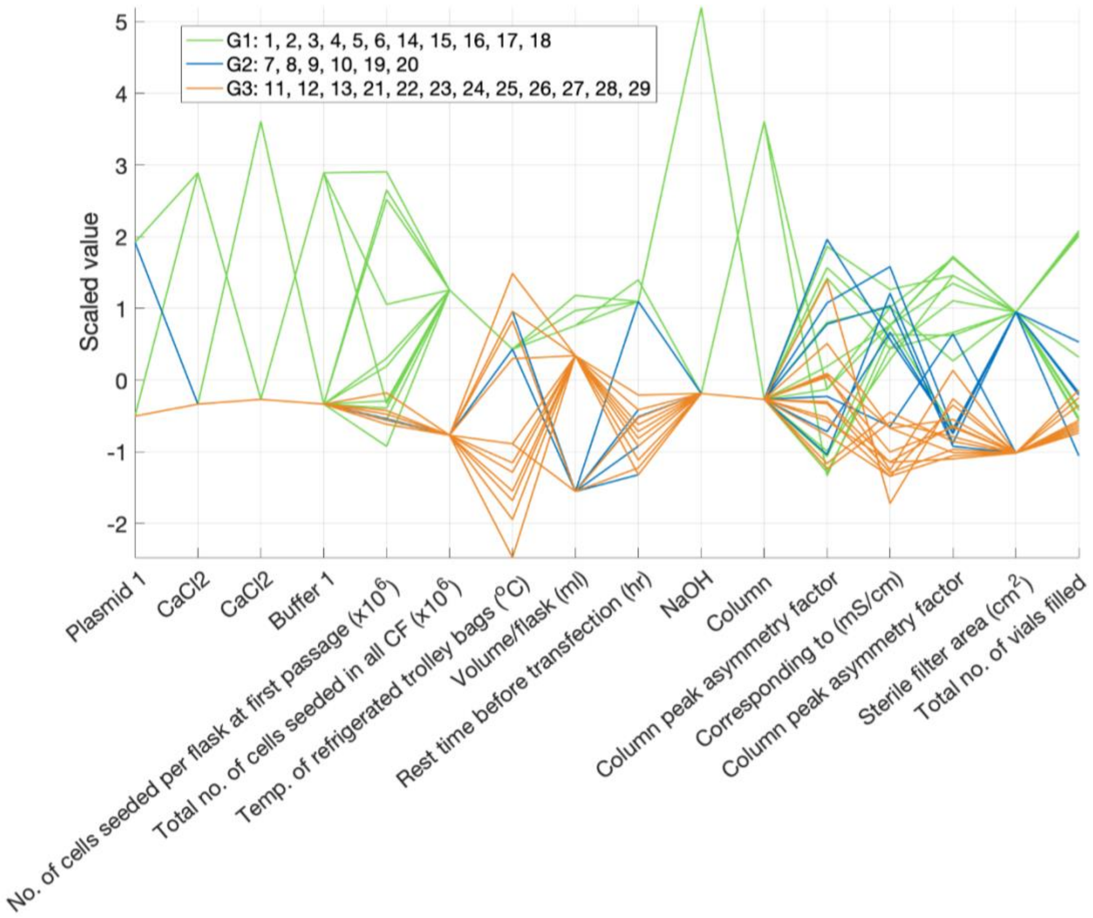


Figure 5-3: Parallel coordinates plot of variables with positive loadings on component 1 of the sparse PCA model. The batches have been grouped and coloured based on the clusters observed on components 1 and 2 in Figure **5-2**.

Variables with negative loadings on component 1 are shown in Figure 5-4. In Figure 5-4, the key variables contributing to the separation of the groups on component 1 are material lots, cell concentrations during expansion, volume of the cell suspension prior to transfection and volume of the viral vector supernatant in downstream processing. To highlight the differences between group 1 batches and group 3 batches, which have opposing scores on component 1, group 3 batches are characterized by:

- low total number of cells seeded into cell factories,
- high cell concentration at thaw and first passage,
- average to high total volume of cell suspension per cell factory,
- short rest time before transfection,
- low conductivity and column peak asymmetry factor in ion exchange chromatography,
- average to high viral supernatant volume in downstream processing.

Group 1 batches are characterized by the opposite of the above conditions. Interestingly, group 3 batches achieved high infectious titres while group 1 batches achieved low infectious titres; henceforth, it is likely that at least some of the aforementioned variables had an impact on the infectious titre. Group 2 batches scored close to group 3 on component 1 but differed on a few parameters, such as slightly lower cell concentrations in expansion, higher conductivity in ion exchange chromatography and larger sterile filter area. In addition to the metric variables that were mentioned, there were a number of different materials lots used between the three groups, including $CaCl_2$, buffer 1 and viral vector plasmids.

The above observations are not true for all batches in all instances, as different combinations of the aforementioned variables are influencing the scores. Exact score contributions for each batch and variable can be calculated; however, in this instance it is sufficient to have an overview of the key variables influencing the scores without investigating the scores on a batch by batch basis.

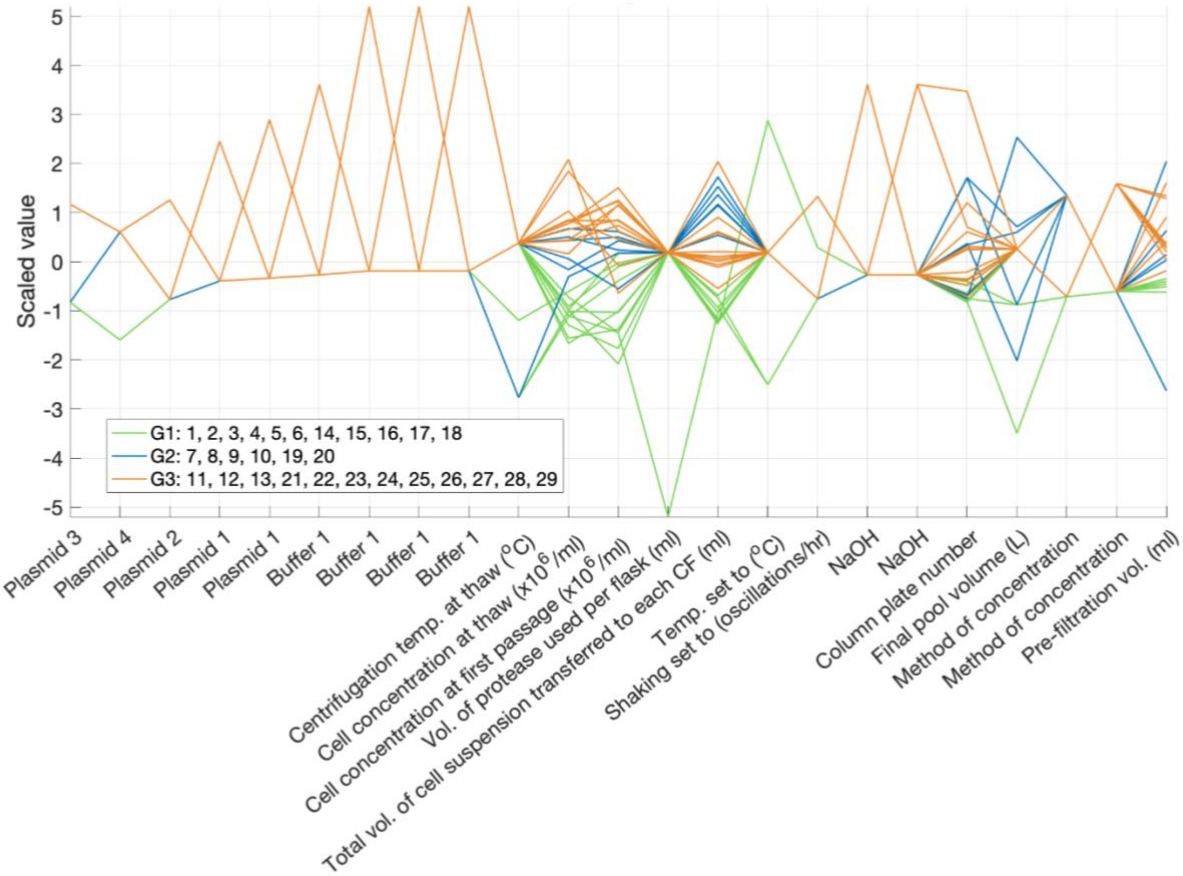


Figure 5-4: Parallel coordinates plot of variables with negative loadings on component 1. The batches have been grouped and coloured based on the clusters observed on components 1 and 2 in Figure 5-2.

Figure 5-5 and Figure 5-6 show parallel coordinates plots for component 2 variables with positive and negative loadings, respectively. Component 2 separated group 2 from groups 1 and 3. Group 2 batches were characterised by the following:

- a low total number of cells during cell expansion,

- high total volume of cell suspension transferred to each cell factory,
- lower cell concentrations than group 3,
- low shaking rate during endonuclease treatment,
- short rest time before transfection in most cases,
- low column peak asymmetry factor in size exclusion chromatography,
- higher than average refrigeration temperature during clarification,
- use of different raw material lots including transfection plasmids, $CaCl_2$, buffer 1 and NaOH buffers.

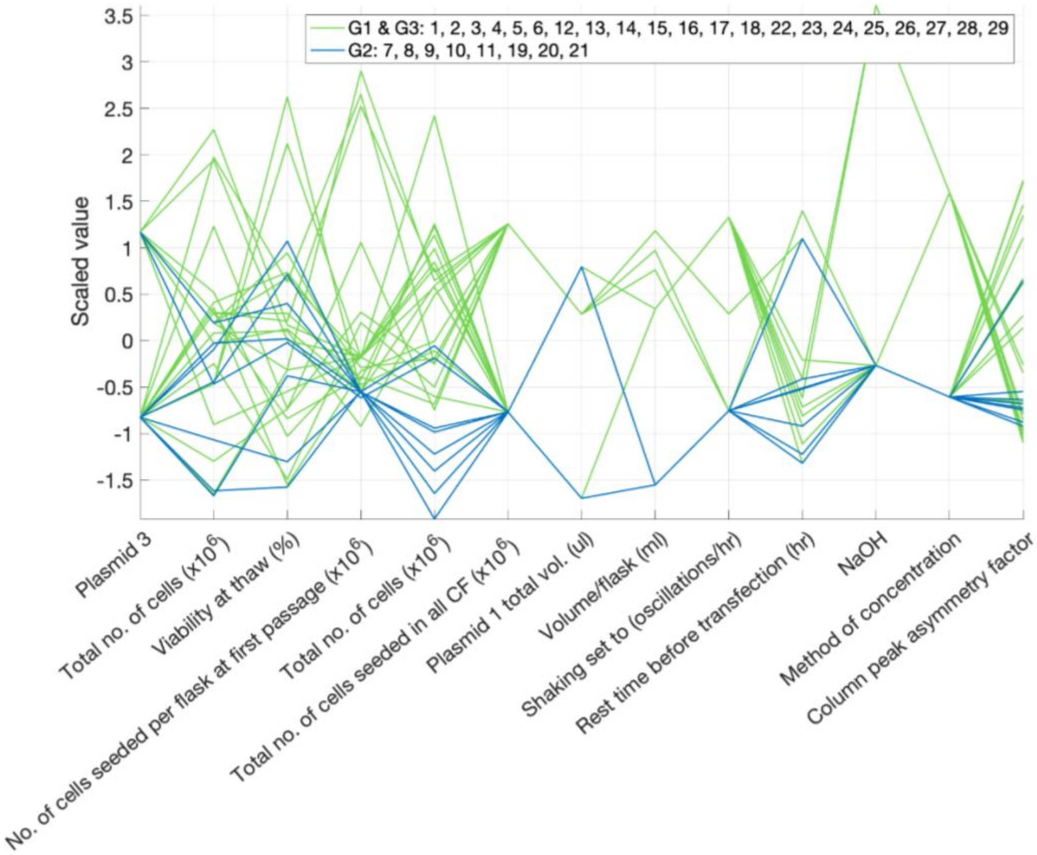


Figure 5-5: Parallel coordinates plot of variables with positive loadings on component 2. Batches in group 1 and 3 in Figure 5 2 score similarly on PC 2 and have therefore been coloured the same (green), whereas group 2 batches scored differently and are coloured in blue. This identifies variables causing differences in scores on PC 2.

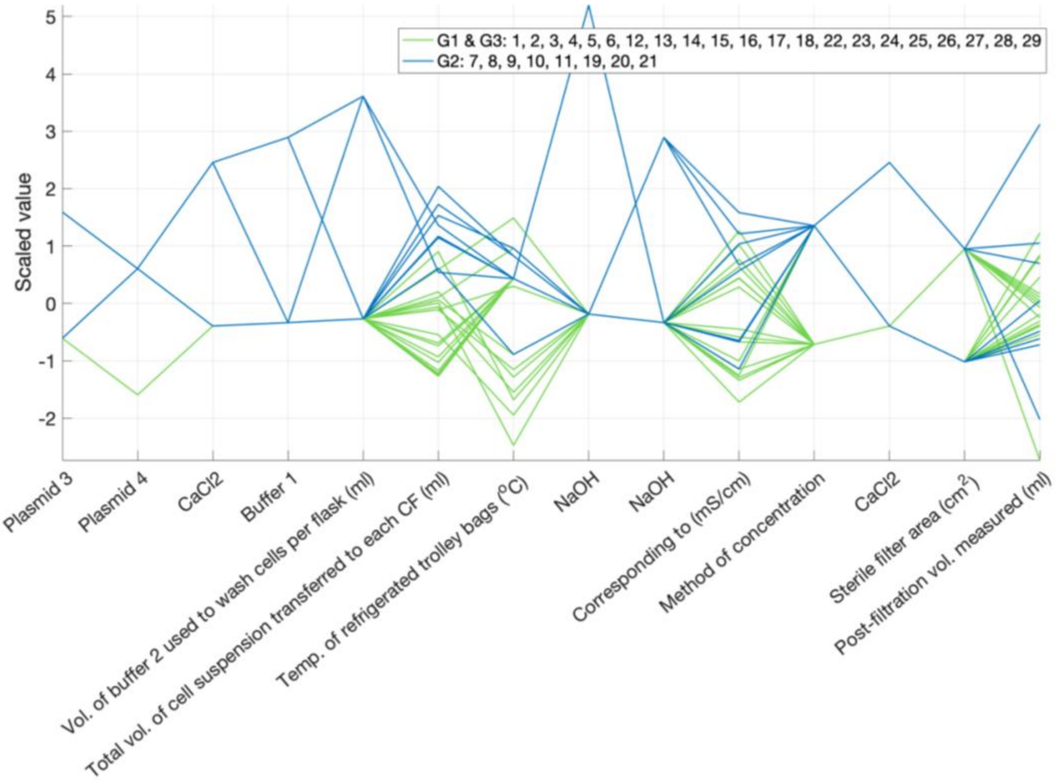


Figure 5-6: Parallel coordinates plot of variables with negative loadings on component 2. Batches in group 1 and 3 in Figure 5 2 score similarly on PC 2 and have therefore been coloured the same (green), whereas group 2 batches scored differently and are coloured in blue. This identifies variables causing differences in scores on PC 2.

5.3.1.2 Components 3 and 4

Figure 5-7 below shows the scores for components 3 and 4 for PCA (Figure 5-7a) and sparse PCA (Figure 5-7b). Comparing sparse PCA and standard PCA, component 3 results in very similar separation of the batches with 7, 10 and 11 scoring low and 19, 20, 24 and 25 scoring high. Similarly, component 4 pulls out batch 21 in both cases, however component 4 provides less separation of the remaining batches in the sparse PCA model, due to the fewer variables contributing to the scores.

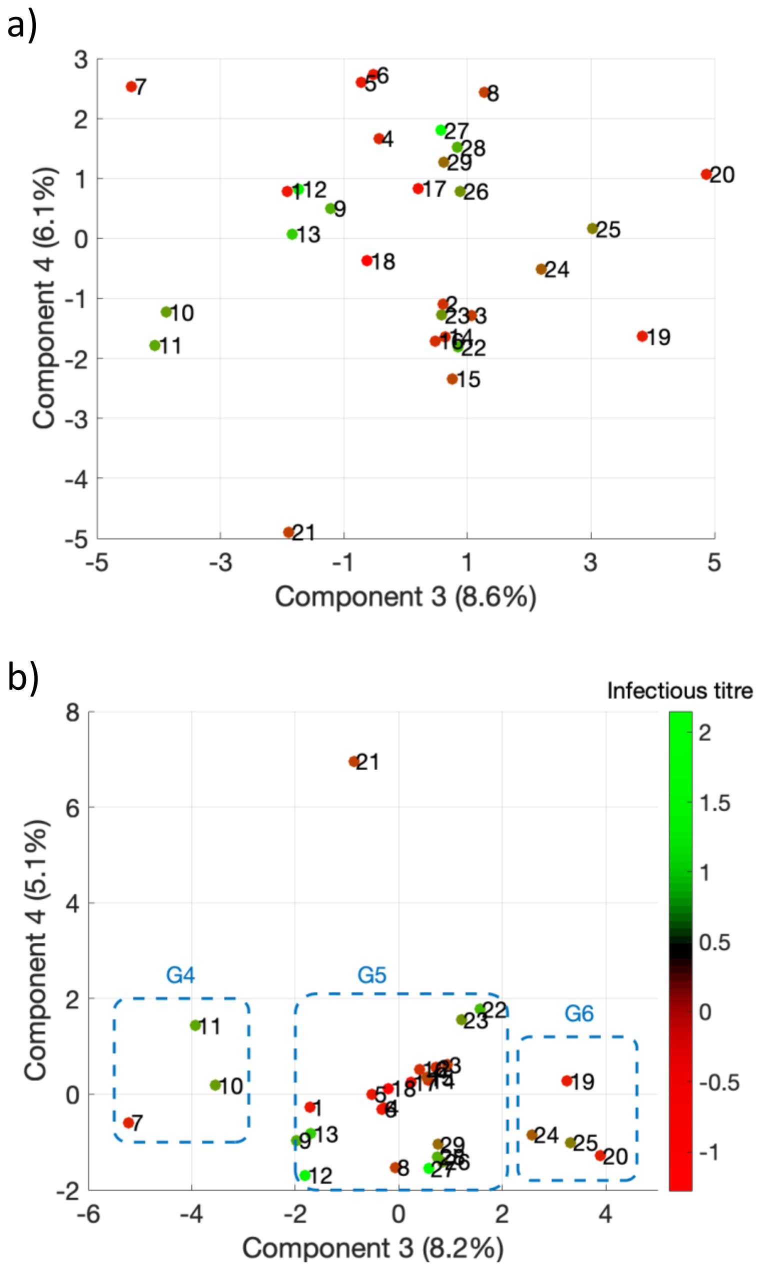


Figure 5-7: PCA (a) and sparse PCA (b) scores plots for components 3 and 4.

To examine the variables causing the separation on component 3, the variables were split into three groups as shown on Figure 5-7, and these groups were plotted on parallel coordinates diagrams in Figure 5-8 and Figure 5-9 below.

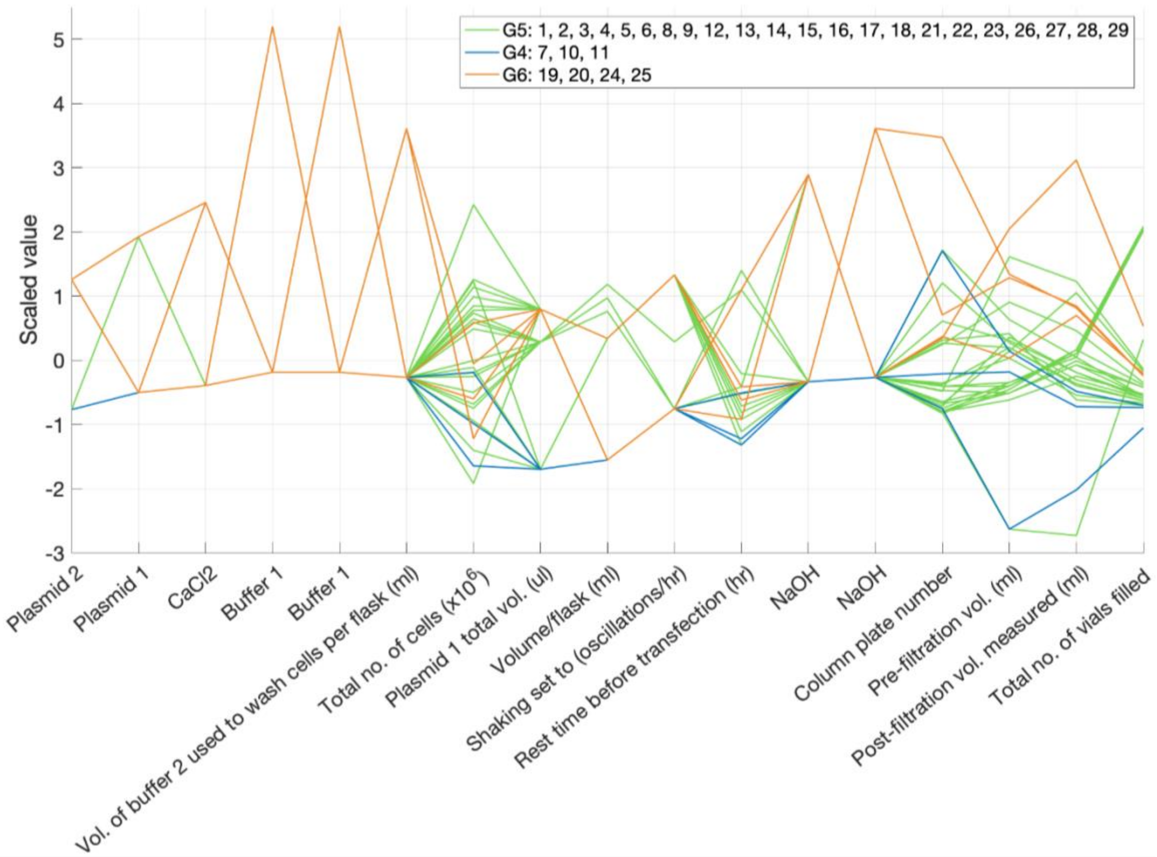


Figure 5-8: Parallel coordinates plot of variables with positive loadings on component 3. The batches have been coloured according to groups 4, 5 and 6 in Figure **5-7** to highlight the variables responsible for differences in the group scores.

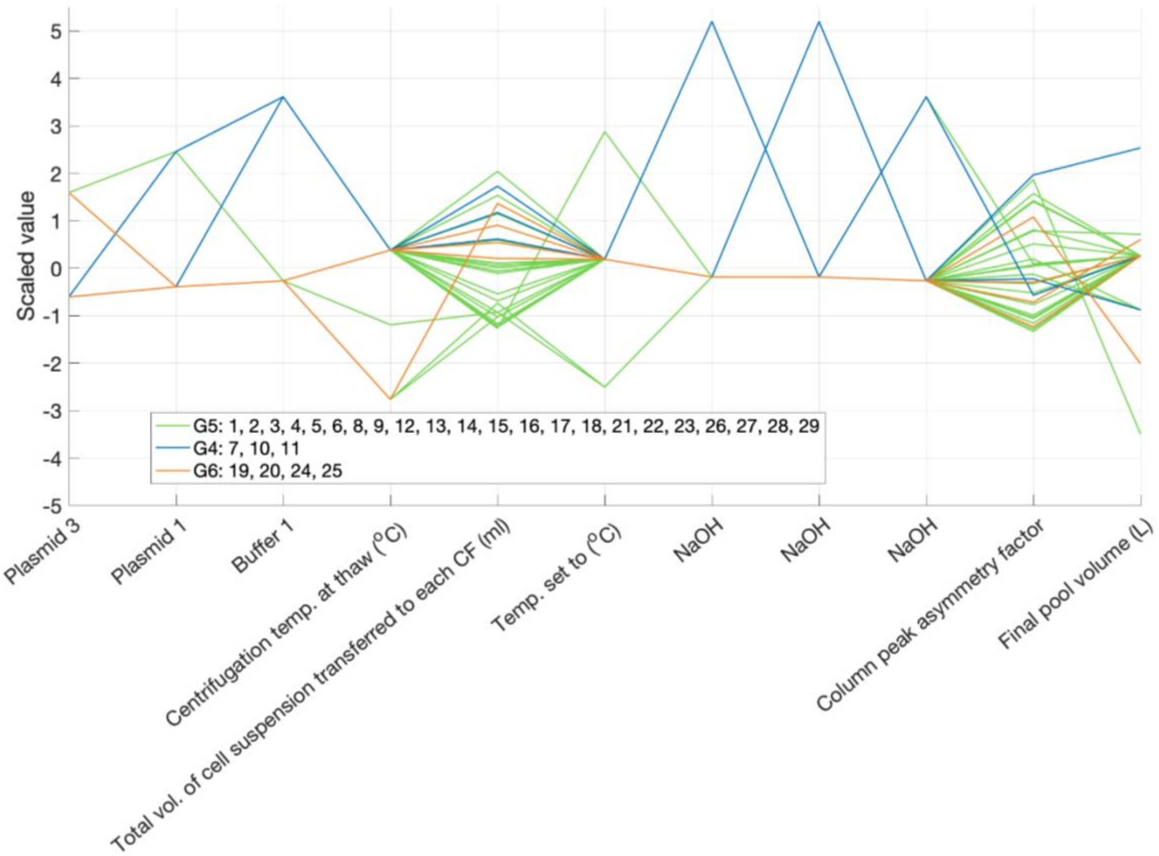


Figure 5-9 Parallel coordinates plot of variables with negative loadings on component 3. The batches have been coloured according to groups 4, 5 and 6 in Figure **5-7** to highlight the variables responsible for differences in the group scores.

Group 4 (G4) and group 6 (G6) on Figure 5-7, represent different process conditions, in several cases the process conditions are at opposite ends of the range, hence the opposing scores. To better understand the process conditions leading to the separation on component 3, group 4 is characterized by:

- average to low total number of cells during expansion,
- low total plasmid 1 volume,
- low oscillation rate during endonuclease treatment,
- short rest time before transfection,
- low volume of viral vector supernatant,
- use of different material lots including plasmids, buffer 1 and NaOH.

Group 6 (G6) scores high on component 3 largely due to the use of different material lots, high therapeutic plasmid volume and a large volume of viral vector supernatant in downstream processing. In general, component 3 has less clearly defined clusters than components 1 and 2, which is indicative of unstructured process variability that occurs batch-to-batch, whereas many of the features observed on components 1 and 2 were deliberate changes to process conditions.

Component 4 separated batch 21 from the rest, the parallel coordinates plot for component 4 are in the appendix chapter 9, section 9.1. Batch 21 was separated on component 4 due to the use of different raw material lots including plasmids and buffers, the total number of cells during expansion was low, the total volume of cell suspension transferred to cell factories was high and the plasmid 1 volume was high. Interestingly batch 21 was close to group 3 on components 1 and 2, which featured high infectious titres, yet batch 21 had a comparatively low titre. Component 4 variables may have contributed to the difference in infectious titre, which will be investigated in the next section on predictive modelling.

Components 5 and 6 capture around 5% of the variance each, and subsequent components capture less. Much of the variability in the later components is the within process variability that is unstructured and not introduced deliberately. The variables on the later components are a mixture of variables from cell expansion, including volumes of cell suspension, cell concentration, cell numbers and volumes reagents, and variables from downstream processing. The scores for component 5 and 6 are available in the appendix, chapter 9, section 9.1, along with a parallel coordinates plots showing the key contributing variables.

Six components were sufficient to explain around 61% of the variance in the data and components 1 to 4 gave a good overview of the main features in the data, attributed to systematic variation of process conditions and within process variability. The amount of variance explained by the components of the PCA or sparse PCA model is influenced by the features of the dataset, but also by several factors that practitioners tend to vary in their approach. These include the choice of categorical variable representation, the type of scaling applied and the extent to which redundant variables or highly correlated variables are removed in pre-processing. The interpretation of results can be the same as long as these factors are taken into consideration. Another choice for this analysis was the inclusion or exclusion of categorical variables in the PCA model and thus the following section briefly compares features of the sparse PCA model with and without categorical variables.

5.3.1.3 Comparison of sparse PCA with and without categorical variables

Figure 5-10 shows a comparison of the variance explained by the first 10 components of the sparse PCA model when metric variables are used exclusively versus when categorical variables are also included. One feature that stands out is the reduced variance explained by each comopnent of the model when categorical variables are included. This is caused by the categoric variable representation whereby, for a categoric variable with n categories, n-1 dummy variables are required to represent this single categoric variable. This partioning of the variance of a single categoirc variable into multiple dummy variables leads to PCA models requiring a greater number of components to capture the variance in the data.

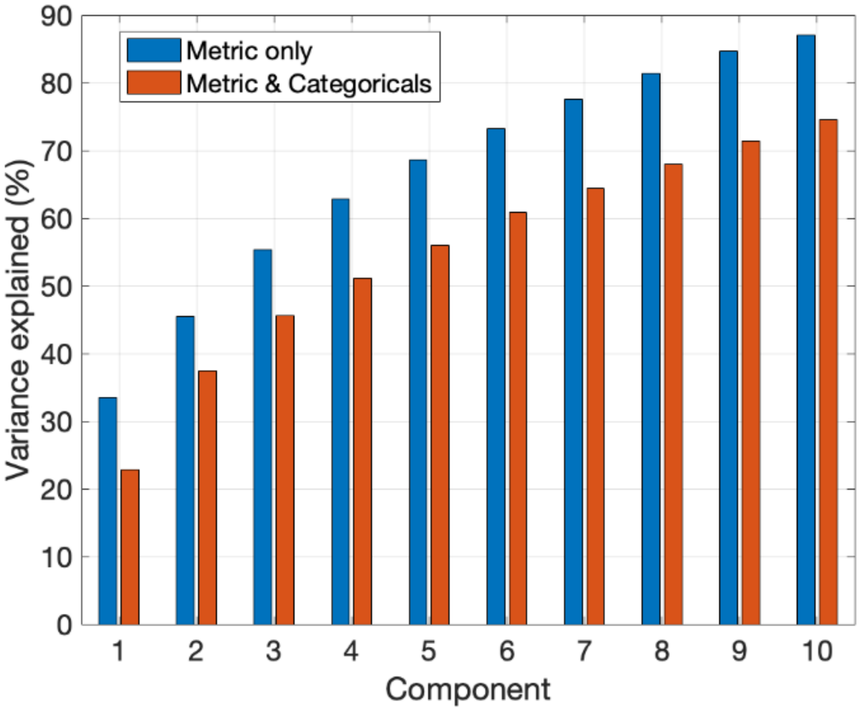


Figure 5-10: Comparison of sparse PCA with and without categorical variables: bar chart of cumulative variance explained for components 1 to 10.

Figure 5-11 shows the scores on the first two components of the sparse PCA models, with (Figure 5-11a) and without (Figure 5-11b) categorical variables included. The negative and positive loading variables happen to be inversed due to a rotational effect; however, it is still feasible to compare the two set of loadings and it can be seen that the clusters formed are very similar. In both plots there are three main clusters present and the batches cluster in the same groups with minor shifts in scores for some batches. Two reasons for this observation are as follows: some of the changes to the material lots used are correlated to changes in the metric variables, hence they influence the scores in the same direction. Another reason is that the metric variables are more influential on the score contributions due to their greater number. The latter may also be influenced by the choice of categoric variable representation and scaling.

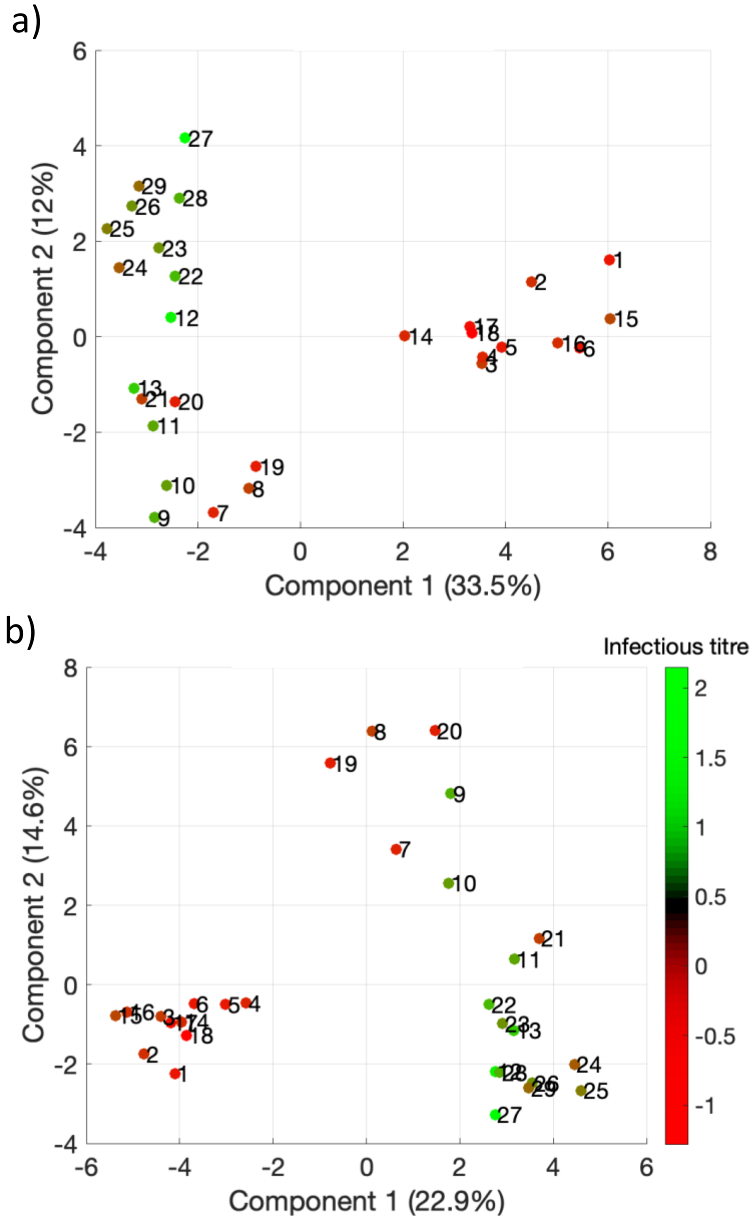


Figure 5-11: Comparison of sparse PCA scores for components 1 and 2 with (a) and without (b) categorical variables included.

