Bayesian hierarchical modelling for biomarkers with applications to doping detection and prostate cancer prediction

Dimitra Eleftheriou

Supervisor
Dr Tereza Neocleous

A thesis submitted in fulfilment of
the requirements for the degree of
Doctor of Philosophy

School of Mathematics and Statistics
College of Science and Engineering
University of Glasgow

August 2022

Declaration of Authorship

I, Dimitra Eleftheriou, hereby declare that this thesis titled, “Bayesian hierarchical modelling for biomarkers with applications to doping detection and prostate cancer prediction” and the work presented in it is entirely my own. Where I have consulted the work of others, this is always clearly attributed.

“The known is finite, the unknown infinite; intellectually we stand on an islet in the midst of an illimitable ocean of inexplicability. Our business in every generation is to reclaim a little more land.”

Thomas Henry Huxley

Abstract

Anabolic androgenic steroids (AAS) are frequently detected doping substances in competitive sports. In order to detect AAS doping with pseudo-endogenous steroids, i.e. steroids that are produced in the human body, such as testosterone (T), urinary concentrations of the athlete's steroid profile are measured over time in the steroidal module of the Athlete Biological Passport (ABP). Monitoring the urinary levels of anabolic steroids can be highly challenging since the distinction between their natural production and exogenous administration is difficult to ascertain. Current methods for monitoring AAS are based on a univariate Bayesian model applied on a single biomarker at a time. The first part of this research work focuses on extending the current univariate Bayesian model to a multivariate adaptive model, able to accommodate repeated measurements from various sensitive biomarkers and their concentration ratios. The developed methodology was applied on data from urine samples obtained from professional athletes. Among these samples, normal, atypical, and abnormal values were identified. An anomaly detection technique based on a one-class classification (OCC) algorithm was carried out to detect the abnormal values within the athletes' steroid profiles, either due to AAS misuse, samples' exchange or other confounding factors. In a Bayesian context, the main idea is to construct adaptive decision boundaries around normal concentration values as new data come, and differentiate them from the abnormal ones (also called outliers or anomalies). Improved prediction performance was obtained when using the same data applied on the proposed model and compared to standard methodologies. Higher values of evaluation metrics suggest that the proposed approach can be used to improve the accuracy of standard techniques for doping detection. The proposed model was implemented in an Rshiny app for doping testing purposes. The BioScan App is a web application which constitutes a user-friendly software for anti-doping laboratories to use for athletes' evaluation in real-life casework.

AAS also have the potential to identify metabolic imbalance and pathological conditions such as benign prostatic hyperplasia and prostatic carcinoma. The second research part focuses on developing novel methodology in statistical modelling to improve prostate cancer diagnosis by analysing a variety of urinary steroids. The proposed approach constitutes a non-invasive, low cost and an improved screening method compared to the widely used PSA test. The thesis uses the Dirichlet process (DP) models for a mixture of Gaussian distributions in a Bayesian framework as an improved classification tool. This parameter-free model can be applied to both univariate and multivariate data sets providing the flexibility of unknown and possible infinite number of components. The models introduced by Görür and Rasmussen (2010) have been extended to models with covariates, which account for possible patterns within them. The main features of the DP mixture models with and without covariate information are highlighted in this dissertation. Emphasis is given to the model structure when covariates are included in the model using a technique to reduce the number of model parameters. This technique also constitutes an elegant way to deal with high-dimensional predictors, providing a significant contribution in dimensionality reduction. The main goal is to compare their predictive performance versus model complexity and computational effort. Given the mathematical and practical convenience, the DP models are defined by specifying conditionally conjugate priors for their base distributions. Markov chain Monte Carlo (MCMC) methods, based on the Gibbs sampling and Adaptive Rejection Sampling (ARS), are the required methods for each variable to generate samples from its conditional distribution given the rest variables in the system. Clustering and classification performance of the models are examined on simulated and real data. We focus on the applications carried out on real clinical data regarding prostate cancer using this methodology as an aim to classify prostate cancer conditions. The implementation of DP-GMM using biomarkers only with age as a covariate increases the prediction accuracy as compared to the corresponding covariate-free model. Finally, the proposed classification model proved to be superior compared to the standard methods of support vector machines (SVM) and linear discriminant analysis (LDA) on three out of four applications on different data sets, including prostate cancer data.

Keywords: Adaptive rejection sampling, Anomaly detection, Bayesian nonparametrics, Biomarkers, Dirichlet processes, Doping, Gaussian mixtures, Markov chain Monte Carlo, Multivariate Bayesian multilevel model, One-class classifier, Predictive models, Prostate cancer.

Acknowledgements

First and foremost, I would like to express my sincere gratitude and respect to my supervisor Dr Tereza Neocleous for her constant support, motivation, and availability throughout my PhD studies. I would also like to thank Dr Ludger Evers, who had a critical influence on my research focus, as well as Dr Theodore Papamarkou for the helpful discussions in my first PhD years. Their valuable guidance in combination with our enjoyable discussions helped me to successfully conduct my PhD dissertation.

I would like to thank Professor Christina Bamia and Professor Dimitris Karlis for our fruitful collaboration on parallel research projects and for being a source of inspiration. I would also like to thank Dr John Maclay for the wonderful collaboration.

In addition, I would like to acknowledge the EPSRC for their financial support during my doctoral studies. I am also grateful for the support provided by the School of Mathematics and Statistics at the University of Glasgow and for providing me the opportunity to attend high-quality doctoral training, conferences and master-classes. I would also like to thank Professor Claire Miller and Professor Janine Illian for giving me the chance to teach and share my passion with students.

A special thanks to Dr Thomas Piper and the Institute of Biochemistry of the German Sport University for our collaborative research on anti-doping. I am also grateful to our colleague Dr Eugenio Alladio at the Department of Chemistry of the University of Turin for sharing laboratory data to carry out prostate cancer research.

I would also like to thank my officemates and my friends in Glasgow for a cherished time spent together, sharing our thoughts and having a lot of fun. Especially Jafet, Mihaela, Yoana, Anna, Efthymios and our music band “Los Guacamoles”.

Last but not least, I would like to express my very profound gratitude to my family, Giannis, Chrysoula, Danai and Vasiliki, and all my friends in Athens, especially Chara, Nikos, Mata, Lina, Kostas Tr. and the “Highlanders”, for providing me with unfailing support and continuous encouragement throughout my years of PhD. A very special thanks to my beloved Dimitris for always believing in me, encouraging me until the last moments of my work, and for adding “dust of happiness” in my years in Glasgow. This accomplishment would not have been possible without them.

To my family

Contents

List of Figures

List of Tables

Abbreviations

A	**A**ndrosterone
AAS	**A**nabolic **A**ndrogenic **S**teroids
ABP	**A**thlete **B**iological **P**assport
ARS	**A**daptive **R**ejection **S**ampling
BPH	**B**enign **P**rostatic **H**yperplasia
CRP	**C**hinese **R**estaurant **P**rocess
CV	**C**oefficient of **V**ariation
DHEA	**D**ehydroepiandrosterone
DHT	**D**ihydrotestosterone
DP	**D**irichlet **P**rocess
DP-GLM	**D**irichlet **P**rocess mixtures of **G**eneralised **L**inear **M**odel
DP-GMM	**D**irichlet **P**rocess - **G**aussian **M**ixture **M**odel
DP-GMMx	**D**irichlet **P**rocess - **G**aussian **M**ixture **M**odel with **C**ovariates
DPMM	**D**irichlet **P**rocess **M**ixture **M**odel
E	**E**pitestosterone
EAAS	**E**ndogenous **A**nabolic **A**ndrogenic **S**teroids
ESS	**E**ffective **S**ample **S**ize
ETIO	**Etio**cholanolone
GC-MS	**G**as **C**romatography - **M**ass **S**pectometry
GLM	**G**eneralised **L**inear **M**odel
GLMM	**G**eneralised **L**inear **M**ixed **M**odel
GMM	**G**aussian **M**ixture **M**odel
HPD	**H**ighest **P**osterior **D**ensity

IQ	**I**nter-**Q**uartile
LDA	**L**inear **D**iscriminant **A**nalysis
MCMC	**M**arkov **C**hain **M**onte **C**arlo
MGMM	**M**ultivariate **G**aussian **M**ixed **M**odel
MH	**M**etropolis - **H**astings
MWG	**M**etropolis-**W**ithin-**G**ibbs
OCC	**O**ne-**C**lass **C**lassification
PCa	**P**rostate **Ca**ncer
PSA	**P**rostate **S**pecific **A**ntigen
ROC	**R**eceiver **O**perating **C**haracteristic
SP	**S**teroid **P**rofile
SVM	**S**upport **V**ector **M**achines
T	**T**estosterone
USADA	**U**nited **S**tates of**A**nti-Doping **A**gency
WADA	**W**orld **A**nti-**D**oping **A**gency

Chapter 1

Introduction

"The important thing in the games is not winning but taking part. The essential thing is not conquering, but fighting well."

by the founder of the modern Olympic Games,
Baron Pierre de Coubertin

1.1 The "Steroid Profile"

The "Steroid Profile" (SP) is a long-established analytical method which can identify and quantify a whole spectrum of steroid metabolites simultaneously in a single analysis, instead of measuring a single analyte at a time. Steroid profiling has a wide application in the study of disorders of human steroid biosynthesis and catabolism, useful in avoiding uncritical and costly molecular diagnostic tests (Wudy et al., 2018). Metabolic assessment is not only a powerful diagnostic tool, but also allows for monitoring and studying a variety of conditions such as obesity, cancer, chronic fatigue syndrome and depression. The technique divides into extraction of free and conjugated steroids, steroid conjugate hydrolysis, free steroid re-extraction, derivatisation and analysis by either gas chromatography (GC), gas chromatography-mass spectrometry (GC-MS) or liquid chromatography tandem mass spectrometry

(LC-MS-MS) (Chapter 17; Wheeler, 2013). All mass spectrometry based methods are very powerful, however, GC-MS has been the golden standard for many years (Krone et al., 2010; Kuuranne et al., 2014). These techniques are complementary and highly sensitive instruments which are employed in analytical chemistry for analysing, separating and categorising compounds of interest based on their mass in a given sample (Chan et al., 2008; Wheeler, 2013). The samples analysed by GC are principally liquids. According to Figure 1.1, reproduced from book of Wheeler (2013) for Molecular Biology, shows an example of compounds with different chemical properties separated using a chromatographic method.

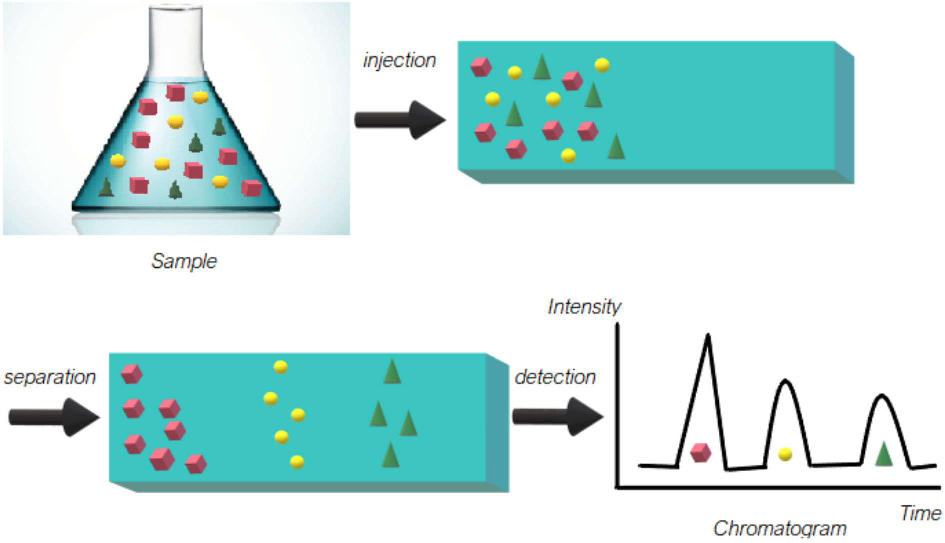


FIGURE 1.1: Chromatographic separation principle (Wheeler, 2013).

In 1968, Horning used the expression “Steroid Profile” for the first time in his publication about urinary steroid analysis by GC-MS. In addition to urine samples, steroids can also be determined by using hair follicles or analysing blood samples. Blood screening is more sensitive and it constitutes the current gold standard for detecting abnormally elevated levels of synthetic hormones that exist in the human body. However, urine tests are frequently used, because the samples they require are easier to obtain and test since they are cheaper, less invasive and have fewer potential complications than blood testing.

A steroid profile includes all major metabolites of steroids. It usually consists of the concentrations of the following markers:

- Testosterone (T),
- Epitestosterone (E),
- Androsterone (A),
- Etiocholanolone (Etio),
- 5α-Androstane-3α, 17β-diol (5α-Adiol),
- 5β-Androstane-3α, 17β-diol (5β-Adiol),
- Dehydroepiandrosterone (DHEA),

together with the specific gravity (SG) of the sample, which is a measure that compares the density of urine to the density of water and provides information on the kidneys' ability to concentrate urine.

1.2 Steroid Profile Components

Testosterone is the primary male sex hormone which has been identified in the mid-1930s. In men, testosterone plays a significant role as it is responsible for the development of male reproductive tissues such as the testes and prostate, as well as in promoting many male characteristics such as growth of muscle mass and body hair. Testosterone is produced naturally by the human body. Its production affects the way men store fat in the body, and even men's mood. Testosterone is primarily produced by the testicles in men. Women's ovaries also produce testosterone, though in much lower levels. Furthermore, testosterone is involved in health and the prevention of osteoporosis. Insufficient testosterone levels in men may lead to abnormalities including frailty and bone loss (Schulze et al., 2008; Society for Endocrinology: Testosterone., 2018).

Epitestosterone, also known as isotestosterone, is an endogenous steroid and the 17-α epimer of the androgen sex hormone testosterone which is excreted in the urine in concentrations similar to T. With respect to its structure, epitestosterone has a similar configuration with testosterone. The only chemical difference is in the configuration of the hydroxy-bearing carbon, on C17 (Catlin et al., 1997).

Androsterone is also an endogenous steroid hormone with a potency that is approximately 1/7 that of testosterone. It is included in the human axilla, skin and in the urine as well. Furthermore, it has been shown that the smell of androsterone may affect human behaviour (Maiworm and Langthaler, 1992).

Etiocholanolone, also known as 5β-androsterone, is an etiocholane (5β-androstane) steroid as well as an endogenous 17-ketosteroid that is produced from the metabolism of testosterone. Etiocholanolone is excreted in the urine and elevated values of it (along with testosterone and androsterone) can be detected in the urine of men with androgenic alopecia (male pattern baldness) (Human Metabolome Database: Etiocholanolone., 2018).

5α-Androstane-3α, 17β-diol (5α-Adiol or dihydroandrosterone) is a major testosterone metabolite. It is the main steroid produced by the immature ovary. Testosterone 5a-reduced metabolites, including dihydrotestosterone are produced in the anterior pituitary and the central nervous system (Human Metabolome Database: 5a-Adiol., 2018).

5β-Androstane-3α, 17β-diol (5β-Adiol) or Etiocholanediol is a major metabolite of dihydrotestosterone and belongs to the class of organic compounds known as androgens and derivatives (Human Metabolome Database: Etiocholanediol., 2018).

DHEA or dehydroepiandrosterone is a precursor hormone produced from cholesterol by the body's adrenal glands, although it is also made by the gonads, the brain, testes and ovaries in smaller amounts. Natural DHEA levels peak in early adulthood and then slowly decline as people get older (Human Metabolome Database: DHEA., 2018).

Sections 1.3 and 1.4 discuss the key role of steroid profile in both doping control analysis and clinical diagnosis, respectively.

1.3 Steroid Profile and Doping

1.3.1 Historical Overview of Doping

The issue of doping in sports has been a concern since the 1920s when restrictions on drug use by athletes were first thought necessary. In 1928 the International Association of Athletics Federations (IAAF) - the athletics' world governing body - became the first international sports federation to forbid the use of specific substances as doping products (Sottas et al., 2011; International Amateur Athletic Federation., 2018). Doping has been widely discussed in recent years and it still remains a hot topic in the athletic world. Assertions of doping have started since 1903 in the Tour de France, an annual men's bicycle race primarily held in France. Early Tour riders used ether and consumed alcohol, among other substances, as a means of diminishing the pain of competing in endurance cycling. A major drug scandal happened in 1999 at the Tour de France with the American former professional road racing cyclist, Lance Armstrong. In 2012, a United States Anti-Doping Agency (USADA) investigation concluded that Armstrong had used performance-enhancing drugs over the course of his career and named him as the ringleader of "the most sophisticated, professionalised and successful doping programme that sport has ever seen" (USADA: Armstrong L., 2018).

Another famous doping case prior to Armstrong's confession, Benjamin Johnson, was probably the world's highest-profile drugs cheat. The Canadian former sprinter tested positive for anabolic steroids at the 1988 Summer Olympics in Seoul. Johnson had won the 100m final, lowering his own world record to 9.79 seconds but was stripped of his gold medal after his positive urine test for the banned anabolic steroid stanozolol (Baron et al., 2007).

1.3.2 Doping

In competitive sports, doping is the use of prohibited substances or methods by athletic competitors in order to allow them to train harder, build more muscle and illegally improve their athletic performance (Van Renterghem et al., 2008). Doping is also used in endurance sports for improved recovery as overtraining can disrupt the balance between anabolic and catabolic states of the hormones of the endocrine system (Snyder and Hackney, 2013). The prevalence of doping in elite sports has been estimated to be greater than 40%. However, this estimate can differ widely in various groups of athletes. Steroids refer to the drugs that are closely associated with the notion of doping (Van Renterghem et al., 2008; Mazzeo and Ascione, 2013; Ulrich et al., 2018). Doping refers to an athlete's use of banned drugs, called doping classes (such as androgens, stimulants, hormones, diuretics, narcotics and cannabinoids) and also the use of forbidden methods (such as blood transfusions or gene doping). These are the types of drugs and methods that are banned in sport by sports' governing bodies and the World Anti-Doping Agency (WADA), established in November 1999, due to the damage they can cause to athletes' health (WADA, 2018b).

Abuse of anabolic steroids may lead to short-term effects such as mental problems (e.g. paranoid jealousy, extreme irritability, delusions etc.). Anabolic steroid abuse may also lead to serious, even permanent, health problems. Some of the most common physiological side effects of anabolic steroid abuse are:

- kidney damage,
- liver problems,
- cardiovascular problems,
- increased aggressiveness,
- male pattern baldness and severe acne,
- depression, and
- insomnia (inability to sleep).

There are also gender-specific effects, such as breast tissue development (gynecomastia), shrinking of the testicles, impotence, reduction in sperm production and increased risk for prostate cancer for men. For women, common effects are the deepening of the voice, cessation of breast development, increased facial hair or excess body hair and enlarged clitoris (USADA: Side-Effects., 2018). Besides the risk of athletes' health, the use of performance-enhancing drugs can also violate the spirit of sport, affecting fairness and equality for athletes worldwide.

1.3.3 The "Athlete Biological Passport"

In 2006, WADA with the support of several international sports federations recruited a group of experts to develop a programme based on longitudinal profiling, or serial analysis of indirect biological markers of doping, that was both scientifically and legally robust. This culminated in the WADA Athlete Biological Passport (ABP) Operating Guidelines and Technical Documents, introduced in Pottgiesser and Schumacher (2012).

A biological passport is an individual electronic document for professional athletes, in which profiles of selected biomarkers of doping, relevant information including training and also the results of doping tests are collated throughout their career. It is worth noting that athletes have their own metabolism and different responses after a drug intake, which creates significant inter-individual variation. The method of ABP steroid profiling fights against doping, overcoming the limitations of population cut-offs. Subsequently, if these biomarkers' levels change significantly within the steroid profile of an athlete, it alerts athlete passport management units (APMUs) that anomalies have been detected that require further testing (Sottas et al., 2010; Sottas et al., 2011; Vernec, 2014; WADA, 2021b). However, the ABP mostly aids in revealing the direct and indirect effects of doping with anabolic agents rather than detecting the prohibited substance itself (Kuuranne et al., 2014; Piper et al., 2021).

1.3.4 Steroid Profile of ABP

The steroid profiling of ABP, also known as steroidal module, is used to denote a follow-up, which essentially is the recording of the concentration levels and ratios of endogenous steroids in urine over time. WADA and other anti-doping laboratories provide harmonised and robust analytical methods for the "steroid profile", which according to their technical document (TD) (WADA, 2021a), is composed by the following endogenous anabolic androgenic steroids (EAAS): testosterone (T), epitestosterone (E), androsterone (A), etiocholanolone (Etio), 5α-androstane-3α, 17β-diol (5α-Adiol or A5), 5β-androstane-3α, 17β-diol (5β-Adiol or B5), as well as the ratios of selected steroids; i.e. T/E, A/T, A/Etio, 5α-Adiol/5β-Adiol (A5/B5) and 5α-Adiol/E (A5/E).

The above mentioned are considered as valuable biological markers for the administration of endogenous steroids and they are calculated to detect multiple forms of steroid doping (Donike et al., 1983; Strahm et al., 2009; Van Renterghem et al., 2010; WADA, 2018a). Specifically, the use of ratios provides a major benefit because it can eliminate the dependence between plain markers and urine volume, but also other factors that affect concentration.

The steroid profile is reported in the Anti-Doping Administration & Management System (ADAMS) by WADA accredited laboratories for all urine samples (Pottgiesser and Schumacher, 2012). It is worth mentioning that monitoring the SP at individual level is very important. The reason is the large inter-individual variability in the urinary steroid concentrations caused by various factors, and thus the reference values based on the population do not always have the sensitivity to track whether anabolic drugs have been administered (Sottas et al., 2008; Van Renterghem et al., 2010).

Furthermore, various confounding factors have been found in the evaluation of an individual ABP steroid profiling. In 2014, Kuuranne et al. pointed out that the factors which affect the levels of the main components of the steroid profile in urine are divided into two categories, i.e. the *endogenous* and the *exogenous* factors. The major endogenous factors, which may lead to physiological variation within the long-term steroid profile, are the athlete's age, gender, ethnicity and genetic polymorphisms. With regard to the urinary steroid profile, environmental conditions, medications and diet are the main external factors which may cause significant variations in the steroids metabolism. Multiple factors of confounding are stored in the ABP for achieving improved decision making.

1.3.5 Anabolic Androgenic Steroids

Anabolic androgenic steroids (AAS) are natural steroidal hormones like the male sex hormone testosterone (the most significant androgenic steroid) as well as synthetic variations of androgens that are structurally related and have similar effects to testosterone. The word "anabolic" refers to muscle building, and the word "androgenic" (i.e. andro = male, genic = formation) refers to the development and maintenance of male sex characteristics such as the growth of facial and body hair. Figure 1.2 is a graphical representation of the chemical structure of the natural AAS testosterone, found in Human Metabolome Database: Testosterone. (2018). According to the "Prohibited List 2018" of the World Anti-Doping Agency, AAS belong to the class S1.1 and are banned in and out of competition (WADA, 2018b).

Anabolic androgenic steroids are divided into "Endogenous" and "Exogenous" AAS. Endogenous anabolic androgenic refers to the administration of steroids that are capable of being naturally produced by the human body, whereas exogenous AAS replacement is when androgens enter directly into the body without being ordinarily produced by it (WADA, 2018b). Exogenous AAS are usually given either orally in a tablet form or injected into the muscles. Some steroids are also applied to the skin in creams, gels or patches. Sometimes steroids are used in medicine, but illegal use of AAS may involve doses 10 to 100 times higher than the normal prescription dose (Van Renterghem et al., 2008; Mazzeo and Ascione, 2013; Andersen and Linnet, 2014; Davis, 2018).

“Cycling”, “stacking”, and “pyramiding” are three common ways that anabolic steroid abusers take their drugs believing that they can avoid unwanted side effects or optimise the drugs’ effects. “Cycling” refers to taking a steroid for a specific period of time, stopping thereafter for some time allowing body to rest, and then restarting again. While “cycling” is associated with taking one type of steroid, “stacking” is when people use more than one type of anabolic steroids at a time in high dosages. There is the belief that combining two or more different types of steroids at a time increases the effectiveness of each, over taking them individually. “Pyramiding” combines “cycling” and “stacking”. The steroid abusers start taking one or more steroids in a low dose which is progressively increased till the peak is reached half way where the amount is maximised and it is then tapered to zero by the end of the cycle (Mazzeo and Ascione, 2013).

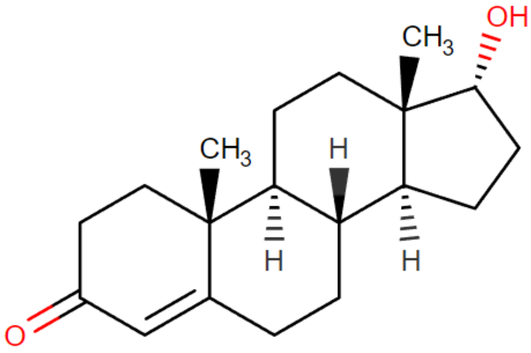


Figure 1.2: Chemical structure of the natural anabolic androgenic testosterone (androst-4-en-17β-ol-3-one), Human Metabolome Database: Testosterone. (2018).

1.3.6 Doping Control Programmes

Doping control programmes have become synonymous with body-derived tests; steroid profiling, Adverse Analytical Findings (AAF) and Atypical Passport Findings (ATPF), where longitudinal profiles of high level athletes are collected to establish population-based comparisons and to test individual endocrine profiles for miscellaneous hormones such as anabolic androgenic steroids, human growth hormone (hGH) and erythropoietin (EPO), through individual reference ranges. These programmes play an important role in revealing the misuse of illegal substances at individual level (Vernec, 2014; Kuuranne et al., 2014).

Some drugs such as steroids, often remain in the body for prolonged periods of time and can be detected by testing urine or blood samples or even using hair follicles. AAS are the class of substances that still gather a high number of AAF and ATPF, due to the recent developments in methods of GC-MS and liquid chromatography with tandem mass spectrometry (LC-MS-MS) that are being applied to detect synthetic anabolic steroids of the biological passport in urine samples (Catlin et al., 1997; Sottas et al., 2010; Mazzeo and Ascione, 2013; Andersen and Linnet, 2014; Parr and Schänzer, 2010; WADA, 2019). Administration of synthetic forms of endogenous anabolic androgenic steroids (EAAS) may lead to alterations of the urinary steroid profile (Mareck et al., 2008; WADA, 2018a). For example, each substance the sample contains has a unique "fingerprint" and as the analysts already know the properties (e.g. the weight) of many steroids they are able to rapidly detect doping. However, it is not always easy to perceive when a sample contains doping substances. Some by-products of doping substances are so small they may not produce a strong enough signal for detection. Consequently, EAAS have become the most popular doping drugs, since the distinction between the endogenous or exogenous origin from the substance remains challenging for the anti-doping laboratories, insomuch that the EAAS is likely the most abundant misused family of substances in elite sports (Van Renterghem et al., 2008).

Furthermore, "masking agents" and "diuretics" are included in the class S5 of the "Prohibited List 2018" (WADA, 2018b). They are designer steroid drugs and constructed to be less detectable removing fluid from the body. It is possible that they may escape detection by giving a false negative test. However, the World Anti-Doping Agency collaborates with many laboratories in order to develop improved detection techniques for controlling the performance-enhancing drugs in the body.

Using blood screening is more likely to detect banned substances and is more difficult to beat through "masking" methods. On the other hand, urine tests are easier to obtain and no needle stick is required. Throughout the years of anti-doping research, several urine tests have been developed to identify steroid abuse, such as the human growth hormone (hGH) urine test. The hGH represents a hormone that is naturally produced by the body. hGH stimulates many metabolic processes in cells and plays an important role in muscle protein synthesis and organ growth. The hGH test uses nanotechnology to bind and amplify hGH in urine so that it may be detectable for a longer period of time. Blood screening can only detect hGH, administered within the previous 24 to 48 hours. Nanotechnology may allow urine detection out to a two-week range (Pottgiesser and Schumacher, 2012; WADA: HGH test., 2018).

Other important screening tests for doping are the tests for testosterone abuse, conducted in urine samples. Specifically, the most commonly used test over the last years in forensic toxicology uses the ratio of urinary testosterone to epitestosterone (T/E) due to it being considered a stable markers' ratio within an athlete's steroid profile, but also sensitive to the administration of T. Measuring only the testosterone levels has proved inadequate due to the small ratio of intra- to inter-individual variability in the urinary steroid concentrations caused by various factors (Harris, 1974; Brooks et al., 1979; Sottas et al., 2006; Mareck et al., 2008). In 1993, Dehennin and Matsumoto have pointed out that high levels of the ratio of testosterone to epitestosterone in urine may indicate that an individual has taken exogenous testosterone, since the administration of testosterone decreases the concentration of epitestosterone. Since the 1980s, a T/E ratio of initially >6:1 and after 2004 >4:1 is considered suspicious of steroid doping (Donike et al., 1983; Schulze et al., 2008; Van Renterghem et al., 2010; Andersen and Linnet, 2014). This screening test has become an indispensable tool in anti-doping laboratories for the identification of synthetic AAS in urine samples, despite the fact that the method is not sensitive to indirect androgen doping. It is worth mentioning that a T/E recorded value lower to this critical value does not necessarily mean that an individual has not used testosterone recently. Although, it is reported at the laboratories as a negative result, it is recognised that it may be a false-negative (Catlin et al., 1997).

Other urine ratios such as testosterone over luteinizing hormone (LH) and the introduction of sulfoconjugates with biomarkers, such as the ratio testosterone sulfate/epistestosterone sulfate (Ts/Es), or testosterone glucuronide/testosterone sulfate (Tg/Ts), or (Tg+Ts)/(Eg+Ts) and many others, have been used largely to develop sensitive tests for AAS abuse. The major benefit of using the ratios instead of the plain markers is the independence between ratios and urine volume, and also other factors that affect concentration (Catlin et al., 1997; Sottas et al., 2010).

1.3.7 Limitations and Further Directions

According to the above, fighting against doping is a matter of high importance in order to protect athletes' health and the spirit of sport, but also to beat the sophisticated network of black market doping programmes that follows the modern sports industry. Given that the reference values based on the population are not always reliable to track doping misuse, it is important to develop more robust methods to monitor the steroid profile at individual level (Sottas et al., 2008; Van Renterghem et al., 2010). Current approaches receive new measurements of a single biomarker or ratio and, under a Bayesian framework, progressively adapt population-derived limits, which are initially used when there are no recorded measurements, to individual normal boundaries as the number of measurements increases (Sottas et al., 2006). Multivariate statistical approaches have also been proposed for this purpose, which are able to combine population information with individual longitudinal monitoring of multiple biomarkers (Alladio et al., 2016; Amante et al., 2019). In contrast to the large number of statistical models which analyse a sequence of biomarkers, the development of robust methods for the detection of abnormal variations of multiple longitudinal biomarkers has remained limited.

1.4 Steroid Profile for Prostate Cancer Diagnosis

1.4.1 Prostate Cancer and Benign Prostatic Hyperplasia

Prostate cancer (PCa) is the uncontrolled development of cancer cells in the prostate, which is currently the most frequently diagnosed non-skin cancer in men and a frequent cause of morbidity and mortality. In fact, it is the fifth leading cause of death worldwide among men over the age of 65 (Rawla, 2019). Benign prostatic hyperplasia (BPH) is also very common urologic condition in older men and can cause the urinary tract to be obstructed. Unlike prostate cancer, BPH is a non-cancerous increase in size of the prostate. As men age, their testosterone levels increase, which in turn causes their prostate to grow in size (Schenk et al., 2011). Early prostate cancer is often asymptomatic, however the most common early symptoms of PCa and BPH include difficulty with urination, increased urinary frequency, dribbling, and frequent night-time urination.

The etiology of prostate cancer remains unknown and there is still no proven prevention strategy for it. Research has shown that African American men, Caribbean men, Black men in Europe and men with a family history in prostate cancer are higher risk groups. Their higher prostate cancer incidence and mortality rates are possibly due to certain genes which are more sensitive to mutations in prostate cancer (Kheirandish and Chinegwundoh, 2011; Rawla, 2019). Except for age, ethnicity and genetic factors, other risk factors that are related to PCa are physical activity and diet. According to various studies, it is possible to reduce the risk of prostate cancer by making healthy choices, such as exercising and eating healthy (Kolonel et al., 1999; Fenton et al., 2018; Rawla, 2019).

1.4.2 Screening Programmes and Overdiagnosis

Serum prostate specific antigen (PSA) level and digital rectal examination (DRE) constitute the major screening tests for PCa diagnosis. The PSA test measures the level of PSA (a specific prostate marker) in a man's blood. For this test, a blood sample is sent to a laboratory for analysis. The results are usually reported as nanograms of PSA per milliliter (ng/mL) of blood, where elevated PSA levels may indicate cancer of prostate (usually for PSA $>$ 4 ng/mL) (Litwin and Tan, 2017). During the last decades since the PSA testing is used, the mortality rate due to prostate cancer has declined. However, it is not clear whether this is caused by PSA screening or by improved cancer treatments (Etzioni and Feuer, 2008). It is noteworthy that PSA levels can also be raised by other non-cancerous conditions such as BPH (Meigs et al., 1996; Mechergui et al., 2009). In such cases, a tissue biopsy is the standard of care to diagnose prostate cancer. According to the American Cancer Society (ACS), which systematically reviewed the literature assessing PSA performance, the estimated sensitivity of a PSA cut-off of 4.0 ng/mL was 21% for detecting any prostate cancer. Using a cut-off of 3.0 ng/mL increased the sensitivity to 32%. The estimated specificity was 91% for a PSA cut-off of 4.0 ng/mL and 85% for a 3.0 ng/mL cut-off. Due to its high false negative error rate, the widely used PSA testing has poorer discriminating ability, especially in men diagnosed with symptomatic BPH, and remains somewhat controversial (Wolf et al., 2010).

Another important health problem regarding early detection of PCa is the potential overdiagnosis of cases that would not have caused clinical consequences during a man's lifetime if left untreated. Overdiagnosis directly results to overtreatment which can be extremely aggressive and might cause unnecessary side effects (Wolf et al., 2010). According to the systematic review of Fenton et al. in 2018, PCa overdiagnosis estimates range from 20.7% to 50.4%.

1.4.3 Relationship between Steroids and Prostate Cancer

Steroid hormones have a potentially important role in the pathogenesis of both prostate cancer and benign prostatic hyperplasia, since the prostate function depends on the hormonal physiology, which subsequently depends on several factors, such as genetic, lifestyle and dietary factors (Kolonel et al., 1999; Albini et al., 2018).

Especially, prolonged presence of androgens and estrogens in tissue and serum may also induce significant alterations in the metabolism, which eventually may be a crucial factor in the prostate enlargement that can play some part in the development of BPH and PCa (Carruba, 2007). Mass spectrometry (MS) is currently the technique of choice for the analysis of steroids in various biological samples and its applications could lead to a diagnostic approach (Yeap, 2014; Adaway et al., 2015; Albini et al., 2018). It is also very important to evaluate the quantitative ratio of steroids due to their complex biological pathways and stoichiometry (Zhang et al., 2017). However, the exact role of steroid hormonal factors has been poorly understood since the precise mechanisms by which factors affect the process of prostatic carcinogenesis are unknown.

1.4.4 Limitations and Further Directions

The emphasis should be, therefore, put on the need to investigate the association between BPH and PCa with the levels of steroids. It is crucial to examine whether other sensitive biomarkers, more accurate than PSA, can be applied with robust statistical methods and provide a less invasive addition to the management, diagnosis and prognosis of prostate cancer, reducing unnecessary prostate biopsies (Carruba, 2007; Kelly et al., 2016).

1.5 Thesis Objectives

This thesis focuses on evaluating urine metabolomic profiles in professional athletes and patients with prostate cancer and benign prostate hypertrophy for anti-doping and clinical purposes, respectively. The main objective is to extend standard statistical methodologies on density estimation and taxonomy problems to propose novel classification tests, designed for doping control analysis and prostate cancer detection.

In anti-doping research, univariate and multivariate statistical approaches have been previously conducted, but Bayesian hierarchical modelling is still under-explored. The first part of the thesis presents a Bayesian multilevel model to analyse multivariate longitudinal data from athletes. The main goal is to construct an adaptive model which defines personalised threshold values for the various biomarkers of each athlete as new observations become available. Semi-supervised classification algorithms based on prior knowledge for the majority class of healthy individuals are implemented to determine whether a value from a new urine sample belongs to the normal class or behaves as an outlier that needs further investigation.

In prostate cancer research, Bayesian multivariate methods remain unexplored in studying the utility of potential metabolite biomarkers, which could enhance the prediction performance of prostate cancer. The second part of this work aims to explore new target biomarkers based on pattern recognition methods to address challenges in distinguishing patients with PCa from patients with BPH and healthy individuals. The developed models are able to deal with high-dimensional and highly correlated datasets and are intended to be used as an improved diagnostic tool in terms of accuracy for reducing the morbidity and mortality from prostate cancer.

1.6 Thesis Structure

Chapter 1 gave an extensive description of the notions of the Steroid Profile and its components, the Athlete Biological Passport and the Anabolic Androgenic Steroids within the context of doping detection. It also collocated the importance of analysing steroids for both prostate cancer and benign prostatic hyperplasia diagnosis. In addition, it introduced the motivation behind using statistical modelling to delineate the mechanism of metabolite biomarkers in the areas of forensic toxicology and cancer research. Emphasis is given on the limitations and the need for robust statistical methods, which can contribute significantly in these areas.

Chapter 2 constitutes the first part of this work, which focuses on the development of an improved non-invasive test for classifying athletes' urine samples into *suspicious* and *non-suspicious* classes. Recordings from athletes with normal concentration values are easily available compared to doped athletes. Since the imbalance in size between the two classes is unavoidable, the classification technique we used is based on a one-class classification method (Khan and Madden, 2014). The AAS concentration levels from non-doped athletes define the "target" class for which adaptive decision boundaries are constructed to separate them from abnormal data. The emphasis is on the flexibility of the proposed model, which takes into account prior information on inter- and intra-individual variations, but also on potential correlation between the EAAS markers. The multivariate Bayesian model is gradually built under the following stages; a) the univariate Bayesian model, b) the univariate Bayesian multilevel model, and c) the multivariate Bayesian multilevel model, which are compared to other statistical methods previously used for the same purpose. Inference and computations, such as the out-of-sample predictive distribution, were either based on samples from the known joint posterior distributions or from a non-closed form distribution using the Markov chain Monte Carlo techniques, depending on the model.

Chapter 3 presents the applications on real athletes' urinary samples over time consisting of confirmed normal and abnormal observations, which are carried out to assess the performance of the models of Chapter 2. Due to its small sample size, our dataset makes it challenging to estimate the distribution of the abnormal class. Therefore, the strategy is to estimate the characteristics of the normal class only and test every athlete's measurement for deviations. A one-class classification method is implemented in learning an efficient Bayesian classifier to discriminate between *non-suspicious* and *suspicious* samples, either due to doping misuse or due to other confounding factors, the results of which are discussed in this chapter. An Rshiny application, called BioScan App, has been developed to apply the proposed methodology specifically for athletes' testing purposes.

Chapter 4 focuses on the second part of this research work, in which the steroidal biomarkers are used for constructing a machine learning Bayesian model to describe prostate cancer's behaviour. Novel classification methods are presented based on Dirichlet Process Gaussian mixture models with and without covariates in a Bayesian nonparametric framework.

Chapter 5 presents the performance of density estimation, clustering and classification methods implementing the Dirichlet Process Gaussian mixture models on simulated and real clinical and non-clinical data. Model comparisons among other methods suitable for prostate cancer detection demonstrate the superiority of the developed methodology.

Chapter 6 concludes with a discussion of the research findings presented in this thesis and provides proposals for future work.

Chapter 2

Adaptive models and anomaly detection techniques for doping

2.1 Introduction

Many statistical methods have been developed for modelling the steroid profile of athletes over the last years. Most of them are models based on measurements of certain biological markers or ratios, such as testosterone (T), T/E, A/Etio, DHT/E, which have been considered to be the most informative and indicative markers for monitoring the endogenous steroid abuse (Van Renterghem et al., 2008; Van Renterghem et al., 2013). Moreover, until recently doping control laboratories used threshold values to characterise a control sample as suspicious. One of the major problems with this approach is that these thresholds are based on population measurements, while physiological values of the urinary module may vary considerably among the individuals due to many different reasons, such as ethnical variations. This indicates that there is room for improvement on the methods to detect administration of endogenous steroids (Saudan et al., 2006; Van Renterghem et al., 2008).

One of the basic tools of the ABP is the statistical model of Sottas et al., published in 2006, which allows an optimised evaluation of longitudinal data. A major benefit of this approach is that as we take individual measurements into account we move towards a subject-specific threshold that is obtained from a combination of the population distribution and the individual subject's distribution. When the number of measurements from the subject is small, this threshold will be closer to the population threshold, but with an increasing number of measurements, it will shift towards the subject-specific distribution. This is the main idea behind the work of Sottas et al. (2006), that is obtaining thresholds for a single and certain marker, i.e. the T/E ratio, for the detection of abnormal variations. However, multivariate statistical methods in a Bayesian framework have not yet been attempted.

2.2 Objectives

In this chapter, the three stages of the proposed Bayesian model are presented. First, the univariate model, secondly, the univariate multilevel model, and lastly, the multivariate Gaussian multilevel model. Emphasis is put on the process to extract information from one or higher-dimensional probability distributions to classify the new "unlabelled" data from athletes either to the normal class (*non-suspicious*), if they adapt smoothly in the distribution of the "normals", or to the abnormal class (*suspicious*). In cases where the posterior density functions are not known, the parameters of the models of normal athletes are estimated by using Markov chain Monte Carlo (MCMC) sampling methods, which are described in Section 2.3.5. A semi-supervised learning technique using a one-class classification (OCC) method is described in Section 2.3.8 (Khan and Madden, 2014). The OCC method was applied to train the models using training sets only from the normal class, i.e. the majority class. Throughout this chapter, all vector quantities are denoted by bold-faced characters.

2.3 Materials and Methods

2.3.1 The Univariate Bayesian Model

This section introduces an expanded version of the univariate Bayesian modelling for T/E ratio proposed by Sottas et al. (2006) to a univariate Bayesian hierarchical method to model any steroidal component or ratio of the ABP. Under a Bayesian framework, the model receives new measurements and progressively adapts population-derived limits, when the number of measurements n is zero, to individual normal boundaries, when n is large.

Let $\boldsymbol{y} = (y_1, y_2, ..., y_n)$ represent the vector with the n log-transformed recorded EAAS values collected from the same athlete. It is important to mention that the period between two sequential samples of an athlete is assumed long enough for them to be considered independent. Hence, we assume the logarithm of EAAS values to be a vector of n independent and identically distributed draws from a Gaussian distribution with mean μ and variance σ^2, i.e. $\boldsymbol{y} \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. Note that we focus on modelling the logarithm of these EAAS concentrations due to the fact that there are physical constraints on the measurement values (i.e. all markers are positive) and taking the logarithm allows us to use a Gaussian distribution to model the log-transformed markers. Furthermore, we chose the Gaussian distribution allows us to generalise the model of Sottas et al. (2006), which is also based on a Gaussian distribution for the T/E ratio, to model any biomarker, as well as to easily extent to a multivariate model using the Multivariate Gaussian distribution.

To describe our prior knowledge about the unknown parameters, i.e. the mean μ and the precision $\tau = 1/\sigma^2$, we specified the joint prior distribution as the product of a conditional and a marginal distribution expressed as $p(\mu, \tau) = p(\mu|\tau)p(\tau)$. Given that we have limited prior information on the parameters of the model regarding the six available biomarkers and their five ratios except for the T/E ratio, we discuss the case of specifying less informative conditionally conjugate priors on these model parameters. The T/E ratio is excluded from the semi-informative prior setting because there is adequate population information regarding its characteristics in the paper of Sottas et al. (2006). However, we present its results in both cases of using more and less informative priors for comparison purposes.

2.3.1.1 The Gaussian-Gamma Conjugate Family

This section is an introduction to the conjugate Gaussian-Gamma family of distributions where the posterior distribution is in the same family as the prior distribution and leads to a marginal Student-t distribution for posterior inference for the mean of the population. This model constitutes the simplest form of the proposed hierarchical models and can be applied to any single marker or ratio at a time.

The Likelihood function

The probability density function of a generic draw y_i is $f(y_i|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2}$. Hence, the likelihood for all observations is

$$\mathcal{L}(\mu,\sigma^2;\boldsymbol{y}) = \prod_{i=1}^{n} f(y_i|\mu,\sigma^2) = (2\pi\sigma^2)^{-n/2}\exp\Big\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\Big\}, \tag{2.1}$$

for $y,\mu \in \mathbb{R}$ and $\sigma > 0$. With some algebraic computations (see Appendix A; A.1), the likelihood function for the mean, μ, and the precision, $\tau = 1/\sigma^2$, can be expressed as

$$\mathcal{L}(\mu,\tau;\boldsymbol{y}) \propto \tau^{n/2}\exp\Big\{-\frac{\tau}{2}\sum_{i=1}^{n}(y_i-\bar{y})^2\Big\}\exp\Big\{-\frac{\tau}{2}n(\bar{y}-\mu)^2\Big\}, \tag{2.2}$$

where $\bar{y}$ is the sample mean. In equation (2.2), the likelihood is proportional to a product of two parts; that is a term that is a function of τ and the data, and a term that involves μ,τ and the data.

Conditional conjugate priors for μ and τ

As both the mean μ and the precision $\tau = 1/\sigma^2$, are unknown parameters, we need to specify a joint prior distribution, $p(\mu, \tau)$, to describe our prior knowledge about them. Based on the factorised likelihood in equation (2.2), and the fact that any joint distribution can be expressed as the product of a conditional and a marginal distribution, the joint distribution for μ and τ can be expressed as

$$p(\mu, \tau) = p(\mu|\tau)p(\tau), \tag{2.3}$$

which is the product of the conditional distribution for μ given τ and the marginal distribution for τ. Therefore, the prior setting is hierarchical. We first specify a prior to the mean conditional on the inverse variance. As the conjugate prior distribution for μ given τ is a Gaussian distribution, we will rely on this to assign the Gaussian distribution to the mean conditional on the precision as

$$\mu|\tau \sim \mathcal{N}(\mu_0, 1/(\kappa_0\tau)), \tag{2.4}$$

with hyper-parameters μ_0 (prior mean) and κ_0 (prior sample size). The prior sample size κ_0 determines how tight is the prior, that is, how probable we deem μ to be very close to the prior mean μ_0. As the prior sample size becomes larger, $1/\kappa_0\tau$ becomes smaller, which indicates that we know the mean with more precision (relative to the variability in observations). On the other hand, smaller prior sample size indicates less precision or more uncertainty. Hence, a non-informative prior can be specified when $\kappa_0 = 0$.

Since the precision τ is also unknown, we assign a prior distribution to describe the uncertainty about it before seeing the data. This parameter is non-negative, continuous, and with no upper limit. It turns out that the inverse of the variance has a conjugate Gamma prior distribution expressed as

$$\tau \sim \mathcal{G}a(\alpha_0, \beta_0), \tag{2.5}$$

with probability density function

$$p(\tau) \propto \tau^{\alpha_0 - 1} \exp\{-\beta_0\tau\}, \tag{2.6}$$

where α_0 and β_0 determine the *shape* and *rate* parameters, respectively. From equation (2.3), the multiplication of both priors gives us the joint distribution for the pair (μ, τ), which is called Gaussian-Gamma family of distributions

$$(\mu, \tau) \sim \mathcal{NG}a(\mu_0, \kappa_0, \alpha_0, \beta_0), \tag{2.7}$$

with hyper-parameters μ_0, κ_0, α_0 and β_0. The joint probability density function is written as

$$p(\mu, \tau) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \left(\frac{\kappa_0}{2\pi}\right)^{\frac{1}{2}} \tau^{\alpha_0 - 1/2} \exp\left\{ -\frac{\tau}{2}[\kappa_0(\mu - \mu_0)^2 + 2\beta_0]\right\}. \tag{2.8}$$

This suggests that the posterior distribution will be the product of two conjugate distributions, as we can see in equation (2.9).

Conjugate posterior distribution

As a conjugate family, the joint posterior distribution for the pair of parameters (μ, τ) is in the same family as the prior distribution when the sample data arise from a Gaussian distribution, i.e. the posterior is also a Gaussian-Gamma distribution (see Appendix A; A.2)

$$\mu, \tau | \boldsymbol{y} \sim \mathcal{NG}a(\mu_n, \kappa_n, \alpha_n, \beta_n), \tag{2.9}$$

where the index n on the hyper-parameters indicates the updated values after seeing the n observations from the sample data. One utility of conjugate families is in the relatively simple updating rules, which can be used for obtaining the new hyper-parameters as

$$\kappa_n = \kappa_0 + n \tag{2.10}$$

$$\mu_n = \frac{\kappa_0 \mu_0 + n\bar{y}}{\kappa_n} \tag{2.11}$$

$$\alpha_n = \alpha_0 + n/2 \tag{2.12}$$

$$\beta_n = \beta_0 + \frac{1}{2}\sum_{i=1}^{n}(y_i - \bar{y})^2 + \frac{\kappa_0 n}{2\kappa_n}(\bar{y} - \mu_0)^2, \tag{2.13}$$

where the updated hyper-parameter μ_n is the posterior mean for μ; it is also the mode and median. The posterior mean μ_n is a weighted average of the prior mean μ_0 and sample mean $\bar{y}$ with weights $w_1 = \frac{\kappa_0}{\kappa_0+n}$ and $w_2 = \frac{n}{\kappa_0+n}$ that are proportional to the prior sample size, κ_0, and the sample size, n, respectively. The posterior sample size κ_n is the sum of the prior sample size κ_0 and the sample size n, representing the combined sample size after seeing the data. The posterior shape parameter α_n is also increased by adding a half of the sample size $(n/2)$ to the prior shape parameter α_0. Finally, the posterior sum of squares, β_n, combines the prior sums of squares, β_0, and the sample sum of squares $\sum_{i=1}^{n}(y_i - \bar{y})^2$. Similarly, we can interpret the third term as the discrepancy between the prior mean and sample mean.

The joint posterior distribution is a hierarchical model, where in the first stage of hierarchy the precision marginally follows a Gamma distribution

$$\tau|\boldsymbol{y} \sim \mathcal{G}a(\alpha_n, \beta_n), \tag{2.14}$$

and in the second stage, μ given τ has a conditional Gaussian distribution

$$\mu|\boldsymbol{y}, \tau \sim \mathcal{N}(\mu_n, (\kappa_n \tau)^{-1}). \tag{2.15}$$

This representation of the model is convenient for generating samples from the posterior distribution.

Marginal distribution for μ

The marginal inference requires the unconditional or marginal distribution of μ that “averages” over the uncertainty in τ. To compute the marginal distribution of μ this averaging is performed by integration, and as we can see below, it leads to a Student-t distribution with density

$$p(\mu|\boldsymbol{y}) \propto \int_0^\infty p(\mu, \tau|\boldsymbol{y}) d\tau$$

$$\propto \overbrace{\int_0^\infty \tau^{\overbrace{\frac{2\alpha_0+n+1}{2}}^{a}-1} \exp\Big\{ -\tau\underbrace{(\beta_0 + \frac{(\kappa_0+n)(\mu-\mu_n)^2}{2} + \frac{\sum_{i=1}^n (y_i-\bar{y})^2}{2} + \frac{\kappa_0 n(\bar{y}-\mu_0)^2}{2(\kappa_0+n)})}_{b}\Big\}}^{\mathcal{G}a(a,b)} d\tau$$

$$\propto \frac{\Gamma(a)}{b^a} \propto \left(\beta_0 + \frac{(\kappa_0+n)(\mu-\mu_n)^2}{2} + \frac{\sum_{i=1}^n (y_i-\bar{y})^2}{2} + \frac{\kappa_0 n(\bar{y}-\mu_0)^2}{2(\kappa_0+n)}\right)^{-\frac{2\alpha_0+n+1}{2}}$$

$$\propto \left(1 + \frac{1}{2(\alpha_0+\frac{n}{2})} \frac{(\alpha_0+\frac{n}{2})(\kappa_0+n)(\mu-\mu_n)^2}{\beta_0 + \frac{1}{2}\sum_{i=1}^n (y_i-\bar{y})^2 + \frac{\kappa_0 n}{2(\kappa_0+n)}(\bar{y}-\mu_0)^2}\right)^{-\frac{2\alpha_0+n+1}{2}}. \tag{2.16}$$

If we write $t = (\mu - \mu_n)/\sqrt{\beta_n/(\alpha_n\kappa_n)}$, then we get a density for t proportional to

$$\left(1 + \frac{t^2}{2\alpha_n}\right)^{-(2\alpha_n+1)/2}, \tag{2.17}$$

and this is proportional to the density of a standard Student's-t distribution which is is centered at μ_n (the location parameter), scaled at $\beta_n/(\alpha_n\kappa_n)$ (squared scale parameter) like in a standard Gaussian, and with $2\alpha_n$ degrees of freedom. We can thus express the distribution of the parameter μ given the data as

$$\mu|\boldsymbol{y} \sim t_{2\alpha_n}(\mu_n, \beta_n/(\alpha_n\kappa_n)). \tag{2.18}$$

The parameters μ_n and β_n play similar roles in determining the centre and dispersion of the distribution as in the Gaussian distribution, however, as Student-t distributions with degrees of freedom less than 3 do not have a mean or variance, the parameter μ_n is called the location of the distribution and the $\beta_n/(\alpha_n\kappa_n)$ is the scale.

The Student-t and Gaussian distributions are both symmetric about the centre and unimodal (i.e., single-peaked). The difference between them is that the Student-t distribution is less concentrated around its peak and its tails are fatter. Figure 2.1 displays the probability density function of Student's-t distribution with various degrees of freedom in comparison to the Gaussian distribution.

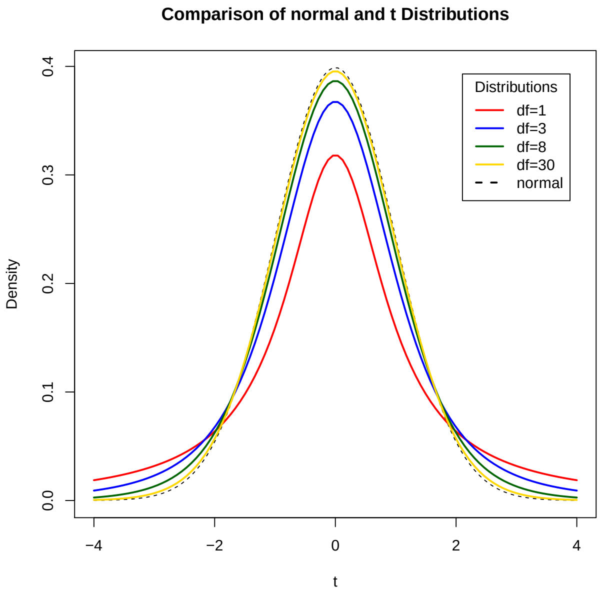


FIGURE 2.1: Standard Gaussian (or Normal) and various Student-t densities.

Posterior approximation

Here, we use a trivial example to examine the sampling performance of Gibbs sampler compared to sampling from the exact posterior distribution. Therefore, we overlook the conjugacy for the joint posterior as we described earlier in Section 2.3.1.1, as well as the information that the distribution of τ given the data is not dependent on μ, as it is shown in equation (2.14). Consequently, through the introduction of the data we hierarchically pass from the prior evidence to the revised knowledge (see Appendix A; A.3), expressed in the posterior density

$$p(\mu,\tau|\boldsymbol{y}) \propto \tau^{\frac{2\alpha_0+n+1}{2}-1} e^{-\frac{\tau}{2}[\beta_0+(\kappa_0+n)(\mu-\frac{\kappa_0\mu_0+n\bar{y}}{\kappa_0+n})^2+\frac{\kappa_0 n(\bar{y}-\mu_0)^2}{\kappa_0+n}+\sum_{i=1}^{n}(y_i-\bar{y})^2]}. \tag{2.19}$$

Assuming that we cannot observe whether the joint posterior can be stated in a closed form, the Gibbs sampling algorithm (Casella and George, 1992) is applied in order to generate parameter samples from the target distribution. In this case the joint posterior distribution $p(\mu,\tau|\boldsymbol{y})$ can be expressed as

$$p(\mu,\tau|\boldsymbol{y}) = p(\mu|\boldsymbol{y})p(\tau|\mu,\boldsymbol{y}) = p(\tau|\boldsymbol{y})p(\mu|\tau,\boldsymbol{y}). \tag{2.20}$$

To apply the Gibbs sampler and simulate the parameters μ and τ, we need to identify the full conditional posterior distributions for these parameters (see Appendix A; A.4). As we can see below in equations (2.21) and (2.22), both conditional distributions have the same starting point, the full joint posterior distribution. This means that after a number of Gibbs steps, the sampler finally obtains a sample from the target distribution without computing any integrals. The conditional distributions for the mean μ and the variance τ are

$$p(\mu|\tau, \boldsymbol{y}) \propto p(\mu, \tau|\boldsymbol{y}) \propto e^{-\frac{\tau(\kappa_0+n)}{2}(\mu-\frac{n\bar{y}+\kappa_0\mu_0}{\kappa_0+n})^2}, \tag{2.21}$$

which implies that $\mu|\tau, \boldsymbol{y} \sim \mathcal{N}\left(\frac{n\bar{y}+\kappa_0\mu_0}{\kappa_0+n}, \frac{1}{\tau(\kappa_0+n)}\right)$, and

$$p(\tau|\mu, \boldsymbol{y}) \propto p(\mu, \tau|\boldsymbol{y}) \propto \tau^{\frac{2\alpha_0+n+1}{2}-1} e^{-\tau[\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2}+\frac{\kappa_0(\mu-\mu_0)^2}{2}+\beta_0]}, \tag{2.22}$$

where $$\tau|\mu, \boldsymbol{y} \sim \mathcal{G}a\left(\frac{2\alpha_0+n+1}{2}, \frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2} + \frac{\kappa_0(\mu-\mu_0)^2}{2} + \beta_0\right).$$

Under conditional conjugacy, each simulation step is simple and the sampler generates a Markov chain $(\mu^{(t)}, \tau^{(t)})$, where t takes values from 1 up to T, according to the following steps of Algorithm 2 in Section 2.3.5.2. If y_{n+1} is the next measurement, we would subsequently like to compute the probability function of that new test result given the previous recorded values. This can be found by computing the following predictive density

$$p(y_{n+1}|y_1, y_2, ..., y_n) = \int\int f(y_{n+1}|\mu, \tau)p(\mu, \tau|y_1, y_2, ..., y_n)d\mu\, d\tau, \tag{2.23}$$

which is exactly equivalent to the following density after using Bayes' theorem and under the assumption of independent measurements

$$p(y_{n+1}|y_1, y_2, ..., y_n) = \int\int \prod_{i=1}^{n+1} f(y_i|\mu, \tau)\frac{p(\mu, \tau)}{p(y_1, y_2, ..., y_n)}d\mu\, d\tau. \tag{2.24}$$

2.3.1.2 Modelling the T/E Ratio

This section presents the Bayesian model applied on the raw T/E ratio, an approach proposed by Sottas et al. (2006). One of the main differences between this approach and the univariate model of section 2.3.1 is that now we model the raw instead of the log-transformed concentration values of the specific marker ratio T/E. A second difference is that in this section the prior setting is informative since there is adequate population information regarding the T/E ratio.