The inclusion or exclusion of the categorical variables is a choice of the practitioner, and it depends on what aspects of the data the researcher is interested in. In many cases, it is worthwhile comparing the results with and without categoric variables to see what different features are observed. In this chapter, the data with categoric variables, which were mostly material lots, were explored because it helped to identify materials that were sources of variability. The next section explores the relationship between the sources of variability that were identified here with key process outputs, such as the infectious titre of the viral vector product.

5.3.2 Predictive modelling of lentiviral vector infectious titre

In this section, the results from modelling the infectious tire of the lentiviral vector product are presented. Initially a PLS model was developed using the metric variables exclusively. The methodology for the development of the PLS model was explained in section 5.2. Figure 5-12 shows the fit of the optimised PLS model that was initially developed for the infectious titre when the model is fitted to the entire training dataset. The model uses as 18 predictor variables selected by the VIP selection method (see chapter 2, section 3.2.2) and 4 latent variables determined through repeated K-fold CV. Specifically, Figure 5-12a shows the predicted infectious titre versus the measured infectious titre and Figure 5-12b shows the residuals against the fitted values.

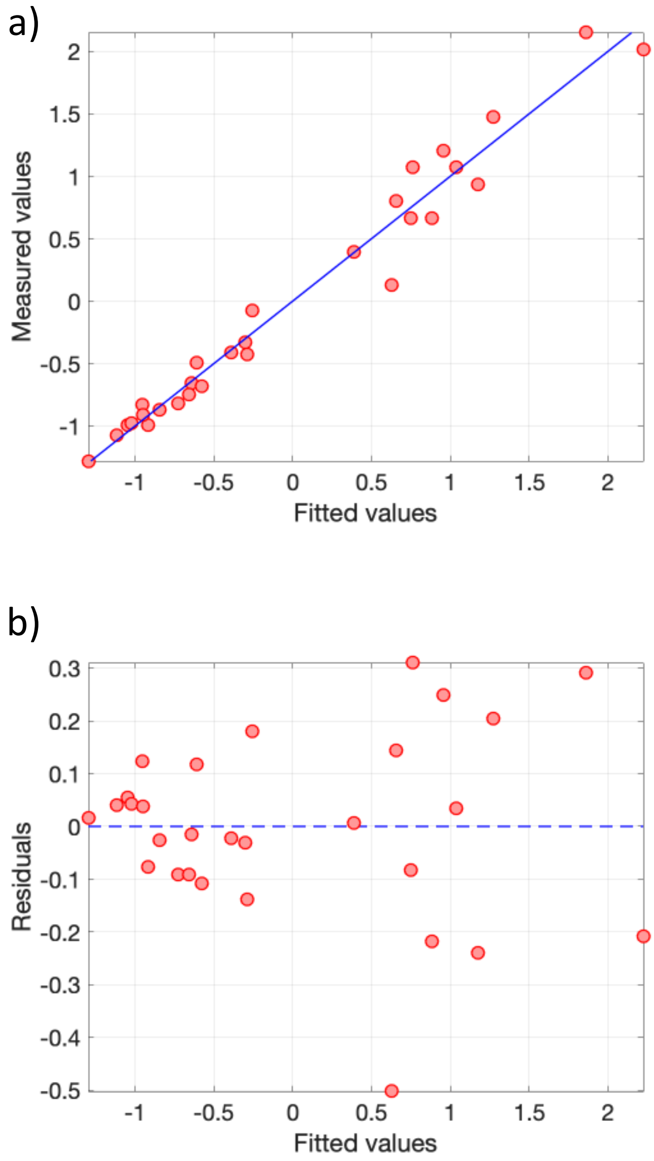


Figure 5-12: PLS model for the infectious titre using metric variables only, without transformations. The model uses 18 variables selected by the VIP selection method, with 4 latent variables determined from repeated 5-fold CV. The graphs show measured versus fitted infectious titre (a) and the residuals versus fitted values (b), for training data fit.

The adjusted R^2 value for the training fit is high at 0.94% and the model appears to fit the data well with the exception that there is noticeable heteroscedasticity in the residual plot. Heteroscedasticity can arise due to measurement error that is a function of the value being measured i.e. it becomes larger or smaller depending upon where the measured value is within the range of the test or instrument (Hair, 2014). Additionally, it can be indicative of extra factors or nonlinear effects, which are not being captured by the model. Model parameters and performance measures for the overall model fit and cross validation are detailed in Table ***5-3***.

The next step in model development was to investigate the potential for nonlinear terms to improve the model fit. To do this, the basic PLS model that was identified above was used as a starting point and the variables that were selected for this model were used to create new variables which were nonlinear transformations of the originals, including quadratic terms and moderator effects (see chapter 3, section 3.2.3). The new set of variables were reduced through the same VIP variable selection procedure. Figure 5-13 shows the fit of the optimised PLS model with the inclusion of nonlinear transformations.

Visually in Figure 5-13, the nonlinear transformations improved the model fit and resulted in more normally distributed residuals. Furthermore, the cross-validation performance metrics, CV R^2 and CV MAE, were improved from 0.74 and 0.34 for the basic model to 0.88 and 0.23 for the model with nonlinear transformations, respectively.

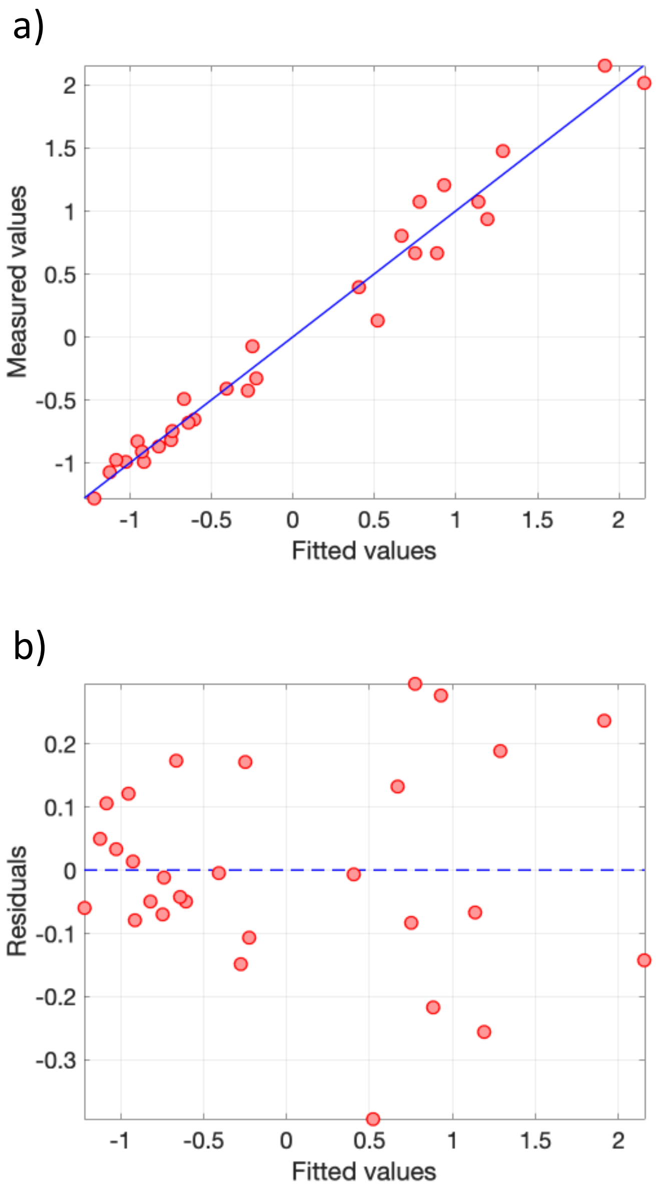


Figure 5-13: PLS model for the infectious titre using metric variables only and nonlinear transformations of the original variables. The model uses 29 variables selected by the VIP selection method, with 5 latent variables determined from repeated 5-fold CV. The graphs show measured versus fitted infectious titre (a) and the residuals versus fitted values (b), for training data fit.

Statistics and performance measures for both the base model and the model with transformations are provided below in Table *5-3*. These results indicate that there are nonlinear effects present in the manufacturing variables, which are correlated to the infectious titre.

Table 5-3: PLS models for prediction of infectious titre in adherent viral vector production: parameters and performance measures for the PLS models with and without transformations.

	Model with no transformations	Model with transformations
R^2 adj.	0.85	0.96
F-ratio	34.5	53.5
No. of latent variables	4	5
Residual degrees of freedom (RDOF)	24	23
CV R^2	0.74	0.88
CV MAE	0.340	0.229
No of variables	18	29

To shed light on the influential process parameters in the optimised PLS model with transformations, Table *5-4* and Table *5-5* below provide details of the variables used and their standardised regression coefficients. Table *5-4* shows variables without interactions, ordered according to the sequence of the unit operations they belong to. Table *5-5* shows interaction terms that were identified, ordered according to the sequence of the unit operation of the first variable that is stated. Bootstrap estimates of the standardised regression coefficients are given in the tables (Table *5-4* and Table *5-5*) along with bootstrap estimates of the 90% confidence interval. A full list of the variables and coefficients for the basic PLS model is given in the appendix, chapter 9, section 9.2. To aid discussion and to provide perspective on the relative importance of each variable in the model, the effect size has been categorised as small, medium or large based on magnitude of the standardised regression coefficient:

- Large effect size – beta > 66th percentile
- Medium effect size – 33rd percentile < beta < 66th percentile
- Small effect size – beta < 33rd percentile

Table 5-4: PLS model for prediction of infectious titre: this table contains a list of variables in the PLS model with nonlinear transformations, no interaction terms are presented in this table. The mean beta coefficient from bootstrap testing is presented along with its 95% and 5% confidence intervals.

Variable	Variable position	Mean beta	Upper 95% confidence interval (CI)	Lower 5% CI
Cell concentration (cells/ml) at thaw	Cell thaw	0.076	0.132	0.022
Viability (%) at thaw	Cell thaw	-0.080	-0.047	-0.117
Viability (%) at thaw squared	Cell thaw	-0.074	-0.043	-0.111
Total number of cells	Seeding cell factories	0.085	0.122	0.038
Total number of cells squared	Seeding cell factories	0.134	0.208	0.063
Total volume of plasmid 1 (µl)	Transfection	-0.066	-0.031	-0.101
Total volume of plasmid 1 (µl) squared	Transfection	-0.066	-0.031	-0.101
Temperature set to (ºC)	Endonuclease treatment	-0.053	-0.013	-0.095
Temperature set to (ºC) squared	Endonuclease treatment	-0.050	-0.015	-0.089
A %B	Ion exchange chromatography	-0.085	-0.049	-0.119
A %B squared	Ion exchange chromatography	-0.083	-0.043	-0.124
Δ conductivity squared	Ion exchange chromatography	-0.030	0.031	-0.074
Sterile filter area (cm^2)	Sterile filtration	-0.079	-0.049	-0.108
Pre-filtration volume (ml)	Sterile filtration	0.049	0.091	0.000
Sterile filter area (cm^2) squared	Sterile filtration	-0.079	-0.049	-0.108
Pre-filtration volume (ml) squared	Sterile filtration	0.047	0.101	-0.011

Starting with cell thaw, it was found that a higher cell concentration was correlated to higher infectious titre, however, the higher the viability of the cells at thaw, the lower the infectious titre. The magnitude of these effects was categorised as medium. It is not immediately clear why the viability of the cells at thaw would negatively correlate to the viral titre. This is something that warrants further investigation into the possible explanations. The higher the initial cell concentration, the easier it is to grow the cells to the desired number necessary for the transfection procedure. It was observed in feature extraction, that the cell concentration at thaw is correlated positively with the cell concentration in the cell factories and the total number of cells present. Henceforth, the impact of these parameters on the infectious titre is likely to be closely linked.

Supporting this, the total number of cells used to seed the cell factories during expansion was correlated positively to the infectious titre with a large effect size. It is known that under normal conditions, the larger the number of cells present at transfection; the more viral vector product can be produced (Petiot et al., 2015; Shen and Kamen, 2012). However, there is an exception to this, as it has been reported in the literature that there is a cell density limit, that when exceeded leads to lower yield of viral vector product (Petiot et al., 2015). This ‘cell density effect’ has been observed to change depending on factors such as the type of cells used and the type of virus (Petiot et al., 2015). In this case, the data indicates that no cell density effect is occurring and a positive correlation between cell count and infectious titre is observed. From an operational perspective, this indicates that the higher cell count should be targeted to achieve consistently higher titres.

The total volume of plasmid 1 used during transfection was negatively correlated to the infectious titre with medium effect size. In a study by Bauler *et al.* (2019), lentiviral vectors were produced by transient transfection of SJ293TS cells - a cell line derived from HEK293T cells that is adapted for growth in suspension culture. In their experiments, two different therapeutic plasmid doses were tested, 0.55 and 1.1 µg of DNA per 10^6 cells, and it was observed that the highest titres were produced when the volume of therapeutic plasmid was low. They also observed that for the lower plasmid DNA concentration, the infectious titre was significantly impacted by the cell seeding density. In order to better to understand the effect of the plasmid 1, a transformed variable was added to the model, which was equal to the total mass of plasmid 1 divided by the cell concentration after seeding the cell factories. This variable had a small positive beta coefficient of 0.124. The results indicate that the plasmid mass per cell was correlated positively to the infectious titre with small effect size, yet the volume of plasmid 1 was negatively correlated to the infectious titre with medium effect size. The results may indicate that the plasmid solution contains a chemical that impacts the process, the magnitude of this effect warrants further investigation. Other variables highly correlated to plasmid 1 volume could help explain this observation, however none were found within the dataset.

During endonuclease treatment to break down DNA contaminants, the temperature setpoint was negatively correlated to the infectious titre with medium effect size. The viral vectors are known to be sensitive to temperature (Higashikawa and Chang, 2001), the data indicates that higher temperatures in endonuclease treatment may be detrimental to the infectious titre. The temperature range observed in the data is small, between 4 and 6 ^{o}C, however these small deviations that were permitted are being pulled out as influential. The known temperature sensitivity of the viral vectors and their increased stability at lower temperatures reported in the literature, both support the findings of the model (Higashikawa and Chang, 2001). The appropriate action after validating this finding would be to tighten the temperature control of the refrigeration unit and to operate at the lower temperature setting. The control of this parameter is already tight so this may be difficult to improve, however, this parameter was also found to be important in the models developed in chapter 6 where some additional considerations are discussed.

Ion exchange chromatography appears to be an important step in downstream processing with two variables negatively correlated to infectious titre, namely the A %B setting and the change in conductivity. Ion exchange chromatography is used to separate the negatively charged virus particles from positively charged proteins. It is known that the choice of solvents used for flushing the column is highly influential on column performance and the infectious titre (Rout-Pitt et al., 2018). A %B is the ratio of buffers used for column elution, which is controlled based on an optimised elution profile. The significant regression coefficient of this parameter in the PLS model may indicate that control of the elution process could be further improved. It may also be the case that variation in the process upstream of the ion exchange unit operation is contributing to variability that Is observed in ion exchange column parameters. In any case, the A %B parameter and the change in conductivity are an early indication of process performance, as they are correlated to the infectious titre. From an operational perspective, perhaps this early indication could be used to take informed actions to improve the recovery of viral vectors from the column.

In the penultimate unit operation, the sterile filter area was found to have a critical influence on the infectious titre. Specifically, sterile filter area was correlated negatively to the infectious titre, indicating that a larger area resulted in greater loss of infectious virions. Sterile filtration has been reported in the literature to be a critical process step where the sterility of the product is enhanced at the expense of a loss infectious particles (Merten et al., 2010). It is known that the choice of sterile filter and its mode of operation impacts the viral vector recovery percentage (Merten et al., 2016). The data here indicates that the smaller filter is the preferential filter for sterile filtration, so long as there are no issues with the sterility of the product. Additionally, the pre-filtration volume was correlated positively with the infectious titre. A larger volume is likely to be correlated to a greater number of virions present, furthermore there may be an effect where a larger volume of media helps to flush the virions through the filter, improving the recovery percentage.

Table 5-5: PLS model for prediction of infectious titre: this table contains a list of the interaction terms that were identified for the PLS model with nonlinear transformations. The mean beta coefficient from bootstrap testing is presented along with its 95% and 5% confidence intervals.

Variable	**Process position**	**Mean beta**	**Upper 95% CI**	**Lower 5% CI**
(Viability (%) at thaw) **x** (Volume of plasmid 1)	Cell thaw	-0.107	-0.079	-0.138
(Volume of buffer 2 used to wash cells per flask (mL)) **x** (Volume of plasmid 1)	First passage	-0.046	-0.001	-0.106
(Total number of cells) **x** (Conc1)	Seeding cell factories	0.150	0.214	0.075
(Total number of cells) **x** (Volume of plasmid 1)	Seeding cell factories	0.076	0.109	0.036
(Plasmid 1 conc. (µg/µl)) **x** (Conc1)	Transfection	0.075	0.126	0.027
(Temperature set to (ºC)) **x** (Volume of therapeutic plasmid)	Benzonase treatment	-0.078	-0.045	-0.114
(A %B) **x** (Volume of plasmid 1)	Ion exchange chromatography	-0.101	-0.067	-0.133
(Final pool volume (L)) **x** (Volume of plasmid 1)	Ion exchange chromatography	-0.102	-0.020	-0.200
(Sterile filter area (cm^2)) **x** (Conc1)	Sterile filtration	-0.068	-0.004	-0.119
(Pre-filtration volume (ml)) **x** **(**Conc1)	Sterile filtration	0.072	0.115	0.035
(Sterile filter area (cm^2)) **x** (Conc2)	Sterile filtration	-0.071	-0.027	-0.114
(Sterile filter area (cm^2)) **x** (Volume of plasmid 1)	Sterile filtration	-0.077	-0.047	-0.106
(Pre-filtration volume (ml)) **x** (Volume of plasmid 1)	Sterile filtration	0.021	0.059	-0.024

Several interaction effects were identified, and these are shown in Table 5-5. The interactions identified were of the form $\beta \boldsymbol{X}_1 \boldsymbol{X}_2$. The interpretation of the beta coefficient in this instance changes, as explained in chapter 3, section 3.2.3. The coefficients for the interaction effects indicate the change in the effect of $\boldsymbol{X}_1$ on the response variable as $\boldsymbol{X}_2$ changes. The sign of the regression coefficient still indicates the direction of the effect on the infectious titre. Several of the process variables were found to have a moderator effect with the plasmid 1 volume, meaning that their relationship with the infectious titre changes depending on the plasmid 1 volume. The cell viability at thaw interacted with the plasmid 1 volume and resulted in a negative correlation with a large effect size. The total volume of buffer 2 used to wash cells interacted with the plasmid 1 volume, featuring a negative correlation to infectious titre with medium effect size. The interaction between concentration of the plasmid 1 and the cell concentration and total number of cells during expansion, resulted in positive correlations to the infectious titre with large effect sizes.

Other interactions present include the total number of cells during seeding of cell factories which interacted with the concentration of cells and correlated positively to the infectious titre. This interaction suggests that a greater number of cells at higher density is beneficial for the infectious titre, which supports an earlier observation that no ‘cell density effect’ is occurring. The sterile filter area and the concentration of the cells during cell expansion also interacted. Higher total number of cells and cell concentration were found to be correlated positively with the infectious titre with a medium to large effect size. High cell concentration in cell expansion should therefore correspond to high concentration of viral vectors in the supernatant in downstream processing. This may explain the interaction effect between the cell concentration and the filter area because the higher the concentration of infectious particles entering the sterile filter, the higher the losses will be during the filtration step, as there is less media to flush the virus particles through. This concludes the relationships that were found between the process parameters and the infectious titre for adherent viral vector manufacturing.

5.4 Conclusions

This chapter explored data from the production of two lentiviral vector products for use in CGT treatments. The data was explored first of all by carrying out feature extraction, using a programming approach to sparse PCA that was developed in chapter 4, followed by predictive modelling with PLS regression to relate process parameters to product quality attributes. The sparse PCA algorithm performed well and extracted the same key features as the standard PCA model on the first 3 components i.e., the principal component scores and clustering was very similar, and the interpretation of the loadings was consistent with standard PCA. It achieved this while using fewer variables in the model, making the loadings vector and the variable contributions easier to evaluate.

The feature extraction work revealed that there were three main clusters present in the data, two of which contained high infectious titre batches and one of which exclusively containing batches with a low infectious titre. This was an early indication that the process parameters contained variation that was correlated to the infectious titre of the virus product. Many of the variables contributing to the separation in scores on components 1 and 2 were from expansion of the producer cells. Variation was observed as early as cell thaw in the total number of cells and the cell concentration, differences in the total number of cells and their concentration were in many cases translated through the whole of the upstream process, i.e. cell concentration at thaw correlated to cell concentration at first passage and seeding of the cell factories. There was also variation in the amounts of buffers and reagents used, such as the amount of buffer 2 used to wash the cells when passaging. In the transfection procedure, there were different plasmid 1 concentrations and volumes used.

In downstream processing, the most significant areas of variability were in the ion exchange chromatography, endonuclease treatment and sterile filtration. Several process parameters pulled out from each of these unit operations in the first three principal components, which explained 45% of the variation in the data. Most of the categorical variables that were highlighted on components 1 to 3 were material lots for viral plasmids and buffers including buffer 1, NaOH, buffer 2 and $CaCl_2$. Principal components 4 to 6 explained less significant portions of the variation in the data and separated single batches or small clusters of batches from the rest, due to comparatively small changes to process conditions.

In predictive modelling of the infectious titre, numerous correlations were identified and the PLS model that was developed demonstrated good predictive capability with cross validation R^2 and MAE of 0.88 and 0.229, respectively. The risk of identifying chance correlations was minimized through the cross validation and variable selection procedure and chance correlations within the model were tested by permutating the ***y*** data and refitting the model repeatedly. The degree of chance correlation was deemed to be low with an average training R^2 of 0.21 for the permutated data, see chapter 3, section 3.3.5 for more details on the ***y***-randomization.

The process parameters with the most significant positive correlations to the infectious titre were the total number of cells and their concentration in the cell factories, the concentration of the plasmid 1 solution combined with the concentration of the cells, and the pre-filtration volume. The most influential process parameters with negative correlations to the infectious titre were the sterile filter area, the volume of plasmid 1 used in transfection, the A%B buffer ratio in ion exchange chromatography, viability of cells at thaw and the temperature of refrigeration during endonuclease treatment.

The model indicates that to maximise the infectious titre, the smaller sterile filter area should be used, smaller of the two plasmid 1 doses, and the cell concentration and total number of cells in the cell factories should be controlled to achieve the upper levels that were observed in this dataset. Going beyond the limits that were observed within the dataset may lead to further improvements, however this would need to be explored with experiments. The A% B buffer ratio in ion exchange chromatography was negatively correlated to the infectious titre. If this parameter is already under control based on an optimal elution profile, then it may be worthwhile exploring the parameters upstream that are responsible for the variation in the buffer ratio. The temperature of the refrigerators during endonuclease treatment should be controlled more tightly at the lower temperature setting.

To summarise the sponsors' (GSK) feedback on the modelling results, the sterile filter area was already identified as likely having a significant impact on infectious titres and the models confirmed this. The positive correlation between cell concentration and infectious titre is also a known and explainable effect and consistently achieving high cell concentrations is the goal of the cell expansion phase of the process. The models highlighted that the ion exchange column settings, plasmid volumes and temperature of refrigeration were important parameters with significant effects on infectious titre. With the exception of the refrigeration temperature, these were seen as potentially valid effects, although the findings would need to be verified through further experimentation. The temperature of refrigeration is already under tight control and studies on virus stability at GSK indicated that the temperatures observed in refrigeration should not significantly impact the product CQAs.

Chapter 6

Analytics for cell drug product manufacturing

This chapter focuses on the manufacturing of the cell drug product. In this process, the patient’s cells are transfected with the viral vectors *ex vivo* and prepared as an injectable solution for grafting back to the patient. Many of the challenges faced are in common with viral vector production, such as high process variability and a lack of advanced process knowledge. MVDA was used with a similar approach to chapter 4 in order to derive useful insights into the cell drug product manufacturing process.

6.1 Introduction

The cell drug product (CDP) manufacturing process is where the patient's extracted cells are processed *ex vivo* and transduced with the viral vectors containing the therapeutic gene of interest, before being prepared as an injectable solution for transplantation back to the patient. The steps involved in CDP manufacturing were discussed in chapter 2, section 2.1.5. The overall process includes procedures which are necessary for the extraction of the patient's cells and for their transplantation back to patient; however, the CDP manufacturing dataset only contains information from the manufacturing process that transforms the extracted cells to an injectable cell drug product.

Many of the challenges of CDP manufacturing are shared in common with viral vector manufacturing, including high levels of variability in the complex raw materials, numerous unit operations involving a high degree of manual handling and a lack of advanced expertise; particularly with respect to knowledge of relationships between process parameters and CQAs of the CDP. In this chapter, data from CDP manufacturing was explored using the sparse PCA algorithm that was developed in chapter 4. This feature extraction work identified the most important sources of variability in the manufacturing process and showed how it is reflected in the process parameters. It also provided information on the correlation between process variables, which provided insights into the translation of variance through the process, in some cases across multiple unit operations. The scores of the sparse PCA model showed how the key features of variance in the manufacturing data impacted different batches.

Following on from feature extraction, the CDP manufacturing data was combined with the data from LV manufacturing to form a single dataset, which was investigated with predictive modelling techniques. The task was to model the relationship between process parameters and CQAs of the CDP, from both LV manufacturing and CDP manufacturing. The LV manufacturing data was included because the CQAs that were modelled are quality measurements of the final CDP, which was transduced with the LVs prior to the measurements being taken. Henceforth, it was of interest to explore any potential links between LV manufacturing and the CQAs. The overall aim was to identify critical process parameters and quantify the relationships between process parameters and the CQAs. The enhanced process understanding that this may provide would be beneficial for the development of CGT manufacturing processes.

6.1.1 Data

The data that was available for analysis of the cell drug product manufacturing process was from the production of treatment 1, the same treatment 1 that was analysed in Chapter 5. 23 batches of historical CDP manufacturing data were available for the analysis, each of which involved the processing of cells from a different patient. The CDP manufacturing data consisted of 225 manufacturing process variables i.e. input parameters and 13 dependent variables, which were recorded on the certificate of analysis for the final cell drug product. Of the 225 process variables, 55 were categorical and 170 were metric. The LV manufacturing data used in this chapter were from the 13 batches of LVs for treatment 1 that were studied in chapter 5, see chapter 5, section 5.1.1 for more details on this dataset. Of the 13 batches in total, 8 were used for transduction of the cells from 23 patients, henceforth 8 of the LV batches were included in the analysis. More specifically, the 8 batches were used to generate 23 samples, some of which were duplicates, to be matched up with the 23 samples of CDP manufacturing data. The LV manufacturing data was aligned with the CDP manufacturing data using the traceable batch numbers that were supplied with both datasets.

6.2 Method

6.2.1 Pre-processing of process data

The following section details the pre-processing steps that were taken to obtain datasets for the PCA, sparse PCA and PLS analyses that were conducted in chapter 6.

6.2.1.1 Processing of cross-sectional process data

The pre-processing of the cross-sectional process data was the same as described in chapter 5, section 5.2.1, the reader is referred to 5.2.1 for details.

6.2.1.2 Processing of product quality data

The product quality data that was analysed in chapter 6 included the number of viral vector copies integrated per cell (LV copy number) and the percentage of CD34+ cells in the cell drug product. The response variables required aligning to the correct input data and autoscaling was applied to the response variables, consistent with the scaling applied to the input variables.

6.2.2 PCA and sparse PCA

The sparse PCA algorithm that was developed in chapter 4 was applied to the cross-sectional process data to carry out feature extraction. Standard PCA was carried out to compare the variance explained by each principal component for the sparse and non-sparse PCA models. For details on the sparse PCA algorithm, the reader is referred to the nonlinear programming approach to sparse PCA, which was described in chapter 4, section 4.2.2. For more details on standard PCA, see chapter 3 section 3.1.1.

6.2.3 Development of PLS regression models

In this chapter, PLS models were developed to predict the LV copy number and CD34+ cells percentage in the cell drug product using the process parameters from LV manufacturing and cell drug product manufacturing. The following list details the steps taken to develop the PLS models:

1. Initially predictor variables were pre-selected by ruling out duplicate variables and variables of no relevance to the analysis. This reduced the original 225 process variables down to 166 potential predictor variables to be selected. The pre-selection was kept to a minimum to allow the data to reveal key correlations and the variable selection procedure to remove redundant variables of low predictive power.
2. The scaled predictor variables in the matrix $\boldsymbol{X}$, and the scaled response variable in the vector $\boldsymbol{y}$, from the training dataset, were then passed to a MATLAB script which carried out forward variable selection on random subsets of the data to generate numerous alternative models (more details of the forward variable selection methodology are provided in chapter 3, section 3.2.2).
3. The models identified in the previous step were then put through repeated 5-fold cross validation, with 2000 repeats, for models with n latent components, where n was varied from 1 to the maximum number of variables in the model.
4. The top performing models in cross validation were selected based on the mean cross validation R^2, the percentage of cross validation R^2 values less than 0.6 and the mean R^2 with $\boldsymbol{Y}$ permutated.
5. Using the bootstrap technique detailed in chapter 3, section 3.3.4, the stability and distribution of the model regression coefficients were evaluated with 2000 bootstrap samples. This provided confidence intervals for the standardised regression coefficients (beta coefficients).

6. Finally, the model was applied to the test set that was held-out of model development and the performance was evaluated.

6.2.3.1 PLS model performance evaluation

Model performance was evaluated through repeated K-fold cross validation (CV MAE and CV R^2), model fit to the whole training data (visual inspection, MAE and R^2) and bootstrapping (stability of regression coefficients).

6.3 Results and discussion

6.3.1 Feature extraction with sparse PCA results

In this section, the results from applying the sparse PCA algorithm to data from the CDP manufacturing process are presented. Figure 6-1 shows the cumulative variance explained for each variable added to the sparse PCA model. Results from standard PCA are shown in Figure 6-1 to provide reference points for the variance explained by non-sparse principal components. Figure 6-1 shows that 5 components explain over 60% of the variance in the data and subsequent components explain less than 5% each. Additionally, the variance captured decreases in a linear fashion from component 5 onwards. Based on this observation and applying the methodology discussed in chapter 3 section 3.1.1, 5 components were selected for analysis.

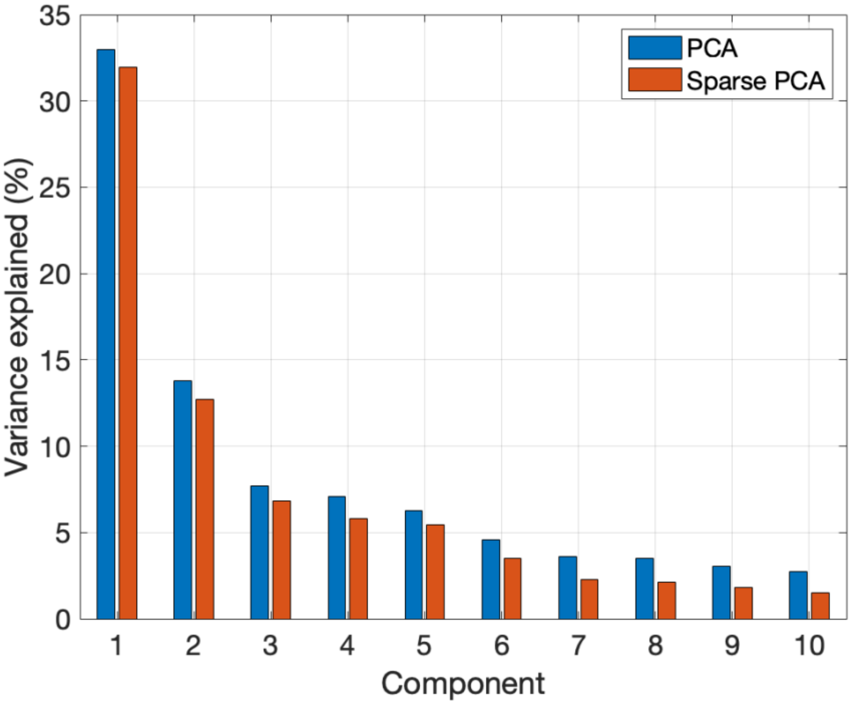


Figure 6-1: Variance explained by each component of the sparse PCA model compared to standard PCA, for the process variables in cell drug product manufacturing.

Figure 6-2 shows the scores for principal components 1 and 2, which explain 32% and 12.7% of the variance, respectively. Components 1 and 2 do not show well defined clusters. Instead the batches are spread out, indicating that the process conditions vary on a batch-to-batch basis and that the process variables vary over a continuous range as opposed to discrete levels. This is in contrast to principal components 1 and 2 for viral vector manufacturing in chapter 5, where there were three distinct groups characterized by a unique set of operating conditions.

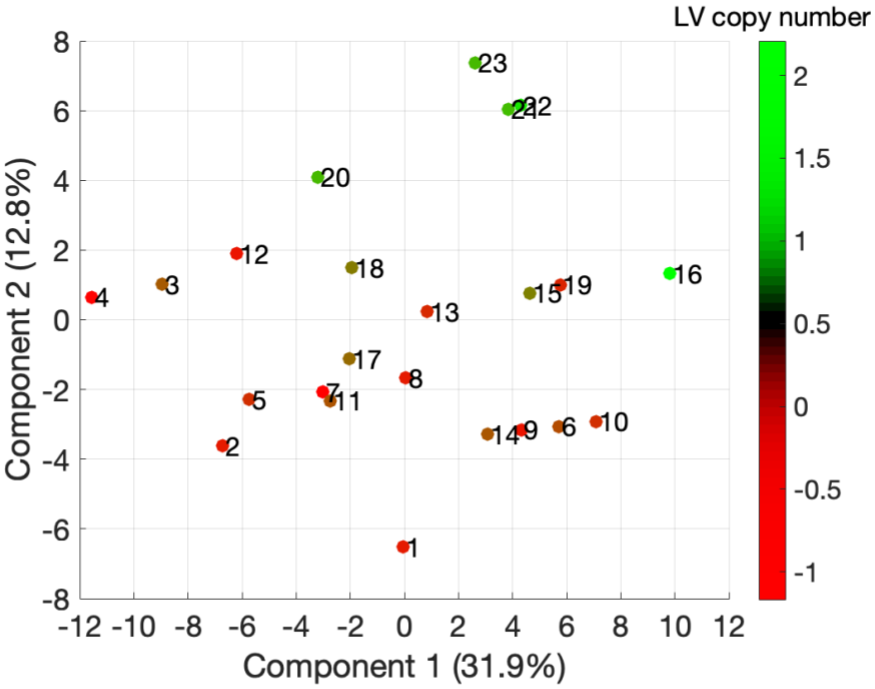


Figure 6-2: Scores for components 1 and 2 of the sparse PCA model for CDP manufacturing. The batches are coloured from red to green representing the scaled LV copy number; red indicates a low copy number while green indicates a high copy number.