Suppose now that $\boldsymbol{y} = (y_1, y_2, ..., y_n)$ represents the observable vector with the n recorded values of the raw ratio Y = testosterone/epitestosterone, measured on the same individual. Let us assume that $y_1, y_2, ..., y_n$ are conditionally mutually independent and each y_i comes from a Gaussian distribution, $y_i|\mu, \sigma \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2), \ \ i = 1, ..., n$. Therefore, the likelihood function for all observations is given by

$$\mathcal{L}(\mu, \sigma; \boldsymbol{y}) = \prod_{i=1}^{n} f(y_i|\mu, \sigma) = \frac{1}{\sigma^n (2\pi)^{\frac{n}{2}}} \exp\Big\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \Big\}, \tag{2.25}$$

for $y, \mu \in \mathbb{R}$ and $\sigma > 0$. We also take into account prior beliefs on the inter-individual variability, which is expressed by the joint distribution $p(\mu, \sigma)$. We use the joint prior distribution of μ and σ in equation (2.27), after correcting the formula in Sottas et al. (2006)

$$p(\mu, \sigma) = p(\mu)\, p(\mathrm{C})\, \mu \ , \tag{2.26}$$

where C is the coefficient of variation ($C = \frac{\sigma}{\mu}$). We have also used the same values for hyperparameters, which have been provided in Sottas et al. (2006). It has been found that there is a strong correlation between μ and σ, but no correlation between μ and C. Consequently, we could use this knowledge accompanied by the transformation theorem in order to build the joint distribution of μ and σ as (see Appendix A; A.5)

$$p(\mu, \sigma) = p(\mu)\, p(\mathrm{C})\, \frac{1}{\mu} \ . \tag{2.27}$$

We then consider the following prior log-Gaussian distributions

$$p(\mathrm{C}) = \frac{1}{\sqrt{2\pi v^2}\mathrm{C}} e^{-\frac{(\log \mathrm{C} - m)^2}{2v^2}} \tag{2.28}$$

or

$$\mathrm{C} \sim \mathcal{LN}(m = -1.74, v^2 = 0.39^2), \tag{2.29}$$

and

$$\mu \sim \sum_{k=1}^{2} \pi_k \cdot \mathcal{LN}(m_k, v_k^2), \tag{2.30}$$

where $\boldsymbol{m} = (-1.96, 0.34)$ and $\boldsymbol{v^2} = (0.36^2, 0.59^2)$ are the mean and the scale parameters of the two-component log-Gaussian mixture. Therefore, we have the following prior distributions for the coefficient of variation expressed on the logarithmic scale for convenience, and for μ, respectively

$$\log p(\mathrm{C}) = -\frac{(\log \mathrm{C} - m)^2}{2v^2} - \log \mathrm{C} + c \tag{2.31}$$

$$p(\mu|\pi_1, \pi_2) = \pi_1 \cdot p_1(\mu) + \pi_2 \cdot p_2(\mu) = \pi_1 \cdot \frac{e^{-\frac{(\log \mu - m_1)^2}{2v_1^2}}}{v_1\sqrt{2\pi}\mu} + \pi_2 \cdot \frac{e^{-\frac{(\log \mu - m_2)^2}{2v_2^2}}}{v_2\sqrt{2\pi}\mu}, \tag{2.32}$$

where π_k is the probability that a unit belongs to subpopulation k of the biomarkers distribution, with $\sum_{k=1}^{2} \pi_k = 1$, and c includes all the constant terms. The mixing proportions are given in Sottas et al. (2006) which are $\pi_1 = 0.13$ and $\pi_2 = 0.87$. Consequently, we can hierarchically pass from the prior evidence to the revised knowledge, expressed in the log-posterior density, through the introduction of the data

$$\begin{aligned} \log p(\mu, \sigma|\boldsymbol{y}) \propto & -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2 - \frac{1}{2v^2}(\log \sigma - \log \mu - m)^2 - (n+1)\log \sigma - \log \mu \\ & + \log\left(\pi_1 \cdot \frac{e^{-\frac{(\log \mu - m_1)^2}{2v_1^2}}}{v_1} + \pi_2 \cdot \frac{e^{-\frac{(\log \mu - m_2)^2}{2v_2^2}}}{v_2}\right). \end{aligned} \tag{2.33}$$

An alternative and useful way to write out this model is using what is called a graphical representation. To write a graphical representation of our model as a Bayesian network, we reverse the order starting with the priors and finishing with the likelihood. In the graphical representation in Figure 2.2, we draw nodes for μ, C, σ and the data $\boldsymbol{y}$. A parameter in a circle means that it is a random variable which has its own distribution. Once we have the parameters, we can generate the data $\boldsymbol{y}$. The exchangeable data are represented as nodes, which now live in a double circle, indicating that these are also random variables but they are observed. Furthermore, the arrows indicate the dependence of the distribution of the data on μ and σ. This means that σ, which is obtained by the multiplication of the variables μ and C, influences the distribution of the data and subsequently μ influences the distributions of both; σ and data.

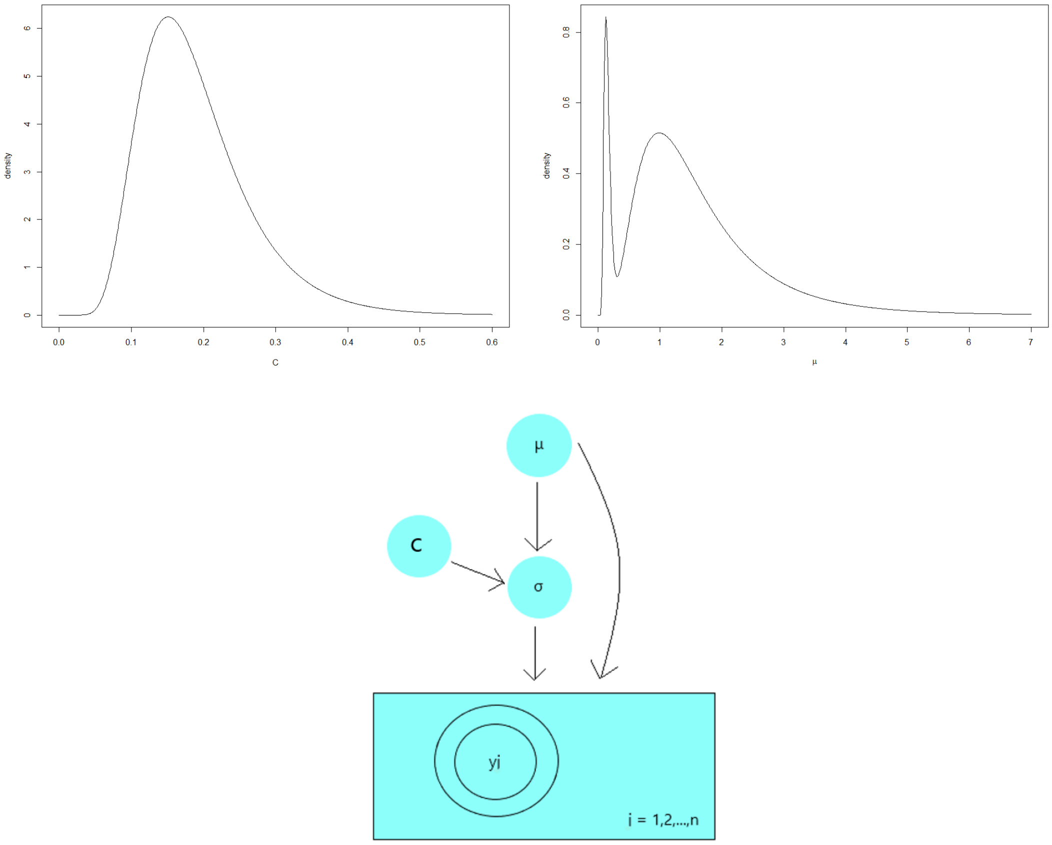


Figure 2.2: Graphical representation of the Bayesian hierarchical model accompanied by the prior distributions for the mean and the coefficient of variation of y_i's based on knowledge regarding the population.

Posterior approximation

As we can observe in equation (2.33), the joint posterior cannot be stated in a closed form. Therefore, we apply MCMC sampling to generate draws from the posterior distribution by constructing a reversible Markov chain that has the target posterior distribution as its equilibrium distribution. A Metropolis-Hastings algorithm (MH) can be used to simulate a Monte Carlo sample from the posterior distribution (Chib and Greenberg, 1995). Algorithm 1, presented in Section 2.3.5.1, samples candidates for all the parameters at once and accepts or rejects all of those candidates simultaneously. A second option is to use a hybrid algorithm that combines Metropolis-Hastings and Gibbs sampling, commonly called Metropolis-within-Gibbs (MWG). We choose to apply both algorithms; MH and the componentwise MWG algorithms for a large number of iterations, T. For the first sampling method, only the joint posterior distribution $p(\mu, \sigma|\boldsymbol{y})$ is needed. However, for the latter algorithm, we need to express the joint posterior as the product of the conditional posteriors as

$$p(\mu, \sigma|\boldsymbol{y}) = p(\mu|\boldsymbol{y})p(\sigma|\mu, \boldsymbol{y}) = p(\sigma|\boldsymbol{y})p(\mu|\sigma, \boldsymbol{y}). \tag{2.34}$$

To apply the Gibbs sampler for μ and σ, the full conditional posterior distributions for each parameter needs to be specified. In equations (2.35) and (2.36), both log-scaled conditional distributions have the same starting point; that is the full joint posterior distribution as

$$\begin{aligned}\log p(\mu|\sigma, \boldsymbol{y}) \propto \log p(\mu, \sigma|\boldsymbol{y}) \propto -\frac{1}{2\sigma^2}n\mu(\mu - 2\bar{y}) - \frac{1}{2v^2}(\log^2\mu - 2\log\sigma\log\mu + 2m\log\mu)\\ + \log\left(\pi_1 \cdot \frac{e^{-\frac{(\log\mu - m_1)^2}{2v_1^2}}}{v_1} + \pi_2 \cdot \frac{e^{-\frac{(\log\mu - m_2)^2}{2v_2^2}}}{v_2}\right) - \log\mu,\end{aligned} \tag{2.35}$$

and

$$\begin{aligned}\log p(\sigma|\mu, \boldsymbol{y}) \propto \log p(\mu, \sigma|\boldsymbol{y}) \propto -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2 - (n+1)\log\sigma\\ - \frac{1}{2v^2}(\log^2\sigma - 2\log\mu\log\sigma - 2m\log\sigma).\end{aligned} \tag{2.36}$$

The MWG sampler generates a Markov chain (μ_t, σ_t) according to the steps of Algorithm 5, presented in Section 2.3.5.3. Similar to the equation (2.24), if y_{n+1} is the next observation of that biomarker, the predictive density is

$$p(y_{n+1}|y_1, y_2, ..., y_n) = \int\int\prod_{i=1}^{n+1} f(y_i|\mu, \sigma)\frac{p(\mu, \sigma)}{p(y_1, y_2, ..., y_n)}d\mu\, d\sigma. \tag{2.37}$$

2.3.2 The Univariate Multilevel Model

This section introduces the Gaussian multilevel model which includes repeated log-transformed measurements from a single marker for each individual. This model separates inter-subject and intra-subject variation by dividing the error term ε_{ij} into two terms

$$y_{ij} = \mu + \varepsilon_{ij} = \mu + b_j + e_{ij}, \quad b_j \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_b^2), \quad e_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_e^2), \tag{2.38}$$

where y_{ij} denotes the logarithm of the ith observation of the jth individual, μ is the fixed effects term denoting the overall mean of the marker, the subject-level b_j is the random effect, e_{ij} is the random variable for other variation of subject j in its ith measurement, where $i = 1, 2, ..., n_j$ (n_j : total number of measuremnets of the jth individual), and $j = 1, 2, ..., J$. The assumptions here are that b_js are independent, identically and normally distributed between subjects, and e_{ij}s are assumed to be independent, identically and normally distributed between and within subjects. Shorthand notation for each of the parameters in this model are μ; $b = \{b_j\}_{j=1}^{J}$; the between-subjects variance σ_b^2; and the within-subjects variance σ_e^2. Suppose we have the following Bayesian hierarchical model for the variable Y

$$\begin{aligned} y_{ij}|\mu_j, \sigma_e^2 &\overset{\text{iid}}{\sim} \mathcal{N}(\mu_j = \mu + b_j,\, \sigma_e^2), \\ \mu &\overset{\text{iid}}{\sim} \mathcal{N}(\mu_0, \sigma_0^2) \\ b_j|\sigma_b^2 &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_b^2) \\ \sigma_e^2 &\overset{\text{iid}}{\sim} \mathcal{IG}a(\alpha_1, \beta_1) \\ \sigma_b^2 &\overset{\text{iid}}{\sim} \mathcal{IG}a(\alpha_2, \beta_2), \end{aligned} \tag{2.39}$$

where μ_0 and σ_0^2 are historical prior information, and α_1, α_2, β_1 and β_2 could be selected such that the priors for inter- and intra-variations will be non-informative. For more details on mixed effects models and Bayesian hierarchical structures on these models see Pinheiro and Bates (2000) and Gelman et al. (2013).

Posterior approximation

Here, the exact joint posterior distribution cannot be calculated, but we can approximate the posterior parameters given the raw data by using Gibbs sampling. To implement the Gibbs sampler (see Algorithm 3 in Section 2.3.5.2), we first need to find the full conditional posterior distribution for all unknown parameters, $\theta = (\mu, \{\mu_j\}_{j=1}^J, \sigma_e^2, \sigma_b^2)$ (see derivations in Appendix A; A.6), which are

$$\mu|\{\mu_j\}_{j=1}^J, \sigma_b^2, \sigma_e^2, \boldsymbol{y} \sim \mathcal{N}\left(\frac{\sum_{j=1}^J \mu_j\, \sigma_0^2 + \mu_0\sigma_b^2}{J\sigma_0^2 + \sigma_b^2}, \frac{\sigma_b^2\sigma_0^2}{J\sigma_0^2 + \sigma_b^2}\right)$$

$$\mu_j|\mu, \mu_{-j}, \sigma_b^2, \sigma_e^2, \boldsymbol{y} \sim \mathcal{N}\left(\frac{\sum_{i=1}^{n_j} y_{ij}\sigma_b^2 + \mu\sigma_e^2}{n_j\sigma_b^2 + \sigma_e^2}, \frac{\sigma_b^2\sigma_e^2}{n_j\sigma_b^2 + \sigma_e^2}\right)$$

$$\sigma_e^2|\mu, \{\mu_j\}_{j=1}^J, \sigma_b^2, \boldsymbol{y} \sim \mathcal{IG}a\left(\text{shape} = \frac{n}{2} + \alpha_1,\ \text{scale} = \frac{\sum_{j=1}^J \sum_{i=1}^{n_j}(y_{ij} - \mu_j)^2 + 2\beta_1}{2}\right)$$

$$\sigma_b^2|\mu, \{\mu_j\}_{j=1}^J, \sigma_e^2, \boldsymbol{y} \sim \mathcal{IG}a\left(\text{shape} = J/2 + \alpha_2,\ \text{scale} = \frac{\sum_{i=j}^J(\mu_j - \mu)^2}{2} + \beta_2\right),$$

where J is the total number of subjects, and the subscript $-j$ denotes all indices except j.

Note that this model constitutes an intermediate stage between the univariate and the multivariate multilevel models. We only intend to present the theoretical background without using any applications on it, since it is not the focus of this thesis.

2.3.3 The Multivariate Bayesian Multilevel Model

In Sottas et al. (2006) approach, the T/E marker is modelled by a univariate Gaussian distribution within a Bayesian context. In this section, we present a multivariate Gaussian multilevel model (MGMM) as a generalisation of the univariate model, which can deal with a wide variety of markers as response variables in their logarithmic scale. The MGMM model is expressed as

$$y_{ijk} = \mu_k + b_{jk} + e_{ijk}, \quad \boldsymbol{b}_j \overset{\text{iid}}{\sim} \mathcal{N}_K(\mathbf{0}, \Omega_b^{-1}), \quad \boldsymbol{e}_{ij} \overset{\text{iid}}{\sim} \mathcal{N}_K(\mathbf{0}, \Omega_e^{-1}), \tag{2.40}$$

where y_{ijk} denotes the logarithm of the ith observation of kth marker for the jth athlete, μ_k is the fixed effect for the overall mean of all observations of kth response marker, b_{jk} is the random effect of athlete j for the kth marker, and e_{ijk} is the random term for other variation in its ith measurement, while $i = 1, 2, ..., n_j$, $j = 1, 2, ..., J$ and $k = 1, 2, ..., K$. The assumptions here are that the random effects b_{jk}s are independent, identically and normally distributed between subjects of the same variable k, but there is a correlation between the K markers in the same athlete. The error terms e_{ijk}s are also independent, identically and normally distributed between and within subjects. Shorthand notation for each of the parameters in the model are $\boldsymbol{\mu} = \{\mu_k\}_{k=1}^K$, $\boldsymbol{b} = \{b_{jk}\}_{j=1}^J{}_{k=1}^K$, and Ω_b and Ω_e are the precision matrices for $\boldsymbol{b}_j$ and $\boldsymbol{e}_{ij}$, respectively. Ω_μ is the unknown precision matrix of the overall mean $\boldsymbol{\mu}$. Suppose we have the following Bayesian hierarchical multiple response model

$$\begin{aligned} \boldsymbol{y}_{ij}|\boldsymbol{\mu_j} &\overset{\text{iid}}{\sim} \mathcal{N}_K(\boldsymbol{\mu_j} = \boldsymbol{\mu} + \boldsymbol{b_j},\, \Omega_e^{-1}) \\ \boldsymbol{\mu}\,|\boldsymbol{\mu_0} &\overset{\text{iid}}{\sim} \mathcal{N}_K(\boldsymbol{\mu_0},\, \Omega_\mu^{-1}) \\ \boldsymbol{b_j} &\overset{\text{iid}}{\sim} \mathcal{N}_K(\mathbf{0},\, \Omega_b^{-1}) \\ \Omega_e &\overset{\text{iid}}{\sim} \mathcal{W}i(d_e, S_e) \\ \Omega_\mu &\overset{\text{iid}}{\sim} \mathcal{W}i(d_\mu, S_\mu) \\ \Omega_b &\overset{\text{iid}}{\sim} \mathcal{W}i(d_b, S_b), \end{aligned} \tag{2.41}$$

where the overall mean $\boldsymbol{\mu}$ and the random effects $\boldsymbol{\mu_j}$ have conjugate multivariate Gaussian priors, while the precision matrices Ω_e, Ω_μ and Ω_b have conjugate Wishart hyperpriors placed on each of them. Historical prior information about all response variables is captured by the prior mean vector $\boldsymbol{\mu_0}$. Moreover, d_e, d_μ and d_b denote the degrees of freedom, and S_e, S_μ and S_b are prior covariance matrices which are selected such that the prior distribution for them will be non-informative. The degrees of freedom of the Wishart distribution need to be greater than the data dimension minus one, i.e. $d_e > K - 1$, where the first moment of the Wishart distribution do exist. Furthermore, the distribution will be non-informative when the hyperparameter for the degrees of freedom is equal to the dimension K (number of parameters). For details related to the Wishart distribution see DeGroot (2005). The prior scale matrices S_e, S_μ and S_b are all set equal to $(1/1000)\mathbb{I}_{\mathbb{K}}$ so that, as we similarly did for $\boldsymbol{\mu_j}$s, posterior inferences would be largely driven by the data. The graphical representation of the MGMM model is depicted in Figure 2.3.

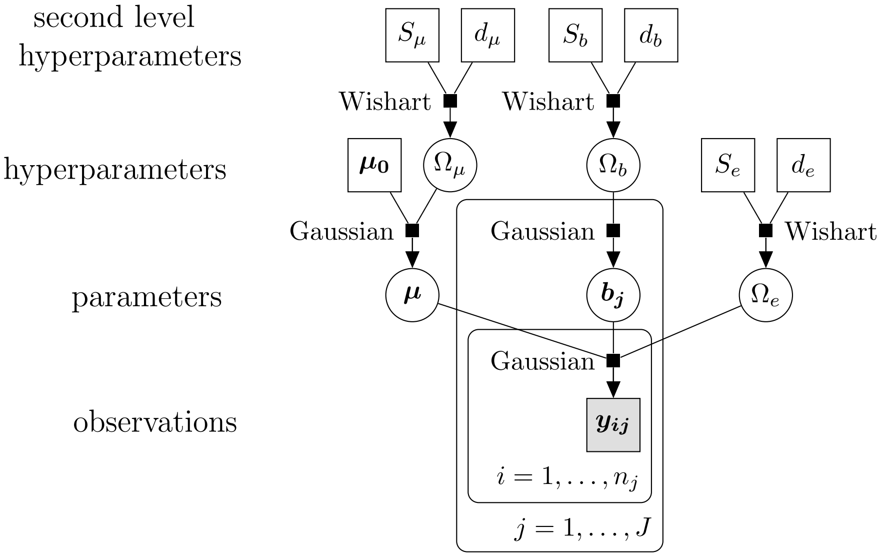


FIGURE 2.3: A graphical representation of the multivariate Gaussian multilevel model (MGMM) with conjugate priors. The overall mean, μ, and the precision matrix, Ω_e are assumed to be independent.

Posterior approximation

We approximate the posterior parameters by using Gibbs sampling. As previously, to apply the Gibbs sampler (see Algorithm 4 in Section 2.3.5.2), the full conditional posterior distributions for all unknown parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}, \{\boldsymbol{\mu_j}\}_{j=1}^{J}, \Omega_e, \Omega_\mu, \Omega_b)$ are calculated as follows (see details in A; A.7)

$$\begin{aligned}
\boldsymbol{\mu}\ |\{\boldsymbol{\mu_j}\}_{j=1}^{J}, \Omega_e, \Omega_\mu, \Omega_b, \boldsymbol{y} &\sim \mathcal{N}_K(A_n^{-1} b_n,\ A_n^{-1}) \\
\boldsymbol{\mu_j}|\boldsymbol{\mu}, \boldsymbol{\mu_{-j}}, \Omega_e, \Omega_\mu, \Omega_b, \boldsymbol{y} &\sim \mathcal{N}_K(A_n^{-1'} b_n',\ A_n^{-1'})\ \forall\ j = 1, ..., J \\
\Omega_e|\{\boldsymbol{\mu_j}\}_{j=1}^{J}, \boldsymbol{\mu}, \Omega_\mu, \Omega_b, \boldsymbol{y} &\sim \mathcal{W}i(d_e', S_e') \\
\Omega_\mu|\{\boldsymbol{\mu_j}\}_{j=1}^{J}, \boldsymbol{\mu}, \Omega_e, \Omega_b, \boldsymbol{y} &\sim \mathcal{W}i(d_\mu', S_\mu'), \\
\Omega_b|\{\boldsymbol{\mu_j}\}_{j=1}^{J}, \boldsymbol{\mu}, \Omega_e, \Omega_\mu, \boldsymbol{y} &\sim \mathcal{W}i(d_b', S_b'),
\end{aligned}$$

where

$$\begin{aligned}
A_n &= J\Omega_b + \Omega_\mu \\
b_n &= \Omega_\mu \boldsymbol{\mu_0} + J\Omega_b \bar{\boldsymbol{\mu}} \\
\bar{\boldsymbol{\mu}} &= \frac{1}{J}\sum_{j=1}^{J} \boldsymbol{\mu_j}
\end{aligned}$$

$$\begin{aligned}
A_n' &= \Omega_b + n_j \Omega_e \\
b_n' &= \Omega_b \boldsymbol{\mu} + \Omega_e \sum_{i=1}^{n_j} \boldsymbol{y_{ij}} \\
d_e' &= d_e + n \\
n &= \sum_{j=1}^{J} n_j
\end{aligned}$$

$$\begin{aligned}
S_e' &= S_e^{-1} + \sum_{j=1}^{J}\sum_{i=1}^{n_j} (\boldsymbol{y_{ij}} - \boldsymbol{\mu_j})(\boldsymbol{y_{ij}} - \boldsymbol{\mu_j})^T \\
d_b' &= d_b + J \\
S_b' &= S_b^{-1} + \sum_{j=1}^{J} (\boldsymbol{\mu_j} - \boldsymbol{\mu})(\boldsymbol{\mu_j} - \boldsymbol{\mu})^T. \\
d_\mu' &= d_\mu + 1 \\
S_\mu' &= S_\mu^{-1} + (\boldsymbol{\mu} - \boldsymbol{\mu_0})(\boldsymbol{\mu} - \boldsymbol{\mu_0})^T.
\end{aligned}$$

2.3.4 The Generalised Linear Mixed Model

The generalised linear mixed model (GLMM) is an extension of the generalised linear model (GLM) to allow response variables from different distributions, such as binary responses (Breslow and Clayton, 1993; Dobson and Barnett, 2018). For comparison reasons we aim to implement GLMMs as a standard method for modelling longitudinal data. GLMMs include both, fixed and random effects, and its general form is denoted as

$$\boldsymbol{Y_j} = \boldsymbol{X_j}\boldsymbol{\beta} + \boldsymbol{Z_j}\boldsymbol{b_j} + \boldsymbol{\epsilon_j},$$

where $\boldsymbol{Y_j}$ in this section is a vector of binary responses for athlete j, which indicates the class of the n_j measurements of the jth athlete ($Y_{ij} = 0$ if the ith measurement of athlete j is normal, or $Y_{ij} = 1$ if it is abnormal). $\boldsymbol{X_j}$ is a design matrix for the fixed effects of athlete j (including the biomarkers and/or ratios), $\boldsymbol{\beta}$ is a vector of fixed effects, $\boldsymbol{Z_j}$ is the design matrix for the random effects of cluster j, $\boldsymbol{b_j}$ is a vector of random effects for athlete j, where $E(\boldsymbol{b_j} = 0)$ and $\text{Cov}(\boldsymbol{b_j}) = G$, and $\boldsymbol{\epsilon_j}$ is a vector of residuals of the observations in the steroid profile of athlete j, where $E(\boldsymbol{\epsilon_j}) = 0$ and $\text{Cov}(\boldsymbol{\epsilon_j}) = R_j = (\sigma^2 \boldsymbol{I_j})$. The parameter vector $\boldsymbol{\beta}$ represents the marginal estimates across all athletes and values do not change among athletes. For this case of binary outcomes, the logit link function is used such that the conditional probability of the outcome being a success is

$$P(\boldsymbol{Y_j} = 1 | \boldsymbol{X_j}, \boldsymbol{Z_j}, \boldsymbol{b_j}) \;=\; \frac{\exp\left(\boldsymbol{X_j}\boldsymbol{\beta} + \boldsymbol{Z_j}\boldsymbol{b_j}\right)}{1 + \exp\left(\boldsymbol{X_j}\boldsymbol{\beta} + \boldsymbol{Z_j}\boldsymbol{b_j}\right)}.$$

2.3.5 Markov Chain Monte Carlo Sampling Methods

In multi-dimensional Bayesian problems, Markov chain Monte Carlo (MCMC) are very powerful methods, well suited to computing via simulation the posterior distributions, which are extremely difficult or impossible to evaluate. This is achieved by constructing a Markov chain, which guarantees the convergence of the chain to the stationary distribution after a burn-in period, and draws samples which are progressively more likely realisations of the distribution of interest; that is the target distribution. There are different ways to implement MCMC. The two most commonly used Markov chain simulation techniques are the Gibbs sampler and the Metropolis-Hastings algorithm. However, it is possible to have a hybrid sampling technique, such as the Metropolis-within-Gibbs (MWG) that uses combinatorial steps of Metropolis-Hastings and Gibbs. We implement one of the appropriate MCMC sampling methods, depending on the model.

2.3.5.1 Metropolis-Hastings Sampling

The Metropolis-Hastings (MH) algorithm is an MCMC algorithm that simulates samples from a generic probability distribution π, which is called "target" distribution, by making use of the full joint density and independent proposal distributions (q) for each of the variables of interest. Algorithm 1 represents the algorithmic rendering of MH in the context of T/E modelling in Section 2.3.1.

Algorithm 1 Metropolis-Hastings algorithm

```
Precondition: Generate an initial state $\theta^{(0)} = (\mu^{(0)}, \sigma^{(0)})$ from $q(\theta)$
for $t \leftarrow 1$ to $T$ do
Propose a new state: $\theta^*$ from $q(\theta^{(t)}|\theta^{(t-1)})$
Calculate the acceptance probability: $\alpha(\theta^*|\theta^{(t-1)}) = \min\left\{1, \frac{q(\theta^{(t-1)}|\theta^*)\pi(\theta^*)}{q(\theta^*|\theta^{(t-1)})\pi(\theta^{(t-1)})}\right\}$
Take $u \sim U(0,1)$                                ▷ simulate a Uniform random variable
    if $u < \alpha$ then
        $\theta^{(t)} \leftarrow \theta^*$                            ▷ accept the proposal
    else
        $\theta^{(t)} \leftarrow \theta^{(t)}_{i-1}$                  ▷ reject the proposal
    end if
end for
```

2.3.5.2 Gibbs Sampling

The Gibbs sampler constitutes a special case of Metropolis-Hastings algorithm. The Gibbs sampling is a Monte Carlo simulation method which obtains samples from the joint posterior distribution, $f(\boldsymbol{\theta}|\boldsymbol{y})$, by successively and repeatedly simulating from the conditional distributions of each component given the other components. Each simulation step is usually straightforward under conditional conjugacy. Algorithm 2 describes the steps of the Gibbs sampler under a specific case of the univariate Gaussian model in Section 2.3.1. Algorithms 3 and 4 refer to the Gibbs sampling for the univariate and multivariate multilevel models from Sections 2.3.2 and 2.3.3, respectively. For a more detailed explanation of the Gibbs sampler and the Metropolis-Hastings algorithm see the papers of Casella and George (1992) and Chib and Greenberg (1995), respectively.

Algorithm 2 Gibbs sampler

```
Precondition: Generate an initial state $\theta^{(0)} = (\mu^{(0)}, \tau^{(0)})$
for $t \leftarrow 1$ to $T$ do
draw $\mu^{(t)} \sim p(\mu \mid \tau^{(t-1)}, \boldsymbol{y})$
draw $\tau^{(t)} \sim p(\tau \mid \mu^{(t)}, \boldsymbol{y})$
end for
```

Algorithm 3 Gibbs sampler

```
Precondition: Generate an initial state $\theta^{(0)} = (\mu^{(0)}, \{\mu_j^{(0)}\}_{j=1}^J, \sigma_e^{2(0)}, \sigma_b^{2\,(0)})$
for $t \leftarrow 1$ to $T$ do
draw $\mu^{(t)} \sim p(\mu \mid \{\mu_j^{(t-1)}\}_{j=1}^J, \sigma_e^{2\,(t-1)}, \sigma_b^{2\,(t-1)}, \boldsymbol{y})$
    for $j \leftarrow 1$ to $J$ do
    draw $\mu_j^{(t)} \sim p(\mu_j \mid \mu^{(t)},\ \mu_{-j}^{(t-1)}, \sigma_e^{2\,(t-1)}, \sigma_b^{2\,(t-1)}, \boldsymbol{y})$
    end for
draw $\sigma_e^{2\,(t)} \sim p(\sigma_e^2 \mid \mu^{(t)}, \{\mu_j^{(t)}\}_{j=1}^J, \sigma_b^{2\,(t-1)}, \boldsymbol{y})$
draw $\sigma_b^{2(t)} \sim p(\sigma_b^2 \mid \mu^{(t)}, \{\mu_j^{(t)}\}_{j=1}^J, \sigma_e^{2\,(t)}, \boldsymbol{y})$
end for
```

Algorithm 4 Gibbs algorithm

```
Precondition: Generate an initial state $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\mu}^{(0)}, \{\boldsymbol{\mu_j}^{(0)}\}_{j=1}^{J}, \Omega_e^{(0)}, \Omega_\mu^{(0)}, \Omega_b^{(0)})$
for $t \leftarrow 1$ to $T$ do
draw $\boldsymbol{\mu}^{(t)} \sim p(\boldsymbol{\mu}\,|\{\boldsymbol{\mu}^{(t-1)_j}\}_{J=1}^{J}, \Omega_e^{(t-1)}, \Omega_\mu^{(t-1)}, \Omega_b^{(t-1)}, \boldsymbol{y})$
    for $j \leftarrow 1$ to $J$ do
    draw $\boldsymbol{\mu_j}^{(t)} \sim p(\boldsymbol{\mu}_j\,|\boldsymbol{\mu}^{(t)}, \boldsymbol{\mu}_{-j}^{(t-1)}, \Omega_e^{(t-1)}, \Omega_\mu^{(t-1)}, \Omega_b^{(t-1)}, \boldsymbol{y})$
    end for
draw $\Omega_e^{(t)} \sim p(\Omega_e\,|\,\boldsymbol{\mu}^{(t)}, \{\boldsymbol{\mu_j}^{(t)}\}_{j=1}^{J}, \Omega_\mu^{(t-1)}, \Omega_b^{(t-1)}, \boldsymbol{y})$
draw $\Omega_\mu^{(t)} \sim p(\Omega_\mu\,|\,\boldsymbol{\mu}^{(t)}, \{\boldsymbol{\mu_j}^{(t)}\}_{j=1}^{J}, \Omega_e^{(t)}, \Omega_b^{(t-1)}, \boldsymbol{y})$
draw $\Omega_b^{(t)} \sim p(\Omega_b\,|\,\boldsymbol{\mu}^{(t)}, \{\boldsymbol{\mu_j}^{(t)}\}_{j=1}^{J}, \Omega_e^{(t)}, \Omega_\mu^{(t)}, \boldsymbol{y})$
end for
```

2.3.5.3 Metropolis-within-Gibbs Sampling

The Metropolis-within-Gibbs (MWG) algorithm is another MCMC algorithm which simulates and updates one parameter at a time from its conditional posterior distribution, as the Gibbs sampler implements. The algorithm under a specific case of the univariate modelling of Section 2.3.1 works as follows in Algorithm 5.

Algorithm 5 Metropolis-within-Gibbs algorithm

```
Precondition: Generate an initial state $\theta^{(0)} = (\mu^{(0)}, \sigma^{(0)})$ (presumably by sampling
    from the prior distributions on the variables).
for $t \leftarrow 1$ to $T$ do
    for $i \leftarrow 1$ to $d$ do                                  ▷ d: # of parameters (d=2)
Propose a new state: $\theta_i^*$ from $q_i(\theta_i^{(t)})$, where $q_i$ is a predefined proposal function
    for parameter $i$.
Calculate the acceptance probability: $C = \min\left\{1, \frac{q_i(\theta^{(t)} \mid \theta_i^{(*)}, \theta_{i-1}^{(t)})\, \pi(\theta_i^{(*)}, \theta_{i-1}^{(t)})}{q_i(\theta_i^{(t)}, \theta_{i-1}^{(t)} \mid \theta^{(t)})\, \pi(\theta_i^{t})}\right\}$
Take  $u \sim U(0,1)$                              ▷ simulate a Uniform random variable
        if $u < C$ then
            $\theta^{(t)} \leftarrow \theta^*$                              ▷ accept the proposal
        else
            $\theta^{(t)} \leftarrow \theta^{(t-1)}$                        ▷ reject the proposal
        end if
    end for
end for
```

2.3.6 MCMC Convergence Diagnostics

There are a variety of MCMC convergence diagnostics to test whether the Markov chains resulting from the MCMC sampling algorithms have converged in distribution to the posterior distribution of interest. The convergence diagnostics of Raftery and Lewis (1991), of Gelman and Rubin (1992), and of Geweke (1992) are currently the most popular diagnostic tests, which have also been used throughout this work and reviewed in this section.

The Gelman-Rubin diagnostic

The Gelman and Rubin (1992) diagnostic evaluates MCMC convergence of a scalar parameter of interest by analysing the difference between multiple independent Markov chains. The convergence is assessed by comparing the estimated within and between variances of the posterior samples for each model parameter using the potential scale reduction factor (PSRF). Suppose we run M independent chains of same length, T, and $\{\theta^{(mt)}\}_{t=1}^{T}$ is the posterior parameter drawn from the tth iteration of the mth chain for the model parameter θ, where $t = 1, ..., T$ and $m = 1, ..., M$. The within and between chain variances are given by

$$W = \frac{1}{M(T-1)} \sum_{m=1}^{M} \sum_{t=1}^{T} (\theta^{(mt)} - \bar{\theta}_{t.})^2 \tag{2.42}$$

and

$$B = \frac{T}{M-1} \sum_{m=1}^{M} (\bar{\theta}_{t.} - \bar{\theta}_{..})^2. \tag{2.43}$$

Under certain stationary conditions, the PSRF is defined to be the ratio of the pooled variance, $\hat{V}$ and W as

$$\hat{V} = \frac{T-1}{T} W + \frac{M+1}{MT} B \tag{2.44}$$

and

$$\hat{R}_{PSRF} = \frac{T-1}{T} + \frac{M+1}{M} \frac{B}{W} \frac{1}{T}. \tag{2.45}$$

If all chains have converged to the stationary distribution, the variability between the chains should be relatively small and $W \approx B$, then the statistic $\hat{R}_{PSRF} \to 1$ when $T \to \infty$.

The Raftery-Lewis diagnostic

The second diagnostic test that we used is the test proposed by Raftery and Lewis in 1992. The test can be applied to parameter samples coming from a single Markov chain, and it focuses on achieving a predefined degree of accuracy of specific quantiles rather than the convergence of the mean. The test reports the T, T_{min}, T_{burn}, and I, where

- T is the total number of iterations that the chain must run.
- T_{min} is the minimum number of iterations required to estimate the quantile of interest with the predefined accuracy assuming independent samples (i.e., with zero autocorrelation).

- T_{burn} is the suggested number of burn-in iterations.
- I is the dependence factor given by $I = T/T_{min}$, which indicates the relative increase of the total sample due to autocorrelations. If $I = 1$, then the generated values are independent. High dependence factors (> 5) are worrisome and may be due to influential starting values, high correlations between coefficients, or poor mixing. In general, I can be considered roughly as an estimate of the required thinning interval.

The Geweke diagnostic

The Geweke diagnostic test (Geweke, 1992) assesses the convergence of the mean of the sampled values for each parameter of interest obtained from a single chain. This diagnostic performs hypothesis testing to check whether the means estimated from two distinct sub-chains of the total MCMC draws are equal by applying a simple Z-test. Usually the comparison is being between samples from the last half (50%) of the chain against some smaller interval in the beginning of the chain, e.g. the first 10% of the draws. A p-value < 0.05 indicates that there is enough evidence to reject the null hypothesis of equal means in favour of the alternative hypothesis that suggests different means implying non-convergence.

2.3.7 The Effective Sample Size

The effective sample size (ESS) (Kass et al., 1998) is a useful measure that estimates the number of independent samples obtained from the MCMC chain. The ESS can quantify how many samples should be taken in a chain to reach a given quality of posterior estimates. The effective sample size of a sequence is defined in terms of the autocorrelations with lag $k \geq 0$, $\rho(k)$, within the sequence at different lags by

$$ESS = \frac{T}{1 + 2\sum_{k=1}^{\infty} \rho(k)}, \tag{2.46}$$

where $\rho(k) = Cov(X_t, X_{t+k})/Var(X_t)$ and $X_1, ..., X_n$ are the MCMC sample draws. Negative autocorrelations may occur due to the noise in the correlation estimates,

$\hat{\rho}(k)$, as k increases. The ESS estimates for each parameter have been calculated based on the paired autocorrelation of Geyer (1992) and Geyer (2011), which is guaranteed to be positive, monotone sequence estimator. Low ESS values indicate small number of independent MCMC draws, implying a poor mixing or lack of convergence of the sampler.

2.3.8 One-Class Classification

One-class classification (OCC) algorithms are used in classification modelling when only one class (known as "target" class) is fully known and the others are either absent or poorly sampled (Minter, 1975; Bishop, 1994; Khan and Madden, 2014). Doping detection constitutes a hot topic in forensic toxicology, which can be framed as a one-class classification problem since measurements from doped athletes can be difficult to obtain, either due to the elaborate techniques that athletes use to avoid testing, or due to the undetectable use of banned substances.

One-Class Classifier

For doping analysis, full information about non-doped athletes who have been voluntarily tested is provided, but limited knowledge is available for athletes who have received doping regimens. Thus in this case, the samples from athletes with normal concentration values are treated as the "target" class. The focus is on answering whether there is evidence that new samples from athletes, whose doping status is unknown, are compatible with the known normal class of samples, or they show abnormal behaviour and should be considered as outliers. A classifier, that is a function which assigns each input data point to a class, accounting for other confounding factors such as gender, cannot be constructed with known standard rules in the case of imbalanced classes. In pattern recognition or machine learning, the main purpose is to infer a classifier from a limited set of training data. Note that, when having *longitudinal data*, the classifier has to deal with the complexities of unbalanced data, their updating nature as well as potential confounders. We approach the one-class classification problem by using a density estimation method, which is described in the following section, for OCC models of any dimensionality.

Note that we control for confounding due to gender, by specifying different prior distributions for male and female athletes based on their sub-population information.

Highest posterior predictive density region as a classification rule

The Bayesian model specification allows us to hierarchically pass from the prior evidence to the revised knowledge, expressed in the posterior density $p(\theta|\boldsymbol{y})$, as the data arrive. Using a variety of sampling techniques for the Bayesian Gaussian models described in previous sections, we can estimate the posterior density function and then we can approximate the predictive density function of a new observable y_{n+1} given the data $\boldsymbol{y} = (y_1, y_2, ..., y_n)$. In a general case the predictive density function is calculated as

$$p(y_{n+1}|\boldsymbol{y}) = \int_{\Theta} f(y_{n+1}|\theta)p(\theta|\boldsymbol{y})d\theta, \tag{2.47}$$

which is formed by weighting the possible values of θ in the future observation $f(y_{n+1}|\theta)$ by how likely we believe they are to occur $p(\theta|\boldsymbol{y})$. We can use the predictive distribution to provide a useful range of plausible concentration values for markers and ratios of a future athlete. Here, $\boldsymbol{y}$ is the training set and consists of the samples from the "target" class; that is the normal concentration values from non-EAAS users. To overcome the curse of dimensionality, we need to ensure a large number of observations in the training set. The main task of the OCC algorithm is to define a classification boundary, such that it accepts as many samples as possible from the normal class, while it minimises the chance of accepting the outlier samples. Hence, the classification is performed by setting a threshold value, γ, on the approximated densities, in such a way that a target (normal) and a non-target (outlier/abnormal) region can be obtained ensuring a low predefined Type I error, i.e. the false positive rate α. Therefore, the $(1-\alpha)\%$ prediction interval for y_{n+1} is the region of the form

$$C_\alpha = \{y_{n+1} : p(y_{n+1}|\boldsymbol{y}) \geq \gamma\}, \tag{2.48}$$

where γ is chosen to ensure that $P(Y_{n+1} \in C_\alpha|\boldsymbol{y}) = 1-\alpha$, as shown in Figure 2.4. A new test result, y_{n+1}, is considered to be an outlier if it is not included in the $(1-\alpha)\%$ highest posterior density (HPD) interval of the conditional probability distribution $p(y_{n+1}|y_1, y_2, ..., y_n)$, and normal otherwise. An indicator variable $\mathcal{X}$ is used to classify each data point i as

$$\mathcal{X}_i = \begin{cases} 1 & , \text{if } y_i \in C_\alpha \\ 0 & , \text{otherwise.} \end{cases} \tag{2.49}$$

Based on the threshold, γ, the lower and upper limits of the HPD interval are obtained, which are used to define the normal boundaries of the EAAS concentration values or ratios at an individual level. These boundaries are used in detecting any steroids misuse that may cause abnormal high or low concentration values of biomarkers or ratios as well as in revealing urine samples replacement or the impact of other confounding factors. For detecting an abnormal sample at time t with the multivariate approach, we examine whether at least two sample values exceed their corresponding HPD intervals at time t.

It is worth mentioning that there is the usual trade-off in choosing an appropriate α, since low values of α will give large intervals. High α values give narrower intervals implying that a new measurement y_{n+1} has a low probability of lying in it. Furthermore, note that testing the first measurement of an athlete is based on the population thresholds only, since $n = 0$. Population thresholds are presented in Table C.5 obtained by Van Renterghem et al. (2010). Population threshold information has been considered also in work of Rauth (1994) and Kicman et al. (1995).

Continuity assumption

In this semi-supervised learning process, we assume that the continuity assumption holds. This is a general assumption in pattern recognition which supports that points which are close to each other are more likely to share a label. For this purpose, when the model suggests an outlier, then this observation is automatically excluded from the set of recordings that are used to compute the HPD intervals. If we do not discard the proposed outliers, we expect to learn the noise. Any noise measurements which are considered as normal measurements have a significant impact on the personalised accepted limits. Hence, we cannot expect to infer a good classification in such a case.

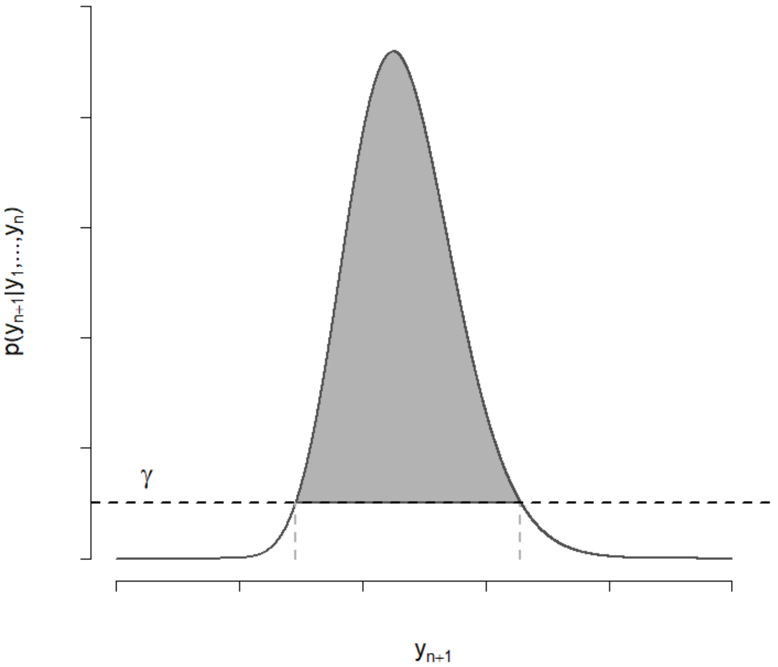


Figure 2.4: Highest posterior predictive density region of a unimodal posterior distribution $p(y_{n+1}|y_1, y_2, ..., y_n)$. The grey shaded area denotes the region C_α where normal measurements from athletes' steroid profile are expected to lie with probability $1 - \alpha$. Observations which lie outside the $(1 - \alpha)\%$ HPD interval are treated as outliers.

Random Oversampling

To treat imbalanced datasets for binary classification problems, we can use several sampling methods such as undersampling, oversampling, synthetic data generation and cost sensitive learning. These methods are used to maintain a balance between the different classes. In our case, we focus on the oversampling, which is a technique that generates synthetic balanced samples. This method works with the minority class, where it replicates the observations from this class to balance the data. It is also known as upsampling, which can also be divided into two types: Random Oversampling (ROSE, Random Over-Sampling Examples) and Informative Oversampling. Random oversampling can balance the data by randomly oversampling the minority class based on a bootstrap sampling. The random sampling method generates new samples using the conditional density estimate of the two classes, while Informative oversampling uses a pre-specified criterion and synthetically generates minority class observations (Kotsiantis et al., 2006; Lunardon et al., 2014).

Chapter 3

Multilevel adaptive models and anomaly detection: applications on doping control analysis

3.1 Introduction

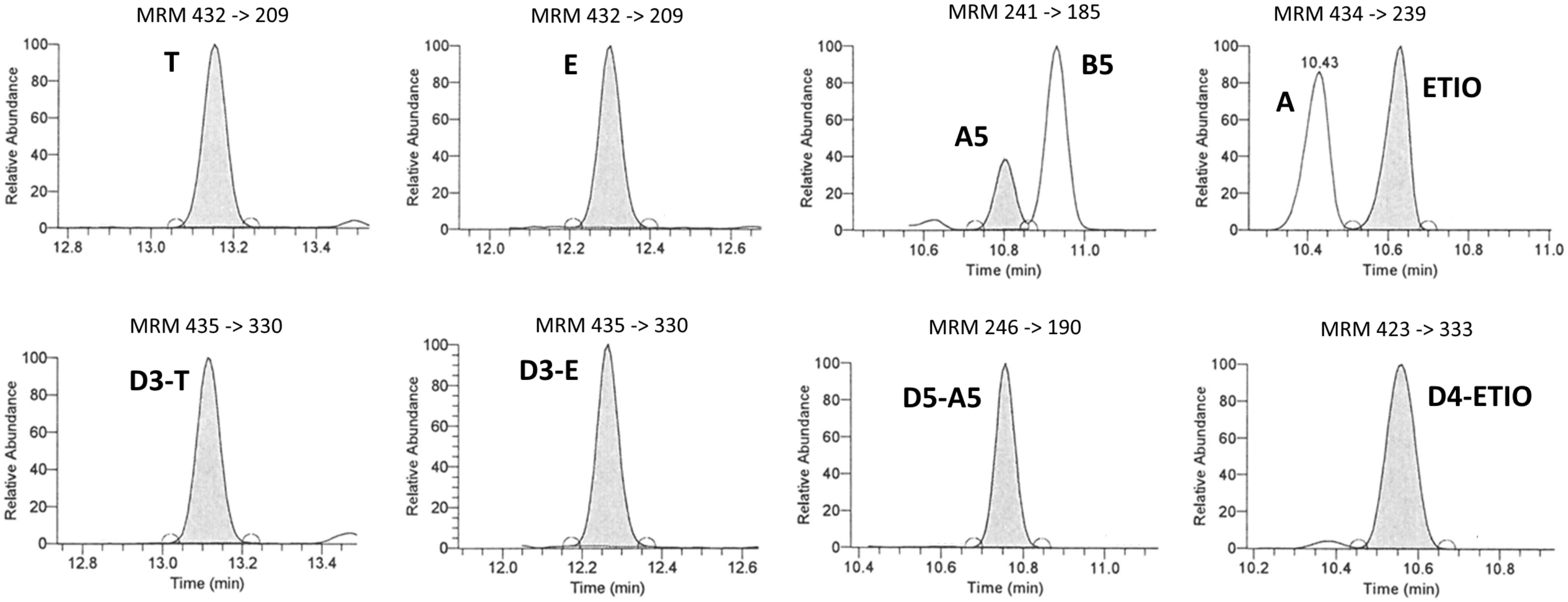
This chapter contains the application of the developed multilevel adaptive models to analyse the urine steroid profile of athletes. The steroid profile of an athlete includes multiple biomarkers, and/or ratios of them. Each biomarker has been repeatedly measured over time. We first apply the univariate model to analyse a single biomarker at a time, as described in Section 2.3.1, and then the multivariate model suitable for a set of selected biomarkers, as presented in Section 2.3.3. Thus, the models provide updates for identifying any outlying observations, which are either abnormal and merit further investigation, or may be suspicious in the likely scenario of sample exchange between athletes or due to other confounding factors. Through the proposed method, we establish personalised thresholds within a multivariate context, which can distinguish "normal" from "abnormal/anomalous" samples of "non-doped" and "doped" athletes. The implementation of the analysis has been conducted using the `R` statistical software (`R` 4.1.1). A user-friendly software was developed to implement the methodology, so that it can be used by practitioners as widely as possible. The latest version of the application can be found at `https://dimitraelegla.shinyapps.io/doping_shiny_app/`.

3.2 Data Summary

The application of the methodology was based on athletes' longitudinal steroid profile data extracted from their ABP. The datasets have been collected by the Institute of Biochemistry of the German Sport University Cologne, an accredited laboratory of Anti-Doping Administration and Management System (ADAMS) of WADA, following all the appropriate ethical approval procedures. Individual steroid profiles were analysed according to established methods including gas chromatography-mass spectrometry. Figure 3.1 represents a real GC-MS multiple reaction monitoring chromatogram produced by an unsuspicious urine sample.

3.2.1 Longitudinal Data

The longitudinal dataset includes six endogenous androgenic steroid concentrations and five concentration ratios proposed by WADA (testosterone (T), epitestosterone (E), androsterone (A), etiocholanolone (Etio), 5α-androstane-3β, 17β-diol (5αAdiol or A5), 5β-androstane-3α, 17β-diol (5βAdiol or B5), T/E, A/T, A/Etio, A5/B5 and A5/E), which were repeatedly collected from each athlete in or out-of-competition. A GC-MS analysis, fulfilling all requirements as per TD EAAS (WADA, 2021a), was initially used to detect the six markers and the five ratios, which compose the urinary steroid profile of the ABP of 229 athletes. Table 3.1 includes the definitions we used to describe the three sample categories of athletes' steroid profiles. Fifty male and fifty female athletes were *negative* from each group with *normal* and *atypical* samples, while 15 male and 14 female athletes were *positive* with at least one confirmed *abnormal* sample in their steroid profile, employing Isotope Ratio Mass Spectrometry (IRMS) in line with WADA regulation (WADA, 2021c).



(a)

Figure 3.1: Extracted chromatograms obtained for an unsuspicious urine sample. Shown are the multiple reaction monitoring (MRM) ion transitions employed for the quantification of endogenous steroids (upper part) and their deuterated analogues (lower part).

Table 3.1: Description of possible outcomes of urine samples in the longitudinal steroid profile (SP).

Urine sample	Description
Normal	EAAS value in the SP classified as normal.
Atypical	EAAS value in the SP classified as atypical.
Abnormal	EAAS value in the SP classified as abnormal.

A total of 1433 spot normal urine samples were obtained from 100 athletes, while 2504 spot urine samples obtained from 100 athletes whose longitudinal steroid profiles contain values classified as atypical, and 462 spot urine samples from 29 athletes with at least one confirmed abnormal value in the steroid profile. This indicates a significant imbalance in the information provided between "negative" athletes (with normal and atypical samples) and "positive" athletes (with abnormal samples) as also shown in Figure 3.2. Specifically, athletes belonging to the group with normal samples were tested between 6 and 47 times (14 times on average), athletes with atypical values were tested between 6 and 69 times (25 times on average), and between 3 and 35 times (16 times on average) for athletes with abnormal samples. Sample calibration was carried out prior to the analysis according to the estimated real limits of the applied methodology. The limit of detection (LOD) values and the limit of quantification (LOQ) values within the steroid profiles of the athletes have been replaced by commonly accepted minimum cut-off values for all markers; i.e. all <LOQ and <LOD values in testosterone and epitestosterone were replaced by 1 ng/mL and 0.1 ng/mL respectively, while for <LOQ and <LOD values in the -diols were replaced by 5 ng/mL and 1 ng/mL, respectively.

Figures 3.3 and 3.4 depict the variation of the values from the six biomarkers and their five ratios by gender against the sampling time (red: female, blue: male). Every trajectory represents measurements from one athlete, the curves are separated into two classes; negative athletes, where no abnormal measurements are included in their steroid profiles, and positive athletes, whose steroid profile includes at least one abnormal measurement (see Table 3.1). Promising biomarkers and ratios can already be noticed from distinguishable trends in their trajectories. For example B5, T, T/E, A/T, and A5/E show a distinctive behaviour between the two classes.

However, there are markers and ratios for which the difference between negative and positive athletes, either simply doesn't exist or is less obvious.