Figure 6-3 shows a plot of the loadings for variables with nonzero loadings on component 1. Loading plots have been used instead of parallel coordinates because there are no distinct clusters of batches for which to interpret the process conditions. The loading plot allows correlations between variables to be observed, where loadings that are similar in value indicate positive correlation between them and loadings with opposite signs indicate that the variables are negatively correlated.

The significant variables on component 1 (Figure 6-3) come from unit operations across the whole process, including mononuclear cells (MNCs) separation, separation of the CD34+ cells, first and second transduction and preparation for infusion. Figure 6-3 shows a large group of variables that are positively correlated to a high degree (Pearson's R close to 1). This highlights a feature in the data where the number of

CD34+ cells for seeding is positively correlated with multiple volume related variables throughout the rest of the process. These volume related variables are mostly volumes of growth media and cytokines used in the resuspension of the cells after washing or medium changes, prior to and in between the two transduction steps. Furthermore, the number of CD34+ cells is positively correlated with several variables that reflect earlier stages in the process; these are the temperature of the centrifuge during MNC separation, the volume of MNCs recovered and the volumes of CD34 reagent and cell wash buffer used in separation of CD34+ cells. Additionally, the concentration of cells expressed as cell count before second transduction was correlated positively with the number of CD34+ cells that were seeded. Other variables present on component 1 with smaller loadings include the temperature and speed of centrifugation during several of the washing steps

.

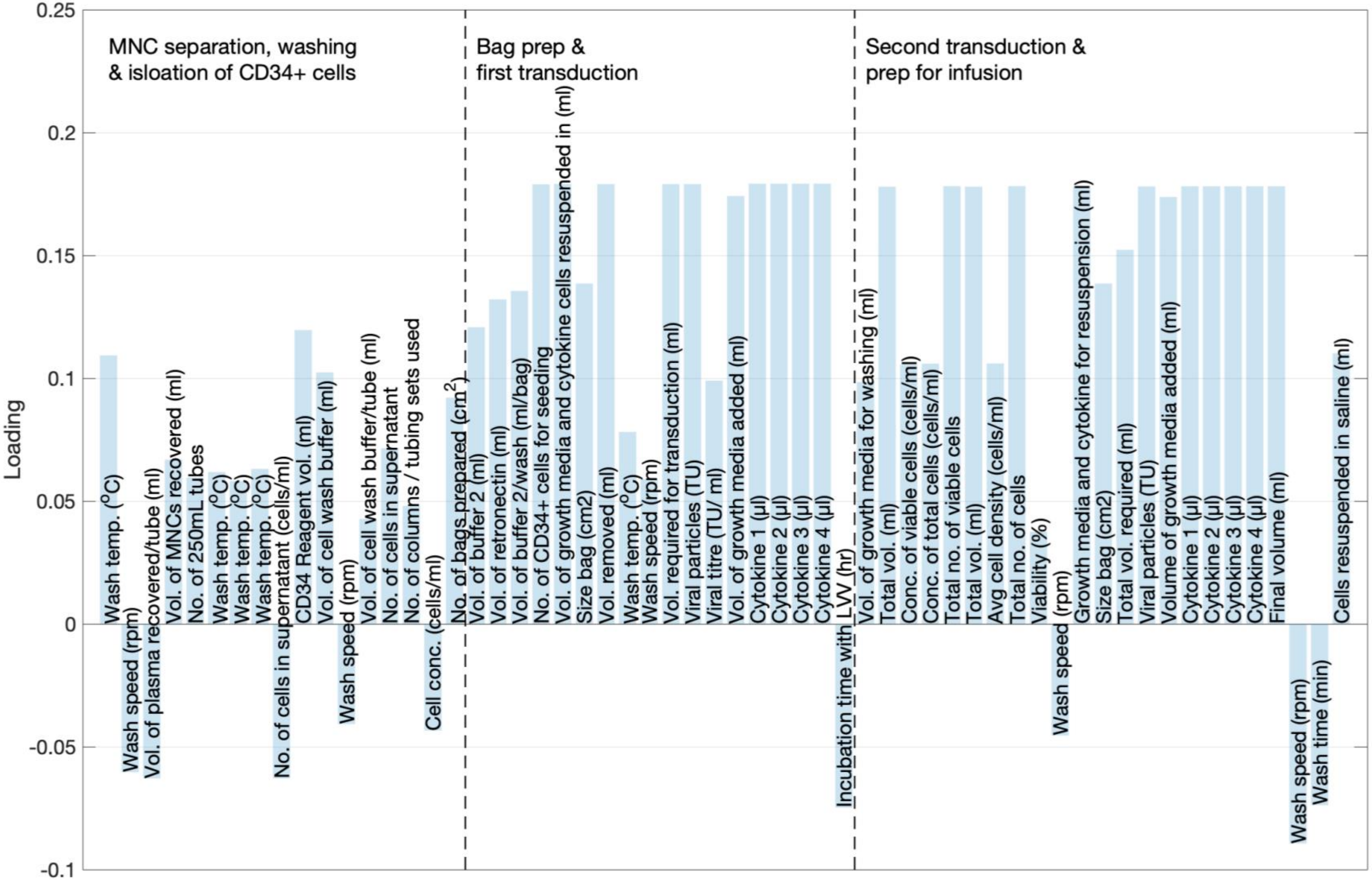


Figure 6-3: Bar plot of principal component 1 loadings showing each of the variables with nonzero loadings on component 1 in the sparse PCA model. The input variables are process variables from CDP manufacture.

Figure 6-4 shows the loadings for component 2 variables, which again cover a wide range of unit operations. Component 2 displays a degree of correlation with the LV copy number, as the high copy number batches in green scored higher than most other batches. The variables with the largest magnitude loadings on component 2 include the volumes of cell wash buffer used in MNCs separation and isolation of CD34+ cells, and the concentration of CD34+ cells after washing to remove antibodies introduced by the CD34 reagent. The loadings indicate that a larger volume of cell wash buffer in CD34+ separation corresponds to lower cell concentration after washing. The volume of LV supernatant and the viral titre also feature on component 2 and are negatively correlated to one another; these parameters are controlled to reach a desired multiplicity of infection and concentration of LV, so the correlation is dictated by the operating regime. Other prominent variables on component 2 include the speed and temperature of centrifugation during the numerous washing steps throughout the process. Figure 6-4 shows that the centrifugation speed and temperature are negatively correlated to one another.

Component 2 shows a degree of correlation with the LV copy number, as batches 20 to 23 all have high LV copy numbers and score high on component 2. It is likely that one or more of the variables on component 2 are correlated to the LV copy number. This was investigated further in the predictive modelling, which follows in section 6.3.2.

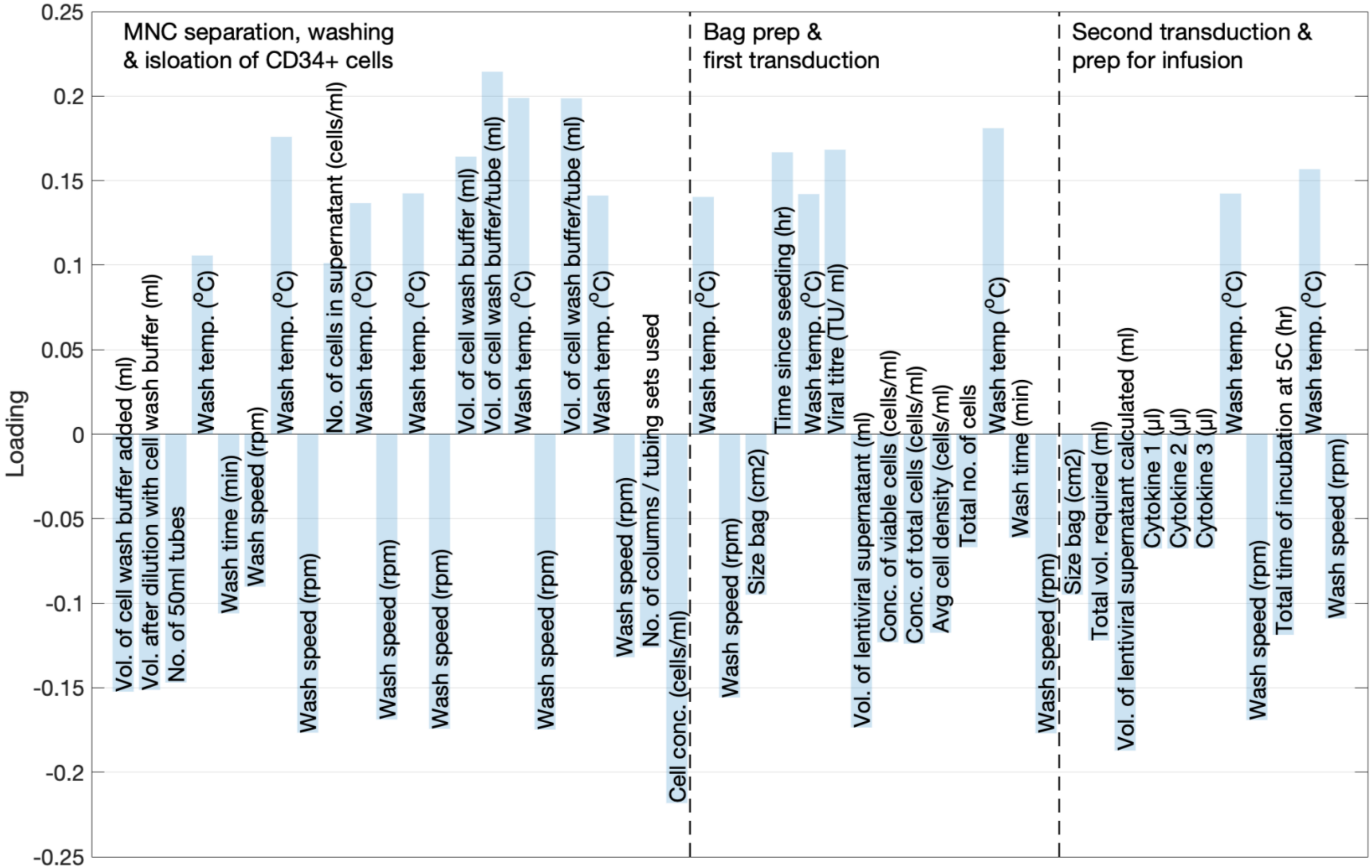


Figure 6-4: Bar plot of principal component 2 loadings showing each of the variables with nonzero loadings on component 2 in the sparse PCA model. The model inputs include process variables from CDP manufacture.

Figure 6-5 shows the scores for components 3 and 4, which explain around 7% and 6% of the variance respectively. Component 3 captures variance which spreads the batches out in a well-distributed manner, as with components 1 and 2. Component 4 on the other hand, singles out batch 7, which scores significantly lower than the rest of the batches.

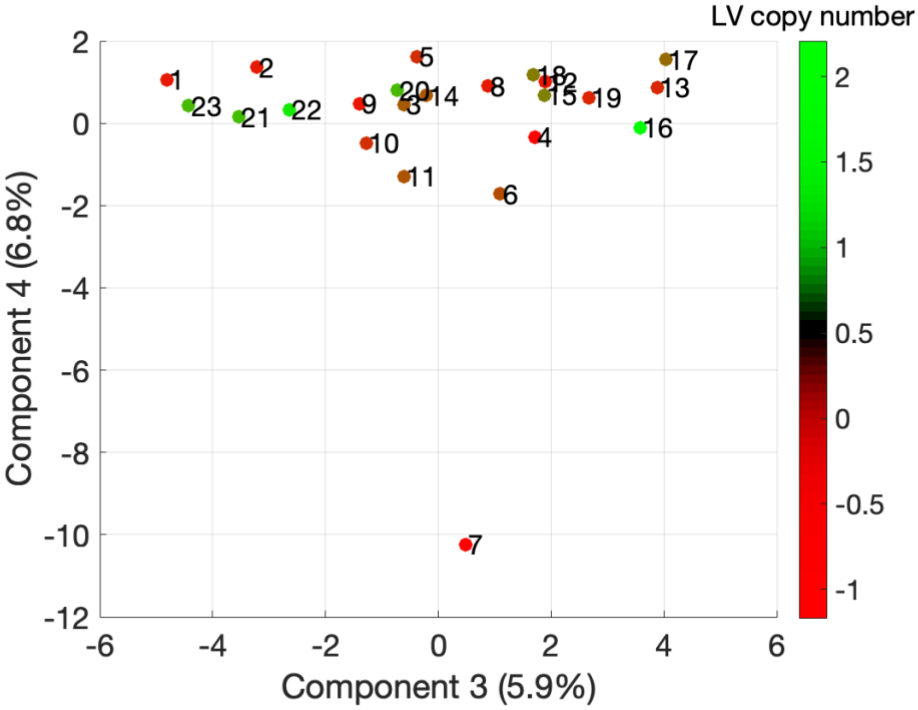


Figure 6-5: Scores for components 3 and 4 of the sparse PCA model, for the CDP manufacturing data. The batches are coloured from red to green representing the scaled LV copy number; red indicates a low copy number while green indicates a high copy number.

The scores for components 3 and 4 which explain around 7% and 6% of the variance, respectively, can be seen in Figure 6-6. Variables that stand out with the largest magnitude loadings on component 3 are the volume of MNCs recovered, the number of tubes used in the antibody washing steps, the CD34 reagent volume and the number of columns/tubing sets used for separation of CD34+ cells. The loadings indicate that the volume of MNCs recovered is positively correlated with the total number of MNCs in the supernatant after washing and the volume of CD34 reagent used for separation of CD34+ cells. Other variables present include the concentration of CD34+ cells once separated, the volumes of cell wash buffer used in the isolation of MNCs and CD34+ cells, the speed of centrifugation during washing steps, the volume of buffer 2 and retronectin used in preparation of bags for transduction, rest times after first and second transduction, viability of the cells after first transduction and the time of incubation at 5^{o}C after second transduction.

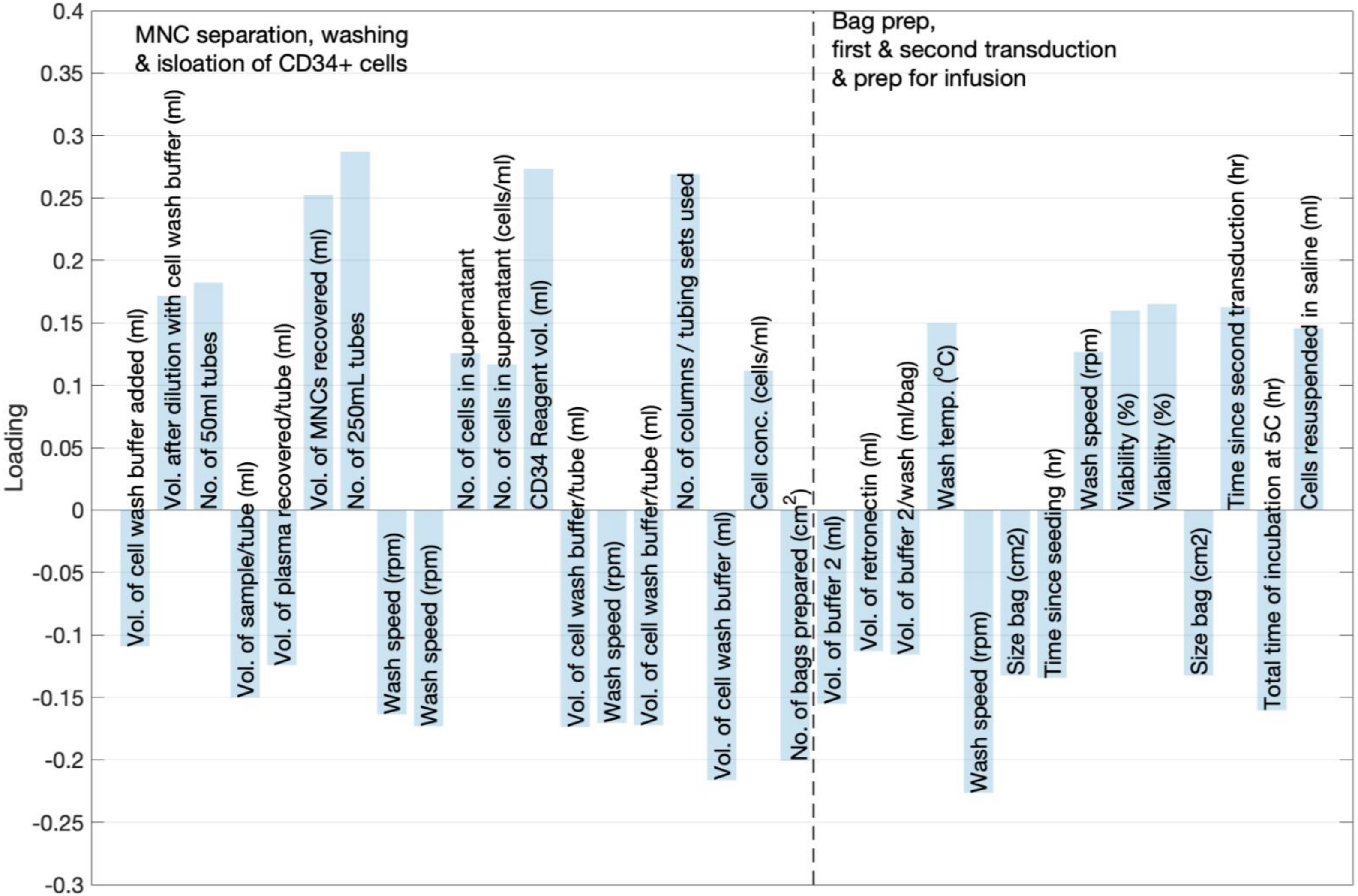


Figure 6-6: Bar plot of principal component 3 loadings showing each of the variables with nonzero loadings on component 3 in the sparse PCA model. The model inputs are process variable from CDP manufacture.

Component 4 (Figure 6-5) separated batch 7 from the rest of the batches; the characteristics of batch 7 which separate it from the rest of the batches are a low volume of buffer 2 used in preparation of bags for transduction, a large volume of cell wash buffer per tube in separation of CD34+ cells and long centrifuge durations at high temperature during several of the washing steps. Some of these effects are likely compounded as a greater volume of wash buffer will require more centrifugation. Visually, components 3 and 4 do not display a correlation with the LV copy number.

To summarize the sparse PCA results, the first three principal components captured approximately 62% of the variance in the data and the score plots revealed that the variation on these components occurred on a batch-to-batch basis. The fact that the variance is not concentrated in the first few principal components, rather it is split over many, indicates that there are many uncorrelated features in the CDP manufacturing data. This observation may be expected because the key input material at the start of CDP manufacturing is the patient's cells. Each of the batches in the dataset were concerned with the processing of cells from a different patient, henceforth the cells were a key source of batch-to-batch variability. The number of cells in the bone marrow sample and their physiological condition are known to vary significantly from patient to patient, depending on multiple genetic and environmental factors, including the cell extraction process (Stroncek et al., 2010). Component 1 captured variation in the volumes of MNCs recovered and the number of CD34+ cells that were seeded for transduction; these parameters are likely impacted by the number of cells present in the bone marrow sample and their physiological condition.

Components 1 to 3 showed that there was correlation between variables across the whole process for example the volume of MNCs recovered was correlated to the volume of growth media and cytokines that were used during both transductions. Many of the key variables were volume related variables including volume of cells, LV supernatant, buffers, cytokines and regents. In many cases these volume related variables varied on a continuous basis, as opposed to featuring a small number of discrete values. For example, for the volume of MNCs recovered, there were 23 unique values, whereas for centrifuge temperature there were two or three unique values between the 23 batches. These volume related variables explain why the scores are well distributed on components 1 to 3, with a lack of clearly defined clusters. In addition to the volume related variables, the numerous washing steps were a key source of variability, with the temperature, speed and duration of centrifugation appearing on the first three principal components.

The fact that the variance was distributed across numerous components meant that it was difficult to summarize key features of variance in the CDP manufacturing data because rather than having a few key features, it has many uncorrelated features. PCA was useful in identifying this characteristic of the dataset and it could also be very useful in characterising batches of interest. For example, if a particular batch behaved atypically in processing or resulted in unusual (good or bad) CQAs, PCA could be used to quickly characterise the unique features. In this work, effort was spent developing regression models to achieve the same goal – understanding the process conditions leading to good and bad CQAs.

6.3.2 Predictive modelling of the lentiviral vector copy number and the percentage of CD34+ cells

In this section, the results from modelling the Lentiviral vector copy number and the percentage of CD34+ cells are presented. A full description of the methodology that was used to develop the predictive models was outlined in section 6.2.3. The approach involved generating multiple models by using random subsets of the data to select variables in a forward variable selection procedure. This process led to the identification of close to 100 models for each of the output variables. The performance of each of the models was then evaluated using repeated K-fold cross validation and the presence of chance correlations in the model was tested by fitting the models to random permutations of the data.

Top performing models for the LV copy number and the CD34+ cells percentage can be seen in Table ***6-1*** and Table ***6-2***, respectively. The tables show the variables included in the model and their beta coefficients, alongside key performance measures including the cross validation R^2, the percentage of test R^2 that were lower than 0.6 and the average training R^2 when the Y data was randomly permutated. These tables (Table ***6-1*** and Table ***6-2***) are provided to show examples of high performing models that were identified for each output. In order to narrow down the models that were identified to a smaller subset of the best performing models, a number of performance criteria were specified. Table ***6-3*** shows the performance criteria that were specified for the models; different criteria were specified for each of the output variables based on differences in overall performance.

Table 6-1: Top multivariate regression models for the lentiviral vector copy number, identified by a forward variable selection procedure repeated on random subsets of the data. The table shows two alternative models and lists the variables with their standardised regression coefficients. On the right of the table, key performance metrics are displayed including the mean cross validation R^2, the percentage of test R^2 that were lower than 0.6 during repeated 10-fold cross validation and the mean training R^2 from repeated Y permutations.

Var 1	Var 2	Var 3	Var 4	Var 5	Var 6	Var 7	Var 8	Mean R^2_{test}	R^2_{test} < 0.6 (%)	Mean R^2_{Yperm}
Temp. of refrigerated trolley bags (ºC)	Vol. of MNCs recovered (ml)	Time since seeding (min)	Vol. of growth media added (ml)	Vol. of Ca/Mg solution added (ml)	Concentration of total cells (cells/ml)	Wash temp. (ºC)	None	0.92	0.8	0.35
-1.1	-0.84	-0.77	0.49	-0.7	0.36	0.19	N/A			
Temp. of refrigerated trolley bags (ºC)	Vol. of growth media added (ml)	Vol. of MNCs recovered (ml)	Viability (%)	Time since seeding (min)	Volume of Ca/Mg solution added (ml)	Average cells density (cells/ml)	Wash temp. (ºC)	0.93	1.1	0.4
-1.11	0.5	-0.82	-0.12	-0.72	-0.64	0.36	0.15			

Table 6-2: Top multivariate regression models for the CD34+ cells percentage, identified by a forward variable selection procedure repeated on random subsets of the data.

Var 1	Var 2	Var 3	Var 4	Var 5	Var 6	Mean R^2_{test}	R^2_{test} < 0.6 (%)	Mean R^2_{Yperm}
Vol. after dilution with cell wash buffer (ml)	Speed (rpm)	Conc. of cells	No. of ells in supernatant	Vol. of buffer 2 per wash (ml/bag)	Centrifuge time (min)	0.903	0.8	0.3
-0.91	-0.41	0.36	-0.3	-0.18	-0.13			
Vol. after dilution with cell wash buffer (ml)	Speed (rpm)	Conc. of cells (cells/ml)	No. of cells in supernatant	Vol. of buffer 2 per wash (ml/bag)	Wash temp. (ºC)	0.897	0.9	0.3
-0.92	-0.52	0.4	-0.25	-0.15	-0.1			

Table 6-3: Maximum and minimum performance criteria used to select the best performing models from the full list of models that were identified with the forward variable selection procedure.

	Minimum mean CV R^2	Maximum $R^2_{test} < 0.6$ (%)	Maximum mean R^2_{Yperm}
LV copy no.	0.6	10	0.4
CD34+ cells percentage	0.6	12	0.4

After filtering the models with the above performance criteria, the model parameters were evaluated by compiling tables of the variables that were featured in the high performing models. One way to assess variable importance is to compare the number of times each variable was selected for the model. Variables that were selected more frequently are likely to be of greater importance. Another method for evaluating the importance of the variables is to compare the beta coefficients and to observe how they change across different models. Both methods were utilised in this work.

Table ***6-4*** and Table ***6-5*** show the variables selected for the prediction of the LV copy number, with positive and negative regression coefficients, respectively. Similarly, Table ***6-6*** and Table ***6-7*** show the variables selected for the CD34+ cells percentage. These tables display all of the influential variables that were identified for the respective outputs. Displayed in the columns is the mean, maximum and minimum of the beta coefficient, across all of the models where the variable was selected. The sign on the mean beta coefficient indicates the direction of the relationship with the output variable and the magnitude indicates its importance relative to the other variables. The maximum and minimum beta coefficients across models are insightful because it is important to check that the sign on the regression coefficient does not change, as this would put the models in disagreement and decrease confidence in the relationship between the process parameter and the CQA. Additionally, the maximum and minimum beta coefficients across all folds of repeated K-fold cross validation are displayed in the tables 6.4 to 6.7. The minimum and maximum across folds is a check on the stability of the regression coefficient, which is more thorough than the maximum and minimum across models.

As with the maximum and minimum beta coefficient across models, the maximum and minimum across folds should both have the same sign. However, for some of the variables shown in tables 6.4 to 6.7, the sign switches across the folds. While this does reduce confidence in the regression coefficient for the variables concerned, they should not be completely disregarded, especially where the sign only changes by a small margin. Due to the low number of batches available for the analysis and conditions that are not repeated or repeated very few times, the partitioning of the data in cross validation can greatly influence the regression coefficients. This is one reason why some instability in the regression coefficients may be tolerated and may not be indicative of a false relationship.

Table 6-4: List of variables with negative regression coefficients in the CD34+ cells percentage regression models. The variables are listed in order of descending regression coefficient magnitude, meaning that the most important variables are at the top of the table. Alongside the variable name is the manufacturing process step during which the variables are recorded, the mean, maximum and minimum of the regression coefficient across all models, and the maximum and minimum across all folds of the repeated K-fold cross validation.

Variable	Process step	Mean across models	Max across models	Min across models	Max across folds	Min across folds
Final volume after dilution with cell wash buffer (ml)	MNCs separation	-0.94	-0.9	-1.03	-0.55	-1.25
Total volume of MNCs recovered (ml)	MNCs separation	-0.57	-0.53	-0.62	-0.27	-0.85
Speed (rpm)	Washing	-0.52	-0.37	-1.03	-0.12	-1.69
Number of cells in supernatant (cells/ml)	3rd wash	-0.41	-0.41	-0.41	-0.09	-0.68
Total time of incubation at 5C (h)	Prep for infusion	-0.37	-0.37	-0.37	-0.04	-0.96
Viability (%)	Release cell count	-0.3	-0.3	-0.3	-0.19	-0.42
Temperature (ºC)	1st transduction washing	-0.29	-0.29	-0.29	-0.16	-0.59
Cells remaining in supernatant	Recovery washing transduced cells	-0.26	-0.19	-0.32	0.14	-0.52
Number of cells in the supernatant	1st wash (antibody elimination)	-0.25	-0.22	-0.29	0.03	-0.67
Volume of buffer 2 per wash (ml/bag)	Washing bags and plates	-0.19	-0.15	-0.23	0.12	-0.47
Time since seeding	1st transduction washing	-0.17	-0.15	-0.2	0.11	-0.75
Temperature (ºC)	Recovery washing transduced cells	-0.16	-0.1	-0.23	0.08	-0.4

Table 6-5: List of variables with positive regression coefficients in the CD34+ cells percentage regression models. The variables are listed in order of descending regression coefficient magnitude, meaning that the most important variables are at the top of the table. Alongside the variable name is the manufacturing process step during which the variables are recorded, the mean, maximum and minimum of the regression coefficient across all models, and the maximum and minimum across all folds of the repeated K-fold cross validation.

Variable	Process step	Mean across models	Max across models	Min across models	Max across folds	Min across folds
Concentration of total cells (cells/ml)	Cell count (transduced)	0.43	0.53	0.34	0.77	-0.03
Concentration of viable cells (cells/ml)	Cell count (transduced)	0.39	0.43	0.35	0.59	0.08
Incubation time with LV	2nd transduction washing	0.39	0.39	0.39	0.54	0.22
Column plate number	'Vector process'	0.36	0.36	0.36	0.69	0.16
Temperature (ºC)	Prep for infusion	0.21	0.21	0.21	0.46	0.05
Method LV DEAE started at (h)	'Vector process'	0.18	0.25	0.11	0.45	0.01

Table 6-6: List of variables with negative regression coefficients in the LV copy number regression models. The variables are listed in order of descending regression coefficient magnitude, meaning that the most important variables are at the top of the table. Alongside the variable name is the manufacturing process step during which the variables are recorded, the mean, maximum and minimum of the regression coefficient across all models, and the maximum and minimum across all folds of the repeated K-fold cross validation.

Variable	Process step	Mean across models	Max across models	Min across models	Max across folds	Min across folds
Temperature of refrigerated trolley bags (ºC)	Vector process	-1.11	-1.1	-1.14	-0.79	-1.51
Total volume of MNCs recovered (ml)	MNCsseparation	-0.85	-0.82	-0.93	-0.56	-1.21
Time since seeding	1st transduction washing	-0.75	-0.72	-0.77	-0.42	-1.19
Volume of Ca/Mg solution added (ml)	'Vector process'	-0.66	-0.58	-0.7	-0.05	-0.98
Number of cells in the supernatant	1st wash (antibody elimination)	-0.16	-0.16	-0.16	0.1	-0.35
Viability (%)	Cell count (transduced)	-0.12	-0.12	-0.12	0.03	-0.28

Table 6-7: List of variables with positive regression coefficients in the LV copy number percentage regression models. The variables are listed in order of descending regression coefficient magnitude, meaning that the most important variables are at the top of the table. Alongside the variable name is the manufacturing process step during which the variables are recorded, the mean, maximum and minimum of the regression coefficient across all models, and the maximum and minimum across all folds of the repeated K-fold cross validation.

Variable	Process step	Mean across models	Max across models	Min across models	Max across folds	Min across folds
Volume of growth media added (ml)	1st transduction washing	0.49	0.5	0.47	0.72	0.31
Average cells density ($x10^6$/ml)	Cell count (transduced)	0.36	0.36	0.36	0.53	0.17
Concentration of total cells ($x10^6$ cells/ml)	Cell count (transduced)	0.34	0.36	0.26	0.6	0.01
Temperature (^oC)	1st wash (antibody elimination)	0.18	0.19	0.15	0.34	-0.03
Number of cells in supernatant ($x10^6$/ml)	3rd wash	0.1	0.15	0.04	0.5	-0.25

Tables 6.4 to 6.7 present information about the stability of the regression coefficients, which is important for understanding the reliability of the relationships that were identified. The variables are listed in descending order of regression coefficient magnitude to make it easy to identify the most important variables with positive and negative correlations to the respective output variables. In Figure 6-7, Figure 6-8 and Figure 6-9 that follow, the variables are listed alongside process flow diagrams for LV manufacturing and CDP manufacturing. The purpose of this is to make it easier to follow the position of the variable relative to the whole manufacturing process. The regression coefficients are coloured green where the regression coefficient is positive and red where the coefficient is negative.

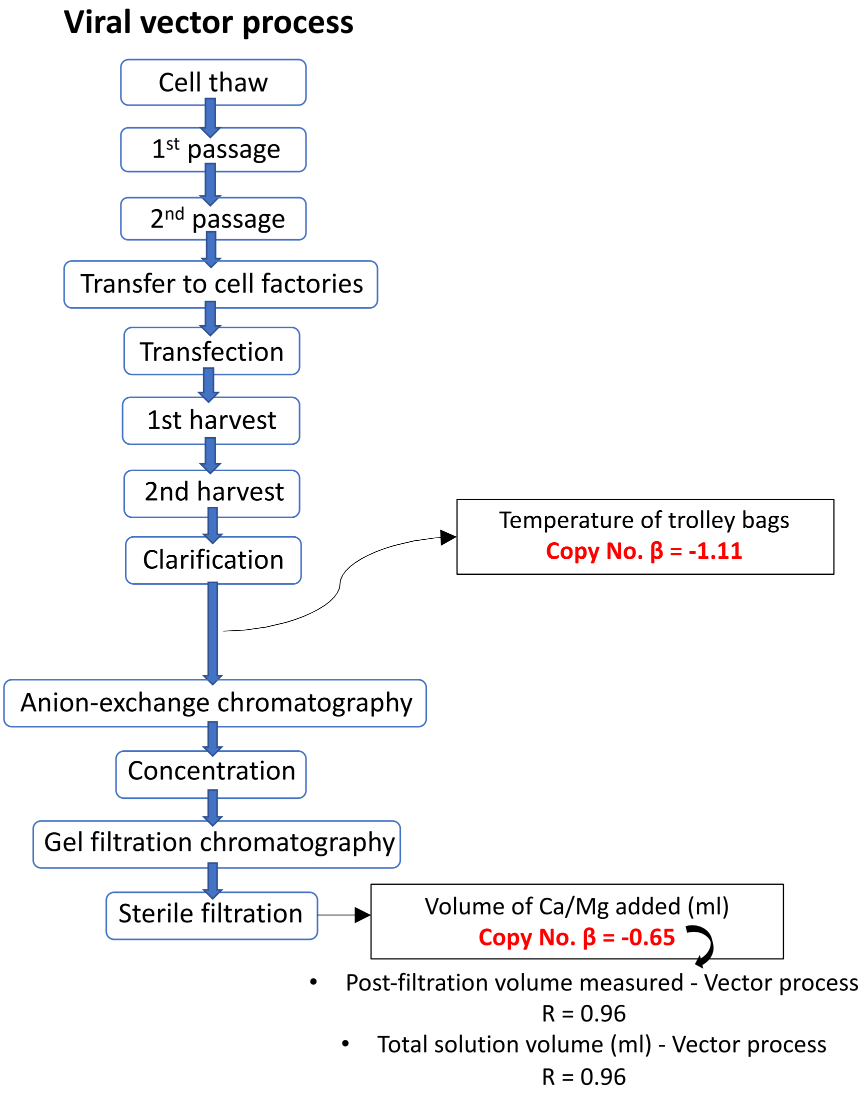

Viral vector process
Cell thaw
1st passage
2nd passage
Transfer to cell factories
Transfection
1st harvest
2nd harvest
Clarification
Anion-exchange chromatography
Concentration
Gel filtration chromatography
Sterile filtration
Temperature of trolley bags
Copy No. β = -1.11
Volume of Ca/Mg added (ml)
Copy No. β = -0.65
• Post-filtration volume measured - Vector process
R = 0.96
• Total solution volume (ml) - Vector process
R = 0.96

Figure 6-7: Process parameters predicting the lentiviral vector copy number: process parameters from viral vector manufacturing that were found to be predictive of the lentiviral vector copy number, placed alongside their respective unit operations.