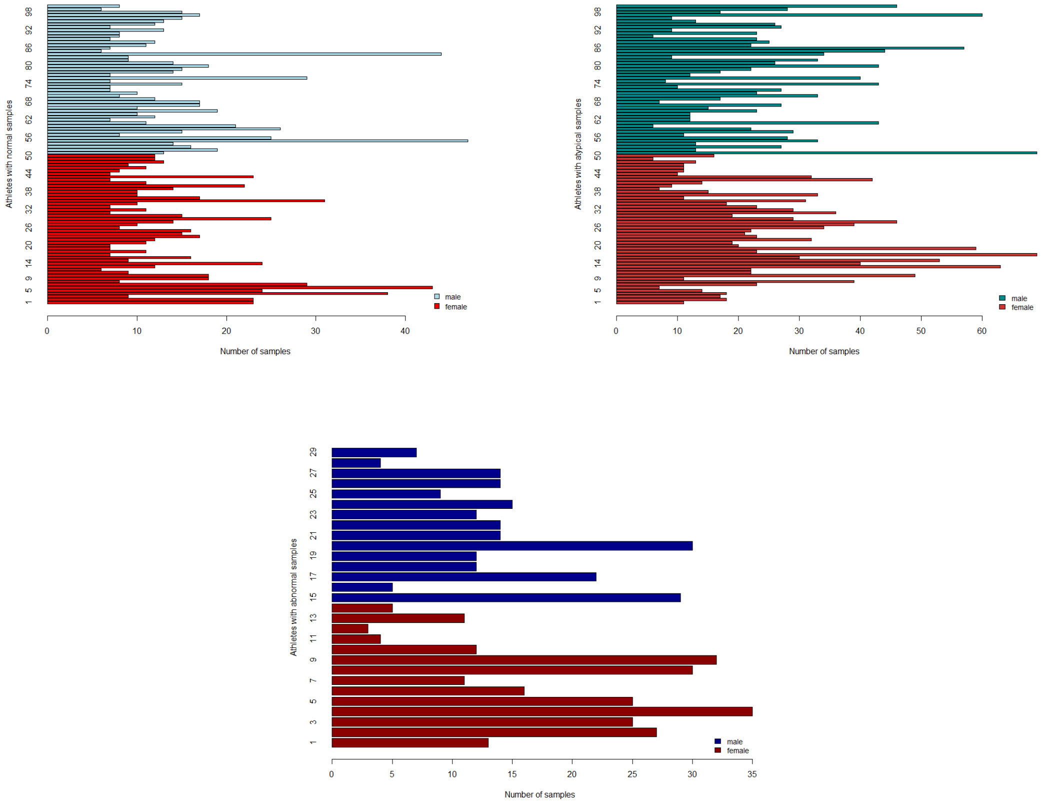
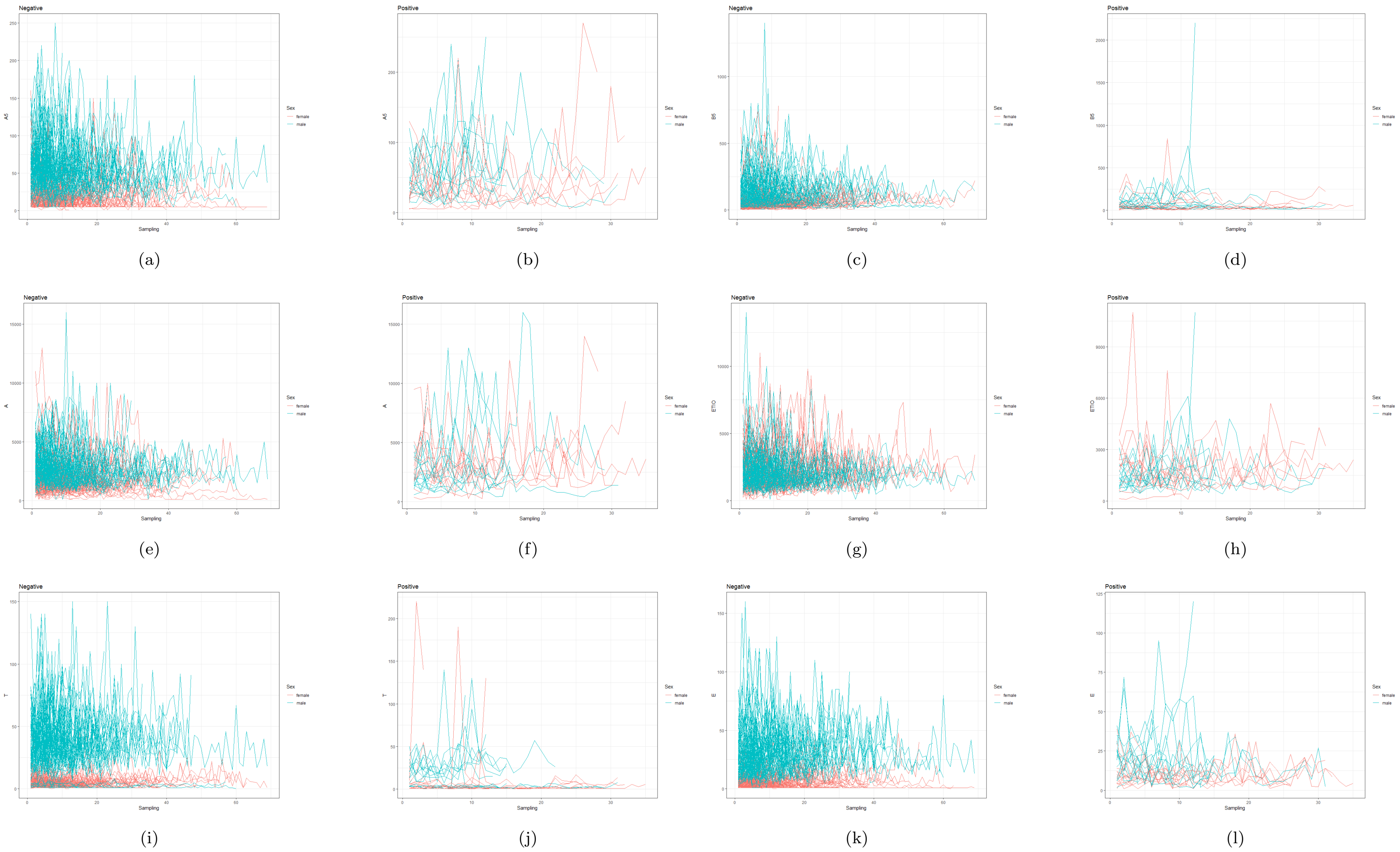


Figure 3.2: Number of samples collected from athletes with normal (top left), atypical (top right), and abnormal (bottom) samples. The majority of these samples were recorded out-of-competition (3164 urines), while the rest were recorded throughout competition events (1253 urines).

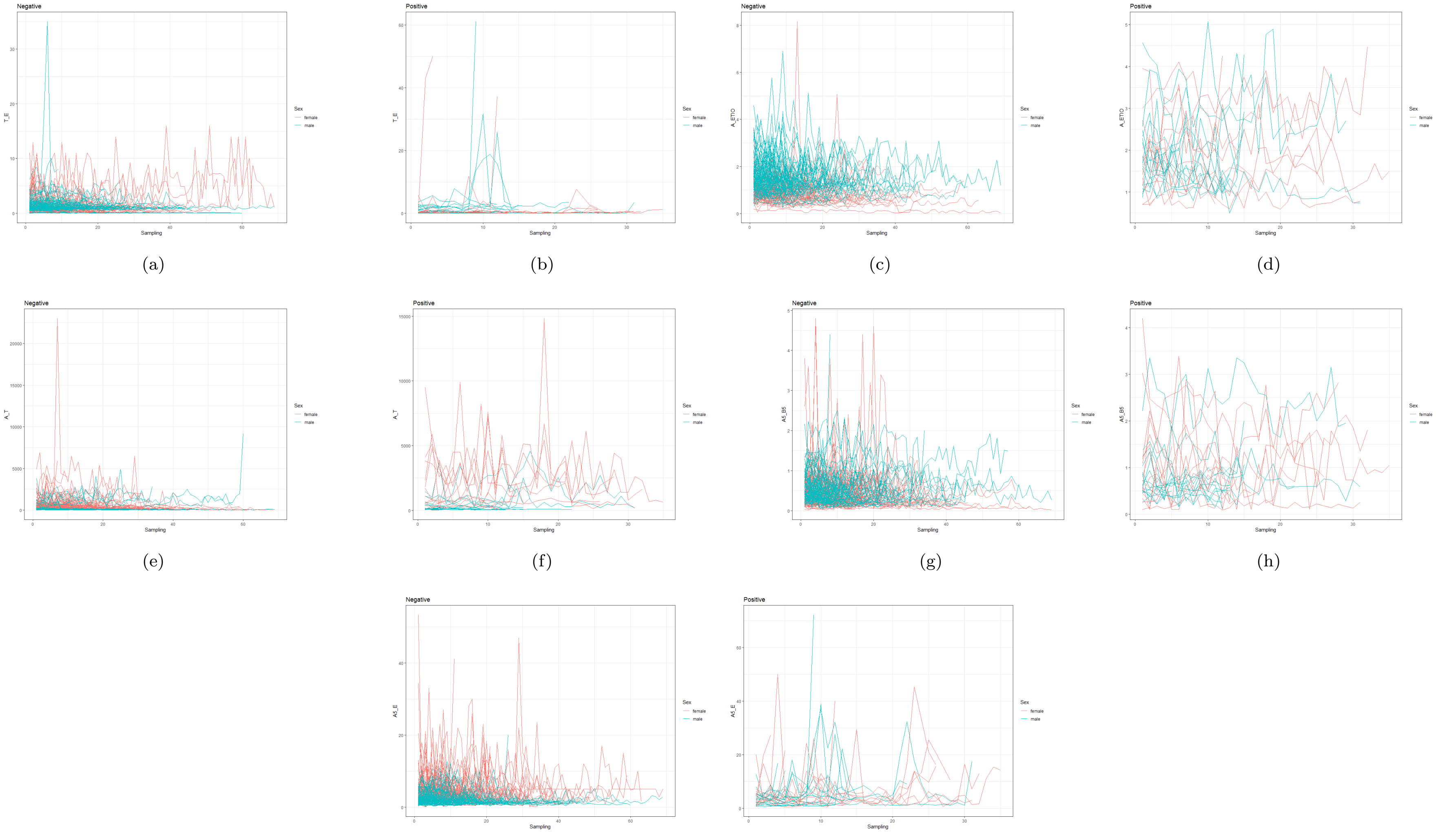


(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

FIGURE 3.3: Biomarkers by gender over time, separated into two classes; negative athletes (i.e. no abnormal measurements in their steroid profiles) and positive athletes (i.e. at least one abnormal measurement is included in their steroid profiles). The markers' profiles of athletes are presented as follows: for negative athletes (a) A5, (c) B5 (e) A, (g) ETIO, (i) T, (k) E and for positive athletes (b) A5, (d) B5 (f) A, (h) ETIO, (j) T, (l) E.



(a) (b) (c) (d)

(e) (f) (g) (h)

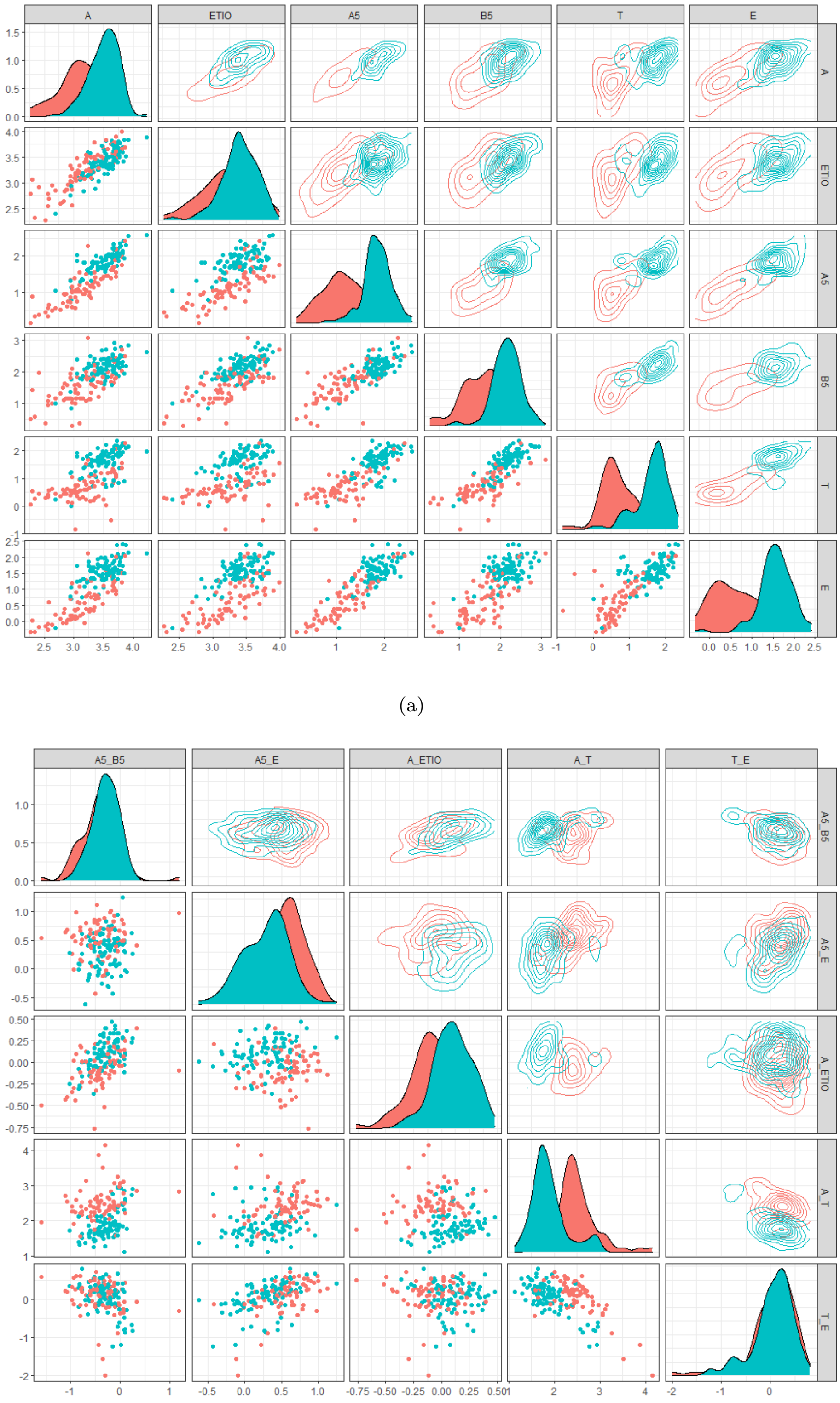
(i) (j)

Figure 3.4: Biomarkers' ratios by gender over time, separated into two classes; negative athletes (i.e. no abnormal measurements in their steroid profiles) and positive athletes (i.e. at least one abnormal measurement is included in their steroid profiles). The markers' profiles of athletes are presented as follows: for negative athletes (a) T/E, (c) A/ETIO (e) A/T, (g) A5/B5, (i) A5/E, and for positive athletes (b) T/E, (d) A/ETIO (f) A/T, (h) A5/B5, (j) A5/E.

3.2.2 Cross-sectional Data

From the cross-sectional study, single EAAS and ratios measurements of 164 healthy individuals have been provided representing a baseline population of which 91 were men and 73 women with age between 18 and 54. The graphs in Figure 3.5 present the scatter, density and contour plots for men (light blue) and women (coral) for all available markers and ratios. The levels of markers seem to be slightly separable between men and women. However, this is not true for the ratios, where the distributions of both genders seem similar, except for the A/T ratio. The correlation between the various markers and ratios is presented in Figure 3.6, which shows that plain markers are more highly correlated compared to the ratios. This also constitutes a reason of preferring ratios, in addition to their higher sensitivity as discussed in Section 1.3.6.

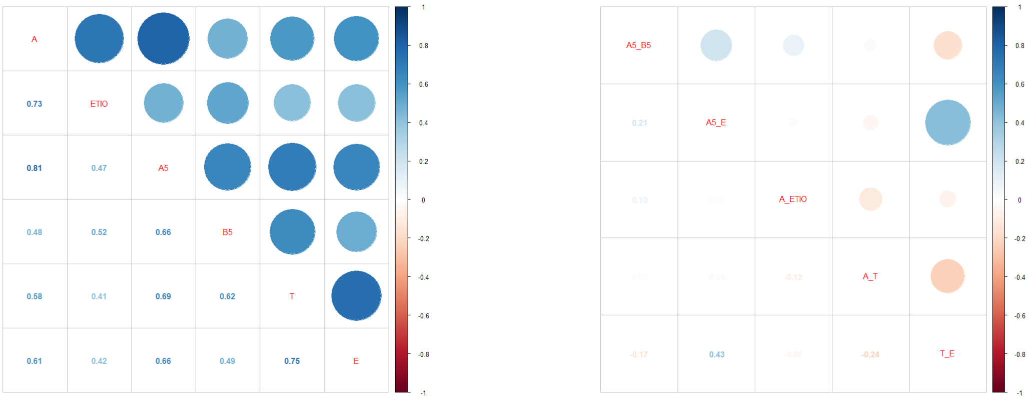
Table 3.2 summarises the baseline population statistics (minimum, inter-quartile range; IQ1 and IQ3, mean, median, maximum EAAS values and standard deviation). The mean values of the available metabolites and ratios obtained from both datasets with 4399 and 164 urine samples respectively, are reported in Figure 3.7. Gold diamonds signify the WADA threshold limits (lower and upper), which are available for some metabolites and ratios. According to the descriptive statistics of the longitudinally-monitored athletes in Tables C.2, C.3 and C.4 in Appendix C, the maximum values for A, T/E as well as for the A/Etio ratio exceeded WADA's thresholds for the three categories, while the maximum for Etio of athletes with atypical and abnormal samples exceeded WADA's threshold. With respect to the levels of epitestosterone (E), all recordings from the three categories were lower than the relevant upper limit, while higher levels of testosterone compared to WADA's threshold limit were found only in athletes with abnormal values.



(a)

(b)

FIGURE 3.5: Pairs plot by gender including the scatter, density and contour plots for (a) the six markers, and (b) their five ratios. Red-coloured distributions correspond to females, while blue-coloured ones to males.



(a)

(b)

FIGURE 3.6: Correlograms of (a) the six markers, and (b) the five ratios accompanied by the correlation coefficients.

TABLE 3.2: Descriptive summaries (minimum, inter-quartile range; IQ1 and IQ3, mean, median, maximum and standard deviation) of the metabolites and ratios of the baseline healthy population (91 men and 73 women). The names of the target metabolites are abbreviated as presented in Table C.1.

Target Metabolite	Min	IQ1	Mean	Median	IQ3	Max	SD
	(ng/mL)	(ng/mL)	(ng/mL)	(ng/mL)	(ng/mL)	(ng/mL)	(ng/mL)
A5	1.46	14.49	61.7	46.62	84.56	388.14	64.68
B5	1.88	40.46	139.92	89.14	180.92	1, 232.95	164.27
A	187.7	1,315.9	2, 997.3	2, 517.2	4,428.1	16, 674.1	2, 169.41
E	0.47	4.31	33.71	20.52	42.22	252.96	42.94
ETIO	187.5	1,379.8	2, 719.4	2, 337.9	3,614	9, 819.9	1, 803.23
T	0.14	4.17	39.37	20.39	61.43	229.03	46.27
A5/B5	0.03	0.3	0.64	0.46	0.74	15.57	1.22
A5/E	0.24	1.47	3.29	2.73	4.23	17.72	2.54
A/ETIO	0.17	0.78	1.17	1.05	1.47	3	0.53
A/T	13.13	53.97	374.49	125.08	282.92	14, 154.24	1, 269.35
T/E	0.01	0.76	1.63	1.44	2.19	6.48	1.16

3.2.3 Pre-model Testing

Before any modelling procedures, we carried out independence and normality hypothesis testing at level of significance 5% on the data coming from 229 athletes on their logarithmic scale. We initially computed the sample autocorrelations of all

EAAS series from athletes and we tested whether their series are white noise, alternatively to test the independence using the method of Li (2003). The independence hypothesis was retained for 94% of the tests, which implied no correlation within the series. Subsequently, Jarque-Bera tests were used for checking normality (Jarque and Bera, 1980), where 85% of the tests showed no evidence against the normality assumption.

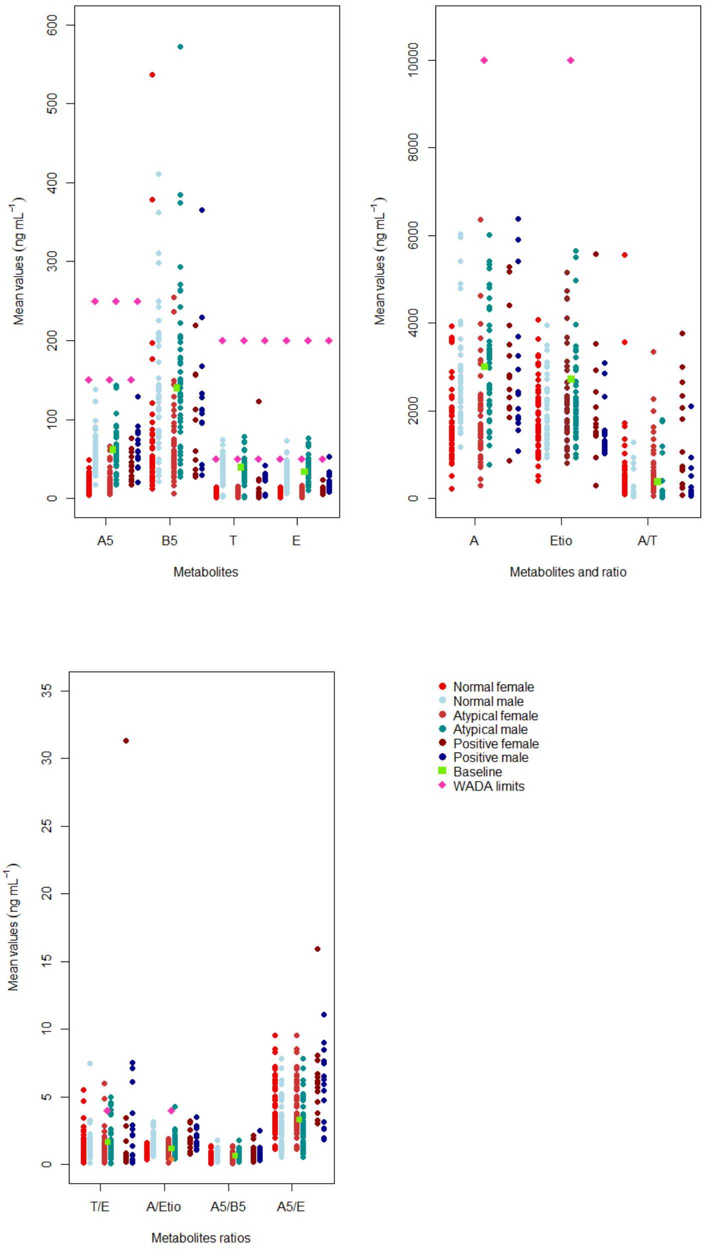


Figure 3.7: Mean values of six metabolites concentration levels (ng mL^{-1}) and five ratios from 229 athletes and the baseline population by gender and doping status (normal, atypical and abnormal/positive). WADA's limits are tagged by the gold diamonds when they are available from Table C.5. Androsterone and Etiocholanolone share the same population thresholds among females and males.

3.3 Results

Figure B.6 shows there is evidence of severe imbalance between normal samples (0: pastel green) and non-normal samples (atypical or abnormal) (1: pastel red), as well as an unstructured nature of the minority class. Normal samples specify the majority class (or the "target" class.), while the minority class (or "non-target" class) consists of atypical and abnormal samples. Out of 4399 urine samples across all athletes, only 327 (7.43%) were true positive/abnormal values (275 from athletes with atypical samples and 52 from athletes with abnormal samples). Since we deal with a severely skewed class distribution, we apply the majority impact learning technique, known as one-class classification, to all models presented in Sections 2.3.1, 2.3.2 and 2.3.3 as discussed in Section 2.3.8.

3.3.1 Univariate Bayesian Model

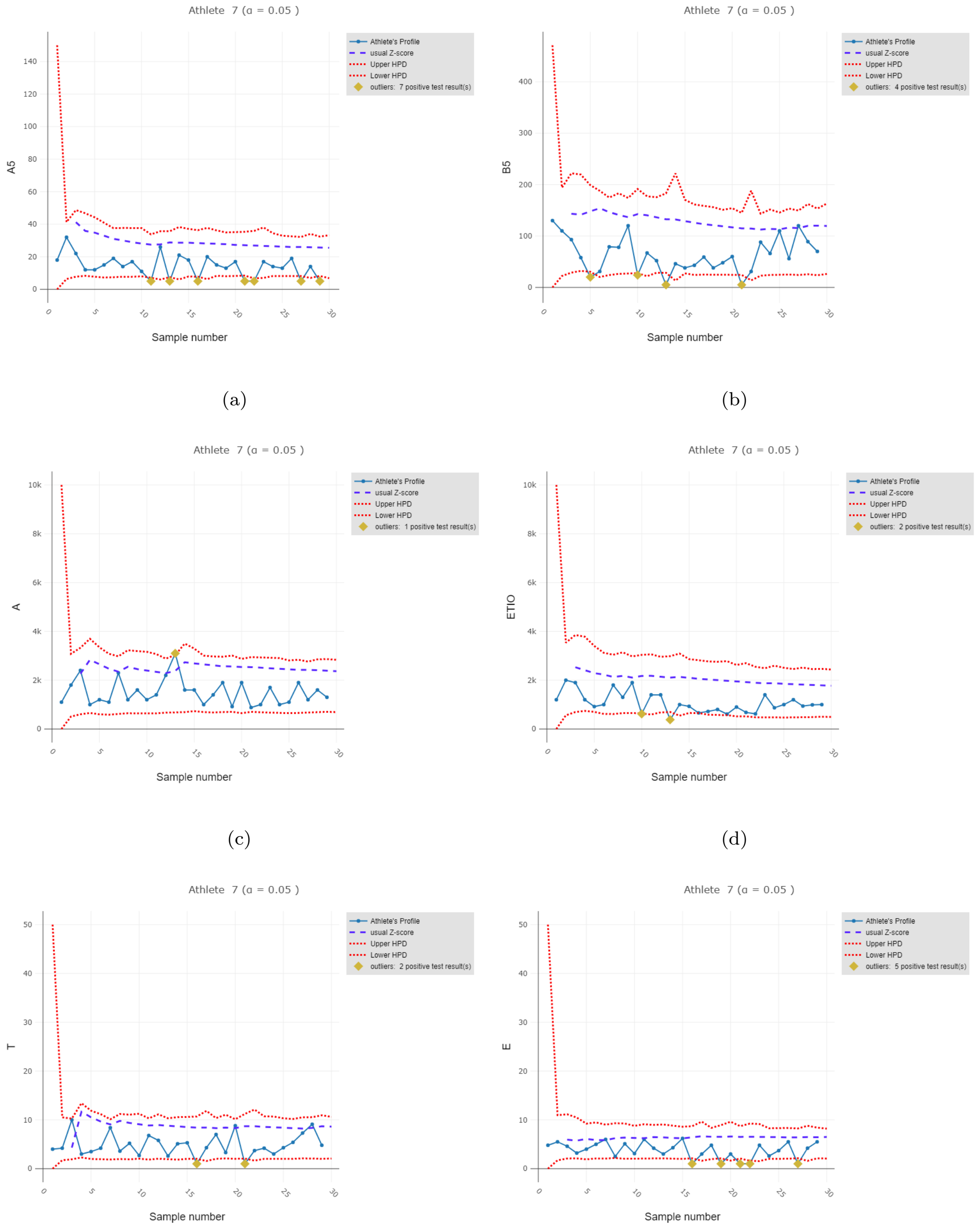
To apply the univariate Bayesian model on each athlete and for each biomarker of the ABP separately, we started by specifying the prior distributions for the model parameters μ and τ as $\mu|\tau \sim \mathcal{N}(\mu_0, 1/\tau\kappa_0)$ and $\tau \sim \mathcal{G}a(\alpha_0, \lambda_0)$, respectively. The correlation between the empirical mean and precision from all athletes' values for each marker was in the range (-0.33, 0.32) denoting a weak relationship. Nevertheless, from the tests performed on the 11 markers and ratios, we found statistically significant correlation, hence we consider a priori dependence between the parameters. To specify non-informative priors with large variances, we set the hyperparameters $\kappa_0 = 1$, $\alpha_0 = 10$, $\lambda_0 = 1$, and $\mu_0 = \mu_{js}$, that is the mean of the jth marker in its logarithmic scale for male subjects if $s = 0$, and for females if $s = 1$. Then, 5000 draws were sampled for each parameter from the known joint posterior distribution (Gaussian-Gamma). The out-of-sample predictive distribution of a new test result y_{n+1} given previous recordings $(y_1, y_2, ..., y_n)$ is

$$p(y_{n+1}|\boldsymbol{y}) = \int\int p(y_{n+1}|\mu,\tau)p(\mu,\tau|\boldsymbol{y})d\mu d\tau \approx \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\tau^{(t)}}{2\pi}\right)^{1/2} e^{-\frac{\tau^{(t)}}{2}(y_{n+1}-\mu^{(t)})^2}, \tag{3.1}$$

where $n \geq 0$, $(\mu^{(t)}, \tau^{(t)})$ is the pair of tth draw obtained through the sampler with total number of iterations $T = 5000$. In practice, the integration averaging is performed using an empirical average based on samples from the posterior distribution. At first, we simulate many replicates of new data, y_{n+1}, from the posterior predictive distribution and we derive the 95% HPD interval. The model is applied on the 4399 EAAS concentrations and ratios of 229 athletes. Since the proportion of non-normal EAAS for the three groups of athletes is known (0/1433 for normal, 275/2504 for atypical, and 52/462 for abnormal) we can estimate the predictive accuracy of the method.

In Figures 3.8 and 3.9 (a-k), the EAAS and ratios series of a non-doped and a doped athlete are depicted, respectively, with the blue-solid lines. The red dotted lines are the 95% HPD intervals of the predictive distribution, which denote the posterior normal boundaries at each time point. Before observing any data, the upper limits are defined by WADA's population thresholds, when they are available (see Table C.5). For marker B5 we used the maximum value obtained by the Caucasian population in Van Renterghem et al. (2010), while for the remaining ratios; i.e. A5/B5, A5/E and A/T we chose the values 4, 10 and 10,000, respectively, as reasonable starting thresholds. The purple dashed lines indicate the usual Z-score's upper limits as presented in Sottas et al. (2006). The gold diamonds symbolise the abnormal values in an athlete's profile suggested by the model, which need further investigation. For the T/E ratio, there are two additional green dashed-dotted lines. These indicate the upper and lower limits of the T/E model with informative priors introduced by Sottas et al. (2006).

Note that if the model suggests an outlier, we automatically exclude it from the set of recordings which are used to compute the HPD intervals, because it might have an impact on the following personalised accepted limits. For example, in Figure 3.8(e), there are two testosterone samples which are lower than the lower boundaries, and in Figure 3.8(g) Sottas' model identifies six abnormal T/E tests, while the general univariate model suggests five. In general, knowing that athlete 7 has not received any doping substances, there were many false positives leading to a weak classification performance. Regarding athlete 202, whose 21^{st}, 22^{nd} and 23^{rd} sample tests are confirmed as abnormal, only E, T/E and A5/E were sensitive enough to detect these anomalies in their SP.



(a)

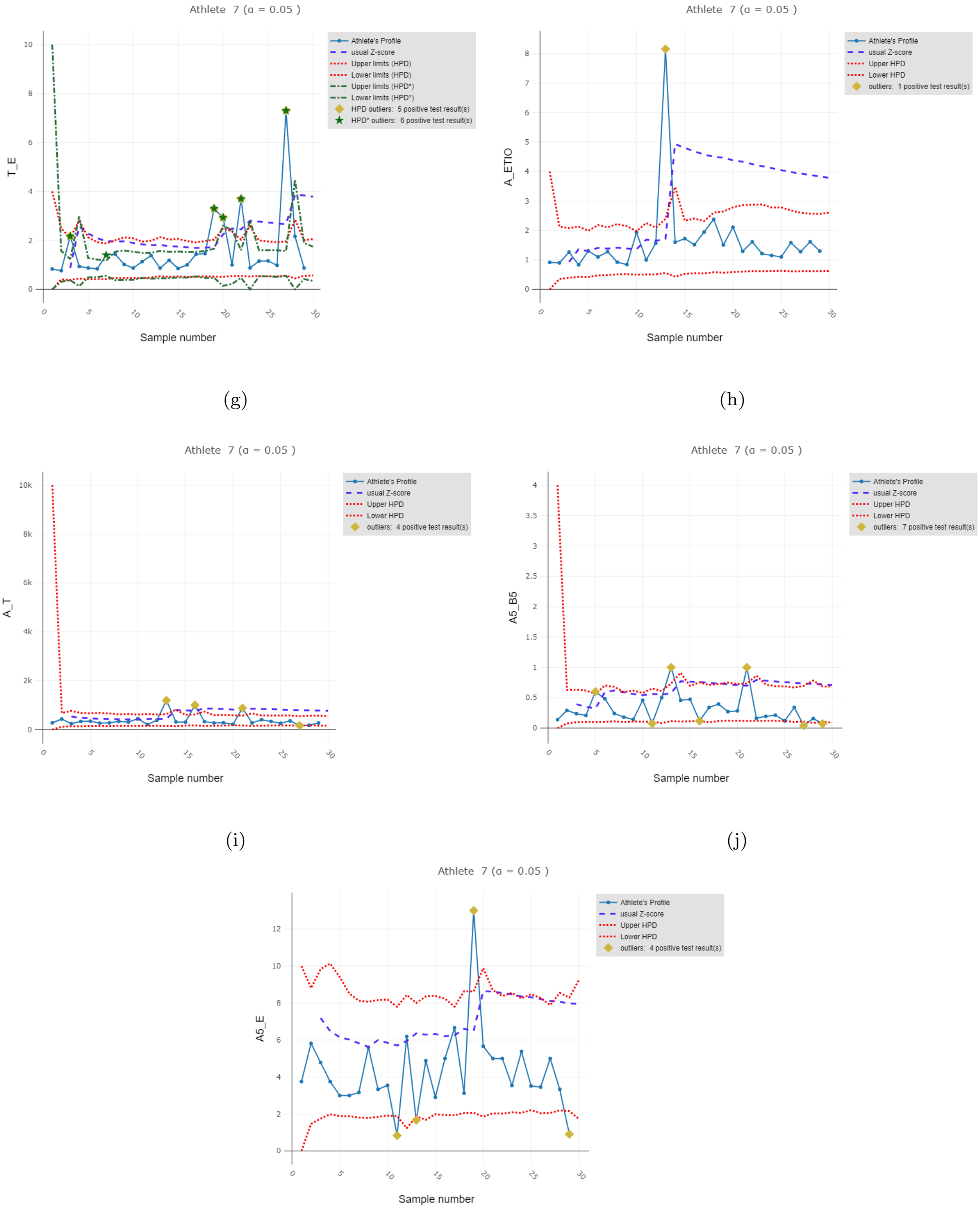
(b)

(c)

(d)

(e)

(f)

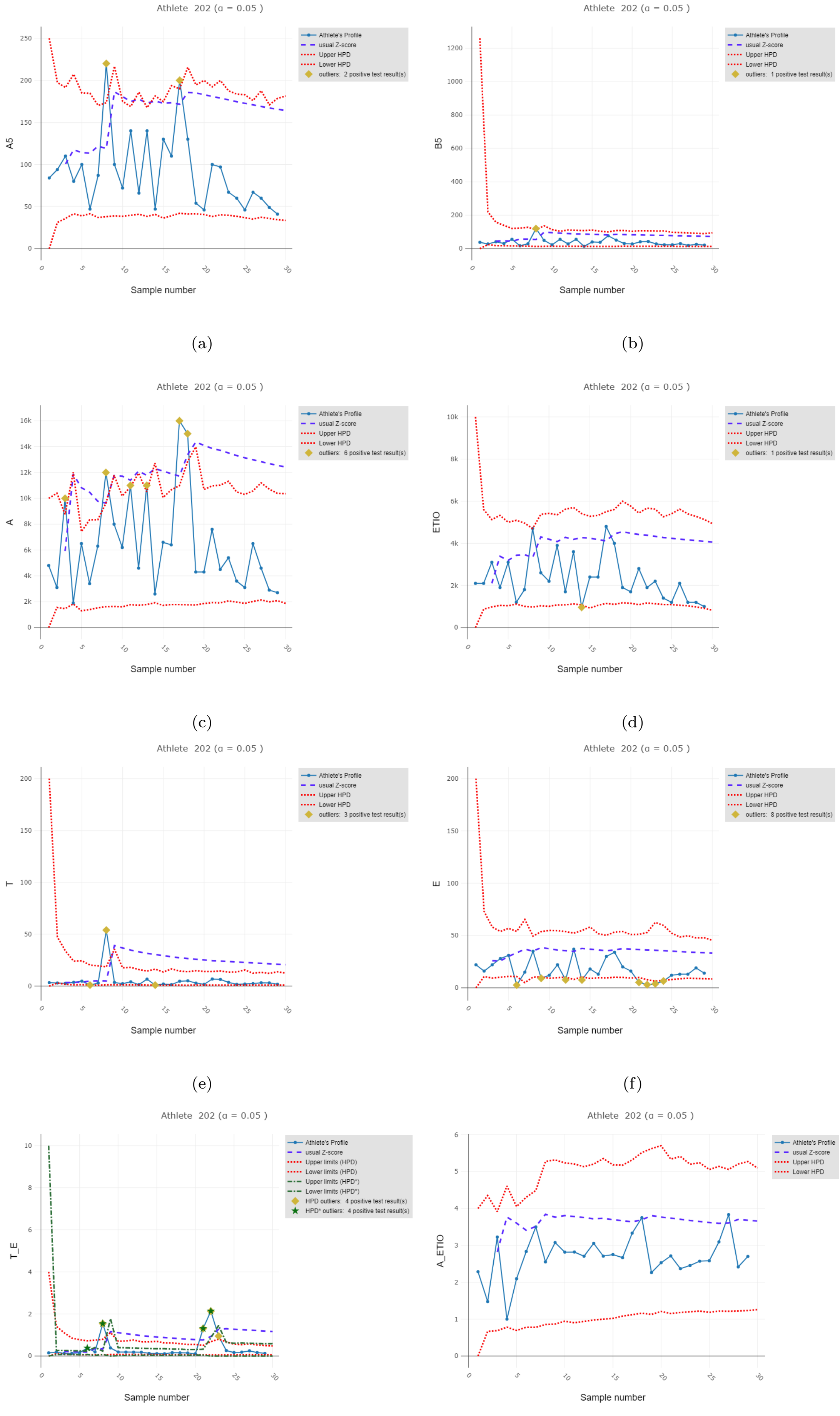


(g) (h)

(i) (j)

(k)

Figure 3.8: (a-k) A series of 29 longitudinal values of the six EAAS and their five ratios (blue solid-dotted line) obtained from a non-doped athlete; upper and lower limits (red dotted lines) are calculated using the 95% HPD intervals of the predictive distribution from the univariate Bayesian model; upper limits assuming a usual Z-score (purple dashed line); suggested abnormal values are denoted by the gold diamonds. (g) Upper and lower limits (green dashed-dotted lines) are calculated using the 95% HPD interval of the predictive distribution from the T/E model of Sottas et al. (2006); suggested abnormal values based on the T/E model are denoted by the green stars.


Athlete 202 (α = 0.05)
A5
Sample number
Athlete's Profile
usual Z-score
Upper HPD
Lower HPD
outliers: 2 positive test result(s)
Athlete 202 (α = 0.05)
B5
Sample number
outliers: 1 positive test result(s)
Athlete 202 (α = 0.05)
A
Sample number
outliers: 6 positive test result(s)
Athlete 202 (α = 0.05)
ETIO
Sample number
outliers: 1 positive test result(s)
Athlete 202 (α = 0.05)
T
Sample number
outliers: 3 positive test result(s)
Athlete 202 (α = 0.05)
E
Sample number
outliers: 8 positive test result(s)
Athlete 202 (α = 0.05)
T_E
Sample number
Athlete's Profile
usual Z-score
Upper limits (HPD)
Lower limits (HPD)
Upper limits (HPD*)
Lower limits (HPD*)
HPD outliers: 4 positive test result(s)
HPD* outliers: 4 positive test result(s)
Athlete 202 (α = 0.05)
A_ETIO
Sample number
Athlete's Profile
usual Z-score
Upper HPD
Lower HPD

(a)

(b)

(c)

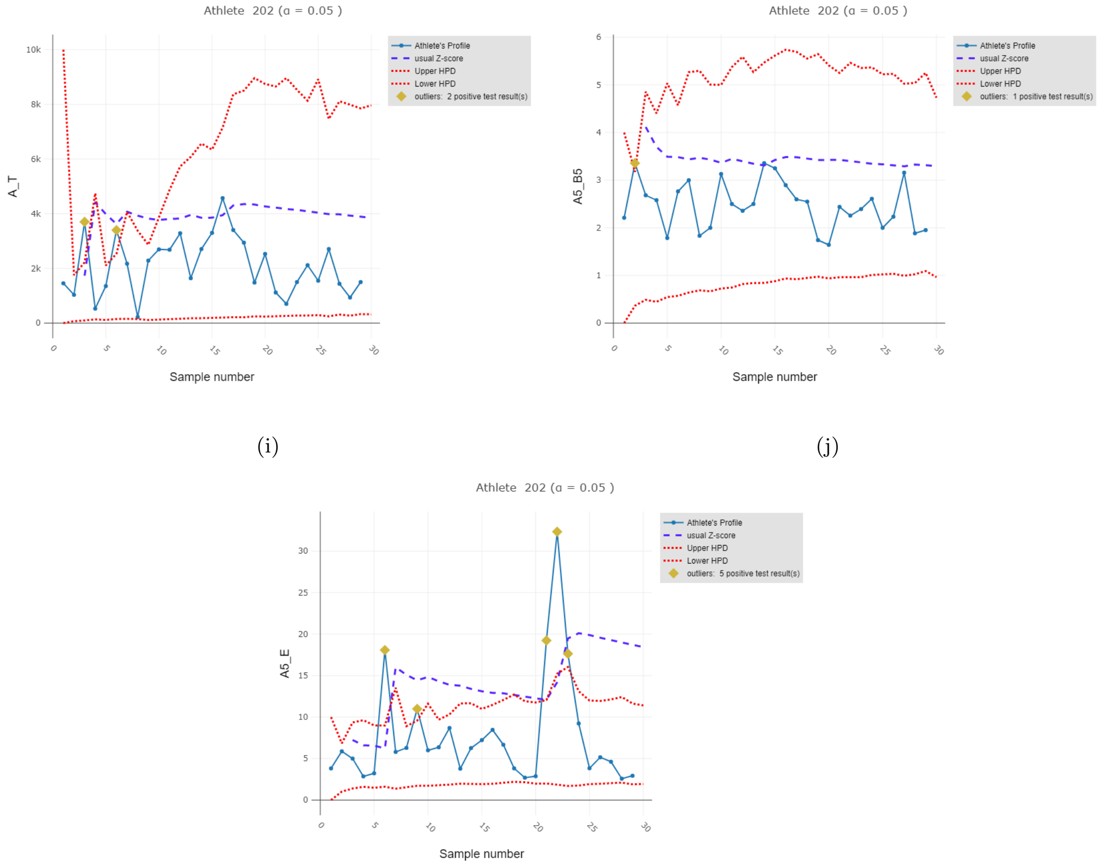
(d)

(e)

(f)

(g)

(h)



(i)

(j)

(k)

FIGURE 3.9: (a-k) A series of 29 values of 6 EAAS and 5 ratios (blue solid-dotted line) obtained from a doped athlete; upper and lower limits (red dotted lines) are calculated using the 95% HPD intervals of the predictive distribution from the univariate Bayesian model; upper limits assuming a usual Z-score (purple dashed line); suggested abnormal values are denoted by the gold diamonds. (g) Upper and lower limits (green dashed-dotted lines) are calculated using the 95% HPD interval of the predictive distribution from the T/E model of Sottas et al. (2006); suggested abnormal values based on the T/E model are denoted by the green stars.

3.3.2 Multivariate Bayesian Multilevel Model

To apply the multivariate Gaussian multilevel model, we initially specified the prior distributions for the model parameters $\boldsymbol{\theta}$ as described in Section 2.3.3. The prior covariance matrices S_e, S_μ and S_b are all set equal to $(1/1000)\mathbb{I}_{\mathbb{K}}$, and the degrees of freedom are $d_e = d_\mu = d_b = K - 1$, where K is the dimensionality of variables of the data. We use historical prior information obtained from the baseline cross-sectional dataset of 164 non-doped athletes (91 men and 73 women), which is captured by the prior mean vector μ_0. Note that the model accommodates different prior mean vectors for men and women. Then, 3000 draws were sampled for each parameter from the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$ using the Gibbs sampler (Algorithm 4), while the first 1/3 was discarded. Given the remaining set of samples $\{\boldsymbol{\theta}^{(t)}\}_{t=1}^{T=2000}$, our estimate for the predictive distribution is

$$p(y_{n+1}|\boldsymbol{y}) \approx \frac{1}{T}\sum_{t=1}^{T} p(y_{n+1}|\boldsymbol{\theta}^{(t)}). \tag{3.2}$$

We first simulate from the posterior predictive distribution many replicates of the new data, y_{n+1}, and thus we derive the 95% HPD interval. The model is applied on the 4399 EAAS and ratios of 229 athletes (100 athletes with normal samples, 100 athletes with atypical samples and 29 athletes with abnormal samples). Data from athletes, whose samples were all normal, were used to train the model, while atypical and abnormal samples were used as a test set. The idea is to train the model with normal data from non-doped athletes, by estimating $p(\boldsymbol{\theta}_{\text{normal}}|\boldsymbol{y}_{\text{normal}})$ and then test how likely is a future unlabelled observation to be generated by this model. Table 3.3 includes the classification performance of the proposed multivariate model applied to: a) all markers and ratios; b) markers only; and c) ratios only.

3.3.3 Classification Performance

In this section we present the classification performance of the models applied on the same dataset for detecting doping cases within athletes' steroid profiles. In forensic toxicology, high specificity is important, thus a very low false positive rate is required in order to prevent the accusation of an innocent athlete. However, the classification accuracy values and measures regarding the majority class such as specificity, tend to be pretty high because they are computed under the assumption of balanced class distributions. Consequently, we need to use appropriate metrics for evaluating the classification performance of the models which can deal with the imbalance of the dataset.

Dealing with the "Accuracy paradox"

Classification accuracy is the most commonly used metric for evaluating classification models as the number of correct predictions divided by the total number of predictions.. When the class distribution is slightly skewed, accuracy can still be a useful metric. However, when the skew in the class distributions is severe as happens with the current dataset, accuracy can become an unreliable measure for assessing the model performance; this point is explored in detail by Glavin and Madden (2009). Achieving very high classification accuracy as shown in Table 3.3 is trivial when dealing with imbalanced classes. In this problem, the majority class represents "normal" urine samples, while the minority class represents "abnormal" urine samples. In cases like this, where there is a large class imbalance, the simple univariate models can predict the value of the majority class for most predictions and achieve a high specificity and classification accuracy, but the problem is that this is not sufficient in order to select the optimum model for detecting potential doping abuse. This is because there is no enough information about the minority class to make an accurate prediction, and the classifiers get biased towards the majority class. Due to this phenomenon, called *"Accuracy Paradox"*, we evaluate the models using the F-Measure (F1 score), precision and sensitivity (Nguyen et al., 2009; Valverde-Albacete et al., 2013)[1]. Precision and sensitivity/recall are the two most common metrics that take into account class imbalance, which are also the foundation of the F1 score (i.e. the harmonic mean of precision and recall). These evaluation metrics take into account not only the number of prediction errors that

[1]

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}},$$

$$\text{F1} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision}+\text{recall}},$$

$$\text{precision} = \frac{\text{TP}}{\text{PP}},$$

$$\text{recall} = \text{sensitivity} = \frac{\text{TP}}{\text{P}},$$

$$\text{specificity} = \frac{\text{TN}}{\text{N}},$$

where TP is true positive, TN is true negative, FP is false positive, FN is false negative, PP is predicted positive, P is positive and N is negative.

our model makes, but also look at the type of errors that are made. Hence, a model with a lower accuracy could be selected because it has a greater predictive power.

Dealing with imbalanced classes

There are a variety of methods used to deal with imbalanced binary classification problems, widely known as "Sampling Methods". The main goal of these methods is to convert the imbalanced dataset into balanced distributions by altering the size of the original dataset to provide the same proportion of balance in each class. We chose to work with the method of *"Random Oversampling"*, which randomly replicates the observations from the minority class to balance the data (Kotsiantis et al., 2006). ROSE package in R was used to generate replicates of the data from each athlete (Lunardon et al., 2014).

Table 3.3: Predictive performance of the univariate and multivariate models on the athlete's profiles based on the 95% HPD interval.

Classification						Balanced	Overall
model	Variable	F1 score	Precision	Sensitivity	Specificity	Accuracy[3]	Accuracy (95% CI)
Univariate	A5	0.16	0.11	0.25	0.84	0.55	0.80 (0.79, 0.81)
	B5	0.17	0.12	0.30	0.83	0.56	0.79 (0.78, 0.80)
	A	0.13	0.11	0.17	0.86	0.53	0.83 (0.82, 0.84)
	ETIO	0.13	0.10	0.16	0.89	0.52	0.84 (0.82, 0.85)
	T	0.18	0.13	0.30	0.83	0.57	0.79 (0.78, 0.81)
	E	0.16	0.11	0.34	0.78	0.56	0.75 (0.73, 0.76)
	T/E[1]	0.19	0.15	0.26	0.88	0.57	0.84 (0.82, 0.85)
	A/ETIO	0.24	0.39	0.17	0.98	0.58	0.92 (0.91, 0.93)
	A/T	0.15	0.13	0.17	0.91	0.54	0.86 (0.85, 0.87)
	A5/B5	0.18	0.14	0.25	0.88	0.57	0.84 (0.82, 0.85)
	A5/E	0.23	0.17	0.35	0.86	0.61	0.82 (0.81, 0.83)
Sottas et al.	T/E[2]	0.20	0.15	0.32	0.85	0.59	0.81 (0.80, 0.82)
pre-oversampling	EAAS	0.24	0.18	0.38	0.78	0.58	0.74 (0.72, 0.75)
MGMM	ratios	0.35	0.29	0.44	0.87	0.65	0.82 (0.80, 0.83)
	all	0.30	0.20	0.55	0.73	0.64	0.71 (0.70, 0.73)
post-oversampling	EAAS	0.34	0.50	0.26	0.74	0.50	0.50 (0.49, 0.52)
MGMM	ratios	0.29	0.51	0.20	0.81	0.50	0.50 (0.49, 0.52)
	all	0.40	0.51	0.33	0.69	0.51	0.51 (0.49, 0.52)
GLMM	EAAS	0.03	0.18	0.02	0.99	0.50	0.88 (0.87, 0.89)
	ratios	0.03	0.17	0.02	0.99	0.50	0.88 (0.87, 0.89)
	all	0.03	0.18	0.02	0.99	0.50	0.88 (0.87, 0.89)

[1] This model specifies weakly informative priors.

[2] For this model the priors are set to be strongly informative.

[3] Balanced Accuracy = $\frac{\text{sensitivity}+\text{specificity}}{2}$

In Table 3.3, the classification using the univariate models with false positive rate $\alpha = 0.05$ suggest the T/E, A/ETIO and A5/E ratios as the most sensitive variables with higher F1 scores ($\text{F1}_{T/E^1} = 0.19$, $\text{F1}_{A/ETIO} = 0.24$ and $\text{F1}_{A5/E} = 0.23$) for detecting anomalies in the steroidal profile. The results from the T/E model in the first part of the table are based on semi-informative conjugate priors, whereas the results from the T/E model in the second part are based on the informative priors of Sottas et al. (2006). The latter model achieves a slightly better prediction performance based on the $\text{F1}_{T/E^2} = 0.20$, $\text{sensitivity}_{T/E^2} = 0.32$ and balanced $\text{accuracy}_{T/E^2} = 0.59$.

We have also presented in Figure 3.11 the ROC (Receiver Operating Characteristics) curves and the Precision-Recall curves to measure the accuracy of the classification predictions in the various models. The ROC curve is created by plotting the sensitivity against the false positive rate (FPR) or 1-specificity at various threshold settings, while the precision-recall curve shows the relationship between precision (or positive predictive value) and recall (or sensitivity) for every possible cut-off. According to the ROC curves, the model for the A5/E ratio showed superiority compared to the other univariate models in Figure 3.11(a), while the model for A/ETIO ratio showed superiority in the Precision-Recall plot. The curve of Sottas et al. (2006) model for T/E is higher in Figure 3.11(b), which verifies its higher predictive performance.

The proposed Bayesian model (MGMM) has been applied on the original data and the data after over-sampling using (i) the markers only, (ii) the ratios only, and (iii) the markers and the ratios together. The values of the metrics of the pre-oversampling multivariate models presented in Table 3.3 are overall higher compared to the values corresponding to the univariate models. Comparing further between the three applications of these MGMMs, the five available ratios were found to be the most powerful set of variables with the highest metric values $\text{F1}_{ratios} = 0.35$, $\text{precision}_{ratios}$=0.29, $\text{sensitivity}_{ratios}$=0.44, $\text{specificity}_{ratios}$=0.87 and highest balanced accuracy_{ratios}=0.65. Same conclusions regarding the superiority of the pre-oversampling multivariate model using the ratios are depicted in the diagrams from Figure 3.11, where blue lines show better relationship between sensitivity and

1-specificity, and between precision and recall. In Figure 3.10 all classification metrics are plotted over the various combinations of modelling, i.e. each marker and ratio separately, markers only (EAAS), ratios only, and all markers and ratios together. Again we conclude that an overall best classification performance is achieved by applying the multivariate model (MGMM) on the ratios only without oversampling.

The issue regarding the class skewness of the dataset has been eliminated after applying the oversampling technique, and this is reflected by the values of the balanced accuracy and overall accuracy which have been equivalent. However, in this example it seems that applying the method of random oversampling does not show any better overall classification performance of the model. It is worth mentioning that F1 scores and sensitivity values obtained from the various applications of generalised linear mixed-effects models (GLMMs) were quite low. This happens because GLMMs can be unstable when sample sizes across groups are highly unbalanced.

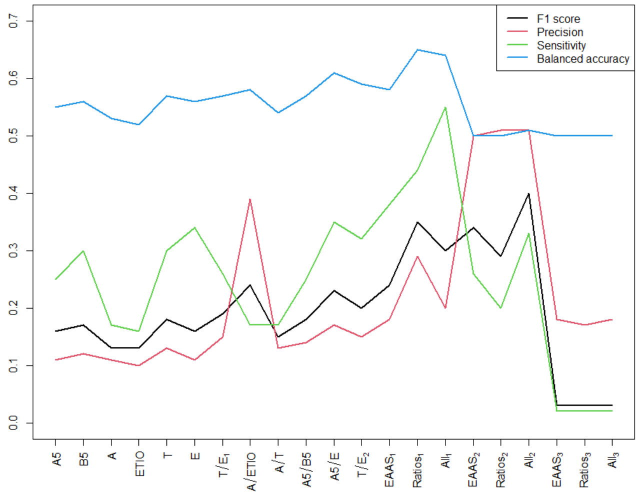
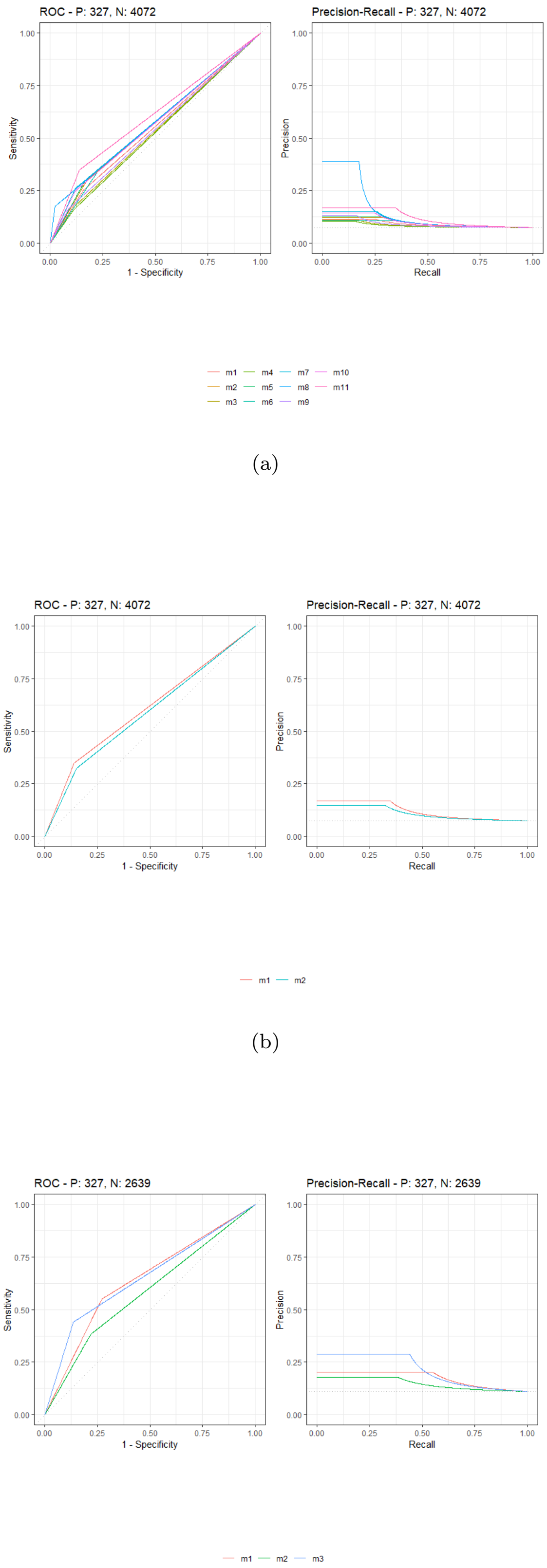


Figure 3.10: Classification metrics per component; i.e. markers and ratios separately, only EAAS, only ratios, all markers and ratios. T/E_1 corresponds to the T/E model from Section 2.3.1, while T/E_2 corresponds to the T/E model by Sottas et al. (2006).



(a)

(b)

(c)

Figure 3.11: ROC curves obtained from (a) the univariate models (m1: A5, m2: B5, m3: A, m4: ETIO, m5: T, m6: E, m7: T/E, m8: A/ETIO, m9: A/T, m10: A5/B5, m11: A5/E), (b) m1: Sottas' model for T/E vs m2: T/E model, and (c) the multivariate models (m1: all markers, m2: EAAS, m3: ratios).