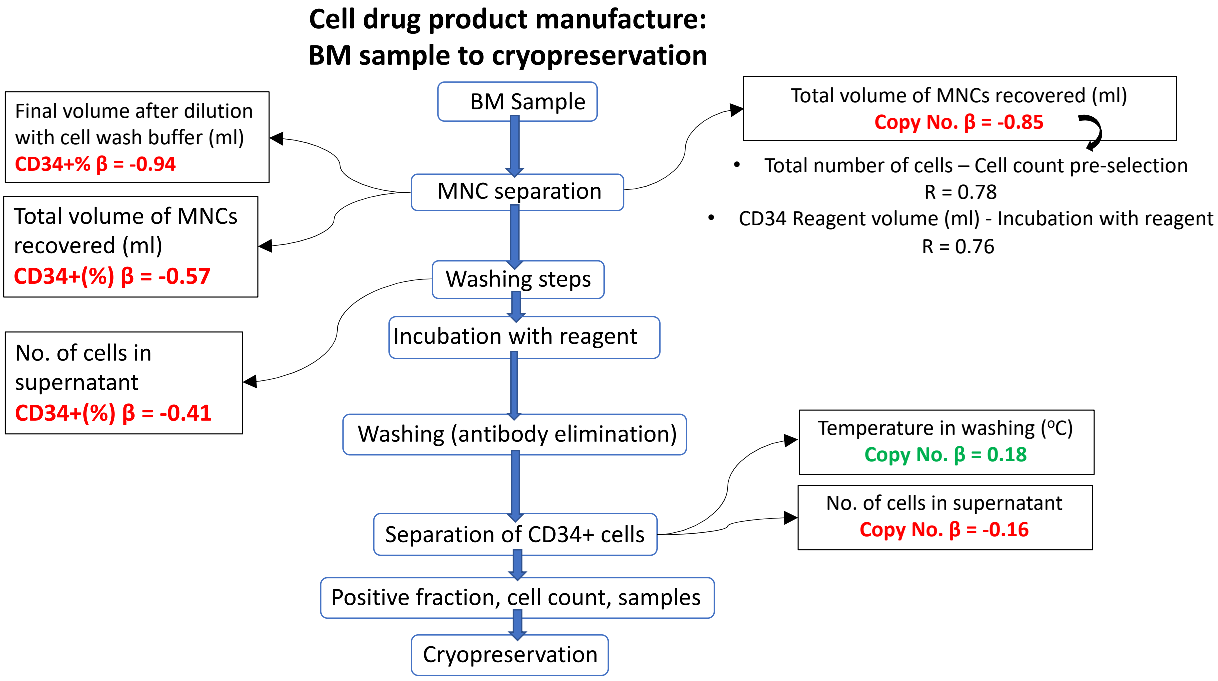

Cell drug product manufacture:
BM sample to cryopreservation
BM Sample
MNC separation
Washing steps
Incubation with reagent
Washing (antibody elimination)
Separation of CD34+ cells
Positive fraction, cell count, samples
Cryopreservation
Final volume after dilution with cell wash buffer (ml)
CD34+% β = -0.94
Total volume of MNCs recovered (ml)
CD34+(%) β = -0.57
No. of cells in supernatant
CD34+(%) β = -0.41
Total volume of MNCs recovered (ml)
Copy No. β = -0.85
• Total number of cells – Cell count pre-selection
R = 0.78
• CD34 Reagent volume (ml) - Incubation with reagent
R = 0.76
Temperature in washing (°C)
Copy No. β = 0.18
No. of cells in supernatant
Copy No. β = -0.16

Figure 6-8: Process parameters predicting the lentiviral vector copy number: process parameters from ex vivo cell processing that were found to be predictive of the lentiviral vector copy number, placed alongside their respective unit operations.

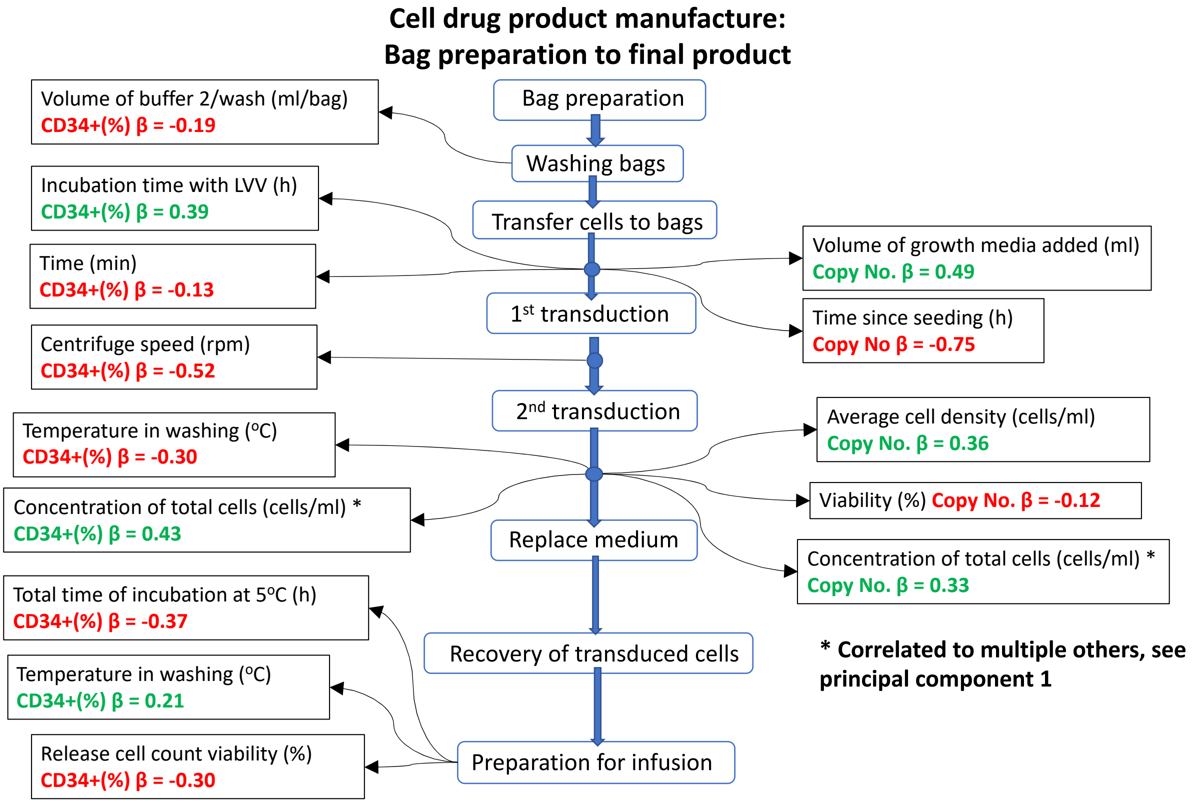

Cell drug product manufacture:
Bag preparation to final product
Bag preparation
Washing bags
Transfer cells to bags
1st transduction
2nd transduction
Replace medium
Recovery of transduced cells
Preparation for infusion
Volume of buffer 2/wash (ml/bag)
CD34+(%) β = -0.19
Incubation time with LVV (h)
CD34+(%) β = 0.39
Time (min)
CD34+(%) β = -0.13
Centrifuge speed (rpm)
CD34+(%) β = -0.52
Temperature in washing (°C)
CD34+(%) β = -0.30
Concentration of total cells (cells/ml) *
CD34+(%) β = 0.43
Total time of incubation at 5°C (h)
CD34+(%) β = -0.37
Temperature in washing (°C)
CD34+(%) β = 0.21
Release cell count viability (%)
CD34+(%) β = -0.30
Volume of growth media added (ml)
Copy No. β = 0.49
Time since seeding (h)
Copy No β = -0.75
Average cell density (cells/ml)
Copy No. β = 0.36
Viability (%) Copy No. β = -0.12
Concentration of total cells (cells/ml) *
Copy No. β = 0.33
* Correlated to multiple others, see principal component 1

Figure 6-9 Process parameters predicting the lentiviral vector copy number: process parameters from ex vivo cell processing that were found to be predictive of the lentiviral vector copy number, placed alongside their respective unit operations.

In order to further analyse the variables that were identified in the predictive models, the following text provides information on the variables and their role in the manufacturing process and considers possible explanations for the process parameter- critical quality attribute (PP- CQA) relationships that were identified.

Volume of Ca/Mg added – Calcium and magnesium ions were added to the lentiviral vector product in the final stage of the lentiviral vector manufacturing process because the manufacturer was aiming to keep the concentration of these ions consistent with a previous cell culture media formulation. This parameter was identified as an important predictor variable for the LV copy number, with a negative beta coefficient of medium to large effect size. It is worth noting that this parameter is also positively correlated with the post-sterile filtration volume of the lentiviral vector product. It has been hypothesised that this volume of Ca/Mg added is negatively impacting the lentiviral vector copy number due to damage caused to the virions by high local concentrations of salt ions when added to the viral vector formulation. The LVs are known to be highly sensitive to high concentrations of salts (Merten et al., 2016), which can cause a loss of infectivity (Zimmermann et al., 2011). If there is ever a need to add salt ions to the viral vector formulation, the model indicates that the addition should be carried out in such a way as to minimise high local concentrations, for example through constant mixing with a slow addition rate.

Temperature of refrigerated trolley bags – The temperature of refrigerated trolley bags was identified as an important predictor variable for the LV copy number with a negative beta coefficient of large effect size. The temperature of refrigerated trolleys bags was also found to be negatively correlated to the infectious titre with small effect size. This result indicates that a high temperature of refrigeration, during clarification of the lentiviral vector supernatant, may be negatively impacting the LV copy number. The lentiviral vectors are known to be temperature sensitive. In a study by Higashikawa *et al.* (2001), HIV-1 derived LVs were incubated at 4, 20, 37, 45 and 50 ^{o}C, for different lengths of time and the vector titres were then determined. They found that with respect to the transduction efficiencies, the half-lives of the LVs decreased rapidly as temperature was increased above 4 ^{o}C. It is therefore plausible that even small increases in storage temperature could cause the LVs to lose potency to a significant degree. The magnitude of the beta coefficient, which was strictly negative across models and folds, instils confidence that this parameter has a reliable negative correlation with the LV copy number. This parameter had a min, max and range of 3.8, 5.4 and 1.6 ^{o}C, respectively. This indicates that the control is already relatively tight. It may be difficult to improve the tightness of this control due to limitations of the equipment and the potential significant capital cost to upgrade.

Concentration of cells after second transduction – The cell concentration, determined in the cell count after the second transduction, was correlated positively to the LV copy number with small effect size. The model indicates that operating at the higher cell concentration is correlated to higher LV copy numbers. The cell concentration is known to be an important parameter in the transduction process and studies in the literature have linked higher cell density with higher transduction efficiency.

In a study by Uchida *et al.* (2019), the effect of cell density on transduction efficiency was investigated by transducing CD34+ cells at a range of concentrations, 5 x10^{4} to 5 x10^{5} cells/ml, with lentiviral vectors expressing an enhanced green fluorescent protein (eGFP). After three days, the transduction efficiency was evaluated by determining the percentage of cells expressing the eGFP. It was found that the higher cell density cultures exhibited significantly higher transduction efficiency. The high-density culture (5 x10^{5} cells/ml) had a 2.7-fold higher eGFP percentage compared to the low-density culture (5 x10^{4} cells/ml). Zhang *et al.* (2004) transduced 293T cells with lentiviral vectors and they also showed that higher cell density led to increased transduction efficiency. It has been hypothesised that cell-to-cell contact increased the efficiency of lentiviral transduction, perhaps due to secondary transduction from exposed to non-exposed cells (Uchida et al., 2019).

Time since seeding – After transferring the cells to the Retronectin coated bags, the cells are incubated for a period before the first transduction takes place. The models indicate that the incubation time is negatively correlated to the lentiviral vector copy number with a large effect size. This parameter was selected for multiple models and had strictly negative regression coefficients across models and folds, increasing confidence in the correlation with the LV copy number. This parameter had a range of 1.75 hours.

Volume of growth media added – After the first transduction procedure the cells are washed and resuspended in fresh medium. The volume of growth media that was added for resuspension of the cells was correlated positively to the LV copy number with medium effect size. The volume of growth media had a min, max and range of 11, 101.95 and 90.95 ml, respectively.

Incubation time with LV – The duration of time that the cells were incubated with the CD34 reagent was found to be an important predictor of the CD34+ percentage, where a positive correlation was identified with medium effect size. The cells are incubated with the reagent to allow time for the antibody conjugates (including dextran microbeads) in the CD34 reagent to bind to the antigen on the CD34+ cells. The percentage of CD34+ cells recovered is dependent on sufficient labelling of the CD34+ cells with the dextran microbeads. This process is not instantaneous because it is subject to rate limitations in mass transfer and bonding between the antibodies and antigens. This provides an explanation for the sign on the beta coefficient, which indicates that a longer incubation period is correlated to greater recovery of CD34+ cells. Studying the mass transfer and bonding process may provide insights into the minimum incubation time that should be used to avoid negatively impacting the percentage of CD34+ cells. Unfortunately, the model does not provide this information, it indicates that longer incubations could be beneficial and that the variation that is currently permitted is impacting the CD34+ cell percentage. The incubation time had a range of 1 hour.

Final volume after dilution with cell wash buffer – the final volume of the cell suspension after dilution with cell wash buffer is negatively correlated to the CD34+ cells percentage with large effect size. The buffer volume is usually scaled in accordance with the number of cells (Kwok et al., 2007). However, it is possible that the cell-to-buffer ratio changes or that the volume of cell wash buffer impacts recovery of the CD34+ cells. The final volume after dilution with cell wash buffer had a min, max and range of 540, 1415 and 875 ml, respectively.

Total volume of MNCs recovered – the total volume of MNCs recovered is negatively correlated to the CD34+ cells percentage with medium effect size. In feature extraction, it was observed that the volume of MNCs recovered was positively correlated with the final volume after dilution with cell wash buffer and the volume of CD34 reagent used. It also correlated positively with the number of cells in the bone marrow sample and its volume. The volume of CD34 reagent was adjusted in proportion to the volume of the cell suspension, meaning that the concentration of CD34 reagent should be approximately equal. Given that the cell number also correlates positively with the volume of MNCs recovered, if there is a physical mechanism negatively impacting the

the CD34+ cells percentage, the data suggests that it is unconnected to the cell density and concentration of CD34+ reagent. The total volume of MNCs may have an impact on another physical mechanism that is related to the recovery of the CD34+ cells. Another unrecorded parameter which is relevant here is the density of the cells in the solution. To further understand and identify a possible mechanism to explain this correlation, a number of hypotheses may need to be tested. The total volume of MNCs recovered had a min, max and range of 250, 600 and 350 ml, respectively.

No. of cells in the supernatant – After MNCs separation, the cells go through several wash cycles. The number of cells remaining in the supernatant after centrifugation was found to be negatively correlated to the CD34+ cells percentage with medium effect size. This parameter is also positively correlated to the volume of MNCs recovered. This parameter is likely to be part of the same effect as the volume of MNCs recovered. The number of cells in the supernatant had a range of $0.1x10^6$ ml.

Washing after first transduction – centrifuge speed, temperature and wash time.

The washing step after first transduction produced several variables that were negatively correlated to the CD34+ cells percentage. The duration of the wash and the temperature of the wash were correlated negatively to CD34+ percentage with small effect size, whilst the speed of centrifugation was correlated negatively with medium effect size.

Total time of incubation at 5ºC – the total time of incubation at 5ºC was negatively correlated to the CD34+ cells percentage with medium effect size. The incubation periods had a range of 11.08 hours.

Temperature in washing – After the second transduction, the cells were incubated at 5ºC before being washed and resuspended in saline. The temperature of the centrifuge during the wash was found to be positively correlated with the CD34+ cells percentage.

Cell viability measured in release cell count – The cell viability percentage, as recorded in the release cell count, was negatively correlated to the CD34+ cells percentage with small effect size. The cell viability had a min, max and range of 98, 100 and 2%, respectively.

6.4 Conclusions

The focus of this chapter was on the CDP manufacturing process, where the LVs are used to transduce the patient’s cells to create the CDP, which is grafted back to the patient. In order to understand the main sources of variability in the CDP manufacturing process and to explore the correlations between process parameters, the CDP manufacturing data was explored using the sparse PCA algorithm developed in chapter 4. Following on from this, the CDP manufacturing data was combined with the LV manufacturing data to form a dataset for investigation with predictive modelling techniques. The aim of this work was to relate the manufacturing variables from both CDP manufacturing and LV manufacturing to two critical quality attributes of the CDP, namely, the LV copy number and the CD34+ cells percentage.

The first three principal components of the sparse PCA model captured close to 50% of the variance in the data and the score plots revealed that variability in the CDP manufacturing process largely occurred on a batch-to-batch basis, as the scores were evenly distributed on the first three components. Key variables on principal component 1 were the total number of CD34+ cells that were isolated, the volume of MNCs obtained from MNCs separation, the volumes of buffers and reagents, and the volume of growth media and cytokines used for in the cell culture during the transduction steps. Many of these variables were highly correlated and the source of the variability begins with the volume of the bone marrow (BM) sample and its composition, particularly the number of cells and their viability.

Several variables from the numerous washing steps throughout the process featured on principal components 1 to 3, including the temperature, speed and duration of centrifugation, showing that the washing steps are important regions of variability in the process. The temperature of centrifugation was negatively correlated to the speed in many instances and these parameters were correlated to other variables, such as the volume of cell wash buffer that was used. Principal component 2 showed a degree of correlation with the LV copy number, as batches 20 to 23 all scored high on component 2 and featured high LV copy numbers. The most influential variables on principal component 2 included the volume of cell wash buffer used in MNCs separation and isolation of CD34+ cells, and the concentration of the CD34+ cells before seeding into bags for transduction. These process parameters were identified as possible sources of the correlation that was observed with the LV copy number. In addition to the metric variables that were analysed with sparse PCA, there were numerous categorical variables in the original dataset. As with the LV manufacturing data, most of the categorical variables were representing different material lots and pieces of equipment that were used in the process. Material lots included buffers, reagents, cytokines, growth media and retronectin used for bag coatings. Equipment included bags, bag transfer sets and the column and tubing sets used the separation of CD34+ cells.

For the CD34+ cells percentage, the MNCs separation step which is carried out prior to the labelling of the CD34+ cells, produced two key predictive variables. The final volume after dilution with cell wash buffer and the volume of MNCs recovered both correlated negatively with the CD34+ cells percentage and were correlated positively to one another. The model indicated that a larger volume of cell suspension may have negatively impacted the labelling and/or the recovery of the CD34+ cells. The dilution with cell wash buffer may have altered the cell concentration to produce an adverse effect or there may have been another adverse effect associated with the processing of a larger volume of cell suspension. Other important variables with negative correlations to the CD34+ cells percentage included the number of cells present in the supernatant after the third wash of the MNCs, the speed and temperature of centrifugation in the washing step before first transduction, the total time of incubation at 5^{o}C during prep for infusion and the viability of the cells in the release cell count. The most significant variables with positive correlations to the CD34+ cells percentage were the concentration of cells after second transduction, the incubation time with LVs during the first transduction and the column plate number in LV manufacturing.

The most influential variable in the LV copy number models came from LV manufacturing, where the temperature of refrigerated trolley bags after the clarification step was negatively correlated to the LV copy number with large effect size. Another influential variable from LV manufacturing was the volume of Ca/Mg solution added to the final formulation of the LVs. In addition to correlating negatively with the CD34+ cells percentage, the total volume of MNCs recovered also correlated negatively with the LV copy number. The rest time between seeding of bags and first transduction is the final variable that was negatively correlated to the LV copy number. The volume of growth media used for cell suspension before first transduction was positively correlated to the LV copy number with medium effect size and the concentration of cells after second transduction was positively correlated with small effect size.

Overall the models provided insights into PP-CQA relationships and identified several process parameters which are potentially critical process parameters, directly influencing the LV copy number and CD34+ cells percentage. A number of these were supported by literature evidence, or by the observations of process experts. Process parameters that are confirmed to have cause-and-effect relationships with the CQAs should be optimised and placed under improved process control where appropriate and where it is practical to do so. In some instances, the deviations permitted within the process specification may need to be reduced. In cases where the physical mechanism connecting the process parameter to the product CQAs is not clear, possible explanations for the mechanism should be considered by drawing from the knowledge of process experts. It is especially important to consider what the identified process parameters may be representing physically and what other parameters of the system the variable is related to, which perhaps are not explicitly recorded in the dataset. This is particularly important for variables with medium and large effect sizes, as the model indicates that these parameters have a significant influence on the process CQAs.

Chapter 7
Analytics for bioreactor-based viral vector production

Due to the increasing demand for large quantities of viral vectors for use in clinical trials, manufacturers and research laboratories are developing scalable approaches to viral vector production. As discussed in chapter 2, viral vector production using adherent cell lines has limited scale-up potential. One of the alternative approaches for viral vector production is bioreactor-based suspension cultures. In this chapter, historical process data from bioreactor production is analysed using MVDA, focusing solely on the upstream process. The upstream process involves the bioreactor-based expansion of a producer cell culture, followed by transient-transduction and production of the viral vectors.

7.1 Introduction

Chapter 5 focused on the manufacturing of LVs using adherent human embryonic kidney 293 cells (HEK 293) as the producer cell line. A typical adherent cell process, such as the process described by Merten *et al.* (2010), involves growing the cells in flasks before transferring them to cell factory stacks, where they are expanded further before transfection takes place. Although the cell factory stacks offer a large surface for the cells to attach to; they are limited in scalability as their design and requirement for manual handling makes them unsuitable for scale-up. Bioreactors, of which there are many different types, are far more suitable for large-scale applications. Consequently, manufacturers are developing and exploring the potential of bioreactor-based LV manufacturing to meet the increasing demands for high quality (GMP standard) and high-volume LVs.

In addition to scalability, bioreactor LV production alleviates some of the other challenges associated with LV manufacturing through its closed system design and ability to automate some of the manual tasks. One challenge that remains is the high degree of complexity of the process, which means that the relationships between process parameters and product CQAs are still relatively poorly understood. Moving from cell factory stacks to a bioreactor, the new environment features different manipulatable process parameters and design features that present new challenges when trying to understand process behaviour. Additionally, characterization of the process and product is required, and comparability must be demonstrated across the various production scales.

In this chapter, the analysis of process data from the bioreactor production of LVs is presented to identify potential learning opportunities contained within the data. The bioreactor process data consisted of both offline and online process measurements for 2L, 50L and 200L bioreactors, which required different pre-processing compared to the 2D cross-sectional data (sample and variable) that was analysed in chapters 5 and 6. This is due to the additional time dimension resulting in 3D data (sample, variable and time). The details of this pre-processing are explained in the methods section within this chapter. The approach to the analysis was first to explore the data with PCA with the objectives to identify the key features of variance within the data, to assess the variability within each of the process scales and to evaluate process comparability between the scales. The main challenge here was scaling the variables appropriately to make sensible comparisons across different bioreactor scales. After appropriate scaling was applied, which is described in the methodology, PCA was found to be a useful tool for comparing the batches from different scales.

Following data exploration, the online and offline process parameters were used as inputs to a PLS model to predict the infectious titre of the LV product. The objectives of this predictive modelling task were to identify critical process parameters and evaluate their relationship with the infectious titre to increase understanding of the process behaviour and offer useful insights to guide process development and optimisation.

7.1.1 Data

A description of the bioreactor-based suspension culture process for the production of LVs is provided in chapter 2, section 2.1.4. The process data consisted of 22 batches with 10 online and 21 offline variables. The online and offline process variables are listed in Table *7-1*. 16 of the batches were at the 2L scale, while four were 50L and two were 200L. Table *7-2* provides a summary of the 22 batches available for the analysis with information on key parameters, including the batch duration and cell seeding density, and two CQAs for the LV product, namely the infectious titre and the physical titre. The cell seeding densities, infectious titre and physical titre have been scaled to protect this confidential information; however, the scaling preserved the distribution of the data and shows where the cell seeding densities were varied.

Table 7-1: Online and offline process variables available in the bioreactor-based LV production dataset.

	Online	**Offline**
1	Jacket temperature	Cell diameter
2	Vessel temperature	Glucose conc.
3	pH	Glutamate conc.
4	Oxygen partial pressure	Glutamine conc.
5	Stir speed	Lactate conc.
6	Air sparge flowrate	Load cell weight
7	Air overlay flowrate	N_2 sparge flowrate
8	CO_2 flowrate	Air overlay flowrate
9	Substrate A	Air sparge flowrate
10	Substrate B	CO_2 sparge flowrate
12		O_2 sparge flowrate
13		Offline pH
14		Vessel temperature
15		Jacket temperature
16		Osmolality
17		Sample volume

18		Vessel volume
19		Viability
20		Viable cells
21		Volume of base added

Table 7-2: Summary data for the 22 bioreactor-based LV batches.

Batch	Batch volume (L)	Batch duration (h)	Scaled cell seeding density	Scaled infectious titre	Scaled physical titre
1	2	119.4	-0.20	-0.73	-0.34
2	2	119.3	-0.40	-0.83	-0.28
3	2	119.4	-0.33	-0.75	-1.00
4	2	119.4	-0.38	-1.15	2.20
5	2	50.2	2.15	-0.75	-0.34
6	2	50.4	3.70	-0.83	-1.00
7	2	99.0	-0.22	Unavailable	Unavailable
8	2	99.5	-0.18	Unavailable	Unavailable
9	50	99.6	-0.29	-0.36	-0.32
10	2	117.7	0.28	Unavailable	Unavailable
11	2	117.7	0.32	Unavailable	Unavailable
12	50	116.8	-0.37	0.66	1.27
13	2	170.2	-0.27	0.23	-0.26
14	2	170.2	-0.41	-0.47	-0.63
15	50	73.3	-0.42	Unavailable	Unavailable
16	2	95.1	-0.46	Unavailable	-0.81
17	2	95.0	-0.32	Unavailable	-1.46
18	200	94.8	-0.36	0.17	0.23
19	50	73.4	-0.26	Unavailable	Unavailable
20	2	118.5	-0.60	1.50	1.34
21	2	118.5	-0.45	1.55	0.87
22	200	119.0	-0.53	1.77	0.54

The batches varied in duration between 50 hours and 170 hours and featured some variation in the cell seeding densities. Infectious and physical titres were not available for all batches, limiting the selection of batches available for predictive modelling of these two parameters. For more information on the physical and infectious titres, the reader is referred to chapter 2, section 2.1.6, on the CQAs of viral vector products.

7.2 Method

7.2.1 Pre-processing of process data

The pre-processing of the time-series data involved typical steps, such as the handling of missing data, as well as some pre-processing steps which were specific to the application. For example, these included the alignment of batches of different time durations and the derivation of additional process parameters. The following section details the pre-processing steps that were taken to obtain datasets for the PCA and PLS analyses that were conducted.

7.2.1.1 Processing of offline and online process variables

1. Missing data was analysed across all of the samples, and samples containing more than 20% missing data in the online and offline process parameters were removed from the analysis.
2. The variables in the dataset were evaluated for missing data and variables with significant portions of missing data (greater than 20% across more than 20% of the batches) were removed from the analysis.
3. At this point, much of the missing data was removed from the analysis; however, there were small percentages of missing data present in the online variables for some runs. To fill in this missing data, linear interpolation was applied to the online process data.
4. Intrinsic properties such as pH and oxygen concentration were comparable across production scales without requiring scaling factors. However, to make process variables representing extrinsic properties comparable across production scales, appropriate scaling factors had to be determined and applied to the data.
    a. The gas flowrate variables: air sparge, CO_2 sparge and air overlay, were scaled by the reactor volume.
    b. Volume related variables, such as the volume of acid and base added to the reactor, were scaled by the reactor volume.
    c. The working volume of the reactor was converted to a percentage.
    d. The stirrer speed was scaled based on the power number of the vessel.
5. The offline data, which consisted of between 5 to 8 datapoints throughout the duration of each batch, was interpolated using 3rd order polynomials. The purpose of this interpolation was to 1) obtain data at consistent timepoints for all of the batches, 2) obtain data at shorter intervals to align with online data and 3) allow numerical determination of the gradients of some offline parameters, such as the cell and metabolite concentrations. While carrying out the interpolations, each graph was checked to ensure that the profile obtained was in-line with the expected trend. Some datapoints are highly influential in the offline data because there are so few datapoints available and this is a potential

source of error. A graphical example of this interpolation is provided in Figure 7-1.

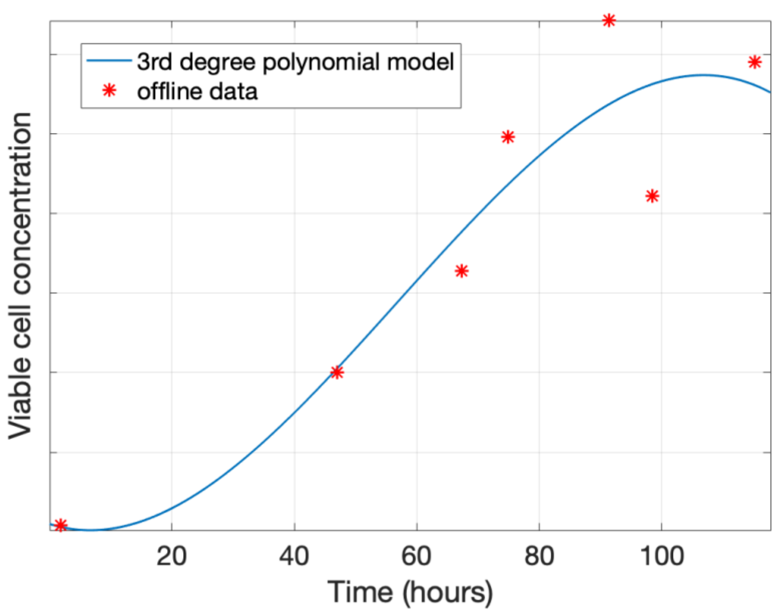


Figure 7-1: Example interpolation of offline data using a 3rd order polynomial.

The practice of fitting 3rd order polynomials to the metabolite and cell concentrations was reported by Le *et al.* (2012). The polynomial model allowed for augmentation of additional datapoints and numerical determination of the gradient of each variable throughout the process. In this work, the polynomial fits were assessed for each of the offline variables and fits with R^2 greater than 0.7 were deemed good enough to represent the data. Other techniques such as splines may have been used; however, the low frequency of the offline data is a limiting factor in the quality of the interpolation. The 3rd order polynomial is relatively constrained in its flexibility which is good for avoiding overfitting the data, yet it is suitable for capturing the smooth curves that are expected in the temporal profiles of the offline process variables, including the osmolality, and the metabolite and cell concentrations (Le et al., 2012).

6. The polynomial fits to the offline data were used to generate new variables including rates determined by the polynomial gradient and variables that were divided by the cell concentration. Appendix (chapter 9, section 9.3) provides an example calculation using fictional data. The following is a list of the new variables that were created:
    a. Specific glucose, glutamine, glutamate and lactate concentrations i.e. on a per cell basis (g/cm^3.cell)

b. Specific glucose, glutamine, glutamate and lactate concentration gradients (g/cm^3.cell.h) i.e. rates of change in concentration on a per cell basis
c. Viable cell concentration gradient (cells/ml.h), i.e. cell growth rate, and viability gradient (%/h)
d. Osmolality gradient (mOsm/h)
e. Vessel volume gradient (L/h)
f. Specific flowrates of air and CO_2 (L^3/h.cell)
g. Specific O_2 partial pressure (Pa/cell)

7. The online, offline and derived process variables were compiled into a single data table for each of the batches.
8. The 22 batches of process data featured a range of batch durations. A requirement for the analysis is that each of the batches features the same number of datapoints. In order to align the batches to equal length, the time of transfection was taken as the central timepoint, to which the temporal profiles of different batches were aligned. Henceforth, the time of transfection was taken as time 0 and data was truncated from both ends of the temporal profiles, so that each batch spanned from *-x* hours to *y* hours, with respect to the time of transfection. Here *x* and *y* are the minimum length of time between inoculation and transfection and the minimum length of time between transfection and the batch endpoint, respectively, for all of the batches included in the analysis. Figure 7-2 illustrates this method of aligning data.

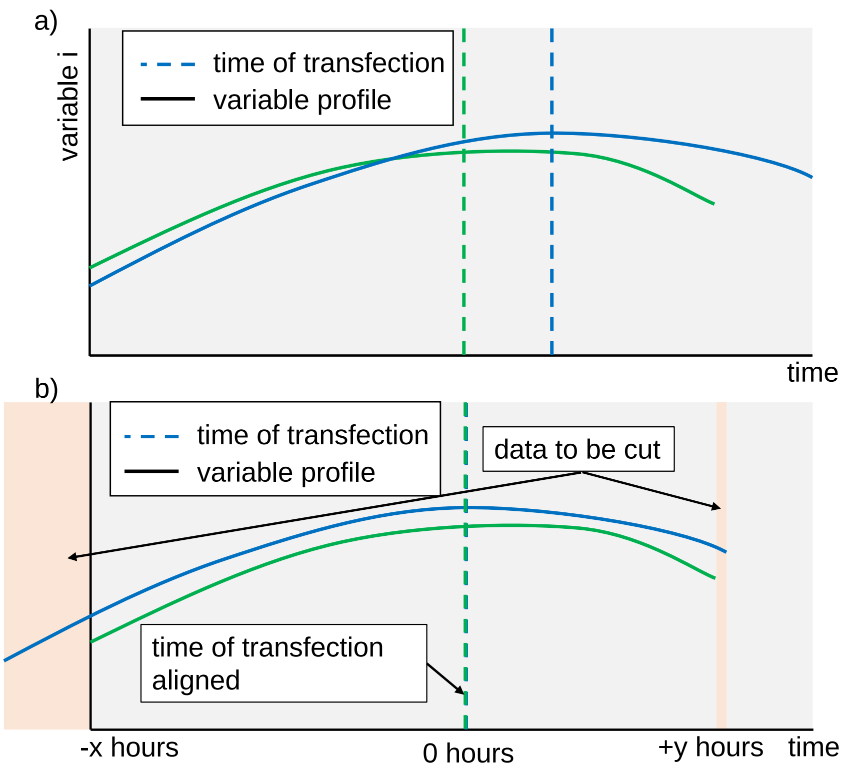


Figure 7-2 Illustration of the batch time alignment method and the cutting of the data required to obtain uniform batch lengths

9. The 3D data, with batch/observation, variable and time dimensions, was unfolded into a 2D matrix, in accordance with the requirements for the multivariate methods that were applied. The unfolding of 3D data can be carried out in various ways depending on which aspect of the data the practitioner wishes to focus on. For process data, it is logical to preserve the batch/observation dimension on one axis and to combine the variable and time instances on the second axis. This means that 'variables' in the traditional sense become variable-time instances, i.e. each variable is one of the original variables at a specific timepoint. Henceforth, this unfolding approach maintains sensible physical meaning. An illustration of the data unfolding is provided in chapter 3, section 3.4.2
10. For PCA and PLS the data inputs were auto-scaled, i.e. each variable was transformed by subtracting the mean and dividing by the standard deviation. Mean centring and scaling were both carried out due to the improved interpretability of the results, which they provide (see chapter 3, section 3.3.4).

7.2.1.2 Processing of product quality data

The product quality data included the infectious titre and the physical titre for each batch, which were the response variables used in this analysis. The response variables only required aligning to the correct input data and scaling. Autoscaling was applied to the response variables, consistent with the scaling applied to the input variables.

7.2.2 PCA

The PCA method used in this chapter was the standard PCA algorithm. The main reason for this is because the data consisted of numerous variables (1000+), which was a result of unfolding the 3D data with variable and time instances along one axis. Unfortunately, a limitation of the sparse PCA algorithm is that the solution time increases exponentially with the number of variables in the problem. However, standard PCA is not overly complex in this situation because the number of original variables included in the analysis was relatively few (less than 20). The approach to PCA was the same as in chapters 5 and 6, the main difference being the use of temporal data. For details on the PCA technique, the reader is referred to chapter 3.

7.2.3 Development of PLS models

In this chapter, PLS models were developed to predict the infectious titre and physical titre using the online and offline process variables as model inputs (predictor variables). The following list details the steps taken to develop the models:

1. The data was partitioned into a training set and a hold-out test set, using the Kennard-Stone algorithm to select the batches for each set. The Kennard-Stone algorithm selects batches for the test set with uniform distribution over the predictor variable space, which is achieved by evaluating the Euclidean distance between observations. More details of the algorithm are in chapter 3, section 3.3.6.
2. After partitioning, autoscaling was applied to the training set and the same scaling parameters were applied to the test set.
3. The scaled predictor variables in the matrix $\boldsymbol{X}$, and the scaled response variable in the vector $\boldsymbol{y}$, from the training dataset, were then passed to the repeated K-fold cross-validation script in Matlab (details of the repeated K-fold cross validation method are provided in chapter 3, section 3.3.3). This script carried out 5-fold cross validation, with 2000 repeats, for models with n latent components, where n was varied from 1 to 10.
4. The model with the smallest cross validation MAE score was selected, thus the number of latent components was determined.
5. Next the PLS model was fitted to the whole of the training data and the VIP selection method (detailed in chapter 3, section 3.2.2) was used to select the

most important predictor variables from the model. The threshold VIP cut-off was varied between 1 and 1.4.

6. The reduced model was put through repeated 5-fold cross validation (2000 repeats) and the number of latent components was varied from 1 to 10.
7. The optimal model (minimum cross validation) was then selected and fitted to the whole of the training data to evaluate model fit and determine model parameters.
8. Using the bootstrap technique detailed in chapter 3, section 3.3.4, the stability and distribution of the model regression coefficients were evaluated with 2000 bootstrap samples. This provided confidence intervals for the standardised regression coefficients (beta coefficients).
9. Finally, the model was applied to the test set that was held-out of model development and the performance was evaluated.

7.2.3.1 PLS model performance evaluation

Model performance was evaluated at four key stages throughout model development; these are summarized in Table *7-3*. See chapter 3, section 3.3.1 for more information on the PLS model performance evaluation.

Table 7-3: Key stages of model development and model performance evaluation methods.

Model development stage	Performance evaluation methods
1. Model fit to the training data as a whole	• Visual inspection of model fit and regression errors • Performance metrics: Model significance (F-stat), R^2, MAE
2. Repeated K-fold cross validation with the training data	• Performance metrics evaluated for latent components from 1 to 10: MAE, mean square error (MSE), R^2
3. Bootstrap sampling	• Distribution of the beta coefficients (standardised regression coefficients) was evaluated and confidence intervals determined • Visual inspection of beta coefficient confidence intervals across temporal profiles of predictor variables

4. Model fit to the test data that was held out of model development	• Visual inspection of model fit and regression errors • Performance metrics: MAE, MSE, R^2

7.3 Results and discussion

7.3.1 Feature extraction with PCA

When conducting feature extraction with the bioreactor-based viral vector production data, there were a few choices to be made with respect to the batches that should be included in the analysis. There were two key factors that were varied in the available data; these were the scale of the bioreactor and the cell seeding density. Table *7-2*, in section 7.1.1 provides details on the bioreactor scales and cell seeding densities. In the data that was available, there were batches at the 2L, 50L and 200L scales and a few of the batches were conducted with higher cell seeding densities. One point of interest for the analysis was the comparability of the batches across different scales, henceforth, it was decided that the cell seeding density should be kept as consistent as possible and the batches from different scales should be included in the analysis. Batches 5, 6, 10 and 11 were therefore excluded from the analysis (see Table *7-2*, section 7.1.1).

The variance explained by each component in the PCA model is shown in Figure 7-3. Six components were selected for the analysis, as beyond six components the variance explained decreases below 5% and the signal-to-noise ratio becomes low. See chapter 3, section 3.3.2 for more details on selection methods used to determine the number of components. Collectively the first six components explain 84.2% of the variance in the data. This is a greater amount of explained variance compared to the PCA models for the adherent LV and CDP manufacturing data. When the same number of principal components captures a larger percentage of the variation, this indicates that there are a greater number of correlated features. This is expected with the bioreactor data because each variable has auto-correlation, where observations around a similar time-point are correlated.

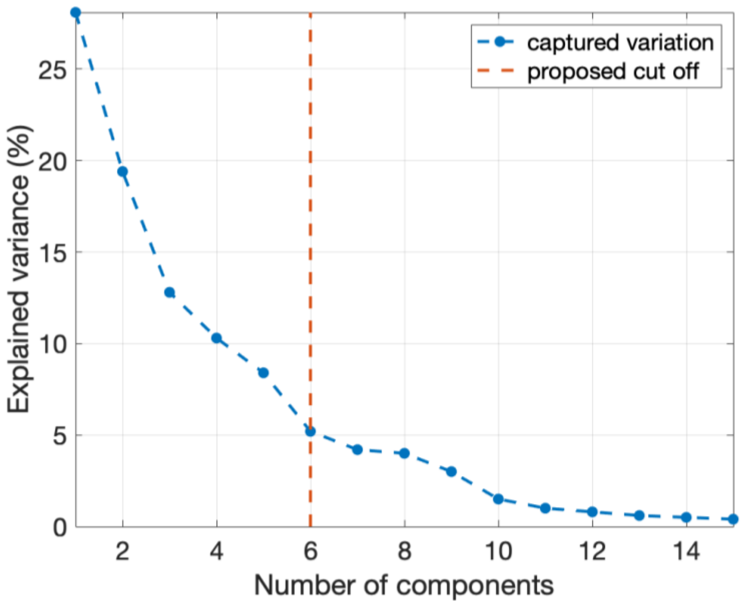


Figure 7-3: Explained variance against number of principal components in the PCA model. The proposed cut off point shows the number of components that were selected for the analysis.

Principal components 1 and 2 are plotted on Figure 7-4, explaining 28.1% and 19.4% of the variance, respectively. Components 1 and 2 capture variation within each scale and between the scales. The 2L batches are spread out on components 1 and 2 with score ranges of around 150 and 120, respectively. The 200L batches score low on components 1 and 2, separating them from the rest of the batches. Similarly, the 50L batches are separated from the rest due to high scores for component 2. The distance between the 50L batches and 2L batches is approximately 30 on component 2, which is less than range of scores observed within the 2L batches. The distance between the 200L batches and the rest of the batches on component 1 is around 40, which is small compared to the range of scores observed on component 1. The separation between the 2L, 50L and 200L batches, shows that there are some differences between the scales that are captured by components 1 and 2. However, the small distances between the different scales compared to the overall variation in scores shows that these differences are not big with respect to the process variability that is present irrespective of scale. This observation implies that the process comparability is relatively good and that process variability within each scale is an important consideration.

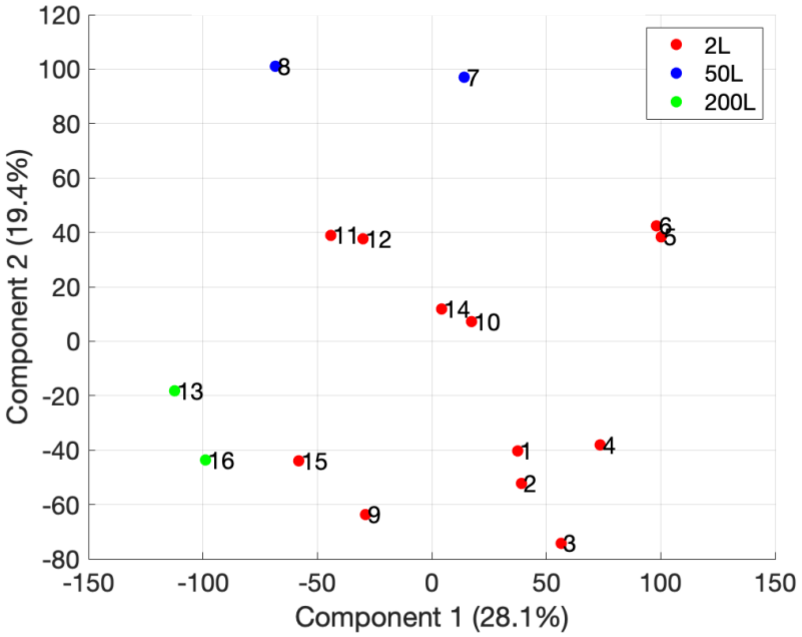


Figure 7-4: Principal component 1 versus principal component 2, explaining 28.1% and 19.4% of the variance respectively. The marker colours indicate the batch volume, which varied between the 2L, 50L and 200L scales.

Figure 7-5 shows the loadings for principal components 1 and 2. Component 1 highlights correlation between multiple variables, as several variables feature large magnitude loadings. The largest contributors are the osmolality, cell viability, number of viable cells, volume of base added, and the specific concentration of O_2, glucose, glutamate, glutamine and lactate. Component 1 highlights positive correlation between osmolality, cell viability in the mid to late stages of the process, and the specific concentration of oxygen, glucose, glutamate and glutamine in the mid to late process. These parameters correlate negatively with the concentration of viable cells mid to late in the process and the specific concentration of glutamine and lactate early in the process. Several variables have loadings that switch sign going from the early stage of the process to the mid to late stages of the process; these include the concentration of cells and the specific concentration of metabolites and oxygen. This implies that a higher concentration early in the process correlates to lower concentration late in the process and vice versa.

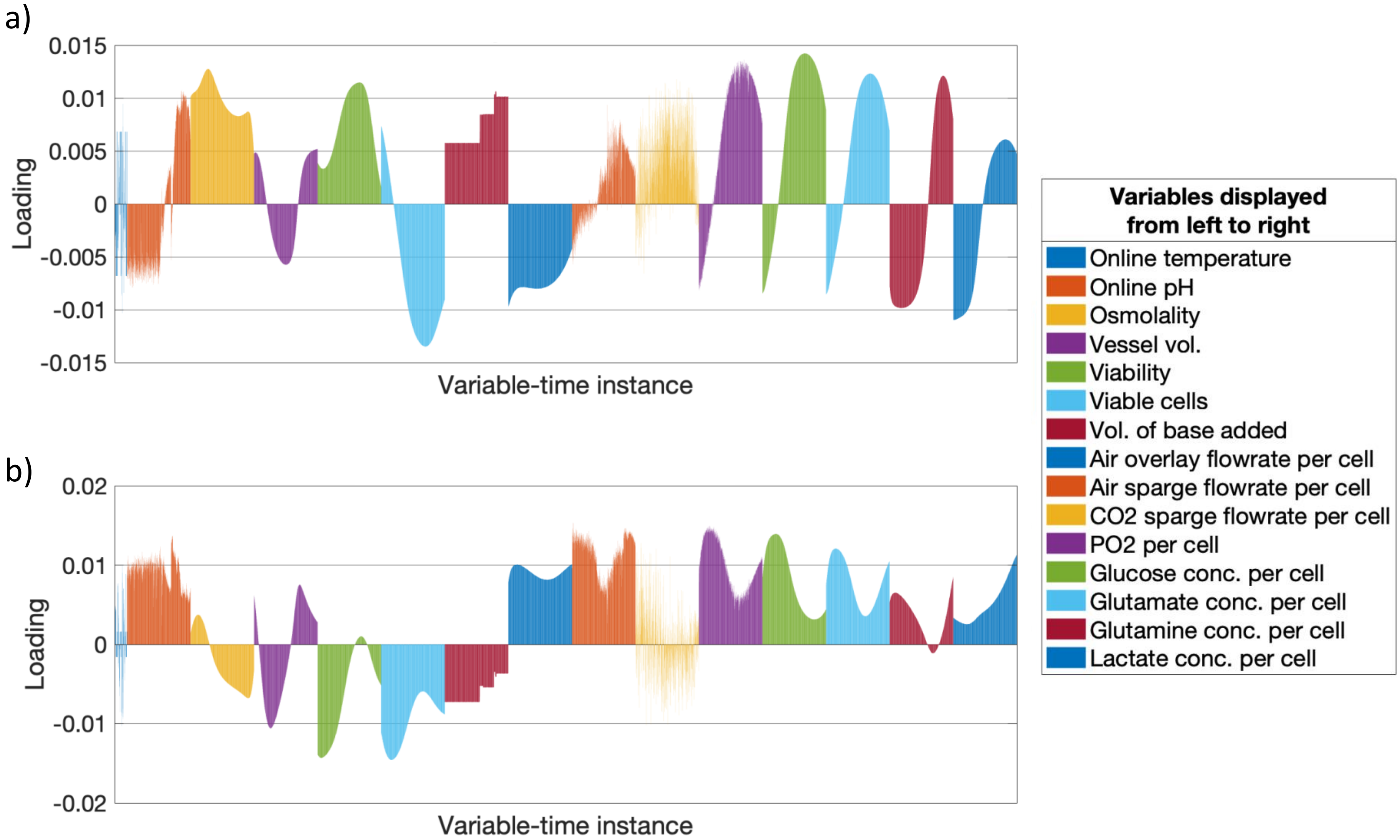


Figure 7-5: Loadings on component 1 (a) and loadings on component 2 (b) for each of the variable-time instances in the unfolded time-series data. Variable-time instances belonging to the same variable have been coloured the same and repeated colours may be distinguished as the variables from the legend appear in the plot from left to right.

The main variables contributing to component 2 scores are the online pH, vessel volume, cell viability and concentration, specific air sparge and overlay flowrates, and the specific O_2, glucose, glutamate and lactate concentrations. The loadings indicate a negative correlation between the pH from the start of the process and the cell concentration and viability. This is also reflected in the variables which are on a per cell basis, such as the specific gas flowrates and some of the specific metabolite concentrations, which correlate positively with the pH. Table *7-4* characterizes the 200L, 50L and 2L batches based on the variables highlighted by components 1 and 2. In order to ensure that the gas flowrates were comparable, the data was put into consistent units and the flowrates were divided by the reactor volume. Table *7-4* shows that despite the scaling, there are some differences in the air overlay and sparge flowrates between the 2L, 50L and 200L scales. The rest of the variables contributing to the separation of the 50L and 200L batches from the 2L batches are also varied within each scale.

Table 7-4: Characteristics of the 50L and 200L batches influencing principal component 1 and 2 scores.

Scale	Observation	Causes
50L	High on component 2	High pH, low viability early on, high air sparge and overlay flowrates, high specific glucose, glutamate and oxygen concentrations early in the process
200L	Low on component 1	High pH, low osmolality, high air overlay flowrate, high viable cell concentration late in the process

The scores for components 3 and 4 explain 12.8% and 10.3% of the variance, respectively, and these are shown on Figure 7-6. The 50L and 200L batches score high on component 3 when compared to most of the 2L batches, whereas component 4 captures variation with no direct link to scale. The variation between each of the scales is small compared to the variation that is present within each scale, as was the case with components 1 and 2. This is further evidence that the processes are comparable in terms of the variables included in this analysis.

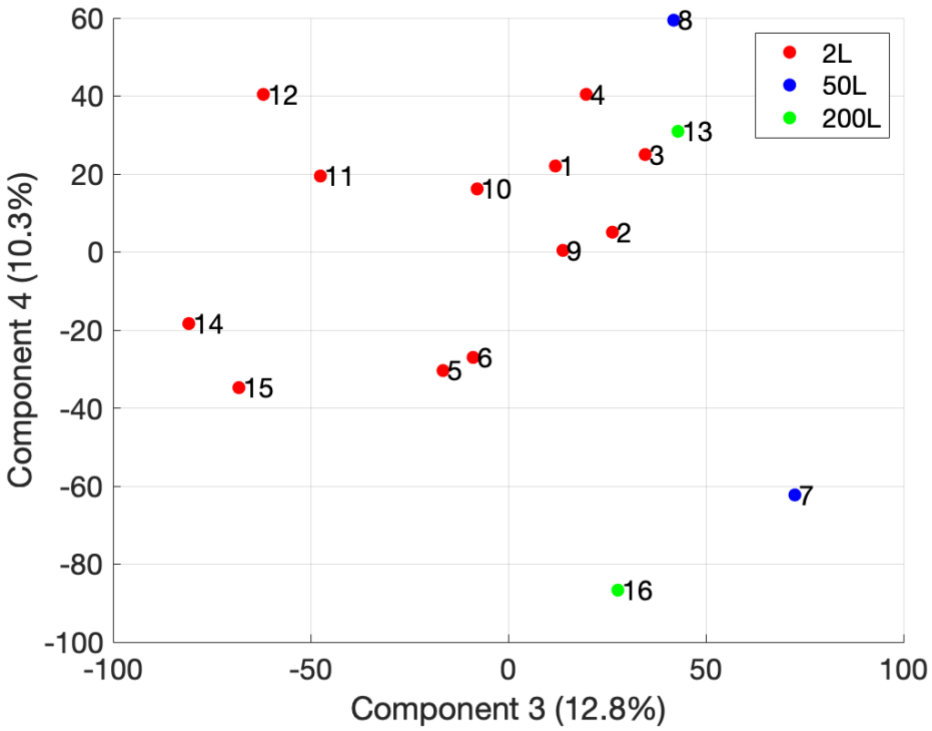


Figure 7-6: Principal component 3 versus principal component 4, explaining 12.8% and 10.3% of the variance respectively. The marker colours indicate the batch volume, which varied between the 2L, 50L and 200L scales.

The largest contributors to component 3, as shown in Figure 7-7, are the air overlay and sparge flowrates, which have large positive loadings and correlate positively to one another. Other key contributors were the osmolality, throughout the process, and the cell viability and specific concentrations of glucose, glutamate and lactate, early in the process. The gas flowrates are the main variables causing the 50L and 200L batches to score high on component 3. These variables were already listed in Table *7-4*, as they were found to be driving the separation of the 50L and 200L batches on components 1 and 2. The large separation of the 2L batches on component 3 shows that there is also significant variation in component 3 key variables within the 2L batches. The main contributors to component 4, as shown in Figure 7-7, were the vessel volume, cell viability, and specific concentrations of glutamine and lactate. It is interesting that the cell viability late in the process correlates positively with the specific glutamate and lactate concentrations at the mid-point of the process. These parameters are a source of variability that is present within each scale.

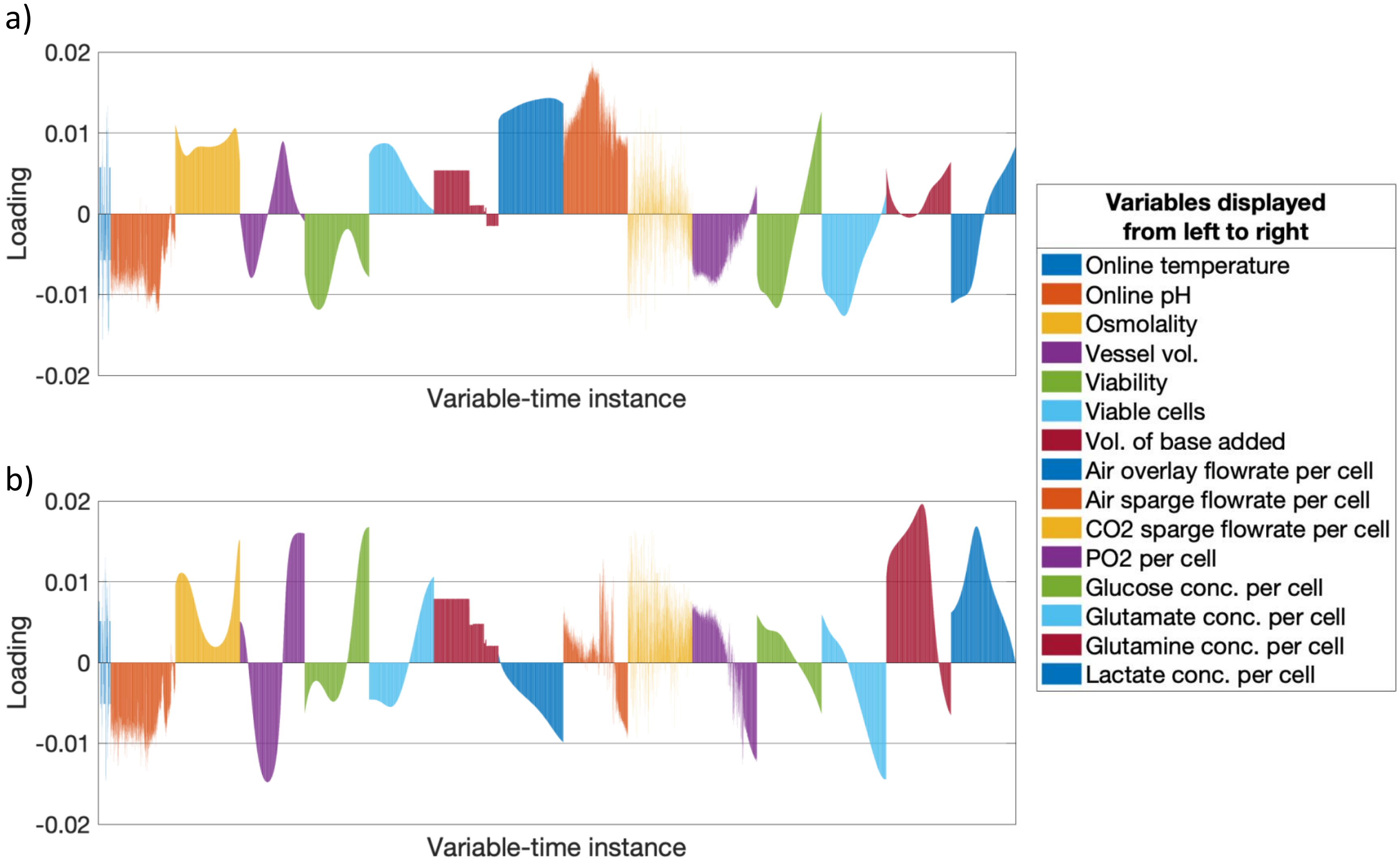


Figure 7-7: Loadings on component 3 (a) and loadings on component 4 (b) for each of the variable-time instances in the unfolded time-series data. Variable-time instances belonging to the same variable have been coloured the same and repeated colours may be distinguished as the variables from the legend appear in the plot from left to right.

The scores for principal components 5 and 6, which explain 8.4% and 5.2% of the variance respectively, are shown in Figure 7-8. The 50L and 200L batches are dispersed within the spread of the 2L batches, indicating that components 5 and 6 capture variance that is independent of scale. Batch 11 is separated from the rest of the batches on component 6, indicating that this batch features some unique behaviour.

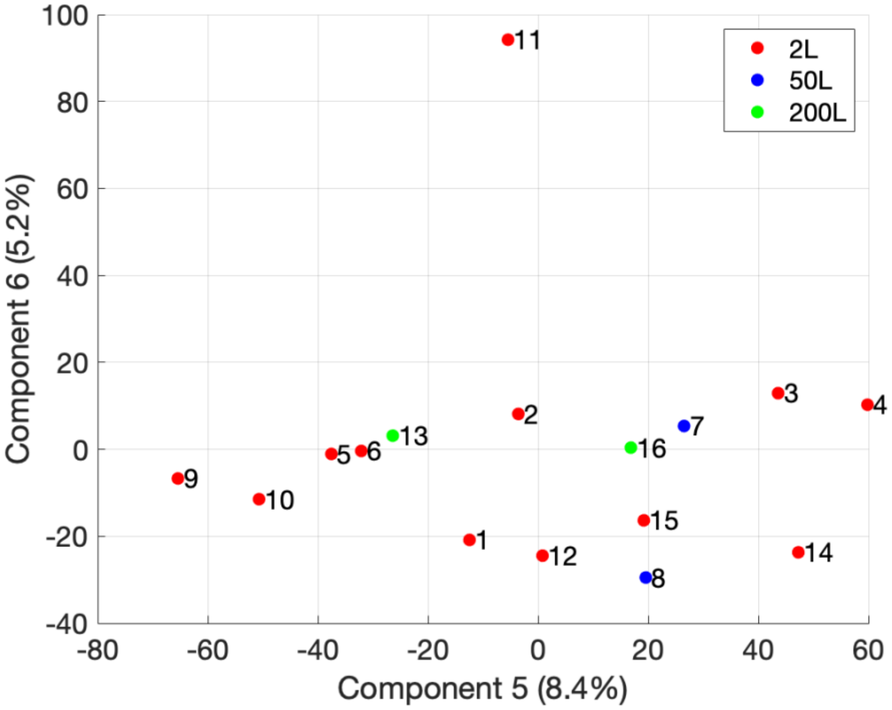


Figure 7-8: Principal component 5 versus principal component 6, explaining 8.4% and 5.2% of the variance respectively. The marker colours indicate the batch volume, which varied between the 2L (red), 50L (blue) and 200L (green) scales.

The loadings for components 5 and 6 are displayed in Figure 7-9. The specific lactate concentration is a key contributor on component 5 and the loadings indicate that the specific lactate concentration late in the process is negatively correlated to the volume of base added. Lactate accumulation lowers the pH of the culture and base addition is sometimes required to maintain the pH at the desired level. Other variables with significant loadings include the vessel volume, which correlated to the specific lactate concentration in the latter half of the process. The lactate accumulation and correlated parameters varied across all batches at all scales. The loadings on Figure 7-8 indicate that batch 11 stands out due to differences in the osmolality, and the CO_2 sparge flowrate and pH late in the process.

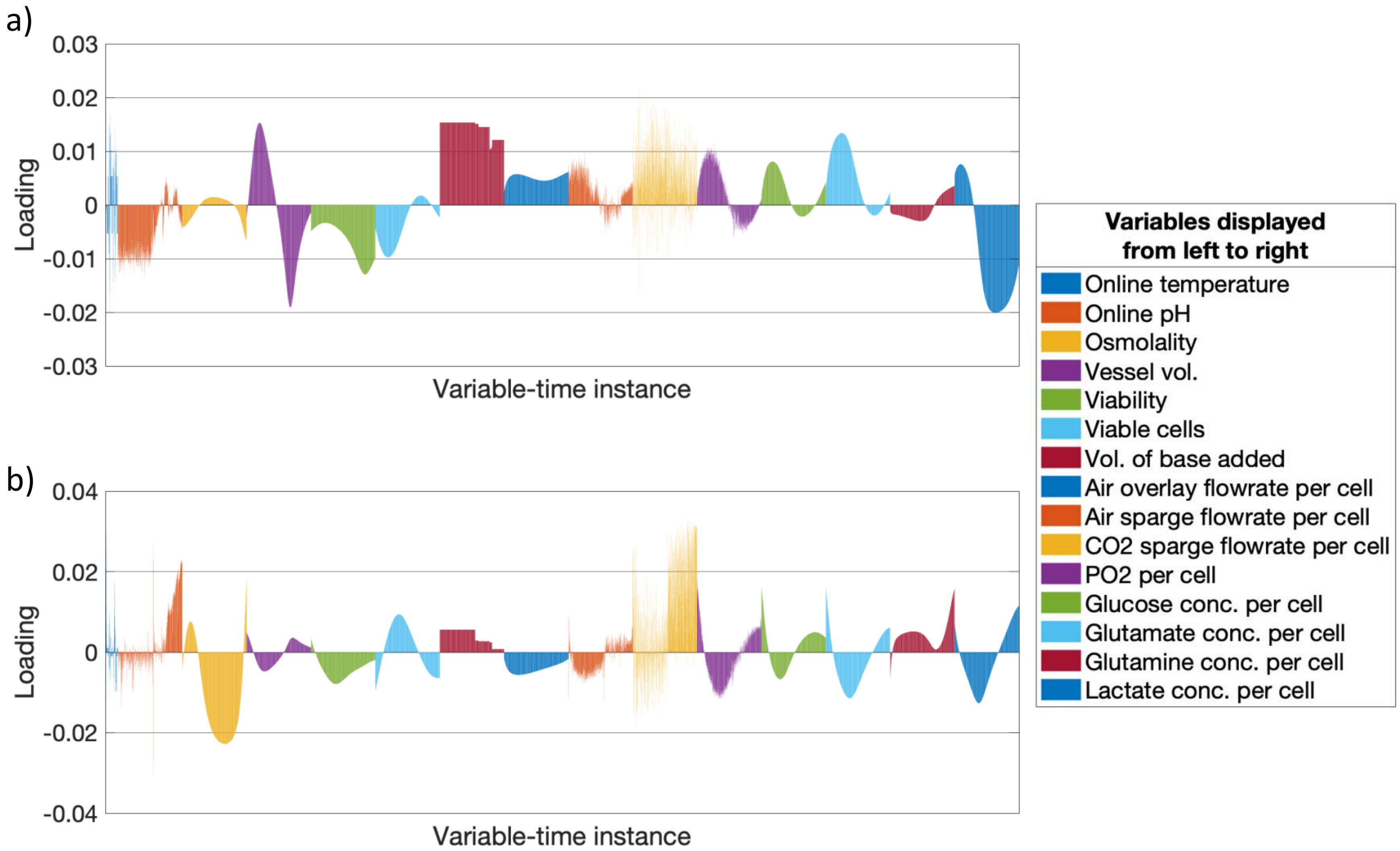


Figure 7-9: Loadings on component 5 (a) and loadings on component 6 (b) for each of the variable-time instances in the unfolded time-series data. Variable-time instances belonging to the same variable have been coloured the same and repeated colours may be distinguished as the variables from the legend appear in the plot from left to right.

7.3.1.1 Summary of PCA

- The first three components explained 60.3% of the variance in the process data. The main sources of variability in the process data, as indicated by the loadings on the first three components, were the culture pH and osmolality, cell viability and concentration, air sparge and overlay flowrates, volume of base added and the specific concentrations of the metabolites. The loadings showed that there were significant correlations between all of these key process variables.
- Components 1 to 3 captured variance that separated the 50L and 200L batches from the 2L batches, mainly due to differences in the air overlay and sparge flowrates, despite the scaling factor that was applied. This may be down to differences in the working volume between the scales; however, the difference was relatively insignificant and the variation present on components 1 to 3 within the 2L scale was more significant than the variation between the different scales, as shown by the spread of scores.
- Components 4 and 5 collectively explained 13.6% of the variance in the process data. Components 4 and 5 captured variance that was independent of scale. Component 4 demonstrated correlation between the vessel volume, cell viability and the specific concentrations of glutamine and lactate. Component 5 showed that there were differences in the accumulation of lactate and consequent base addition.

7.3.2 Predictive modelling of the infectious titre

This section presents the PLS models that were developed for the prediction of the infectious titre of LVs produced in the bioreactor-based suspension culture process. In order to develop models with optimal predictive performance and minimal complexity, model development was carried out using repeated K-fold cross validation to identify the optimal number of latent components and variable selection was implemented using the VIP selection approach. Model performance was validated through the repeated K-fold cross validation and a test on data that was held out of the model development process.

After the generation of transformed variables, the complete list of variables available for the analysis was 41. Initially pre-selection of the variables was carried through examination of the variable profiles and testing of PLS models on a trial-and-error basis. Through this approach it was possible to rule out 20 of the variables, which negatively impacted PLS model performance. The remaining 21 variables were used to generate a PLS model, which demonstrated good performance in cross validation and testing. This model served as a basis from which to improve model performance and reduce model complexity through variable selection using the VIP selection approach. The performance of the base model and two reduced complexity models is displayed in Table 7-5.

Table 7-5: Performance metrics for three PLS models with varying degrees of complexity, as determined by the VIP threshold value used in variable selection. Performance metrics are provided from the repeated K-fold cross-validation and the model fit to the hold-out test set.