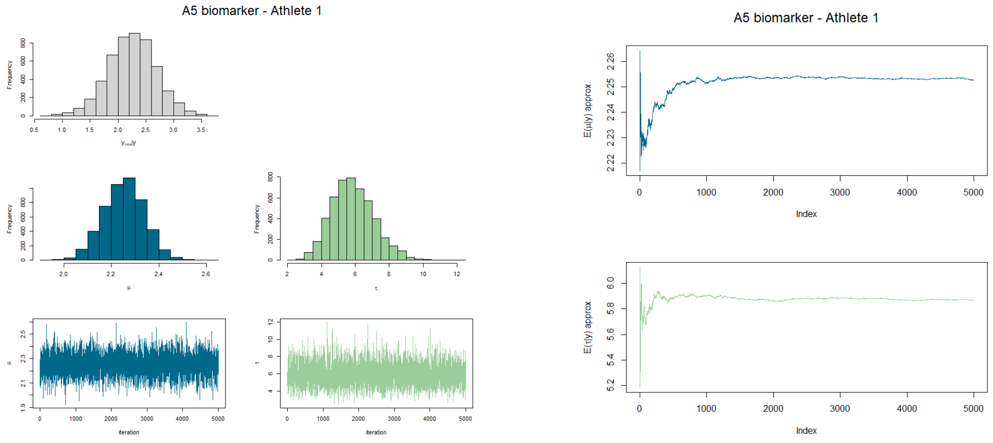
3.3.4 MCMC Results

3.3.4.1 MCMC for the Univariate Bayesian Model

Gibbs vs Exact Sampling

This section constitutes an example where we test the sampling performance of Gibbs sampler compared to the exact sampling as discussed in Section 2.3.1. Both sampling techniques have been applied using the recorded measurements on the logarithmic scale of each biomarker separately for every athlete, as discussed in Sections 2.3.5.2 and 3.3.1.

For implementing the Gibbs sampler, we set the same hyperparameters as for the exact sampling method, that is $\kappa_0 = 1$, $\alpha_0 = 10$, $\lambda_0 = 1$, and $\mu_0 = \mu_{js}$, which is the mean of the jth marker on its logarithmic scale for male subjects if $s = 0$, and for females if $s = 1$. Using the Gibbs algorithm 2, 5,000 draws of the parameters μ and τ were generated from the posterior distribution after a burn-in period of 1,000 iterations. To visualize the draws of the parameters, we present their traceplots obtained from the analysis of the A5 series of a single athlete (see Figure 3.12a). Informally, the convergence of the sampler is achieved since the posterior means for both parameters have settled around a certain value (Figure 3.12b). Figure 3.13 shows the points after complete sweeps through both parameters. Both are valid samples from the posterior distribution. Also, the starting point (red dot) and a two-dimensional 95% credible region with a dark red line are displayed here. Specifically, the latter is a bivariate region of highest marginal posterior density for the two variables μ and τ, given the sample from the posterior distribution. This region has been calculated using the `HPDregionplot` function in R, which uses the two-dimensional kernel density estimation to calculate a bivariate density, then normalizes the plot and calculates the contour corresponding to a contained volume of probability level of the total volume under the surface (a two-dimensional Bayesian credible region).



(a)

(b)

Figure 3.12: (a) Histograms, traceplots and (b) running averages of the parameters μ and τ over the iterations using the Gibbs sampler and semi-informative priors for A5 of an athlete. The posterior means for μ and τ can be approximated by the averages $E[\mu|\boldsymbol{y}] \approx \frac{1}{N}\sum_{i=1}^{N}\mu^{(i)}$ and $E[\tau|\boldsymbol{y}] \approx \frac{1}{N}\sum_{i=1}^{N}\tau^{(i)}$, respectively.

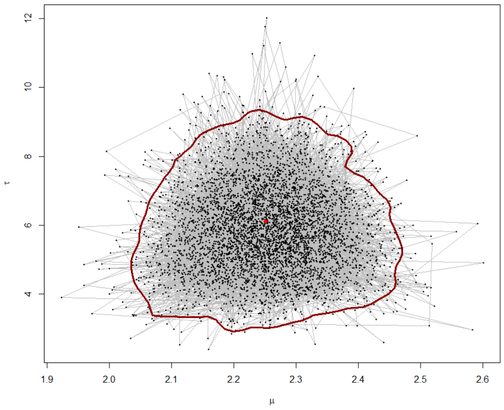


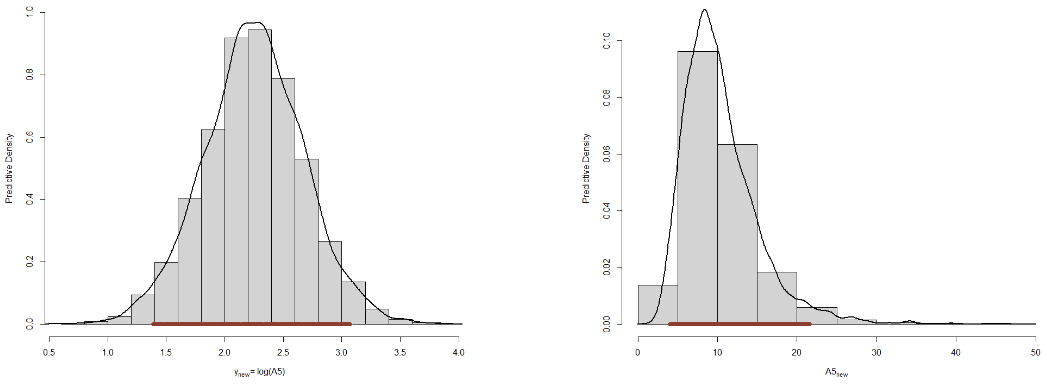
Figure 3.13: Scatter plot of μ and τ with the 95% highest posterior region in red using the Gibbs sampler and semi-informative priors for A5 of one athlete.

Now suppose we want to forecast a new observable $y_{new} = y_{n+1}$ for this certain marker of athlete 1 by using the out-of-sample predictive distribution given by the equation (3.1). Consequently, we compute the $(1-\alpha)\%$ highest predictive density interval (HPD). For each pair $(\mu^{(t)}, \tau^{(t)})$ drawn from the Gibbs sampler, we draw a value of $y_{new} = \log(A5_{new})$ from $\mathcal{N}(\mu^{(t)}, 1/\tau^{(t)})$ distribution. Then, we compute the corresponding HPD intervals as shown in Table 3.4.

Table 3.4: The 95% and 99% HPD intervals using the Gibbs sampler with semi-informative priors for A5 of athlete 1.

	$\log A5$		A5	
	L	U	$\exp(L)$	$\exp(U)$
95% HPD	1.39	3.07	4.0	21.64
99% HPD	1.07	3.36	2.91	28.77

Figure 3.14 represents the histogram of predictive densities of $\log(A5_{new})$ and $A5_{new}$ accompanied by the 95% HPD intervals for each case denoted by the dark red lines. We conclude that the most likely upcoming values of biomarker A5 for this athlete are in the range (4, 21.64). Values outside this range are considered as abnormal observations that should be examined further.



(a)

(b)

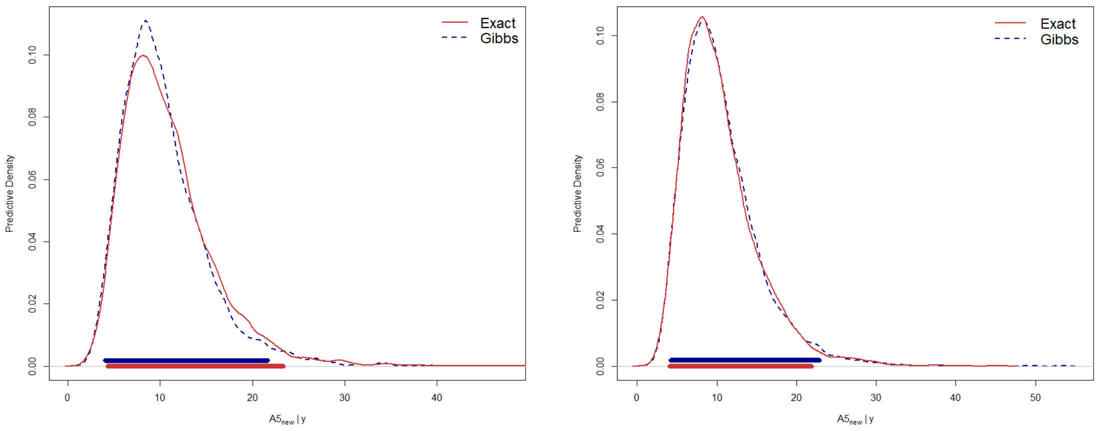
Figure 3.14: Histograms of the predictive densities with the 95% HPD intervals for $y_{new} = \log(A5_{new})$ and $A5_{new}$ symbolised by the dark red lines. The black curves represent the probability densities (kernel densities) of future observations.

Table 3.5 presents the results extracted from the sampling methods used; that is the Gibbs sampler using semi-informative priors, and exact sampling from the known posterior distribution. Estimates of the posterior means for both parameters as well as the 95% HPD intervals for A5 were fairly close. We can finally compare their predictive densities by overlaying the two curves and their 95% HPD intervals in a single plot (see Figure 3.15). Again we derive the same conclusions that the

Gibbs and Exact predictive densities look similar particularly in the right plot where more draws (10,000) have been sampled compared to the left plot (5,000). The Kolomogorov-Smirnov test (Conover, 1971), applied on the 10,000 samples, also supports that there is no statistically significant deviation in distribution between the samples from Gibbs and Exact sampling methods with test statistic $D = 0.0151$ and p-value $= 0.2043$. All the above indicate that a large number of Gibbs sweeps can provide reliable estimates in situations where the posterior distribution is unknown and it cannot be stated in a closed form.

Table 3.5: Posterior estimates and the 95% HPD intervals of A5 obtained from Gibbs sampler and Exact sampling.

	$E(\mu\|y)$	$E(\tau\|y)$	L(A5)	U(A5)	CPU time
Gibbs	2.25	5.87	4.0	21.64	0.28
Exact	2.28	5.62	4.33	23.34	≈ 0



(a)

(b)

Figure 3.15: Predictive distributions of future observations $A5_{new}$ and their 95% HPD intervals from each sampling method (Gibbs and Exact) using a) 5,000 and b) 10,000 draws.

Table 3.6: Two-sample Kolmogorov-Smirnov tests

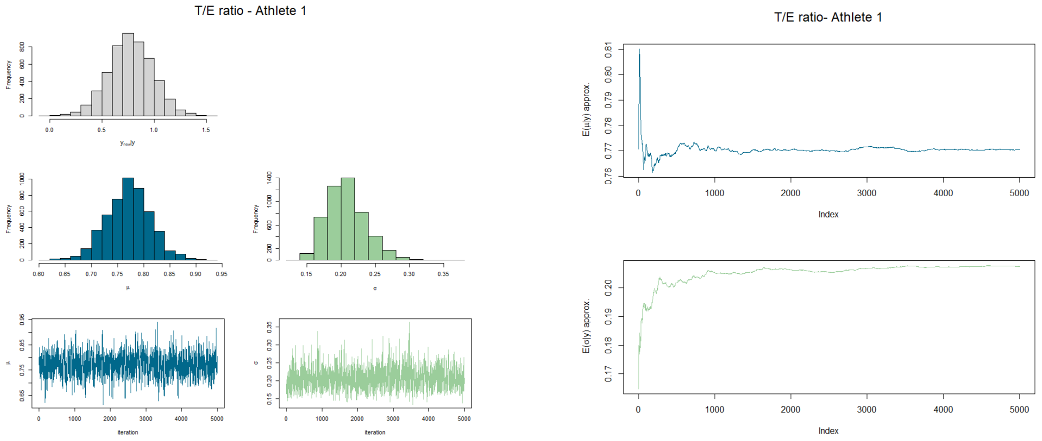
Sample size	D	p-value
5,000	0.0484	1.638×10^{-5}
10,000	0.0151	0.2043

MH vs MWG Sampling

In this section we examine the performance of Metropolis-Hastings (MH, algorithm 1) and Metropolis-within-Gibbs (MWG, algorithm 5) sampling algorithms applied on the T/E ratio for every single athlete as discussed in Sections 2.3.1.2, 2.3.5.1 and 2.3.5.3. To perform both algorithms we first specify initial values for the model parameters, $\boldsymbol{\theta}^0$. Then, new parameter values, $\boldsymbol{\theta}^*$, are proposed at iteration t by adding some noise, s. Since the parameter vector $\boldsymbol{\theta} = (\mu, \sigma)$ is strictly positive, the proposed values are $\boldsymbol{\theta}^* =\mid \boldsymbol{\theta}^{t-1} + s \mid$, where $s \sim \mathcal{N}(0, 0.05)$. Implementing the remaining algorithmic steps of MH and MWG samplers $T = 6,000$ times with a burn-in of 1,000 iterations, we have finally generated 5,000 realisations from the joint posterior distribution. Figures 3.16 and B.2 contain the histograms, the traceplots and the running averages of the sampled draws for both parameters from the analysis performed on the T/E ratios of a single athlete. The running averages plot have settled around certain values for both parameters giving us an informal impression of convergence. Figure B.1 in Appendix B shows the points after complete sweeps through both parameters confirming that both are samples from the posterior distribution. In both algorithms the acceptance rates for μ and σ were the same, i.e. 66% for μ and 53% for σ. The HPD intervals are also computed and presented in Table 3.7.

Table 3.7: The 95% and 99% HPD intervals using the MH and MWG sampling methods with informative priors for the T/E ratio for athlete 1.

	T/E$_{MH}$		T/E$_{MWG}$	
	L	U	L	U
95% HPD	0.31	1.27	0.34	1.16
99% HPD	0.15	1.39	0.21	1.31

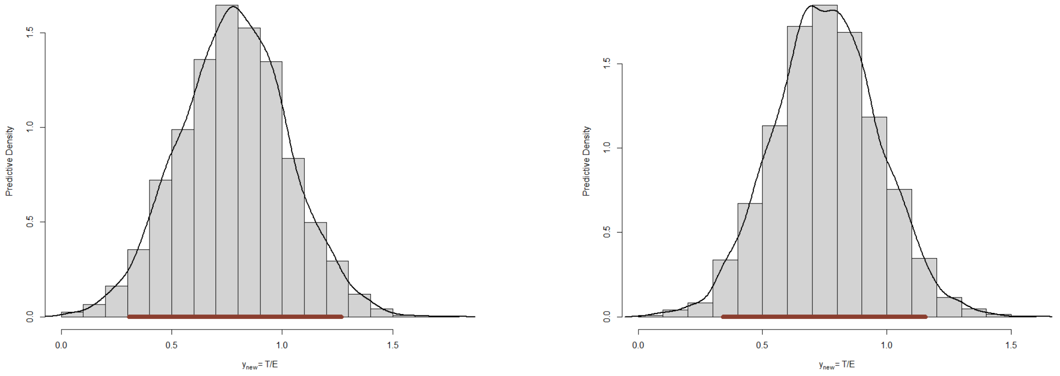


(a)

(b)

Figure 3.16: (a) Histograms, traceplots and (b) running averages of the parameters μ and σ over the iterations using the MH sampling algorithm and Sottas' informative priors (Sottas et al., 2006) for T/E of athlete 1. The posterior means for μ and σ can be approximated by the averages $E[\mu|\boldsymbol{y}] \approx \frac{1}{N}\sum_{i=1}^{N}\mu^{(i)}$ and $E[\sigma|\boldsymbol{y}] \approx \frac{1}{N}\sum_{i=1}^{N}\sigma^{(i)}$, respectively.

Figure 3.17 represents the histogram of predictive densities for y_{new} =T/E for both sampling algorithms, accompanied by the 95% HPD intervals for each case denoted by the dark red lines. Values outside the normal boundaries, which are defined by the 95% HPD intervals, are considered as abnormal observations that should be re-examined.



(a)

(b)

Figure 3.17: Histograms of the predictive densities using a) MH and b) MWG with the 95% HPD intervals symbolised by the dark red lines. The black curves represent the probability densities (kernel densities) of future observations.

Table 3.8 presents the results extracted from both sampling methods; MH and MWG algorithms using informative priors. Estimates of the posterior means for both parameters as well as the 95% HPD intervals for T/E were quite close. We then compare their predictive densities by overlaying their two curves and their 95% HPD intervals in a single plot (see Figure 3.18). The MH and MWG predictive densities do not differ much.

Table 3.8: Posterior estimates and the 95% HPD intervals of T/E obtained from MH and MWG sampling algorithms.

	$E(\mu\|y)$	$E(\sigma\|y)$	L(T/E)	U(T/E)
MH	0.77	0.21	0.31	1.27
MWG	0.77	0.207	0.34	1.16

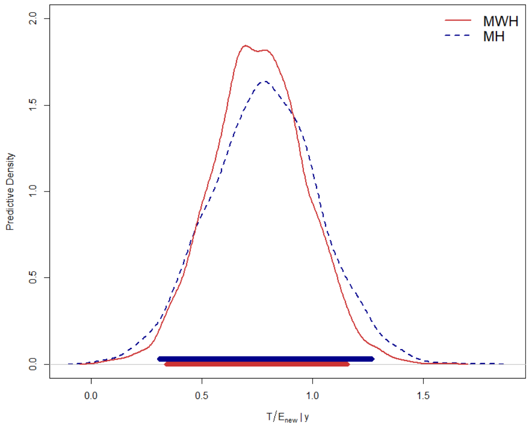


Figure 3.18: Predictive distributions of future observations T/E$_{new}$ and their 95% HPD intervals from each sampling method (MH and MWG).

3.3.4.2 MCMC using the Multivariate Bayesian Model

For the multivariate Gaussian multilevel model, we have estimated multiple parameters via the Gibbs sampler (algorithm 4) of 5,000 iterations. Because of the large number of the model parameters, in the following sections we provide a part of the diagnostics plots produced for the evaluation of the MCMC convergence. All diagnostic plots are available in the supplementary folder titled "Diagnostics" on Github.

3.3.4.3 MCMC Convergence and Efficiency Diagnostics

Traceplots can help to visually assess the mixing of the Markov chains. However, MCMC convergence and efficiency diagnostics were implemented for all sampling methods used for the various models, which all showed good mixing. Table 3.9 presents the results based on the Gelman-Rubin test, the Raftery-Lewis test, the Geweke test, and the Effective Sample Size (ESS) applied on the T/E ratio using the univariate model. The Gelman-Rubin tests return PSRF values below 1.1 for all algorithms using 5 independent chains. All Geweke tests registered a p-value greater than 0.05, suggesting that there was no evidence of a difference between the mean of the first 10% samples and the last 50% samples of the Markov chain. The small estimates of the dependence factors and ESS/T suggest that there is correlation between MCMC draws for MH and MWG sampling methods. However,the dependence factors and ESS/T values are close to 1 for Gibbs sampler. This indicates that there is independence between the MCMC samples and our Markov chain as well as the sampler has converged to the stationary distribution.

Table 3.9: MCMC convergence diagnostics for parameters μ and τ corresponding to the the Gibbs algorithm 2 results: Gelman-Rubin test (PSRF), Raftery-Lewis test (dependence factor, I_{R-L}), Geweke test, and sampler efficiency (ESS normalised by the number of $T = 10,000$ samples) for all parameters.

Sampler	Parameters	$\hat{R}_{PSRF}$	I_{R-L}	P-value$_{Geweke}$	ESS/T
Gibbs [2]	μ	1.035	1.01	0.239	1
	τ	1.097	1.01	0.237	0.93
MH [1]	μ	1	4.03	0.347	0.0053
	σ	1	3.56	0.519	0.0032
MWG [5]	μ	1	3.93	0.34	0.007
	σ	1	3.14	0.45	0.008

The evaluation of the MCMC samples obtained from the multivariate Gaussian multilevel model (MGMM) was based on the “multivariate PSRF” or MPSRF. The MPSRF was 1 which is a strong evidence of convergence. The Gelman plots in

Figures 3.19 and 3.20 of a specific set of model parameters (i.e. the overall means for each biomarker out of the eleven available) verify the good mixing.

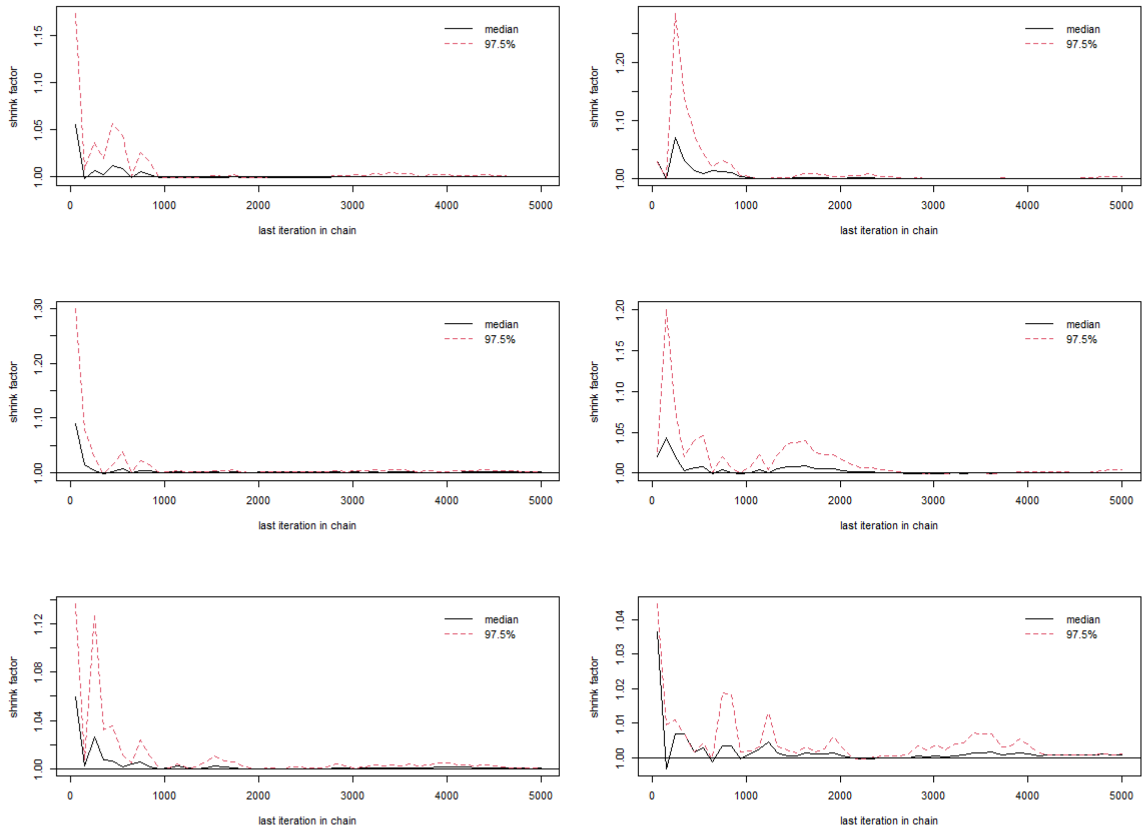


FIGURE 3.19: Gelman plots for the overall means of biomarkers A5, B5, A, ETIO, T, and E.

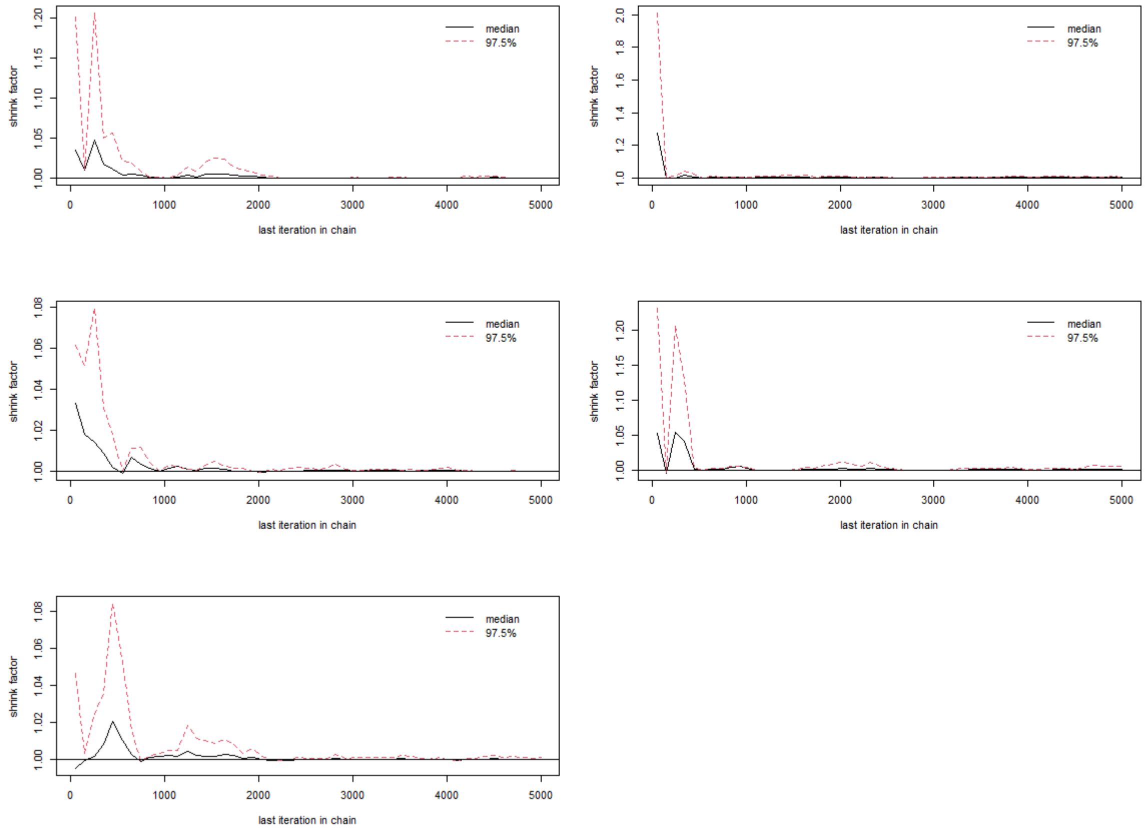


FIGURE 3.20: Gelman plots for the overall means of biomarkers' ratios T/E, A/ETIO, A/T, A5/B5, and A5/E.

3.4 Conclusions

Our major goal in this research work was to develop a methodology which is able to examine the behaviour of a multivariate Bayesian multilevel model by exploiting repeated measurements of several biomarkers and their ratios such as Testosterone, Androsterone, Epitestosterone, the T/E ratio etc. Such methodology can prove helpful as an improved screening test suitable to analyse the selected biomarkers, and to monitor the "level of trust" of an athlete by taking into account: a) the population distribution of these biomarkers, b) the individual's own history; i.e. previous measurements of the athlete's biomarkers and c) other demographic characteristics. Specifically, we aimed to validate measurements of endogenous substances that are able to reveal the presence of toxic substances, drugs of abuse, and/or doping agents. The developed multivariate adaptive model makes a contribution in proposing personalised normal limits of the longitudinal steroid profile of athletes in and out of competition compared to other standard approaches. The tools for applying the proposed model are incorporated in a user-friendly software we have designed for use by anti-doping laboratories.

It is also worth mentioning that the proposed screening method is non-invasive, easily accessible, reliable, quickly quantifiable and reproducible. It also has low financial burden, low risk level, and achieves an improved predictive performance regardless of the imbalance of the data. These characteristics contribute effectively in sports drug testing laboratories for doping detection. Moreover, since the imbalanced classification is not a "solved" problem, we would ideally need more information regarding the class with doping cases. However, to improve the performance of the algorithm one can use other techniques such as undersampling, or a combination of undersampling and oversampling or boosting algorithms that are able to convert weak learners to strong learners (Kotsiantis et al., 2006; Schapire, 2013). In Figure 3.21, a very coarse overview is given which shows a proposal of the applicability of different classification approaches to different situations. For small differences in distribution between majority and minority class, standard classification techniques work well for most of the cases regardless the sample size. For adequate sample sizes and moderate distribution differences simple, one-class classification can work well with respect to the predictive performance. For larger distribution differences, random oversampling in the OCC framework may work better when the sample size is small compare to the random subsampling, which can work better for larger sample sizes.