	No of variables	No of latent variables	CV R^2	CV MAE	Test R^2	Test MAE
All variables	21	2	0.86	0.29	0.97	0.24
VIP > 1	20	1	0.90	0.20	0.95	0.27
VIP > 1.4	10	1	0.90	0.16	0.97	0.24

Table 7-5 shows that all three models performed well. The base model with all 21 variables featured two latent components and achieved an R^2 score of 0.86 and a MAE of 0.29 in repeated K-fold cross validation. Additionally, it predicted the viral titre for the two batches in the test set with high accuracy (test R^2 = 0.97, test MAE = 0.24). Given the known large error in the infectious titre assay (35-40%), the models perform very well. This raises the question of model over-fitting; however, the thorough cross-validation and testing has been implemented to avoid this and the metrics suggest that the models are not over-fitting. This also indicates that despite the measurement error, the models are capturing an underlying physical relationship between the process variables and the infectious titre. Initially a VIP threshold of one was applied, as is common practice because one is the average VIP score for all variables in the model. This model resulted in an improvement to the cross-validation performance metrics compared to the base model (R^2 increased by 0.04 and MAE decreased by 0.09) and featured only one latent component. The model demonstrated high accuracy with the test set (test R^2 = 0.95, test MAE = 0.27), although the base model performed marginally better. The VIP threshold of one reduced the number of variables included by one, although many timepoints of the 20 remaining variables were also excluded from the model, leaving only the portions of the time series profiles that correlated to the infectious titre.

To explore further reduction of model complexity, the VIP threshold was increased above one. In the literature, it has been recommended to do so when the portion of predictive variables is small or when there is a high degree of correlation between predictors (Chong and Jun, 2005). The latter is true in this case. A VIP threshold of 1.4 was found to work well, reducing the number of variables in the model down to ten, while maintaining good performance in cross validation and testing (CV: R^2 = 0.9, MAE = 0.16; Testing: R^2 = 0.97, MAE = 0.24). Visual inspection of the model fit to the training data did not reveal any nonlinearities or trends in the errors to suggest that the model requires further adaptions. The fit to the training data for the 10-variable model is displayed in Figure 7-10.

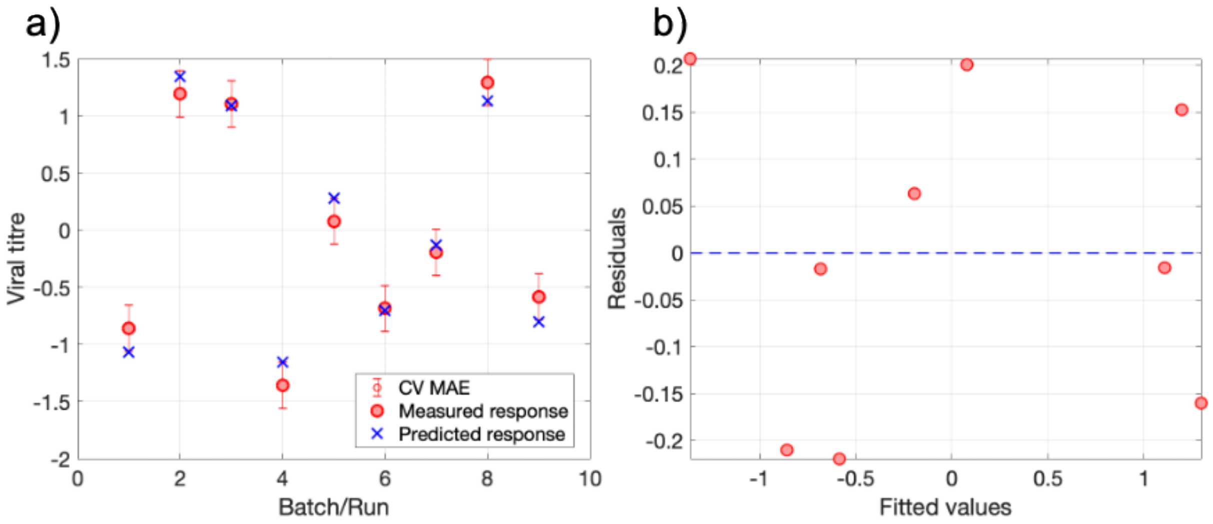


Figure 7-10: The 10-variable reduced complexity model fitted to the training data: a) the predicted viral titre and actual viral titre (scaled) with error bars showing the MAE from repeated K-fold cross validation, b) the regression residuals which were confirmed to be normally distributed.

The fit to the training data gave an R^2 of 0.96 and MAE of 0.14. The errors shown in Figure 7-10 are normally distributed and an Anderson test confirmed that the errors were normally distributed at the 90% confidence level (p = 1x10^{-3}). The 10-variable model retains the most important predictor variables for the infectious titre; although, the model can be reduced further without losing performance as there is a high degree of correlation between the predictors. The 10 variables included in the model are listed in Table 7-6 in order of importance based on average VIP score for each variable, i.e. average VIP across the numerous timepoints that were selected for the model for each variable.

Table 7-6: Variables retained in the reduced model after removing variables with VIP scores less than 1.4. The variables are displayed in descending order of importance based on average VIP score. Note, the average VIP scores shown are less than 1.4 because the reduced model was fitted to the training data and VIP scores were re-calculated.

No.	**Name**	**Avg VIP**	**No.**	**Name**	**Avg VIP**
1	Osmolality	1.06	**6**	Glucose conc. per cell	0.94
2	Specific glucose cons. rate	1.04	**7**	Online pH	0.91
3	Glucose conc.	1.02	**8**	Lactate conc. per cell	0.90
4	Viable cells gradient	0.99	**9**	Glutamate conc.	0.90
5	Specific lactate prod. rate	0.95	**10**	Glutamine conc. per cell	0.89

The PLS model parameters were used to derive standardised regression coefficients for the model. The magnitude of these beta coefficients indicates the relative importance of each variable-time instance and the sign indicates whether the correlation to the infectious titre is positive or negative. Following model development, the models were repeatedly fitted to bootstrap samples to determine confidence intervals for the standardised regression coefficients. Figure 7-11 shows the model coefficient for 9 of the 10 variables listed in Table 7-6. The specific glutamine concentration was left out because it was the least important of the 10. it correlated negatively with the infectious titre late in the process and for the first half of the process the 90% CI on the beta coefficients crossed zero, indicating that correlation during the first half of the process was of low confidence and low importance.

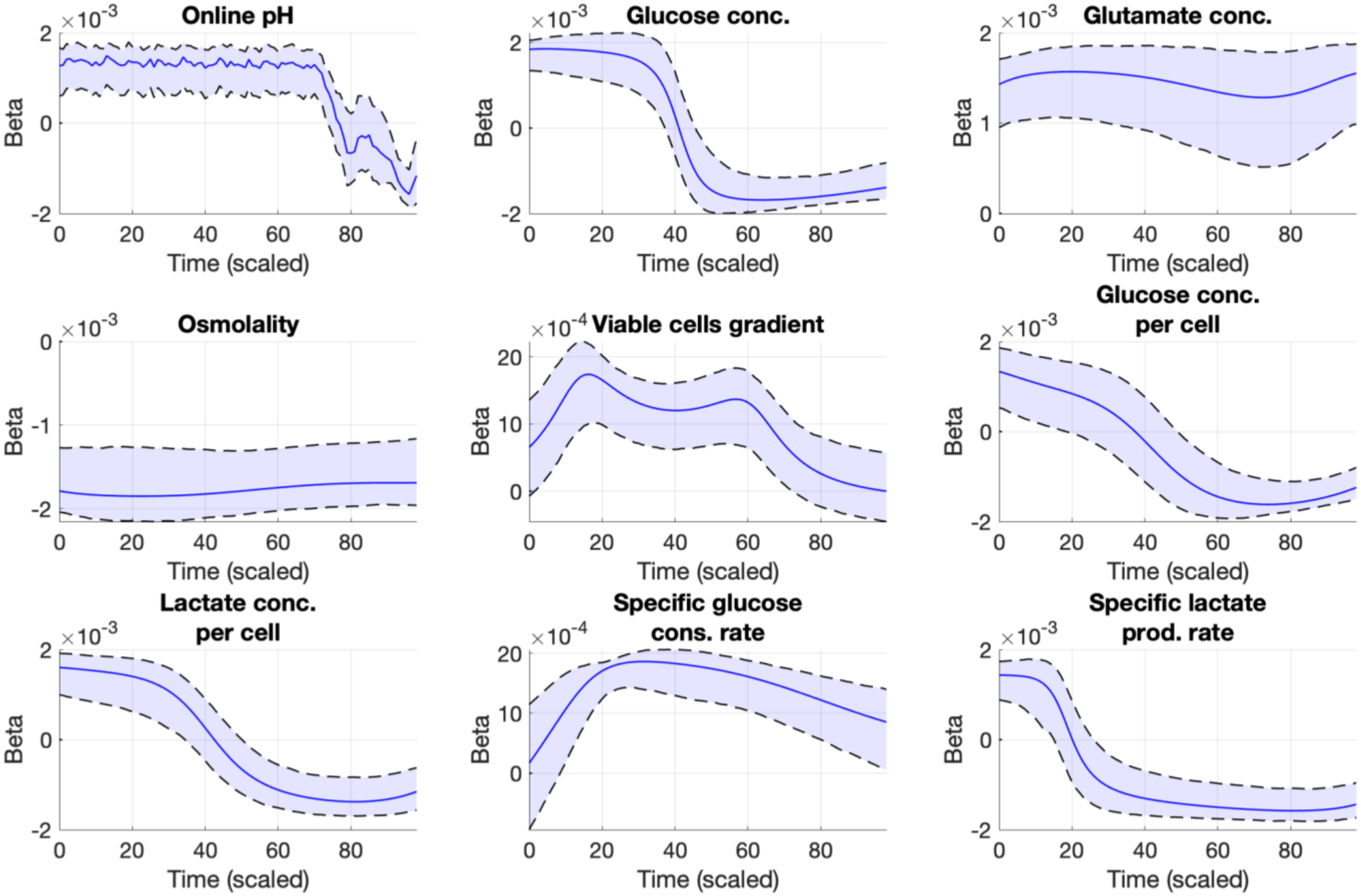


Figure 7-11: Standardised regression coefficients (beta coefficients) for nine key process parameters that were retained in the reduced complexity model after variable selection. The magnitude of the beta coefficent indicates the relative importance of the correlation with the viral titre and the sign indicates whether the correlation is negative or postitive. The beta coefficients for each variable are shown to vary throughout the process due to temporal variation in the original data. The blue central line shows the mean beta coefficient, while the dashed black lines show the upper and lower 90% confidence intervals for the beta coefficients, obtained from fitting the model to 2000 bootstrap samples.

The following section discusses the process variables and their regression coefficients with respect to a scaled timeline, from 0 to 88, in order to keep the real process timings confidential. The online pH correlated positively with the infectious titre from the beginning of the process until around a scaled time of 70, then in the late stages of the process the signs on the coefficients switch and the pH correlates negatively with the infectious titre at the end of the process. The glucose concentration is initially positively correlated to the infectious titre and rapidly becomes negatively correlated to the infectious titre at a scaled time of 38. The glutamate concentration is positively correlated with the infectious titre throughout the process. Conversely, the osmolality of the media is negatively correlated with the infectious titre throughout. The viable cells gradient is an important predictor between the scaled time period 10 and 60. During this period a high cell growth rate correlated positively to the infectious titre.

The beta coefficients for the specific glucose concentration were initially positively correlated to the infectious titre before becoming negatively correlated at a scaled time of 20, similar to the beta coefficients for the glucose concentration. The confidence intervals are wide in places, partly due to the increased error in the specific concentrations as a consequence of combining two offline variables, which were interpolated. The beta coefficients for the specific lactate concentration follow a similar trend. Initially they are positively correlated to the infectious titre before becoming negatively correlated at a scaled time of 40.

The specific glucose consumption rate is positively correlated with the infectious titre throughout most of the process; however, this correlation peaks between a scaled time of 10 and 50. The specific lactate production rate was initially positively correlated to the infectious tire and rapidly becomes negatively correlated at a scaled time of 18. In order to analyse the relationship between the process variables and the infectious titre, it was also important to understand the correlation between the predictor variables. For this purpose, four timepoints were selected across the duration of the process and the correlation between predictor variables at these timepoints was evaluated with a Pearson's correlation matrix. The four scaled timepoints: 5, 35, 55, and 95, were given the labels early, mid-early, mid-late and late, respectively, for convenience in the discussion. To display the correlations between the key predictor variables, Figure 7-12 shows a heatmap of the Pearson's correlation matrix, where the colours blue and red represent negative and positive correlations, respectively.

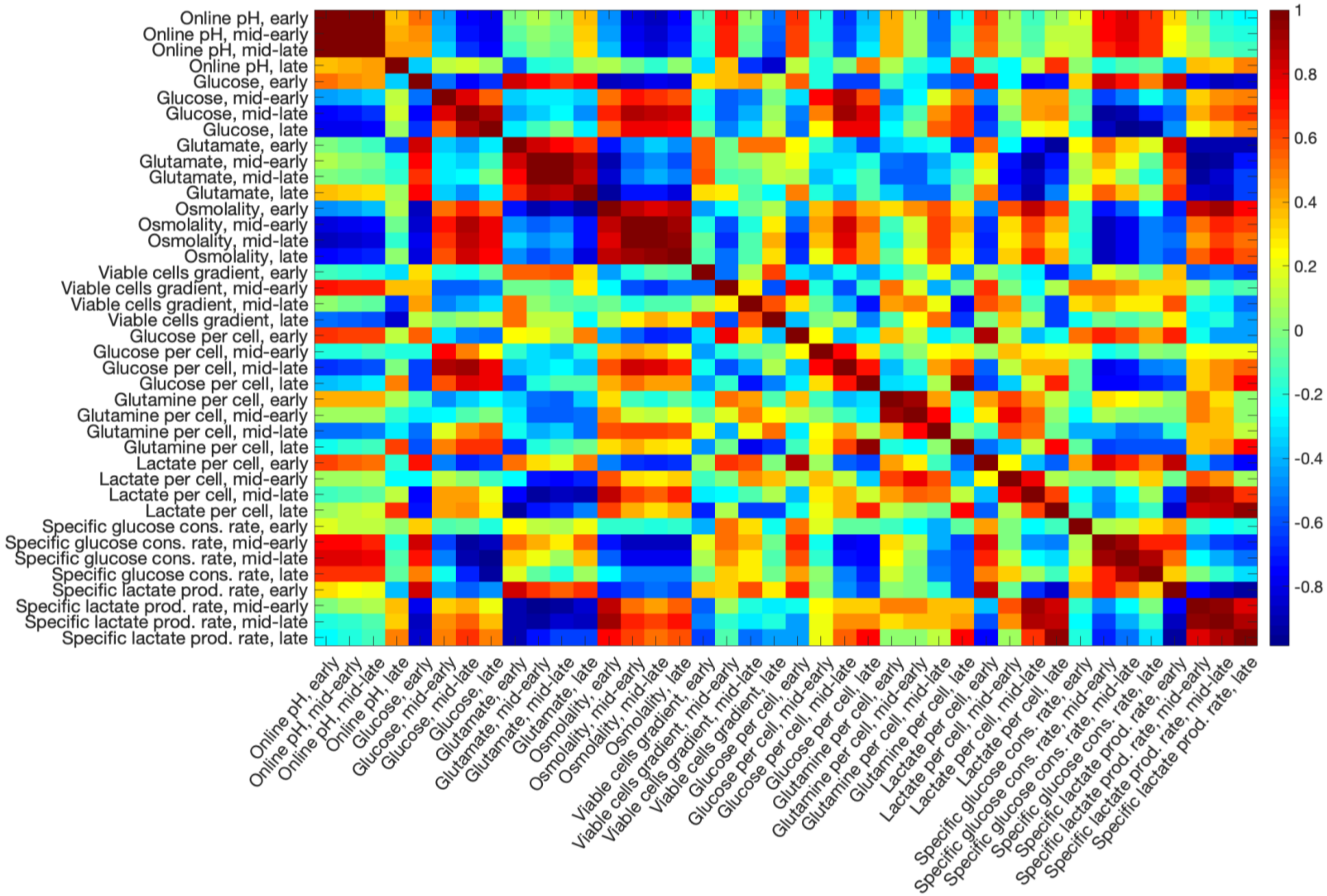


Figure 7-12: Heatmap of a correlation matrix containing the Pearson's linear correlation coefficients between key predictor variables. Each variable has been included at four timepoints throughout the process; early (5), mid-early (35), mid-late (55) and late (95). The colour bar on the right shows that significant negative correlations are coloured blue, while significant positive correlations are coloured red.

The heatmap (Figure 7-12) reveals significant correlations between all of the key predictor variables, in some cases these correlations are time dependent. High glucose consumption rate at the mid-early stage, where it was found to be most important, correlated positively with the specific lactate (R = 0.79) and glucose (R = 0.66) concentrations early, the online pH up to the mid-late stage (R = 0.73) and the glucose concentration early (R = 0.83). Variables negatively correlated to the glucose consumption rate mid-early included the specific lactate production rate mid-late to late (R = -0.66), the specific glucose concentration mid-late to late (R = -0.78), the osmolality throughout (R = -0.87) and the glucose concentration mid-early to late (R = -0.76). The aforementioned correlations represent important correlated features in the process data that are predictive of the infectious titre. Based on the experimental conditions that were observed in the dataset that was provided, the optimal conditions for high infectious titre may be described by the following process features:

- Low osmolality throughout the process
- High specific glucose consumption rate, particularly in the early to middle phase
- High specific lactate production rate initially, up to a scaled time of 20, and low specific lactate production rate thereafter
- High cell growth rate up to the middle point of the process
- High pH up until a scaled time of 70, then pH should drift low
- High specific glucose and specific lactate concentrations early in the process
- Low glucose concentration and low specific glucose concentration late in the process
- High glutamate concentration throughout the process
- Low glutamine concentration at the end of the process

The descriptions ‘high’ and ‘low’ apply to the range that was observed in the data that was used to train the PLS model. The following section highlights observations from the literature relevant to the model findings, including some insights into potential influential variables and mechanisms that could explain the differences in process performance between the batches.

Osmolality – The osmolality is an important parameter of the culture in viral vector production (Coroadinha et al., 2006; Shen and Kamen, 2012). Shen and Kamen (2012) investigated the effect of osmolality on the production of adenoviral vectors using HEK 293 cells. The results showed that the growth of cells under hyperosmotic conditions was favourable and that the osmotic pressure should be reduced for the virus production phase for optimal titres. However, both hypo and hyper osmotic stresses were found to improve viral productivity. Additionally, it was observed that the optimal osmolality for cell expansion and preparation for infection differed from the optimal osmolality for virus production. Higher osmolality (370 mOsm or greater) was favoured for the cell expansion phase and a lower osmolality (290 mOsm) was optimal for the virus production phase.

Coroadinha *et al.* (2006) found that increased osmotic pressure was beneficial for the production yield and the stability of a Moloney murine leukaemia virus, of the retrovirus family. To manipulate the osmolality of the media, the authors compared the use of NaCl, sorbitol and fructose, and it was found that the viral productivity was dependent upon the osmotic agent chosen. The authors concluded that a balance must be struck between cell yield, viral productivity and retroviral stability. They did not investigate the effect of varying the osmolality between the cell expansion and viral production phases.

Another study attributed a loss of viral infectivity in downstream processing to damage caused by high osmotic pressure (Zimmermann et al., 2011). Literature sources refer to the culture osmolality as a critical attribute, and it is measured in quality control (QC) analysis of the viral vector product (Merten et al., 2010). From the literature, it is evident that the culture osmolality is a key parameter in transient virus production processes, which has been found to impact cell expansion, viral productivity and the stability of virions. However, the existing literature indicates that the optimal osmolality is dependent on numerous factors including the type of producer cells, the virus type, the osmotic agents and the phase of the process, as cell expansion and virus production phases often require different conditions (Petiot et al., 2015; Shen and Kamen, 2012).

pH – Holic *et al.* (2014) produced vesicular stomatitis virus- glycoprotein (VSV-G) pseudotyped lentiviral vectors using transient transfection of HEK293T cells and tested the impact of pH on LV production. The pH was varied between 6 and 8, and it was found that infectious and physical titres were increased by two to threefold at pH 6 compared to neutral. Valkama *et al.* (2018) produced LVs with transient transfection of adherent 293T cells in a fixed-bed bioreactor. Referencing the findings of Holic *et al.*, the authors decided to reduce the pH from 7.2 down to 7 after PEI-mediated transfection for some of the runs. The runs with the lowered pH demonstrated the highest titres that were achieved, and it was also found to increase the ratio of functional LVs to p24 protein. In control flasks without pH control, the pH drifted to values lower than 7 by the end of the process, which indicated the plausibility of decreasing the pH in the bioreactor.

In the data that was analysed in this work, the pH was controlled within a dead band with a range of 0.2, and the drift towards low pH at the end of the process occurred due to changes in the media composition. The model indicated that high pH was favourable up to a scaled time of 70 and low pH was favourable thereafter. Literature sources indicate that the pH may be a critical process parameter impacting the LV infectious titre.

Specific glucose consumption and lactate production rates – Glycolysis and glutaminolysis are the main sources for energy production in HEK 293 cells (Petiot et al., 2015). The specific glucose consumption rate is indicative of the metabolic state of the cells, which is known to be highly important for process performance, as the energetic state of cell has been linked to viral productivity (Petiot et al., 2015). With HEK293 SF cells, the efficiency of glucose consumption has been increased using low protein media, which resulted in a 3 to 4-fold increase in adenovirus cell productivity (Nadeau et al., 2002).

The lactate production rate is linked to the glucose consumption rate and depends upon the relative activity of the numerous metabolic pathways. Figure 7-12 showed that the glucose consumption rate and lactate production rate correlated positively to a degree. The activity of the metabolic pathways is also dependent on the composition of the media. The concentrations of metabolites, such as glucose and lactate can influence the metabolic pathways (Merten et al., 2001; Petiot et al., 2015). Some cell lines have been shown to switch from lactate production to lactate consumption during the culture and cells consuming lactate were shown to have up to 6 times greater energy efficiency than lactate producing cells (Petiot et al., 2015). HEK 293 cells used for viral vector production have been reported to shift from lactate production to lactate consumption (Le Ru et al., 2010; Nadeau et al., 2000).

It is clear that the metabolic activity and energetic state of the cells is highly important for viral vector production, and that these factors are impacted by the media composition, including the concentration of key metabolites, such as glucose, glutamine and lactate. The PLS model showed that a high glucose consumption rate was favourable throughout the process, particularly in the early to middle phase, where the regression coefficients peaked. It also showed that lactate production was favourable initially, but quickly transitioned to a negative correlation with the infectious titre; indicating that at the time of transfection high glucose consumption and low lactate production was favourable. The ratio of glucose consumption to lactate production is related to the efficiency of glucose utilization, and higher efficiency has previously been linked to higher cell productivity, which is consistent with the model regression coefficients beyond 10 hours into the process.

Glucose and lactate concentrations – As previously discussed, the composition of the media can significantly impact viral vector production. The concentrations of key metabolites can influence the activity of metabolic pathways and the efficiency of ATP production, which have previously been linked to viral vector productivity. The relationship between metabolite concentrations and the rate of their production or consumption is relatively complex because of their interdependency. Glucose and glutamine are key metabolites utilised for adenosine triphosphate (ATP) production in HEK 293 cells, henceforth, it is important that these metabolites do not become depleted. In the fed-batch bioreactor process, the concentration of glucose is monitored, and glucose is fed into the reactor if the concentration decreases below a pre-defined threshold. Glutamine is provided by a slow release compound added to the media.

Additionally, the accumulation of lactate and ammonia at high concentrations leads to toxic effects, such as inhibition of cell growth, changes to intracellular pH and cell apoptosis (Merten et al., 2001; Petiot et al., 2015). However, small amounts of lactate and ammonia added to cell cultures has previously been reported to increase specific viral productivity compared to cultures with no lactate or ammonia added (Petiot et al., 2015).

Cell concentration and growth rate - The cell growth rate was found to be an important predictor variable and correlated positively with the infectious titre, while the cell concentration was not found to be an important predictor variable. As previously mentioned, cells in the log phase are in the optimal state for viral productivity, which is consistent with the model findings (Petiot et al., 2015). This correlation is not always found with other types of suspension culture and viral productivity in Lentiviral vectors has been shown to be negatively affected if the cell concentration exceeds a critical limit. However it appears that this critical threshold is not crossed in the process data that is under study (Nadeau et al., 2002; Petiot et al., 2015). In the complex model that featured 21 process variables, the cell concentration was found to correlate positively with the infectious titre; however, the VIP scores were below 1. This may be because the cell concentration was within a good range for all of the batches and therefore did not produce a significant effect on the titre. Alternatively, the metabolite concentrations may be acting as an inferential measurement that is representative of the cell concentration, due to the metabolite concentrations being more accurate measurements than the cell concentration. It was reported that determination of the cell concentrations was sometimes affected by aggregation of the cells, thus decreasing the accuracy and reliability of the measurement. This also applies to the cell growth rate, which was determined numerically from the polynomial interpolation of the cell concentration data.

7.3.2.1 Additional sources of variability

There is the possibility that key predictor variables for the LV infectious titre are acting as surrogate variables for other parameters that are not directly present in the dataset. Additional factors that were not part of this analysis include the material lots for cell culture media, substrates, producer cells, plasmids and reagents. Furthermore, there was no information included from the seed train processing steps. Several authors have acknowledged the importance of tightly controlled conditions throughout the seed train before the cells are inoculated into the bioreactor (Glassey et al., 2011; Streefland et al., 2013). This is because the seed train has been shown to represent a key source of variability, where small changes in conditions can impact the condition of the cells, leading to variability that is translated through the rest of the process. The seed train is a likely source of the variability occurring in the bioreactor-based LV production process and in future work it would be beneficial to work include data from the seed train in the analysis.

7.3.2.2 Concluding remarks

Interestingly many of the aforementioned variables have been shown to influence viral productivity in previous studies; however, in the data that was analysed in this work, these parameters were highly correlated to one another, as well as to the infectious titre. It is likely that some of these parameters have a cause-and-effect relationship with the viral titre, although it is difficult to know which parameter(s) is/are the root cause, due to the high degree of correlation. Some of the correlations between the predictor variables may have occurred by chance, while others are known to be physically linked, directly or indirectly. For example, the specific glucose concentration and the non-specific glucose concentration have a simple relationship and can be expected to correlate to a high degree. Additionally, the glucose consumption rate in the middle part of the process can be expected to correlate negatively with the glucose concentration at the end of the process, provided that glucose feeding was consistent between batches. These relationships are well understood by process experts. However, the impact of culture pH and osmolality on the metabolic activity of the cells is less clear and both are linked to the concentration of metabolites in the media. Henceforth, decoupling the physical mechanisms and cause-and-effect relationships within the data is relatively complex.

7.3.2.3 Prediction of 50L and 200L infectious titre using model trained with 2L batches

The models that were previously presented were trained on 2L, 50L and 200L batches, and then tested on 2L batches, which was determined by the Kennard-Stone algorithm used to partition the data. The successful model development using all three scales demonstrated that there is comparability between the scales with respect to the key process variables that influence the infectious titre. To confirm this observation, an experiment was carried out where the 2L batches were used to train the PLS model and it was then used to predict the viral titre of the 50L and 200L batches. Figure 7-13 shows the model fit to the test data, which consists of two 50L batches and two 200L batches. The model explained approximately 82% of the variation in the infectious titre for the 50L and 200L batches (R^2 = 0.82) and the MAE was low at 0.35. The results show that performance at the 2L scale can be used to predict performance at the 50L and 200L scale, which is further evidence that the processes are comparable. Process comparability is highly important for process development and increases the value of experimentation at low scale, which is preferable to conducting experiments at large-scale. Furthermore, if GMP standard LVs are to be produced at a range of scales, it is crucial that the product is consistent irrespective of the production volume.

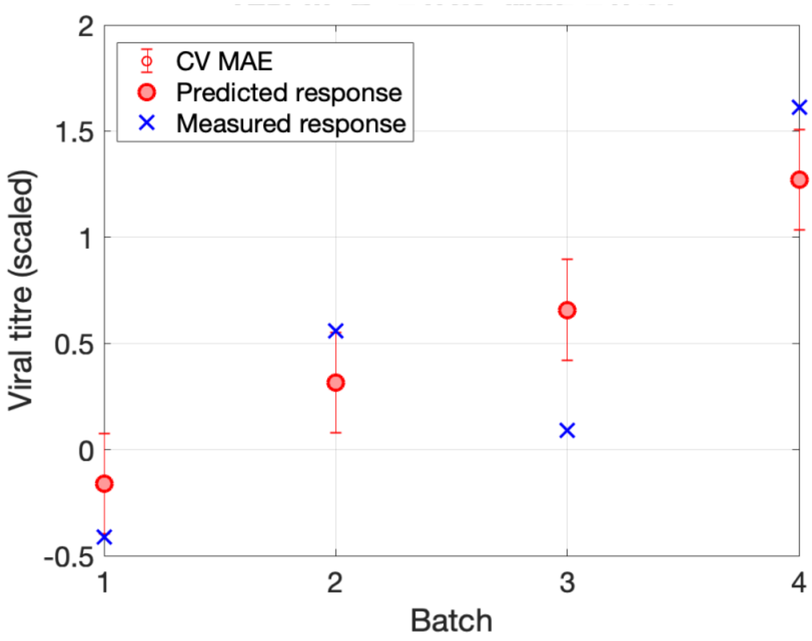


Figure 7-13: PLS model fit to the 50L and 200L batches when trained on the 2L batches. The R^2 and low MAE values were 0.82 and 0.35, respectively, demonstrating that the model can predict the infectious titre of the 50L and 200L batches with high accuracy.

7.4 Conclusions

The bioreactor-based suspension culture process is a scalable manufacturing solution for the production of LVs. In this chapter, the process variability and comparability between the 2L, 50L and 200L scales was investigated with PCA and the relationships between process variables and the infectious titre was explored with PLS regression. The PCA model showed that there was a high degree of similarity between the 2L batches and the 50L and 200L batches and that the most significant variation that was present occurred within each scale. This was most noticeable for the 2L batches, as these were the most numerous. The first 5 principal components captured variation on a batch-to-batch basis, resulting in a relatively even distribution of the batch scores. Small differences observed between the scales was mainly due to differences in air overlay and sparge flowrates. Other key parameters such as the specific metabolite concentrations, culture osmolality, pH and viable cell concentrations were comparable between the three scales. A high degree of correlation was observed between key process parameters including the culture osmolality and pH, the cells concentration and viability and the specific concentration of oxygen, glucose, glutamate, glutamine and lactate.

Predictive models were developed for the infectious titre and the physical titre of the LV product; however, only the infectious titre model demonstrated good model fit and predictive performance. The poor performance of the physical titre PLS model may be due to influential factors that are not present in the dataset or it may be the case that the variation present in the physical titre data is low with respect to the error of the physical titre measurement. Fortunately, the infectious titre is the one that contains more critical information because it is a measure of the number of fully infectious virions that are produced, rather than both noninfective and infective. The infectious titre was modelled successfully with a 10-variable reduced complexity PLS model demonstrating good predictive performance (R^2 = 0.9, MAE = 0.16) in repeated 5-fold cross validation (2000 repeats) and in testing with a hold-out set (R^2 = 0.97, MAE = 0.24). This model was developed using batches from the 2L, 50L and 200L scales, further demonstrating that comparability between the scales was good. Additionally, it was possible to generate the PLS model by training on the 2L batches and to use this model to successfully predict the outcome of the 50L and 200L batches (R^2 = 0.82, MAE = 0.35).

The key predictor variables for the infectious titre were found to be the pH, osmolality, cell growth rate, the specific concentrations of the metabolites and their rates of change in concentration in the media. Literature sources indicate that many of these key process parameters have previously been found to influence the performance of transient viral vector production processes. As indicated by the PCA model and confirmed in the predictive modelling activity, there was significant correlation between these key process parameters. Due to the high degree of correlation in the dataset, it is difficult to gain a clear understanding of which parameters are the root-cause, driving the differences in the infectious titre. Process experts are best placed to understand the correlations between predictor variables and to interpret the relationships between predictor variables and the viral titre.

It is likely that further experiments are necessary to evaluate the relationships between the influential variables identified in the PLS model and the LV infectious titre. Such experiments may prove to be extremely valuable for process and product development and would lead to increased understanding of process behaviour. The findings of this predictive modelling activity have identified likely candidates, which are influencing process performance. This information may be used to guide future experiments to gain a deeper understanding of the physical relationships and to confirm the route to process optimisation. Through leveraging information contained in historical process data, the modelling activity will thereby have provided an efficient pathway for process optimisation and development.

Chapter 8

Conclusions and future work

In this work, multivariate data analysis techniques were applied to historical process data from the manufacture of LVs, in adherent and suspension cell cultures, and from cell drug product manufacturing. The key manufacturing challenges for each of these three processes was described, the main aspects being high levels of variability in materials and production methods, and a lack of advanced process knowledge due to processes and products being in relatively early stages of development. In the chemical and biochemical process industries, MVDA and ML techniques have been widely exploited for the development and optimisation of manufacturing processes, including for increased efficiency and sustainability. To date, there have been very few publications and little publicly available information on the application of MVDA or ML in cell gene therapy manufacturing. Although, several publications have described the need for such techniques in cell and gene therapy manufacturing.

Here MVDA was leveraged to produce beneficial insights into the behaviour of viral vector and cell drug product manufacturing processes. The datasets from the adherent cell culture process for LV production and cell drug product manufacturing were both cross-sectional datasets, with two dimensions: variable versus batch/observation. Both datasets featured high dimensionality with numerous process variables and relatively few batches, which was challenging for model development. The use of latent variable methods was key to overcoming high dimensionality and multicollinearity within the data. In order to further simplify the interpretation of principal components, linear and nonlinear programming approaches to sparse PCA were developed. The techniques presented relatively simple and more accessible alternatives to sparse PCA approaches presented in the current literature. After successfully demonstrating their performance on benchmark datasets, the mixed integer nonlinear programming approach to sparse PCA was applied to the two cross-sectional datasets to carry out feature extraction with simplified principal components.

In viral vector production with the adherent cell culture process, key areas of process variability were identified, for example, the cell expansion phase of the process was found to contribute significantly to process variability with differences in the production volume, total cell count and cell concentrations, which translated to variability in the downstream process. The models provided information on the major sources of variability, which could be targeted in efforts to reduce overall process variability. Predictive modelling identified likely critical process parameters, such as the sterile filter area, which had significant correlations with to the infectious titre and/or infectivity of the LV product.

Process data from manufacturing of the final cell drug product was also analysed, where sources of variability and correlations between process variables were observed. The LV copy number and percentage of CD34+ cells in the cell drug product were modelled using variables from LV production (adherent cell culture) and from cell drug product manufacturing as model inputs. Interestingly, a few parameters from LV manufacture, such as the temperature of refrigeration during clarification and the volume of Ca/Mg added to the LV product formulation, were found to be significant predictors of the LV copy number. This demonstrated the ability to model the relationship between process parameters and CQAs across numerous unit operations. The CD34+ cells percentage and LV copy number were both predicted with high accuracy and likely critical process parameters were identified.

Bioreactor-based suspension culture processes are far more scalable than adherent cell culture processes, henceforth they are likely to become more widely adopted for GMP manufacture of LVs in the future. Here, online and offline process variables from the bioreactor production of LVs were analysed with MVDA. The 3D data, with batch, time and variable dimensions presented new challenges compared to the cross-sectional data. The alignment of temporal profiles of the numerous batches was a key challenge and alignment based on the time transfection was found to produce a coherent dataset suitable for feature extraction and predictive modelling. Through application of PCA, batches at the 2L, 50L and 200L scales were found to be highly comparable in terms of key the process parameters. Predictive modelling revealed significant correlations between 11 key process variables and the infectious titre. The most important predictor variables included the culture osmolality, specific glucose consumption rate, glucose concentration, cell growth rate and the specific lactate production rate. It was found to be possible to train a PLS model on the 2L batches and predict the infectious titre resulting from 50L and 200L batches, based on the 11 key process variables.

Overall the models provided insights into process variability, comparability and the relationships between process parameters and CQAs of the viral vector and cell drug products. Some findings had already been noticed by process experts, for example the sterile filter area used in downstream processing of the LVs was already recognized as a likely critical process parameter, and the modelling work supported this observation, which was a useful confirmation for GSK. Other findings highlight CPP-CQA relationships that were previously not recognized and so here the models have identified relationships that expand process understanding and have potential to be optimised to improve process performance. In the development of predictive models, model validation, testing, ***Y***-permutations and bootstrap sampling were carried out to evaluate the significance of the relationships identified and to test for the presence of chance correlations. This was designed to provide confidence in the models identified; however, in some cases it will still be necessary to validate the findings of the models through further experiments. This is particularly important because the number of batches available was low and there were few repeats of process conditions. Nevertheless, the model findings have provided GSK with direction by highlighting process-parameters that have a potential impact on CQAs, which can be investigated further with a DoE. If the relationships are confirmed, then the process parameters can be optimised leading to impactful improvements to process and product.

In addition to experiments to validate the model findings, in future, it would be interesting to apply predictive modelling techniques to a greater number of product CQAs. For example, it could be beneficial to model the transduction efficiency in the CDP manufacturing process or to investigate whether there is any link between manufacturing variables and the levels of impurities in the viral vector and cell drug products. The modelling approaches carried out in this work could be used to guide the methodology. Furthermore, it would be interesting to investigate the application of a wider range of MVDA/ML techniques to see if there are any additional benefits and insights that can be gained. PCA and PLS were chosen in this work because of their relatively simple model structure, which allows interpretation of variable relationships. Other interpretable models that could be explored include K-means clustering for feature extraction or regression trees for predictive modelling.

The linear and nonlinear programming approaches to sparse PCA that were developed in chapter 4 could be modified and transferred to a predictive modelling technique, such as PLS regression. Regularisation of predictive models is an important area of research, since regularisation offers an alternative to variable selection techniques and simplifies the process by combining the model fitting and variable selection procedures into one process.

Chapter 9

References

Abdi, H., 2010. Partial least squares regression and projection on latent structure regression (PLS Regression). WIREs Comput. Stat. 2, 97–106. https://doi.org/10.1002/wics.51

Aiuti, A., Cattaneo, F., Galimberti, S., Benninghoff, U., Cassani, B., Callegaro, L., Scaramuzza, S., Andolfi, G., Mirolo, M., Brigida, I., Tabucchi, A., Carlucci, F., Eibl, M., Aker, M., Slavin, S., Al-Mousa, H., Al Ghonaium, A., Ferster, A., Duppenthaler, A., Notarangelo, L., Wintergerst, U., Buckley, R.H., Bregni, M., Marktel, S., Valsecchi, M.G., Rossi, P., Ciceri, F., Miniero, R., Bordignon, C., Roncarolo, M.-G., 2009. Gene therapy for immunodeficiency due to adenosine deaminase deficiency. N. Engl. J. Med. 360, 447–458. https://doi.org/10.1056/NEJMoa0805817

Andersen, C.M., Bro, R., 2010. Variable selection in regression—a tutorial. J. Chemom. 24, 728–737. https://doi.org/10.1002/cem.1360

Ansorge, S., Lanthier, S., Transfiguracion, J., Henry, O., Kamen, A., 2011. Monitoring lentiviral vector production kinetics using online permittivity measurements. Biochem. Eng. J. 54, 16–25. https://doi.org/https://doi.org/10.1016/j.bej.2011.01.002

Arbuthnot, P., 2015. Chapter 4 - Viral Vectors for Delivery of Antiviral Sequences, in: Arbuthnot, P. (Ed.), Gene Therapy for Viral Infections. Academic Press, Amsterdam, pp. 95–126. https://doi.org/https://doi.org/10.1016/B978-0-12-410518-8.00004-1

Arlot, S., Celisse, A., 2010. A survey of cross-validation procedures for model selection. Stat. Surv. 4, 40–79. https://doi.org/10.1214/09-SS054

Ausubel, L.J., Hall, C., Sharma, A., Shakeley, R., Lopez, P., Quezada, V., Couture, S., Laderman, K., McMahon, R., Huang, P., Hsu, D., Couture, L., 2012. Production of CGMP-Grade Lentiviral Vectors. Bioprocess Int. 10, 32–43.

Bauler, M., Roberts, J.K., Wu, C.-C., Fan, B., Ferrara, F., Yip, B.H., Diao, S., Kim, Y.-I., Moore, J., Zhou, S., Wielgosz, M.M., Ryu, B., Throm, R.E., 2019. Production of Lentiviral Vectors Using Suspension Cells Grown in Serum-free Media. Mol. Ther. Methods Clin. Dev. 17, 58–68. https://doi.org/10.1016/j.omtm.2019.11.011

Behrens, J.T., 1997. Principles and Procedures of Exploratory Data Analysis. Psychol. Methods 2, 131–160. https://doi.org/10.1037/1082-989X.2.2.131

Bethesda, MD, National Institutes of Health, U.S., 2016. Use of Genetically Modified Stem Cells in Experimental Gene Therapies [WWW Document]. Natl. Institutes Heal.

Biancolillo, A., Marini, F., 2018. Chemometric Methods for Spectroscopy-Based Pharmaceutical Analysis. Front. Chem. 6, 576. https://doi.org/10.3389/fchem.2018.00576

Boeckle, S., Wagner, E., 2006. Optimizing targeted gene delivery: chemical modification of viral vectors and synthesis of artificial virus vector systems. AAPS J. 8, E731-42. https://doi.org/10.1208/aapsj080483

Boztug, K., Schmidt, M., Schwarzer, A., Banerjee, P., Díez, I., Dewey, R., Böhm, M., Nowrouzi, A., Ball, C., Glimm, H., Naundorf, S., Kühlcke, K., Blasczyk, R., Kondratenko, I., Maródi, L., Orange, J., Kalle, C., Klein, C., 2010. Stem-Cell Gene Therapy for the Wiskott-Aldrich Syndrome. N. Engl. J. Med. 363, 1918–1927. https://doi.org/10.1056/NEJMoa1003548

Bradley, P.S., Mangasarian, O.L., 1998. Feature Selection via Concave Minimization and Support Vector Machines, in: Proceedings of the Fifteenth International Conference on Machine Learning, ICML ’98. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 82–90.

Bro, R., Smilde, A.K., 2014. Principal component analysis. Anal. Methods 6, 2812–2831. https://doi.org/10.1039/C3AY41907J

Cai, Y., Rodriguez, S., Hebel, H., 2009. DNA vaccine manufacture: scale and quality. Expert Rev. Vaccines 8, 1277–1291. https://doi.org/10.1586/erv.09.84

Cavazzana-Calvo, M., Payen, E., Negre, O., Wang, G., Hehir, K., Fusil, F., Down, J., Denaro, M., Brady, T., Westerman, K., Cavallesco, R., Gillet-Legrand, B., Caccavelli, L., Sgarra, R., Maouche-Chrétien, L., Bernaudin, F., Girot, R., Dorazio, R., Mulder, G.-J., Polack, A., Bank, A., Soulier, J., Larghero, J., Kabbara, N., Dalle, B., Gourmel, B., Socie, G., Chrétien, S., Cartier, N., Aubourg, P., Fischer, A., Cornetta, K., Galacteros, F., Beuzard, Y., Gluckman, E., Bushman, F., Hacein-Bey-Abina, S., Leboulch, P., 2010. Transfusion independence and HMGA2 activation after gene therapy of human β-thalassaemia. Nature 467, 318–322. https://doi.org/10.1038/nature09328

Charrier, S., Dupré, L., Scaramuzza, S., Jeanson-Leh, L., Blundell, M.P., Danos, O., Cattaneo, F., Aiuti, A., Eckenberg, R., Thrasher, A.J., Roncarolo, M.G., Galy, A., 2007. Lentiviral vectors targeting WASp expression to hematopoietic cells, efficiently transduce and correct cells from WAS patients. Gene Ther. 14, 415–428. https://doi.org/10.1038/sj.gt.3302863

Chen, Z., Lovett, D., Morris, J., 2011. Process analytical technologies and real time process control a review of some spectroscopic issues and challenges. J. Process Control 21, 1467–1482. https://doi.org/https://doi.org/10.1016/j.jprocont.2011.06.024

Chiu, S.-J., Ueno, N.T., Lee, R.J., 2004. Tumor-targeted gene delivery via anti-HER2 antibody (trastuzumab, Herceptin) conjugated polyethylenimine. J. Control. Release 97, 357–369. https://doi.org/10.1016/j.jconrel.2004.03.019

Chong, I.-G., Jun, C.-H., 2005. Performance of some variable selection methods when multicollinearity is present. Chemom. Intell. Lab. Syst. 78, 103–112. https://doi.org/https://doi.org/10.1016/j.chemolab.2004.12.011

Clavaud, M., Roggo, Y., Von Daeniken, R., Liebler, A., Schwabe, J.-O., 2013. Chemometrics and in-line near infrared spectroscopic monitoring of a biopharmaceutical Chinese hamster ovary cell culture: Prediction of multiple cultivation variables. Talanta 111, 28–38. https://doi.org/https://doi.org/10.1016/j.talanta.2013.03.044

Coroadinha, A.S., Silva, A.C., Pires, E., Coelho, A., Alves, P.M., Carrondo, M.J.T., 2006. Effect of osmotic pressure on the production of retroviral vectors: Enhancement in vector stability. Biotechnol. Bioeng. 94, 322–329. https://doi.org/10.1002/bit.20847

d’Aspremont, A., Bach, F., Ghaoui, L. El, 2008. Optimal Solutions for Sparse Principal Component Analysis. J. Mach. Learn. Res. 9, 1269–1294.

d’Aspremont, A., El Ghaoui, L., Jordan, M.I., Lanckriet, G.R.G., 2007. A Direct Formulation for Sparse PCA Using Semidefinite Programming. SIAM Rev. 49, 434–448. https://doi.org/10.1137/050645506

Davis, P.B., Cooper, M.J., 2007. Vectors for airway gene delivery. AAPS J. 9, E11-7. https://doi.org/10.1208/aapsj0901002

Dicker, L., Huang, B., Lin, X., 2013. Variable selection and estimation with the seamless-L 0 penalty. Stat. Sin. 23. https://doi.org/10.5705/ss.2011.074

Efron, B., Gong, G., 1983. A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. Am. Stat. 37, 36–48. https://doi.org/10.1080/00031305.1983.10483087

Efron, B., Tibshirani, R., 1986. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. Stat. Sci. 1, 54–75. https://doi.org/10.1214/ss/1177013815

Emerson, J., Kara, B., Glassey, J., 2020. Multivariate data analysis in cell gene therapy manufacturing. Biotechnol. Adv. 45, 107637. https://doi.org/https://doi.org/10.1016/j.biotechadv.2020.107637

Farcomeni, A., 2009. An exact approach to sparse principal component analysis. Comput. Stat. 24, 583. https://doi.org/10.1007/s00180-008-0147-3

Fushiki, T., 2011. Estimation of prediction error by using K-fold cross-validation. Stat. Comput. 21, 137–146. https://doi.org/10.1007/s11222-009-9153-8

Gajjar, S., Kulahci, M., Palazoglu, A., 2017. Selection of non-zero loadings in sparse principal component analysis. Chemom. Intell. Lab. Syst. 162, 160–171. https://doi.org/10.1016/j.chemolab.2017.01.018

Gálvez, J., Lecina, M., Solà, C., Cairó, J.J., Gòdia, F., 2012. Optimization of HEK-293S cell cultures for the production of adenoviral vectors in bioreactors using on-line OUR measurements. J. Biotechnol. 157, 214–222. https://doi.org/https://doi.org/10.1016/j.jbiotec.2011.11.007

Gándara, C., Affleck, V., Stoll, E.A., 2018. Manufacture of Third-Generation Lentivirus for Preclinical Use, with Process Development Considerations for Translation to Good Manufacturing Practice. Hum. Gene Ther. Methods 29, 1–15. https://doi.org/10.1089/hgtb.2017.098

Gardlík, R., Pálffy, R., Hodosy, J., Lukács, J., Turna, J., Celec, P., 2005. Vectors and delivery systems in gene therapy. Med. Sci. Monit. Int. Med. J. Exp. Clin. Res. 11, RA110-21.

Gaspar, H.B., Bjorkegren, E., Parsley, K., Gilmour, K.C., King, D., Sinclair, J., Zhang, F., Giannakopoulos, A., Adams, S., Fairbanks, L.D., Gaspar, J., Henderson, L., Xu-Bayford, J.H., Davies, E.G., Veys, P.A., Kinnon, C., Thrasher, A.J., 2006. Successful reconstitution of immunity in ADA-SCID by stem cell gene therapy following cessation of PEG-ADA and use of mild preconditioning. Mol. Ther. 14, 505–513. https://doi.org/10.1016/j.ymthe.2006.06.007

Giacca, M., 2010. Gene therapy. Dordrecht , Dordrecht .

Glassey, J., Gernaey, K. V, Clemens, C., Schulz, T.W., Oliveira, R., Striedner, G., Mandenius, C., 2011. Process analytical technology (PAT) for biopharmaceuticals. Biotechnol. J. 6, 369–377. https://doi.org/10.1002/biot.201000356

Golchin, A., Farahany, T.Z., 2019. Biological Products: Cellular Therapy and FDA Approved Products. Stem Cell Rev. Reports 15, 166–175. https://doi.org/10.1007/s12015-018-9866-1

Gottlieb, S., Marks, P., 2019. Statement from FDA Commissioner Scott Gottlieb, M.D. and Peter Marks, M.D., Ph.D., Director of the Center for Biologics Evaluation and Research on new policies to advance development of safe and effective cell and gene therapies.

Graham, F.L., Smiley, J., Russell, W.C., Nairn, R., 1977. Characteristics of a Human Cell Line Transformed by DNA from Human Adenovirus Type 5. J. Gen. Virol. 36, 59–72. https://doi.org/https://doi.org/10.1099/0022-1317-36-1-59

Griffith, F., 1928. The Significance of Pneumococcal Types. J. Hyg. (Lond). 27, 113–159. https://doi.org/10.1017/s0022172400031879

Hacein-Bey-Abina, S., Hauer, J., Lim, A., Picard, C., Wang, G.P., Berry, C.C., Martinache, C., Rieux-Laucat, F., Latour, S., Belohradsky, B.H., Leiva, L., Sorensen, R., Debré, M., Casanova, J.L., Blanche, S., Durandy, A., Bushman, F.D., Fischer, A., Cavazzana-Calvo, M., 2010. Efficacy of gene therapy for X-linked severe combined immunodeficiency. N. Engl. J. Med. 363, 355–364.

https://doi.org/10.1056/NEJMoa1000164

Hair, J.F., 2014. Multivariate data analysis., Pearson ne. ed. Harlow, Essex : Pearson.

Harmouche, J., Delpha, C., Diallo, D., 2014. Incipient fault detection and diagnosis based on Kullback-Leibler divergence using Principal Component Analysis: Part i. Signal Processing 94, 278–287. https://doi.org/10.1016/j.sigpro.2013.05.018

Harold, D., Abraham, R., Hollingworth, P., Sims, R., Gerrish, A., Hamshere, M.L., Pahwa, J.S., Moskvina, V., Dowzell, K., Williams, A., Jones, N., Thomas, C., Stretton, A., Morgan, A.R., Lovestone, S., Powell, J., Proitsi, P., Lupton, M.K., Brayne, C., Rubinsztein, D.C., Gill, M., Lawlor, B., Lynch, A., Morgan, K., Brown, K.S., Passmore, P.A., Craig, D., McGuinness, B., Todd, S., Holmes, C., Mann, D., Smith, A.D., Love, S., Kehoe, P.G., Hardy, J., Mead, S., Fox, N., Rossor, M., Collinge, J., Maier, W., Jessen, F., Schürmann, B., Heun, R., van den Bussche, H., Heuser, I., Kornhuber, J., Wiltfang, J., Dichgans, M., Frölich, L., Hampel, H., Hüll, M., Rujescu, D., Goate, A.M., Kauwe, J.S.K., Cruchaga, C., Nowotny, P., Morris, J.C., Mayo, K., Sleegers, K., Bettens, K., Engelborghs, S., De Deyn, P.P., Van Broeckhoven, C., Livingston, G., Bass, N.J., Gurling, H., McQuillin, A., Gwilliam, R., Deloukas, P., Al-Chalabi, A., Shaw, C.E., Tsolaki, M., Singleton, A.B., Guerreiro, R., Mühleisen, T.W., Nöthen, M.M., Moebus, S., Jöckel, K.-H., Klopp, N., Wichmann, H.-E., Carrasquillo, M.M., Pankratz, V.S., Younkin, S.G., Holmans, P.A., O'Donovan, M., Owen, M.J., Williams, J., 2009. Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. Nat. Genet. 41, 1088–1093. https://doi.org/10.1038/ng.440

Hastie, T., Friedman, J., Tibshirani, R., 2001. Model Assessment and Selection BT - The Elements of Statistical Learning: Data Mining, Inference, and Prediction, in: Hastie, T., Friedman, J., Tibshirani, R. (Eds.), . Springer New York, New York, NY, pp. 193–224. https://doi.org/10.1007/978-0-387-21606-5_7

Herrero, M.J., Sabater, L., Guenechea, G., Sendra, L., Montilla, A.I., Abargues, R., Navarro, V., Aliño, S.F., 2012. DNA delivery to "ex vivo" human liver segments. Gene Ther. 19, 504–512. https://doi.org/10.1038/gt.2011.144

Higashikawa, F., Chang, L.-J., 2001. Kinetic Analyses of Stability of Simple and Complex Retroviral Vectors. Virology 280, 124–131. https://doi.org/https://doi.org/10.1006/viro.2000.0743

Hirai, H., Satoh, E., Osawa, M., Inaba, T., Shimazaki, C., Kinoshita, S., Nakagawa, M., Mazda, O., Imanishi, J., 1997. Use of EBV-based Vector/HVJ-liposome complex vector for targeted gene therapy of EBV-associated neoplasms. Biochem. Biophys. Res. Commun. 241, 112–118. https://doi.org/10.1006/bbrc.1997.7776

Holic, N., Seye, A.K., Majdoul, S., Martin, S., Merten, O.W., Galy, A., Fenard, D., 2014. Influence of Mildly Acidic pH Conditions on the Production of Lentiviral and Retroviral Vectors. Hum. Gene Ther. Clin. Dev. 25, 178–185. https://doi.org/10.1089/humc.2014.027

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. 24, 417–441. https://doi.org/10.1037/h0071325

Hu, W.-W., Wang, Z., Hollister, S.J., Krebsbach, P.H., 2007. Localized viral vector delivery to enhance in situ regenerative gene therapy. Gene Ther. 14, 891–901. https://doi.org/10.1038/sj.gt.3302940

Jeffers, J.N.R., 1967. Two Case Studies in the Application of Principal Component Analysis. J. R. Stat. Soc. Ser. C (Applied Stat. 16, 225–236. https://doi.org/10.2307/2985919

Jessup, M., Greenberg, B., Mancini, D., Cappola, T., Pauly, D.F., Jaski, B., Yaroshinsky, A., Zsebo, K.M., Dittrich, H., Hajjar, R.J., 2011. Calcium

Upregulation by Percutaneous Administration of Gene Therapy in Cardiac Disease (CUPID): a phase 2 trial of intracoronary gene therapy of sarcoplasmic reticulum Ca2+-ATPase in patients with advanced heart failure. Circulation 124, 304–313. https://doi.org/10.1161/CIRCULATIONAHA.111.022889

Jolliffe, I.T., 1995. Rotation of principal components: choice of normalization constraints. J. Appl. Stat. 22, 29–35. https://doi.org/10.1080/757584395

Jolliffe, I.T., 1989. Rotation of Ill-Defined Principal Components. J. R. Stat. Soc. Ser. C (Applied Stat. 38, 139–147. https://doi.org/10.2307/2347688

Jolliffe, I.T., Cadima, J., 2016. Principal component analysis: a review and recent developments. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. 374, 20150202. https://doi.org/10.1098/rsta.2015.0202

Jolliffe, I.T., Trendafilov, N.T., Uddin, M., 2003. A Modified Principal Component Technique Based on the LASSO. J. Comput. Graph. Stat. 12, 531–547. https://doi.org/10.1198/1061860032148

Journée, M., Nesterov, Y., Richtárik, P., Sepulchre, R., 2010. Generalized Power Method for Sparse Principal Component Analysis. J. Mach. Learn. Res. 11, 517–553.

Kaemmerer, W.F., 2018. How will the field of gene therapy survive its success? Bioeng. Transl. Med. 3, 166–177. https://doi.org/10.1002/btm2.10090

Keeler, A.M., ElMallah, M.K., Flotte, T.R., 2017. Gene Therapy 2017: Progress and Future Directions. Clin. Transl. Sci. 10, 242–248. https://doi.org/10.1111/cts.12466

Kelley, J.E., 1960. The Cutting-Plane Method for Solving Convex Programs. J. Soc. Ind. Appl. Math. 8, 703–712.

Kennard, R.W., Stone, L.A., 1969. Computer Aided Design of Experiments. Technometrics 11, 137–148. https://doi.org/10.2307/1266770

Khan, S., Ullah, M.W., Siddique, R., Nabi, G., Manan, S., Yousaf, M., Hou, H., 2016. Role of Recombinant DNA Technology to Improve Life. Int. J. Genomics 2016, 2405954. https://doi.org/10.1155/2016/2405954

Khera, A. V, Chaffin, M., Aragam, K.G., Haas, M.E., Roselli, C., Choi, S.H., Natarajan, P., Lander, E.S., Lubitz, S.A., Ellinor, P.T., Kathiresan, S., 2018. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat. Genet. 50, 1219–1224. https://doi.org/10.1038/s41588-018-0183-z

Kiralj, R., Ferreira, M.M.C., 2009. Basic validation procedures for regression models in QSAR and QSPR studies: theory and application. J. Braz. Chem. Soc. 20, 770–787.

Kirdar, A.O., Conner, J.S., Baclaski, J., Rathore, A.S., 2007. Application of multivariate analysis toward biotech processes: Case study of a cell-culture unit operation. Biotechnol. Prog. 23, 61–67. https://doi.org/10.1021/bp060377u

Kirdar, A.O., Green, K.D., Rathore, A.S., 2008. Application of Multivariate Data Analysis for Identification and Successful Resolution of a Root Cause for a Bioprocessing Application. Biotechnol. Prog. 24, 720–726. https://doi.org/10.1021/bp0704384

Kotin, R.M., 2011. Large-scale recombinant adeno-associated virus production. Hum. Mol. Genet. 20, R2–R6. https://doi.org/10.1093/hmg/ddr141

Kotterman, M.A., Chalberg, T.W., Schaffer, D. V, 2015. Viral Vectors for Gene Therapy: Translational and Clinical Outlook. Annu. Rev. Biomed. Eng. https://doi.org/10.1146/annurev-bioeng-071813-104938

Kourti, T., 2003. Multivariate dynamic data modeling for analysis and statistical process control of batch processes, start-ups and grade transitions. J. Chemom. 17, 93–109. https://doi.org/10.1002/cem.778

Kuhn, M., Johnson, K., 2013. Applied predictive modeling, Applied Predictive Modeling. https://doi.org/10.1007/978-1-4614-6849-3

Kwok, Y.K., Tang, M.H.Y., Law, H.K.W., Ngai, C.S., Lau, Y.L., Lau, E.T., 2007. Maternal plasma or human serum albumin in wash buffer enhances enrichment and ex vivo expansion of human umbilical cord blood CD34+ cells. Br. J. Haematol. 137, 468–474. https://doi.org/10.1111/j.1365-2141.2007.06606.x

Larson, S.C., 1931. The shrinkage of the coefficient of multiple correlation. J. Educ. Psychol. 22, 45–55.

Le, H., Kabbur, S., Pollastrini, L., Sun, Z., Mills, K., Johnson, K., Karypis, G., Hu, W.-S., 2012. Multivariate analysis of cell culture bioprocess data—Lactate consumption as process indicator. J. Biotechnol. 162, 210–223. https://doi.org/https://doi.org/10.1016/j.jbiotec.2012.08.021

Le Ru, A., Jacob, D., Transfiguracion, J., Ansorge, S., Henry, O., Kamen, A.A., 2010. Scalable production of influenza virus in HEK-293 cells for efficient vaccine manufacturing. Vaccine 28, 3661–3671. https://doi.org/https://doi.org/10.1016/j.vaccine.2010.03.029

Leardi, R., 2000. Application of genetic algorithm–PLS for feature selection in spectral data sets. J. Chemom. 14, 643–655. https://doi.org/10.1002/1099-128X(200009/12)14:5/6<643::AID-CEM621>3.0.CO;2-E

Li, B., Ryan, P.W., Ray, B.H., Leister, K.J., Sirimuthu, N.M.S., Ryder, A.G., 2010. Rapid characterization and quality control of complex cell culture media solutions using raman spectroscopy and chemometrics. Biotechnol. Bioeng. 107, 290–301. https://doi.org/10.1002/bit.22813

Li, C., Joiner, J., Krotkov, N.A., Bhartia, P.K., 2013. A fast and sensitive new satellite SO2 retrieval algorithm based on principal component analysis: Application to the ozone monitoring instrument. Geophys. Res. Lett. 40, 6314–6318. https://doi.org/10.1002/2013GL058134

Liu, D., Ren, T., Gao, X., 2003. Cationic transfection lipids. Curr. Med. Chem. 10, 1307–1315. https://doi.org/10.2174/0929867033457386

Lopes, J.A., Costa, P.F., Alves, T.P., Menezes, J.C., 2004. Chemometrics in bioprocess engineering: process analytical technology (PAT) applications. Chemom. Intell. Lab. Syst. 74, 269–275. https://doi.org/https://doi.org/10.1016/j.chemolab.2004.07.006

Lopes, M.B., Calado, C.R.C., 2018. Assessing plasmid bioprocess reproducibility and C-source uptake stage through multivariate analysis of offline and online data. J. Chem. Technol. Biotechnol. 93, 3056–3066. https://doi.org/10.1002/jctb.5666

Lundstrom, K., 2018. Viral Vectors in Gene Therapy. Dis. (Basel, Switzerland) 6, 42. https://doi.org/10.3390/diseases6020042

Ma, Z., 2013. Sparse principal component analysis and iterative thresholding. Ann. Stat. 41, 772–801. https://doi.org/10.1214/13-AOS1097

Malm, M., Saghaleyni, R., Lundqvist, M., Giudici, M., Chotteau, V., Field, R., Varley, P., Hatton, D., Grassi, L., Svensson, T., Uhlen, M., Nielsen, J., Rockberg, J., 2020. Evolution from adherent to suspension – systems biology of HEK293 cell line development. bioRxiv 2020.01.29.924894. https://doi.org/10.1101/2020.01.29.924894

Manceur, A.P., Kim, H., Misic, V., Andreev, N., Dorion-Thibaudeau, J., Lanthier, S., Bernier, A., Tremblay, S., Gélinas, A.-M., Broussau, S., Gilbert, R., Ansorge, S., 2017. Scalable Lentiviral Vector Production Using Stable HEK293SF Producer Cell Lines. Hum. Gene Ther. Methods 28, 330–339. https://doi.org/10.1089/hgtb.2017.086

Marintcheva, B., 2018. Chapter 9 - Virus-Based Therapeutic Approaches, in: Marintcheva, B. (Ed.), Harnessing the Power of Viruses. Academic Press, pp.

243–276. https://doi.org/https://doi.org/10.1016/B978-0-12-810514-6.00009-X

McCarron, A., Donnelley, M., McIntyre, C., Parsons, D., 2016. Challenges of up-scaling lentivirus production and processing. J. Biotechnol. 240, 23–30. https://doi.org/https://doi.org/10.1016/j.jbiotec.2016.10.016

Mehmood, T., Liland, K.H., Snipen, L., Sæbø, S., 2012. A review of variable selection methods in Partial Least Squares Regression. Chemom. Intell. Lab. Syst. 118, 62–69. https://doi.org/https://doi.org/10.1016/j.chemolab.2012.07.010

Melcher, M., Scharl, T., Luchner, M., Striedner, G., Leisch, F., 2017. Boosted structured additive regression for Escherichia coli fed-batch fermentation modeling. Biotechnol. Bioeng. 114, 321–334. https://doi.org/10.1002/bit.26073

Mercier, S.M., Diepenbroek, B., Wijffels, R.H., Streefland, M., 2014. Multivariate PAT solutions for biopharmaceutical cultivation: current progress and limitations. Trends Biotechnol. 32, 329–336. https://doi.org/https://doi.org/10.1016/j.tibtech.2014.03.008

Merten, O.-W., Charrier, S., Laroudie, N., Fauchille, S., Dugué, C., Jenny, C., Audit, M., Zanta-Boussif, M.-A., Chautard, H., Radrizzani, M., Vallanti, G., Naldini, L., Noguiez-Hellin, P., Galy, A., 2010. Large-Scale Manufacture and Characterization of a Lentiviral Vector Produced for Clinical Ex Vivo Gene Therapy Application. Hum. Gene Ther. 22, 343–356. https://doi.org/10.1089/hum.2010.060

Merten, O.-W., Hebben, M., Bovolenta, C., 2016. Production of lentiviral vectors. Mol. Ther. Methods Clin. Dev. 3, 16017. https://doi.org/10.1038/mtm.2016.17

Merten, O.-W., Landric, L., Danos, O., 2001. Influence of the Metabolic Status of Packaging Cells on Retroviral Vector Production BT - Recombinant Protein Production with Prokaryotic and Eukaryotic Cells. A Comparative View on Host Physiology: Selected articles from the Meeting of the EFB Section o, in: Merten, O.-W., Mattanovich, D., Lang, C., Larsson, G., Neubauer, P., Porro, D., Postma, P., de Mattos, J.T., Cole, J.A. (Eds.), . Springer Netherlands, Dordrecht, pp. 303–318. https://doi.org/10.1007/978-94-015-9749-4_22

Milone, M.C., O'Doherty, U., 2018. Clinical use of lentiviral vectors. Leukemia 32, 1529–1541. https://doi.org/10.1038/s41375-018-0106-0

Mishra, V., Thakur, S., Patil, A., Shukla, A., 2018. Quality by design (QbD) approaches in current pharmaceutical set-up. Expert Opin. Drug Deliv. 15, 737–758. https://doi.org/10.1080/17425247.2018.1504768

Morgan, R.A., Gray, D., Lomova, A., Kohn, D.B., 2017. Hematopoietic Stem Cell Gene Therapy: Progress and Lessons Learned. Cell Stem Cell 21, 574–590. https://doi.org/10.1016/j.stem.2017.10.010

Nadeau, I., Gilbert, P.A., Jacob, D., Perrier, M., Kamen, A., 2002. Low-protein medium affects the 293SF central metabolism during growth and infection with adenovirus. Biotechnol. Bioeng. 77, 91–104. https://doi.org/10.1002/bit.10128

Nadeau, I., Sabatié, J., Koehl, M., Perrier, M., Kamen, A., 2000. Human 293 Cell Metabolism in Low Glutamine-Supplied Culture: Interpretation of Metabolic Changes through Metabolic Flux Analysis. Metab. Eng. 2, 277–292. https://doi.org/https://doi.org/10.1006/mben.2000.0152

Naldini, L., 2015. Gene therapy returns to centre stage. Nature 526, 351.

Naldini, L., 2011. Ex vivo gene transfer and correction for cell-based therapies. Nat. Rev. Genet. 12, 301.

National Human Genome Research Institute [WWW Document], n.d. . Rare Genet. Dis.

Nayerossadat, N., Maedeh, T., Ali, P.A., 2012. Viral and nonviral delivery systems for gene delivery. Adv. Biomed. Res. 1, 27. https://doi.org/10.4103/2277-9175.98152

O’Malley, C.J., Montague, G.A., Martin, E.B., Liddell, J.M., Kara, B., Titchener-Hooker, N.J., 2012. Utilisation of key descriptors from protein sequence data to aid bioprocess route selection. Food Bioprod. Process. 90, 755–761. https://doi.org/https://doi.org/10.1016/j.fbp.2012.01.005

Oliynyk, R.T., 2019. Future Preventive Gene Therapy of Polygenic Diseases from a Population Genetics Perspective. Int. J. Mol. Sci. 20, 5013. https://doi.org/10.3390/ijms20205013

Palmer, D.J., Ng, P., 2004. Physical and infectious titers of helper-dependent adenoviral vectors: a method of direct comparison to the adenovirus reference material. Mol. Ther. 10, 792–798. https://doi.org/https://doi.org/10.1016/j.ymthe.2004.06.1013

Pearson, S., Jia, H., Kandachi, K., 2004. China approves first gene therapy. Nat. Biotechnol. 22, 3–4. https://doi.org/10.1038/nbt0104-3

Petiot, E., Cuperlovic-Culf, M., Shen, C.F., Kamen, A., 2015. Influence of HEK293 metabolism on the production of viral vectors and vaccine. Vaccine 33, 5974–5981. https://doi.org/https://doi.org/10.1016/j.vaccine.2015.05.097

Prakash, V., Moore, M., Yáñez-Muñoz, R.J., 2016. Current Progress in Therapeutic Gene Editing for Monogenic Diseases. Mol. Ther. 24, 465–474. https://doi.org/10.1038/mt.2016.5

Prather, K.J., Sagar, S., Murphy, J., Chartrain, M., 2003. Industrial scale production of plasmid DNA for vaccine and gene therapy: plasmid design, production, and purification. Enzyme Microb. Technol. 33, 865–883. https://doi.org/https://doi.org/10.1016/S0141-0229(03)00205-9

Qi, X., Luo, R., Zhao, H., 2013. Sparse principal component analysis by choice of norm. J. Multivar. Anal. 114, 127–160. https://doi.org/https://doi.org/10.1016/j.jmva.2012.07.004

Ramirez, J., 2018. Lentiviral Vectors Come of Age? Hurdles and Challenges in Scaling Up Manufacture. https://doi.org/10.5772/intechopen.81105

Ramos, M., Ascencio, J., Hinojosa, M.V., Vera, F., Ruiz, O., Jimenez-Feijoó, M.I., Galindo, P., 2021. Multivariate statistical process control methods for batch production: a review focused on applications. Prod. Manuf. Res. 9, 33–55. https://doi.org/10.1080/21693277.2020.1871441

Rathore, A.S., Bhambure, R., Ghare, V., 2010. Process analytical technology (PAT) for biopharmaceutical products. Anal. Bioanal. Chem. 398, 137–154. https://doi.org/10.1007/s00216-010-3781-x

Rathore, A S, Mittal, S., Pathak, M., Arora, A., 2014. Guidance for performing multivariate data analysis of bioprocessing data: Pitfalls and recommendations. Biotechnol. Prog. 30, 967–973. https://doi.org/10.1002/btpr.1922

Rathore, Anurag S, Mittal, S., Pathak, M., Mahalingam, V., 2014. Chemometrics application in biotech processes: assessing comparability across processes and scales. J. Chem. Technol. Biotechnol. 89, 1311–1316. https://doi.org/10.1002/jctb.4428

Robertson, E.S., Ooka, T., Kieff, E.D., 1996. Epstein-Barr virus vectors for gene delivery to B lymphocytes. Proc. Natl. Acad. Sci. U. S. A. 93, 11334–11340. https://doi.org/10.1073/pnas.93.21.11334

Rodríguez, J.D., Pérez, A., Lozano, J.A., 2010. Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation. IEEE Trans. Pattern Anal. Mach. Intell. 32, 569–575. https://doi.org/10.1109/TPAMI.2009.187

Rodwell, C., Aymé, S., 2014. Report on the State of the Art of Rare Disease Activities in Europe Part Ii: Key Developments in the Field of Rare Diseases in Europe in 2013.

Roldão, A., Oliveira, R., Carrondo, M.J.T., Alves, P.M., 2009. Error assessment in recombinant baculovirus titration: Evaluation of different methods. J. Virol. Methods 159, 69–80. https://doi.org/https://doi.org/10.1016/j.jviromet.2009.03.007

Rosenberg, L.E., Rosenberg, D.D., 2012. Chapter 17 - Detection and Treatment of Genetic Disorders, in: Rosenberg, L.E., Rosenberg, D.D. (Eds.), Human Genes and Genomes. Academic Press, San Diego, pp. 289–314. https://doi.org/https://doi.org/10.1016/B978-0-12-385212-0.00017-2

Rosenberg, S.A., 1992. Gene Therapy for Cancer. JAMA 268, 2416–2419. https://doi.org/10.1001/jama.1992.03490170088031

Rout-Pitt, N., McCarron, A., McIntyre, C., Parsons, D., Donnelley, M., 2018. Large-scale production of lentiviral vectors using multilayer cell factories. J. Biol. Methods; Vol 5, No 2.

Rücker, C., Rücker, G., Meringer, M., 2007. y-Randomization and Its Variants in QSPR/QSAR. J. Chem. Inf. Model. 47, 2345–2357. https://doi.org/10.1021/ci700157b

Schweizer, M., Merten, O.-W., 2010. Large-Scale Production Means for the Manufacturing of Lentiviral Vectors. Curr. Gene Ther. 10, 474–486. https://doi.org/10.2174/156652310793797748

Seasholtz, M.B., Kowalski, B.R., 1992. The effect of mean centering on prediction in multivariate calibration. J. Chemom. 6, 103–111. https://doi.org/10.1002/cem.1180060208

Sharon, D., Kamen, A., 2018. Advancements in the design and scalable production of viral gene transfer vectors. Biotechnol. Bioeng. 115, 25–40. https://doi.org/10.1002/bit.26461

Shen, C.F., Kamen, A., 2012. Hyperosmotic pressure on HEK 293 cells during the growth phase, but not the production phase, improves adenovirus production. J. Biotechnol. 157, 228–236. https://doi.org/https://doi.org/10.1016/j.jbiotec.2011.11.016

Shen, H., Huang, J.Z., 2008. Sparse principal component analysis via regularized low rank matrix approximation. J. Multivar. Anal. 99, 1015–1034. https://doi.org/https://doi.org/10.1016/j.jmva.2007.06.007

Shi, Y., Cao, Y., Yu, J., Jiao, Y., 2018. Variable selection via generalized SELO-penalized linear regression models. Appl. Math. J. Chinese Univ. 33, 145–162. https://doi.org/10.1007/s11766-018-3496-x

Smoliak, B. V, Wallace, J.M., Stoelinga, M.T., Mitchell, T.P., 2010. Application of partial least squares regression to the diagnosis of year-to-year variations in Pacific Northwest snowpack and Atlantic hurricanes. Geophys. Res. Lett. 37. https://doi.org/10.1029/2009GL041478

Stanton, J.M., 2001. Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors. J. Stat. Educ. 9, null-null. https://doi.org/10.1080/10691898.2001.11910537

Stone, M., 1974. Cross-Validatory Choice and Assessment of Statistical Predictions. J. R. Stat. Soc. Ser. B 36, 111–147.

Streefland, M., Martens, D.E., Beuvery, E.C., Wijffels, R.H., 2013. Process analytical technology (PAT) tools for the cultivation step in biopharmaceutical production. Eng. Life Sci. 13, 212–223. https://doi.org/10.1002/elsc.201200025

Stroncek, D.F., Jin, P., Ren, J., Feng, J., Castiello, L., Civini, S., Wang, E., Marincola, F.M., Sabatino, M., 2010. Quality assessment of cellular therapies: the emerging role of molecular assays. Korean J. Hematol. 45, 14–22. https://doi.org/10.5045/kjh.2010.45.1.14

Suhonen, J., Ray, J., Blömer, U., Gage, F.H., Kaspar, B., 2006. Ex vivo and in vivo gene delivery to the brain. Curr. Protoc. Hum. Genet. Chapter 13, Unit 13.3.

https://doi.org/10.1002/0471142905.hg1303s51

Thomas, A.H., Brown, K., Jr, F.H.P., Veres, G., Ryu., B.Y., 2013. Large Scale Production of Lentiviral Vectors Using Serum Free Suspension Cell Culture System. Mol. Ther. 21, S21. https://doi.org/10.1016/S1525-0016(16)34386-6

Thurstone, L.L., 1931. Multiple factor analysis. Psychol. Rev. 38, 406–427.

Tibshirani, R., 1996. Regression Shrinkage and Selection via the Lasso. J. R. Stat. Soc. Ser. B 58, 267–288.

Uchida, N., Nassehi, T., Drysdale, C.M., Gamer, J., Yapundich, M., Demirci, S., Haro-Mora, J.J., Leonard, A., Hsieh, M.M., Tisdale, J.F., 2019. High-Efficiency Lentiviral Transduction of Human CD34$^+$ Cells in High-Density Culture with Poloxamer and Prostaglandin E2. Mol. Ther. - Methods Clin. Dev. 13, 187–196. https://doi.org/10.1016/j.omtm.2019.01.005

Ulfarsson, M., Solo, V., 2011. Vector l0 sparse variable PCA. Signal Process. IEEE Trans. 59, 1949–1958. https://doi.org/10.1109/TSP.2011.2112653

Ündey, C., Ertunç, S., Çınar, A., 2003. Online Batch/Fed-Batch Process Performance Monitoring, Quality Prediction, and Variable-Contribution Analysis for Diagnosis. Ind. Eng. Chem. Res. 42, 4645–4658. https://doi.org/10.1021/ie0208218

Urthaler, J., Schuchnigg, H., Garidel, P., Huber, H., 2012. Industrial Manufacturing of Plasmid-DNA Products for Gene Vaccination and Therapy BT - Gene Vaccines, in: Thalhamer, J., Weiss, R., Scheiblhofer, S. (Eds.), . Springer Vienna, Vienna, pp. 311–330. https://doi.org/10.1007/978-3-7091-0439-2_16

Valkama, A.J., Leinonen, H.M., Lipponen, E.M., Turkki, V., Malinen, J., Heikura, T., Ylä-Herttuala, S., Lesch, H.P., 2018. Optimization of lentiviral vector production for scale-up in fixed-bed bioreactor. Gene Ther. 25, 39–46. https://doi.org/10.1038/gt.2017.91

van den Berg, R.A., Hoefsloot, H.C.J., Westerhuis, J.A., Smilde, A.K., van der Werf, M.J., 2006. Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics 7, 142. https://doi.org/10.1186/1471-2164-7-142

van der Loo, J.C.M., Wright, J.F., 2015. Progress and challenges in viral vector manufacturing. Hum. Mol. Genet. 25, R42–R52. https://doi.org/10.1093/hmg/ddv451

Vlachakis, J.C.R.E.-D., 2019. Lentiviral Vectors Come of Age? Hurdles and Challenges in Scaling Up Manufacture. IntechOpen, Rijeka, p. Ch. 3. https://doi.org/10.5772/intechopen.81105

Weston, J., Elisseeff, A., Schölkopf, B., Tipping, M., 2003. Use of the Zero Norm with Linear Models and Kernel Methods. J. Mach. Learn. Res. 3, 1439–1461.

Wickham, T.J., 2003. Ligand-directed targeting of genes to the site of disease. Nat. Med. 9, 135–139. https://doi.org/10.1038/nm0103-135

Willis, M.J., von Stosch, M., 2017. L0-constrained regression using mixed integer linear programming. Chemom. Intell. Lab. Syst. 165, 29–37. https://doi.org/https://doi.org/10.1016/j.chemolab.2016.12.016

Wirth, T., Parker, N., Ylä-Herttuala, S., 2013. History of gene therapy. Gene 525, 162–169. https://doi.org/10.1016/j.gene.2013.03.137

Wirth, T., Samaranayake, H., Pikkarainen, J., Määttä, A.-M., Ylä-Herttuala, S., 2009. Clinical trials for glioblastoma multiforme using adenoviral vectors. Curr. Opin. Mol. Ther. 11, 485–492.

Wold, H., 1966. Estimation of principal components and related models by iterative least squares. Multivar. Anal.

Wolschek, M.F., Thallinger, C., Kursa, M., Rössler, V., Allen, M., Lichtenberger, C., Kircheis, R., Lucas, T., Willheim, M., Reinisch, W., Gangl, A., Wagner, E., Jansen, B., 2002. Specific systemic nonviral gene delivery to human hepatocellular carcinoma xenografts in SCID mice. Hepatology 36, 1106–1114. https://doi.org/10.1053/jhep.2002.36372

Worgall, S., Crystal, R.G., 2014. Chapter 34 - Gene Therapy, in: Lanza, R., Langer, R., Vacanti, J. (Eds.), Principles of Tissue Engineering (Fourth Edition). Academic Press, Boston, pp. 657–686. https://doi.org/https://doi.org/10.1016/B978-0-12-398358-9.00034-3

Xiaoshuang, S., Zhihui, L., Zhenhua, G., Minghua, W., Cairong, Z., Heng, K., 2013. Sparse Principal Component Analysis via Joint L2,1-Norm Penalty BT - AI 2013: Advances in Artificial Intelligence, in: Cranefield, S., Nayak, A. (Eds.), . Springer International Publishing, Cham, pp. 148–159.

Xu, L., Pirollo, K.F., Tang, W.H., Rait, A., Chang, E.H., 1999. Transferrin-liposome-mediated systemic p53 gene therapy in combination with radiation results in regression of human head and neck cancer xenografts. Hum. Gene Ther. 10, 2941–2952. https://doi.org/10.1089/10430349950016357

Zhang, B., Metharom, P., Jullie, H., Ellem, K.A.O., Cleghorn, G., West, M.J., Wei, M.Q., 2004. The significance of controlled conditions in lentiviral vector titration and in the use of multiplicity of infection (MOI) for predicting gene transfer events. Genet. Vaccines Ther. 2, 6. https://doi.org/10.1186/1479-0556-2-6

Zhang, Y., Zhang, Y.-F., Bryant, J., Charles, A., Boado, R.J., Pardridge, W.M., 2004. Intravenous RNA interference gene therapy targeting the human epidermal growth factor receptor prolongs survival in intracranial brain cancer. Clin. cancer Res. an Off. J. Am. Assoc. Cancer Res. 10, 3667–3677. https://doi.org/10.1158/1078-0432.CCR-03-0740

Zheng, R., Pan, F., 2016. Soft sensor modeling of product concentration in glutamate fermentation using gaussian process regression. Am. J. Biochem. Biotechnol. 12, 179–187. https://doi.org/10.3844/ajbbsp.2016.179.187

Zheng, X., Levine, D., Shen, J., Gogarten, S.M., Laurie, C., Weir, B.S., 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28, 3326–3328. https://doi.org/10.1093/bioinformatics/bts606

Zimmermann, K., Scheibe, O., Kocourek, A., Muelich, J., Jurkiewicz, E., Pfeifer, A., 2011. Highly efficient concentration of lenti- and retroviral vector preparations by membrane adsorbers and ultrafiltration. BMC Biotechnol. 11, 55. https://doi.org/10.1186/1472-6750-11-55

ZINDER, N.D., LEDERBERG, J., 1952. Genetic exchange in Salmonella. J. Bacteriol. 64, 679–699.

Zou, H., Hastie, T., 2005. Regularization and Variable Selection via the Elastic Net. J. R. Stat. Soc. Ser. B (Statistical Methodol. 67, 301–320.

Zou, H., Hastie, T., Tibshirani, R., 2006. Sparse principal component analysis. J. Comput. Graph. Stat. 15, 265–286. https://doi.org/10.1198/106186006X113430

Chapter 10

Appendix

10.1 Appendix A

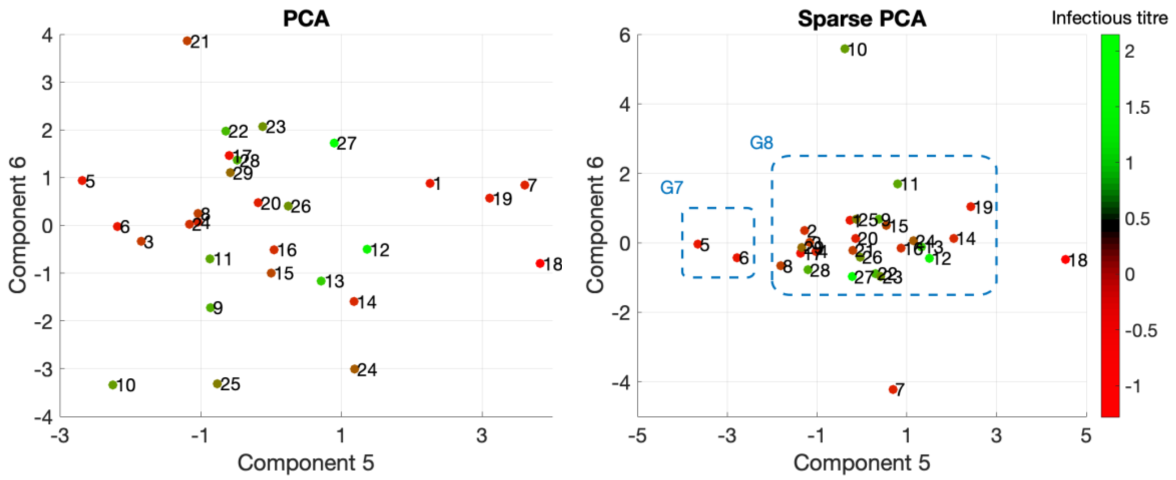


Figure A1: PCA and sparse PCA scores plots for components 5 and 6.

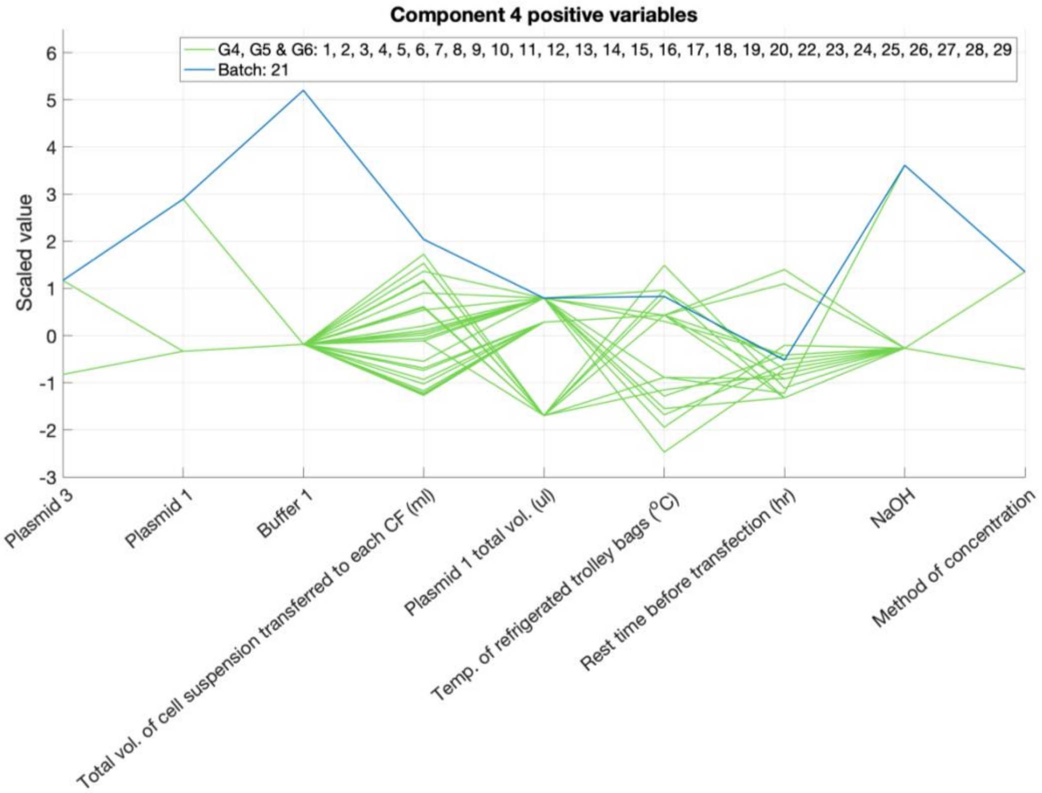


Figure A2: Parallel coordinates plot of variables with positive loadings on component 4

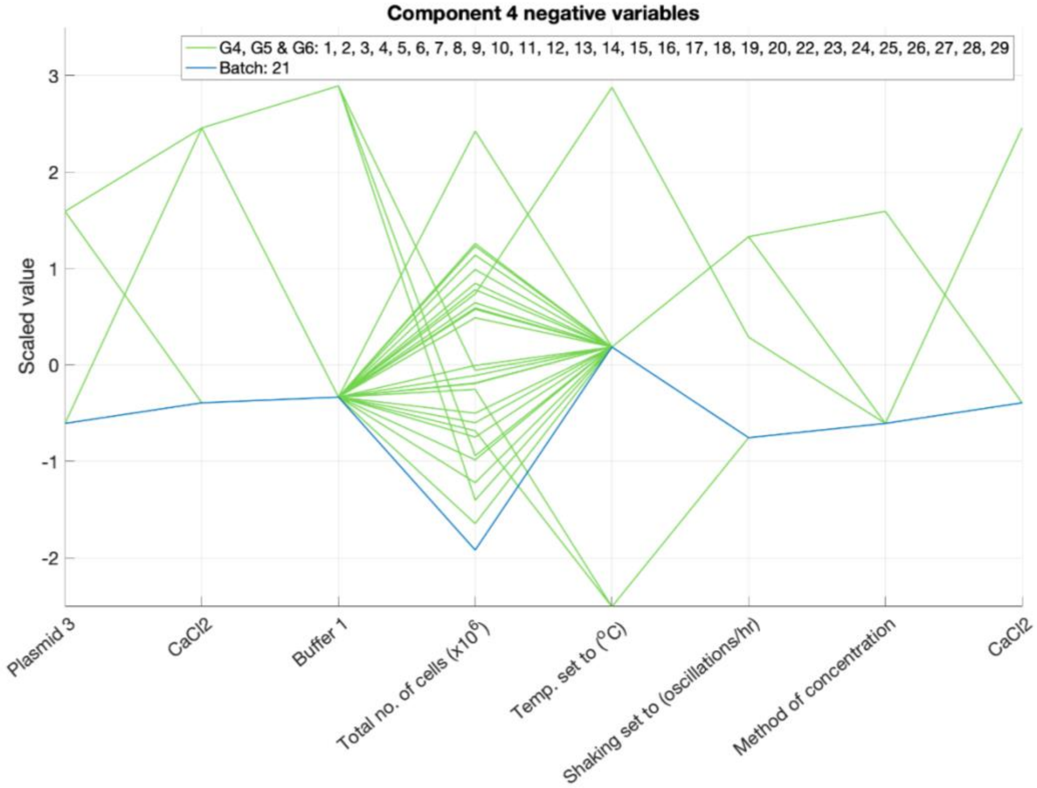


Figure A3: Parallel coordinates plot of variables with negative loadings on component 4

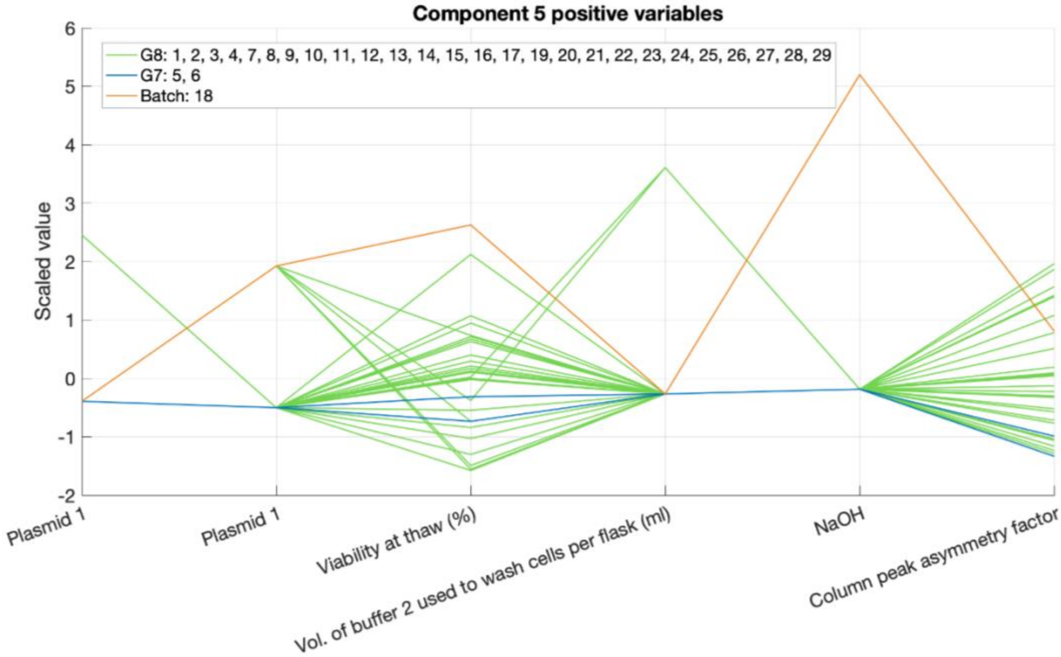


Figure A4: Parallel coordinates plot of variables with positive loadings on component 5

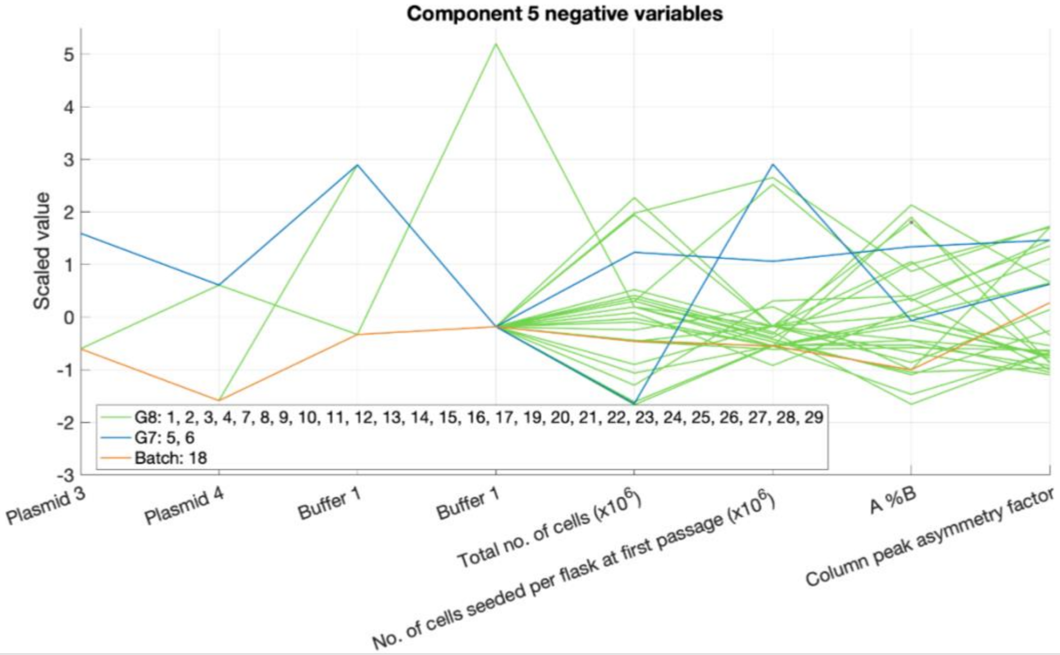


Figure A5: Parallel coordinates plot of variables with negative loadings on component 5

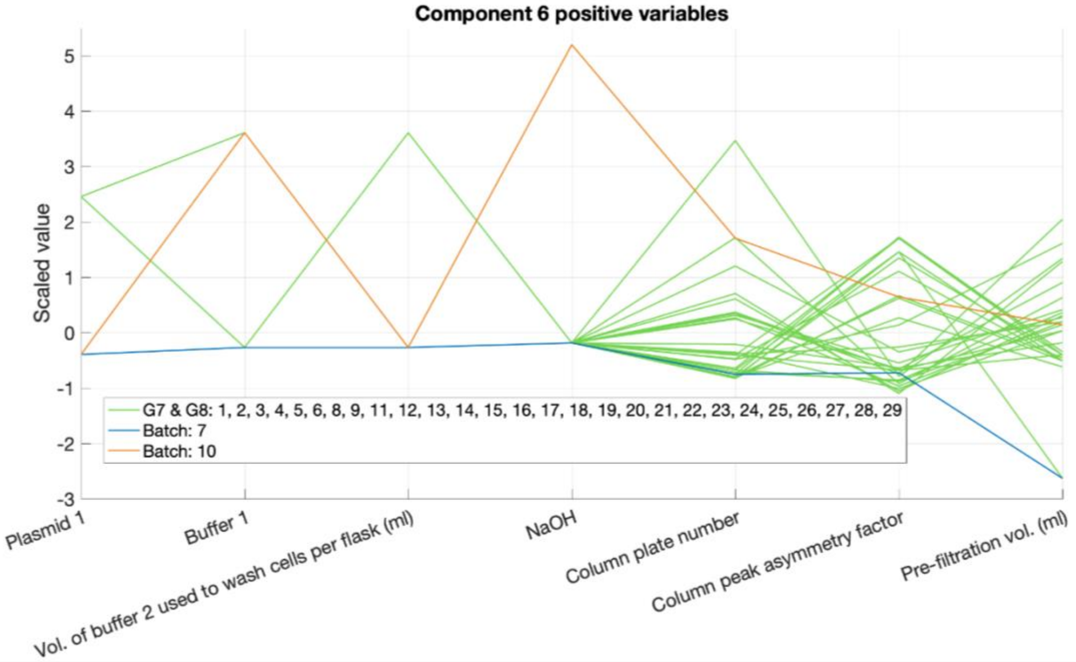


Figure A1: Parallel coordinates plot of variables with positive loadings on component 6

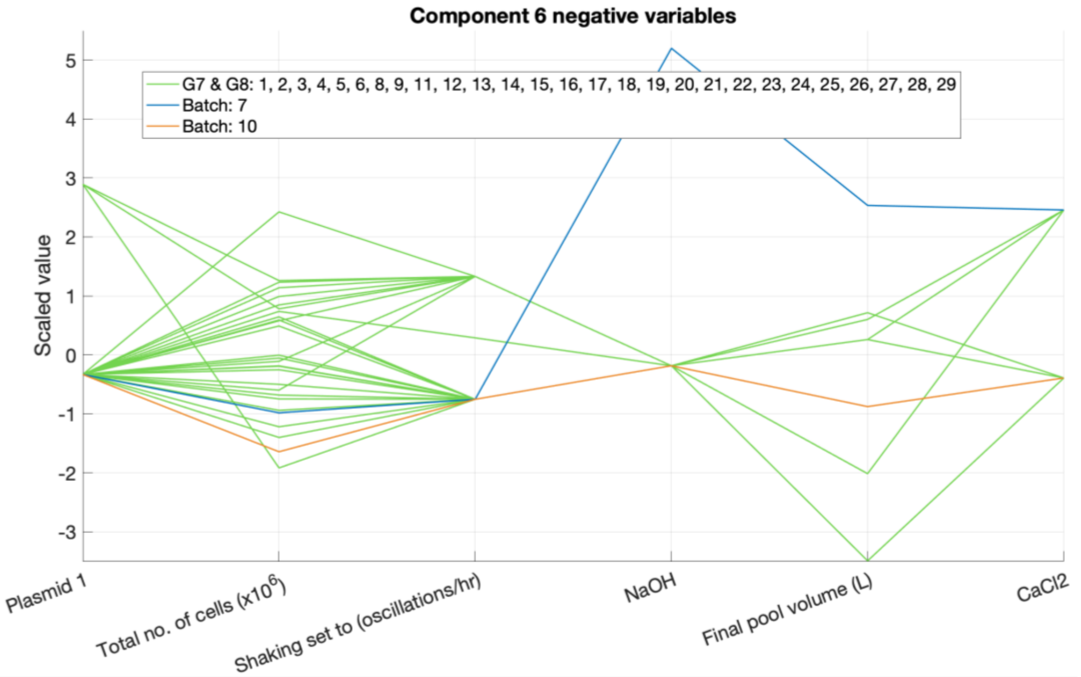


Figure A2: Parallel coordinates plot of variables with negative loadings on component 6

10.2 Appendix B

Table B1: List of variables included in the basic PLS model for the prediction of the infectious titre. The basic model had no transformations applied to the data. In the columns, the mean beta coefficient from bootstrap sampling is presented along with upper (95%) and lower (5%) confidence intervals.

Variable	Variable position	Mean beta	Upper 95% confidence interval (CI)	Lower 5% CI
Cell concentration at thaw	Cell thaw	0.226	0.351	0.071
Total number of cells	Cell thaw	0.114	0.225	-0.026
Viability at thaw	Cell thaw	-0.239	-0.132	-0.338
Number of cells seeded per flask at thaw	Cell thaw	-0.034	0.084	-0.162
Volume of buffer 2 used to wash cells per flask (mL)	First passage	-0.066	0.000	-0.133
Cell concentration at first passage	First passage	-0.043	0.154	-0.221
Total number of cells	Seeding cell factories	0.310	0.419	0.176
Total volume of plasmid 1(ul)	Transfection	-0.288	-0.173	-0.398
Temperature set to (C)	Endonuclease treatment	-0.158	-0.028	-0.264
Plasmid 1 concentration (mg/ml or ug/ul)	Transfection	0.047	0.142	-0.063
Column peak asymmetry factor	Prep for ion exchange	0.099	0.216	-0.031
A %B	Ion exchange chromatography	-0.281	-0.139	-0.422
Δ conductivity	Ion exchange chromatography	-0.164	-0.067	-0.256
Volume (L)	Ion exchange chromatography	-0.090	0.007	-0.205
Final pool volume (L)	Ion exchange chromatography	-0.069	0.073	-0.185
Sterile filter area (cm2)	Sterile filtration	-0.316	-0.220	-0.407
Pre-filtration volume (ml)	Sterile filtration	0.190	0.312	0.049

10.3 Appendix C

This appendix explains how gradients were determined for offline variables in Chapter 7. Initially, the 3rd order polynomials were fitted to the offline data using a least squares fitting algorithm in Matlab 2019. The form of the polynomial is given by (C1) where δ, γ, β and α are the polynomial coefficients, ***x*** represents the variable and ***y*** is a vector of fitted values.

$$\boldsymbol{y} = \delta\boldsymbol{x}^{\boldsymbol{3}} + \gamma\boldsymbol{x}^{\boldsymbol{2}} + \beta\boldsymbol{x} + \alpha \tag{C1}$$

After the coefficients were obtained from the fitting algorithm, the gradient at each point in ***x*** was determined numerically by differentiating the polynomial equation and substituting the model coefficients into (C2).

$$\frac{dy}{dx} = 3\delta\boldsymbol{x}^{\boldsymbol{2}} + 2\gamma\boldsymbol{x} + \beta \tag{C2}$$