Lastly, this research work examines the predictive performance of a multivariate adaptive model applied to eight biomarkers and five ratios. Applications on more, but also sensitive, markers and ratios is another direction which is worth investigating. An attempt of this idea is presented in Wilkes et al. (2018). Including the age of athletes as a covariate in the model might enhance the model performance

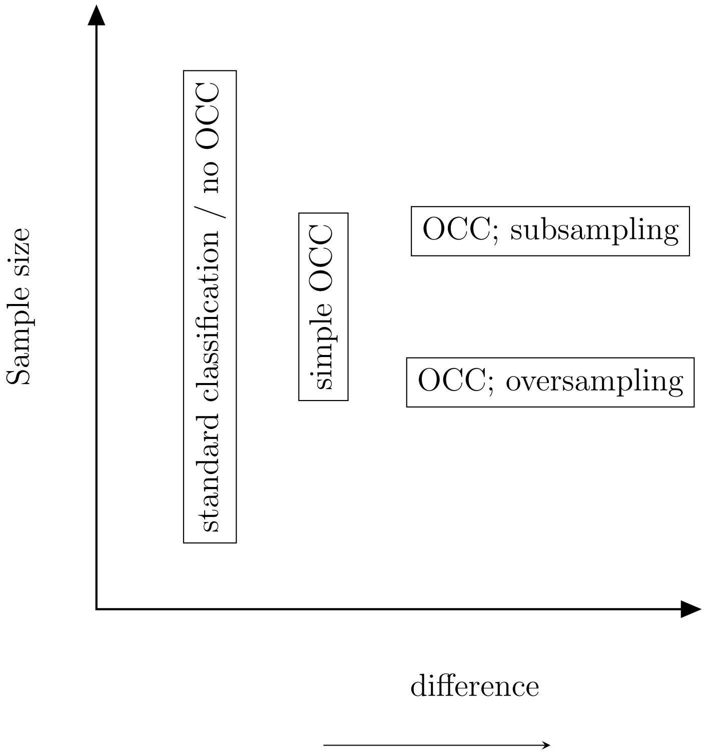


Figure 3.21: Different classification methods for different sample sizes and for low, moderate and high values of difference in distribution between majority and minority class.

since the biosynthesis of endogenous androgenic anabolic steroids (EAAS) varies with age.

Chapter 4

Dirichlet Process Gaussian Mixtures for prostate cancer

"We look to medical research to discover remedial measures to insure better health and more happiness for mankind."

Thomas Hunt Morgan

4.1 Introduction

In cancer biology, advances in high performance technologies have made it possible to study complex multivariate physical and psychological characteristics and their simultaneous associations with high-dimensional biomarker data. This problem can be studied with multi-response regression, where the response variables are potentially highly correlated. Steroid hormones can play a significant role in normal and cancerous prostate physiology, as stated in Section 1.4.3. However, there is little information that steroidal biomarkers clearly delineate their function in the pathogenesis of prostate cancer and benign prostatic hyperplasia. Previous research studies, for example the endogenous androgenic steroids in Amante et al. (2018), various metabolites like sarcosine by Wu et al. (2011), as well as the serum steroid ratio profiles by Albini et al. (2018), have explored male urinary hormones to build a screening tool for early stage prostate cancer and compare its performance with the widely used PSA test.

The steroid metabolites have been previously analysed for cancer detection purposes using both frequentist and Bayesian statistical approaches. However, Bayesian inference is commonly used in cancer biology as a means of updating prior knowledge with new observables in order to construct a better model for our understanding of cancer's behavior. In this work, we are interested in analysing biomarkers and their ratios, which appear to be multimodal data (see Figure B.12). To deal with data multimodality and to provide a useful summary of the data, we can often use mixture models. Therefore, we focused on implementing Dirichlet Process Gaussian Models (DP-GMMs) as a classification tool to model biomarkers for prostate cancer prediction.

The Dirichlet process (DP) is a stochastic process commonly used in Bayesian nonparametrics, particularly in Dirichlet process mixture models (DPMMs) because of its flexibility and computational simplicity. The term *nonparametric* in this context means that the DP model has in principle an infinite number of parameters. Due to the lack of clear knowledge about the data-generating mechanism, we can only make minimal assumptions. Thus, a large part of the mechanism can be left unspecified, meaning that the distribution of the data is not restricted by a predefined finite number of mixture components.

There are various representations of DPs whose formulation differs because they examine the problem from different points of view. The first definition of DPs was provided by Ferguson (1973), who described the Posterior Dirichlet process using the Kolmogorov Consistency Theorem. Blackwell et al. (1973) continued to prove the existence of such a random probability measure based on the Pólya urn scheme, which satisfies the DP properties. In 1982, Sethuraman and Tiwari introduced an additional definition of constructing a DP, which is characterised as the stick-breaking construction. Finally, another representation was provided by Aldous (1983), who introduced the Chinese Restaurant Process (CRP) as an effective way to construct a Dirichlet Process assuming that there is a Chinese restaurant with infinite number of tables. The main idea is that as customers enter the restaurant, they randomly choose any of the occupied tables to sit, or they choose the first available empty table. The CRP is explicitly described in Section 4.3.3.

The hierarchical Bayesian models, which use a DP as a prior distribution over the possible component parameters, are known as Dirichlet process mixtures (Antoniak, 1974; Escobar and West, 1995; Jara, 2017). It has become common to use a countably infinite DPMM for data assumed to come from a finite mixture with an unknown number of components, not only for kernel density estimation and clustering, but also for inference regarding the mixture parameters. For practical reasons, the number of mixture components in this model is usually bounded by a reasonably high upper boundary. Hence, a DPMM can be seen as an infinite-dimensional generalisation of a parametric mixture model. One of its main advantages is that the model selection for defining the optimal number of mixtures is avoided. Another advantage is that the extension to multivariate data is easier and guarantees that the proposed covariance matrices are positive definite compared to the mixture modelling based on the reversible jump Markov chain Monte Carlo (RJMCMC) algorithm due to the difficulty in the split or in the combine moves (Green, 1995; Richardson and Green, 1997; Rasmussen, 2000; Dellaportas and Papageorgiou, 2006; Görür and Rasmussen, 2010).

The DP is a distribution over probability measures which are distributions. The basic DP model has two parameters: the base distribution G_0, which is a random probability measure and serves as a prior mean, and the concentration parameter α, which is a strictly positive real number and serves as a prior inverse variance. A widely used unsupervised machine learning algorithm for Gaussian Mixture Model (GMM) was introduced by Rasmussen in 2000. Görür and Rasmussen (2010) examined both conjugate prior specification and conditionally conjugate prior specification, which were used for the base function of DP, and the paper included a sequence of comparisons between them, showing that the conditionally conjugate case resulted in no or modest amount of complexity loss, and equally good predictive accuracy. Inference for Dirichlet process mixture models is frequently performed using Markov chain Monte Carlo (MCMC) methods. There is a variety of MCMC algorithms developed for conjugate and conditional conjugate base distributions. Conditional methods impute the Dirichlet process and update it as a component of the Gibbs sampler.

This chapter presents the use of DPs to construct infinite Gaussian mixture models (DP-GMMs) with and without covariates for multivariate density estimation problems. Through DP modelling, the inference procedure focuses on identifying the hidden associations of the data points to the infinite number of parameters. The main application of the method has been conducted on two high-dimensional datasets which are composed by urinary biomarkers, as well as their biomarkers' ratios, measured for three classes of individuals; healthy individuals, benign prostatic hyperplasia and prostate cancer patients. Supervised learning techniques are used in order to train the models of each distinct class separately by using training sets from each class. Specifically, the supervised classification, also known as "predictive discriminant analysis" (Huberty, 1984; Huberty, 1994), is used based on DP-GMMs to correctly assign future data to existed and already known classes (Johnson et al., 2002). The DP-GMMs can be used to extract information from the high-dimensional probability distribution and classify the new unlabelled data to the class which corresponds to the most likely model. An empirical Bayes approach is used to evaluate the predictive posterior distributions between the models which describe each class for the data classification (Brown and Greenshtein, 2009; Greenshtein and Park, 2009). However, these models are flexible enough to be used with any countably infinite dimensionality; either univariately or multivariately.

For the DP models, we also assume conditionally conjugate base distribution G_0 by removing the dependency of the distributions between the mean and the covariance parameters. Hence, each mixture parameter can be integrated out given the other.

In the following sections, formulations of the DP models with and without covariates are presented. The hierarchical structure for both models is complete by setting hyperpiors to express uncertainty on the hyperparameters α and G_0 (Rasmussen, 2000).

The computer code for DP-GMM modelling is provided as a completed package in the statistical language R (`R` 4.1.1). In addition, Collapsed Gibbs sampling and Adaptive Rejection Sampling (ARS) are required for drawing realisations from the desired conditional posterior distributions (Geman and Geman, 1984; Gilks and Wild, 1992). It is worth mentioning that the implementation of the sampling algorithms has been carried out with predetermined initial number of Gaussians and fixed priors on the model parameters. After specifying the DP models, we implement the DP models on real prostate cancer data, and inferencing the model parameters. A simulation study is also carried out to assess and compare their clustering performance. Three publicly available datasets have also been used to examine the classification performance of the proposed model; the *iris* data by Fisher (1936), the *crabs* data collected at Fremantle, W. Australia and analysed by Campbell and Mahon (1974), and *breast cancer* data, which can be found here.

4.2 Objectives

The aim of this study is to investigate whether the urinary steroid profile might be an improved clinical tool, suitable for early stage prostate prognosis and predictions. For this purpose, we introduce the Dirichlet Process Gaussian Mixture Models (DP-GMMs) applied on multiple biomarkers and their ratios, capable to capture the heterogeneity of the data and to classify individuals into healthy, benign prostatic hyperplasia or cancer cases. In the multivariate mixture Bayesian framework, as we consider simultaneously multiple responses and several predictors for each mixture component, the analysis of joint associations between multiple correlated response variables becomes challenging. We finally compare the applications of the proposed model on two available datasets of prostate cancer, while including or excluding age as a covariate and using PSA levels as a response variable. We show that convergence and mixing is achieved for the DP-GMMs, while their predictive performance exceeds the one using the test based on the PSA levels. Applications of the methodology on additional datasets have been conducted to enhance the evidence of the good performance of the developed model. All mathematical derivations are presented in detail in Appendix A.

4.3 Materials and Methods

4.3.1 Dirichlet Process Mixtures

The Dirichlet Process is an infinite dimensional generalisation of Dirichlet distributions (Ferguson, 1973; Antoniak, 1974). In practice, it might be unclear how to specify a family of distributions for probability distribution F, which generates the observations $y_1, ..., y_n$. Thus, the Dirichlet Process mixture model is hierarchically defined as

$$\begin{aligned} y_i|\theta_i &\sim F(\theta_i) \\ \theta_i|G &\sim G \\ G &\sim DP(\alpha, G_0), \end{aligned} \tag{4.1}$$

where G is defined as $G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$, $\delta_{\theta_k^*}$ indicates a delta function that takes 1 if $\theta = \theta_k^*$ and 0 elsewhere, and π_k represents the mixing proportions. The argument $\boldsymbol{\theta} = \{\theta_1, ..., \theta_n\}$ are the mixture parameters randomly drawn from G. It basically represents a distribution over finite-dimensional distributions on some probability space Θ. Thus, draws from a DP can be interpreted as random distributions with hyperparameters α and G_0, a positive scalar and a distribution over the same support of G, respectively. The base distribution acts as the prior mean of the DP, where for any finite measurable set $S \subset \Theta$ we have that $E[G(S)] = G_0(S)$. On the other hand, the concentration parameter can be considered as a prior inverse variance (precision), where $V[G(S)] = \frac{G_0(S)(1-G_0(S))}{\alpha+1}$.

Since draws from a Dirichlet Process are discrete distributions with probability one, we can represent this model as a countably infinite mixture of distributions. Using Bayes rules and the conjugacy between Dirichlet and Multinomial distributions we have that $\boldsymbol{\theta} \sim G_0$. Then, the Posterior Dirichlet Process as discussed by Ferguson (1973) is

$$G|\boldsymbol{\theta} \sim DP\biggl(\alpha + n, \frac{\alpha G_0 + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n}\biggr), \tag{4.2}$$

where δ_{θ_i} indicates a point-mass concentrated at θ_i.

4.3.2 Gaussian Mixture Model

The Gaussian mixture model (GMM) is the most widely used modelling for K finite parametric mixtures, but also for nonparametric approaches. In a DP mixture model context, we specify the hierarchical GMM, and then we explore the limit as the number of mixture components K tends to infinity. Throughout the chapter, all vector quantities are denoted by bold-faced characters. Hence, the Gaussian mixture model with K mixture components can be written as

$$p(\boldsymbol{y_1}, ..., \boldsymbol{y_n}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \sum_{j=1}^{K} p(c_i = j|\boldsymbol{\pi}) p_j(\boldsymbol{y_i}|\boldsymbol{\theta_j}, c_i = j) = \prod_{i=1}^{n} \sum_{j=1}^{K} \pi_j \mathcal{N}_d(\boldsymbol{y_i}|\boldsymbol{\mu_j}, S_j^{-1}), \tag{4.3}$$

where $\boldsymbol{\theta_j} = \{\pi_j, \boldsymbol{\mu_j}, S_j\}$ is the *parameter set* for the j^{th} component, $\boldsymbol{\pi}$ denotes the vector with the *mixing proportions* (which are positive and $\sum_{j=1}^{K} \pi_j = 1$), $\boldsymbol{\mu_j}$ is the mean vector for the j^{th} component, and S_j is its *precision* (inverse covariance) matrix. The setting of the allocation variable $c_i = j$ means that the data point $\boldsymbol{y_i}$ was generated from the d-variate Gaussian j. For a DP mixture model construction, we specify jointly a conditionally conjugate prior distribution, G_0, on the component parameters and use allocation variables, c_i, for all observations $i = 1, ..., n$. Therefore, the DP-GMM is a limiting case of the finite GMM model and can be expressed as

$$\begin{aligned} \boldsymbol{y_i}|c_i, \boldsymbol{\theta} &\sim \mathcal{N}_d(\boldsymbol{\mu_{c_i}}, S_{c_i}^{-1}) \\ c_i|\boldsymbol{\pi} &\sim Multinomial(\pi_1, ..., \pi_K) \\ (\boldsymbol{\mu_j}, S_j) &\sim G_0 \\ \boldsymbol{\pi}|\alpha &\sim Dirichlet(\alpha/K, ..., \alpha/K). \end{aligned} \tag{4.4}$$

Analytically, the prior for the *occupation number*, n_j, which represents the number of observations belonging to jth component given the mixing proportions, $\boldsymbol{\pi}$, is Multinomial and the joint distribution of indicators, c_i, is defined by

$$p(c_{1:n}|\boldsymbol{\pi}) = \prod_{j=1}^{K} {\pi_j}^{n_j}, \quad \text{where} \quad n_j = \sum_{i=1}^{n} \delta_{Kronecker}(c_i, j). \tag{4.5}$$

The mixing proportions, $\boldsymbol{\pi}$, are given a symmetric Dirichlet prior, and for mathematical convenience all mixing proportions share the same concentration parameter α/K such that

$$p(\boldsymbol{\pi}|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K} \prod_{j=1}^{K} {\pi_j}^{\alpha/K-1}. \tag{4.6}$$

Using equations (4.5) and (4.6), we can integrate out the mixing proportions, $\boldsymbol{\pi}$, and then obtain the probability of a particular set of assignments, $c_{1:n}$, as follows[1]

$$p(c_{1:n}|\alpha, n) = \int p(c_{1:n}|\boldsymbol{\pi}) p(\boldsymbol{\pi}|\alpha) d\boldsymbol{\pi} = \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)} \prod_{j=1}^{K} \frac{\Gamma(n_j + \alpha/K)}{\Gamma(\alpha/K)}. \tag{4.7}$$

Keeping all the allocation variables fixed except for a single one, the conditional prior for this allocation given the rest is given by[1]

$$p(c_i = j|\boldsymbol{c_{-i}}, \alpha) = \frac{n_{-i,j} + \alpha/K}{n - 1 + \alpha}, \tag{4.8}$$

where the subscript $-i$ denotes all indices except i, and $n_{-i,j}$ is the number of observations, excluding the ith observation, that are associated with component j.

[1] The derivation of equations (4.7) - (4.10) are presented in more detail in Appendix A; A.8.4 - A.8.6.

4.3.3 Chinese Restaurant Process for Infinite Mixture Models

Chinese Restaurant Process (CRP) is a useful analogy for helping to understand the Dirichlet Process introduced by Aldous (1983). Consider a Chinese restaurant with infinite series of tables. In this analogy, every possible cluster corresponds to a “table” in this infinitely large Chinese restaurant. Each observation corresponds to a “customer” entering the restaurant and sitting at a table. According to the notation we used in Section 4.3.2, suppose there are currently $n-1$ customers sitting in the restaurant. Then a new customer i comes in and either randomly chooses to sit at an occupied table (e.g. jth table) with probability proportional to the number of customers they already sit there ${n_{-i,j}}^1$

$$p(c_i = j|\boldsymbol{c_{-i}}, \alpha) = \frac{n_{-i,j}}{n-1+\alpha}, \tag{4.9}$$

or to sit at the first available, currently empty, table (e.g. a new table K + 1) with probability proportional to the concentration parameter α/K^1

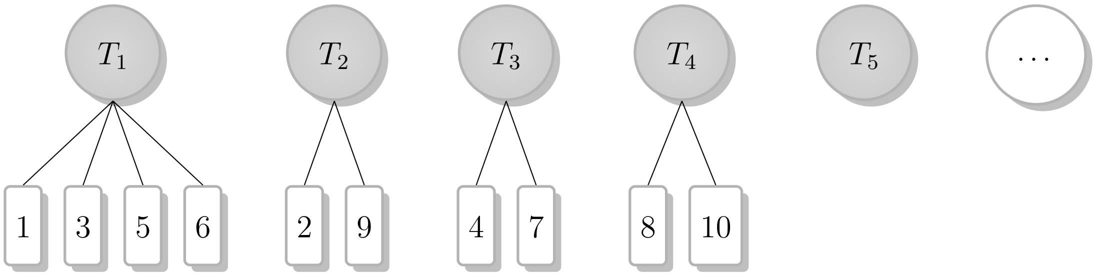
$$p(c_i = K+1|\boldsymbol{c_{-i}}, \alpha) = \frac{\alpha}{n-1+\alpha}. \tag{4.10}$$

Hence, we observe that a greater value of the parameter α tends to encourage the use of a new unoccupied table, especially when the table size (i.e. the number of customers sitting in a table) is small. This observation can alternatively result from the Dirichlet Process in the expression (4.2), where the concentration parameter α specifies how strong the discretisation of clusters is. This means that for large values of α, most data are likely to be distinct and concentrated on the base distribution G_0, but for small values of α, the number of clusters is likely inferred a posteriori from the data. Specifically, tables with many customers become popular and tend to attract more customers. The latent variable, c_i, stores the table number of the ith customer and takes values from 1 to K. K is the total number of currently occupied tables, when $n-1$ customers are in the restaurant, and the constrain that $K < n$ should be also satisfied. Equations (4.9) and (4.10) provide a characterisation of the

CRP derived from the limit of equation (4.8) as $K \to \infty$.

The diagram in Figure 4.1 represents an illustration of the CRP, where the numbered rectangles denote customers. Ten customers enter the restaurant one by one and choose to sit at a table (shown as shaded circles). Note that there are component parameters ($\boldsymbol{\theta_j}$), which are not related in principle to the CRP, but due to they are associated with the jth table (component) belong to the overall mixture modelling. According to the CRP, the first customer of the restaurant is always assigned at the first table. When the second customer arrives, either sits at the first table with probability $\frac{1}{1+\alpha}$, or at the second table with probability $\frac{\alpha}{1+\alpha}$. Similarly for the next customers up to the 10^{th} customer, who has a probability of sitting at the already occupied tables 1 to 4 proportional to the vector $n_{-10,j} = (4, 2, 2, 1)$, respectively.

It is also worth noting that the DP processes are “exchangeable” meaning that the induced distribution does not depend on the order in which the customers arrive or the labels of the tables, but only on the cardinality of the clusters. This results in the customers’ arrivals being independent, which is a useful property for Gibbs sampling in DPs that uses the conditional probabilities in equations (4.9) and (4.10) instead of the joint distribution in equation (4.7).



	T_1	T_2	T_3	T_4	T_5
$p(c_1 = j \mid \boldsymbol{c_{-1}}, \alpha)$	1	0	0	0	0
$p(c_2 = j \mid \boldsymbol{c_{-2}}, \alpha)$	$\frac{1}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	0	0	0
$p(c_3 = j \mid \boldsymbol{c_{-3}}, \alpha)$	$\frac{1}{2+\alpha}$	$\frac{1}{2+\alpha}$	$\frac{\alpha}{2+\alpha}$	0	0
$\vdots$					
$p(c_{10} = j \mid \boldsymbol{c_{-10}}, \alpha)$	$\frac{4}{9+\alpha}$	$\frac{2}{9+\alpha}$	$\frac{2}{9+\alpha}$	$\frac{1}{9+\alpha}$	$\frac{\alpha}{9+\alpha}$

FIGURE 4.1: The Chinese restaurant process where the numbered rectangles represent 10 customers assigned at the attached tables (shaded circles). The illustration of CRP is accompanied by the table of probabilities of each customer (data entry) being placed at either any of the currently occupied or unoccupied (empty) tables.

4.3.4 Conditionally Conjugate DP-GMM

A Dirichlet Process Gaussian Mixture Model (DP-GMM) is equivalent to the infinite limit of the Gaussian mixture model as we described in model (4.4). To entirely delineate this model, the base distribution G_0 needs to be specified. Specifically, G_0 corresponds to the joint distribution of its mixture parameters (the means $\boldsymbol{\mu_j}$s and precision matrices S_js in this case), which for mathematical convenience are usually conjugate. However, we choose to set a conditionally conjugate base distribution, not only because it is free of the property of prior dependency between the mean and covariance, but it has also showed better modelling performance compared to the conjugate base distribution (Görür and Rasmussen, 2010). In this semi-conjugate case, the prior distributions for each parameter are defined independently as

$$G_0(\boldsymbol{\mu_j}, S_j) = \mathcal{N}_d(\boldsymbol{\mu_j}|\boldsymbol{\mu_0}, S_0^{-1})\ \mathcal{W}(S_j|\beta_0, (\beta_0 W_0)^{-1}), \tag{4.11}$$

where the prior of $\boldsymbol{\mu_j}$ is a Gaussian distribution centred at $\boldsymbol{\mu_0}$ with precision matrix S_0, and the prior of S_j is a Wishart distribution with β_0 degrees of freedom and mean W_0^{-1}. Also, the parameter set $\boldsymbol{\phi} = \{\boldsymbol{\mu_0}, S_0, \beta_0, W_0\}$ includes the hyperparameters which are common to all mixture components. This implies that the prior of each model parameter is conjugate to the likelihood conditional on the other. Next, we choose conjugate-style hyperpriors (second level priors) on the hyperparameters of the model to enhance its flexibility and robustness. These are very broad prior distributions in line with the model of Görür and Rasmussen (2010); a Gaussian for mean $\boldsymbol{\mu_0}$ and a Wishart for precision S_0

$$\boldsymbol{\mu_0} \sim \mathcal{N}_d(\boldsymbol{\mu_y}, \Sigma_y), \quad S_0 \sim \mathcal{W}(d, (d\Sigma_y)^{-1}), \tag{4.12}$$

where $\boldsymbol{\mu_y}$ and Σ_y are, respectively, the empirical mean and covariance matrix of the data. For the hyperparameter W_0 a Wishart hyperprior is given, whereas the hyperprior for β_0 is defined indirectly by a Gamma distribution for the parameter $(\beta_0 - d + 1)^{-1}$ with shape and rate parameters $1/2$ and $d/2$, respectively. It is important to note the presence of the restriction $\beta_0 \geq d - 1$. Lastly, a Gamma hyperprior is specified for α^{-1}. Specifically, we specify

$$W_0 \sim \mathcal{W}(d, (d\Sigma_y)^{-1}), \quad (\beta_0 - d + 1)^{-1} \sim \mathcal{G}(1/2, d/2), \quad \alpha^{-1} \sim \mathcal{G}(1/2, 1/2). \tag{4.13}$$

A visual representation of the hierarchical Dirichlet process Gaussian mixture model is depicted in Figure 4.2.

4.3.5 Inference, Gibbs and ARS Sampling

Inference in the hierarchical DP-GMM model is carried out using Collapsed Gibbs sampling, one of the basic forms of MCMC which is based on the conditional conjugacy. The Collapsed Gibbs sampler obtains a sample from the joint posterior

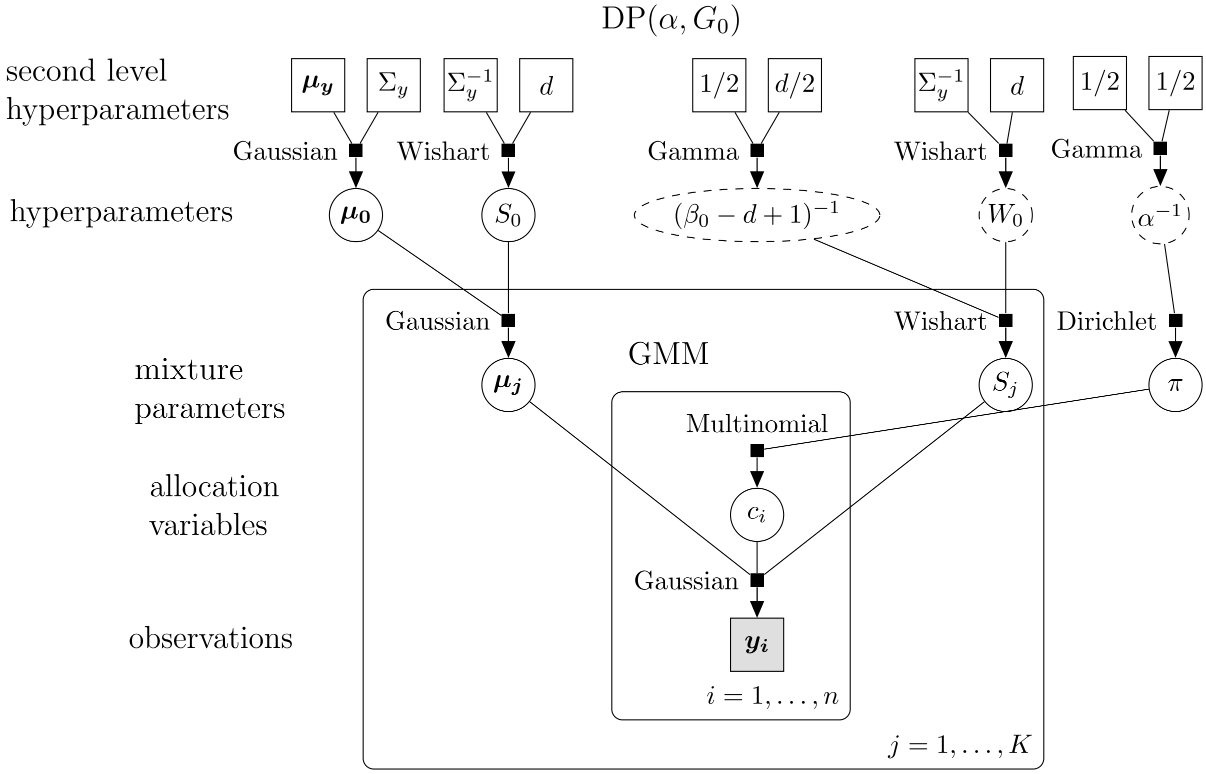


FIGURE 4.2: A graphical representation of the infinite hierarchical DP-GMM model with conditionally conjugate priors, where there is no dependency between the component means and precisions. The dashed nodes denote the indirect prior assignment on some of the model hyperparameters, including the transformed parameters on which a prior has been assigned.

distribution $p(\{\boldsymbol{\mu_j}\}_{j=1}^{K}, \{S_j\}_{j=1}^{K}, \boldsymbol{\mu_0}, S_0, \beta_0, W_0, c_{1:n}, \alpha|\boldsymbol{y_{1:n}})$ by successively and repeatedly simulating from the full conditional distributions of each parameter given the others. The conditional distribution of each parameter has been derived in reference to the likelihood from equation (4.3) and the prior setting in Section 4.3.4[1]. It should be noted that due to the fact that there is no direct prior on the parameter β_0, but for $z = \beta_0 + d - 1$, we obtain the conditional posterior distribution for z and transform its samples to generate realisations from $p(\beta_0|\{S_j\}_{j=1}^{K}, W_0)$.

The DP-GMM model starts with $K^{(0)} = 10$ components and a large number of Gibbs iterations. Due to the fact that not all priors are (conditionally) conjugate, we need to utilise the Adaptive Rejection Sampling method, among the Gibbs sweeps, to generate independent samples from $p(z|\{S_j\}_{j=1}^{K}, W_0)$ and $p(\log(\alpha)|K, n)$ using

[1] The derivation of the conditional posterior distributions are presented in more detail in Appendix A; A.8.1, A.8.2, and A.8.3.

the *log-concavity* property[1] (further details in Appendix A.8.8). Each sampling step is straightforward under conditional conjugacy and Algorithm 6 details a generic summary of the Gibbs sampling process in conjunction with ARS (see Algorithm 7) (Casella and George, 1992; Gilks and Wild, 1992). Sampling from the conditional posterior of a new cluster $K + 1$ under fully conjugate priors is easier because the posterior can be computed analytically. Nevertheless, under partial conjugacy the integral in the posterior cannot be evaluated because it is intractable (see equations (A.1), (A.2)). Consequently, we face this difficulty by implementing numerical integration with Algorithm 8 of Neal (2000), which makes use of auxiliary variables (auxiliary tables in terms of the CRP). We call these auxiliary variables γ and we create them by sampling from the base distribution with means and precision matrices $\boldsymbol{\mu_\gamma}$ and S_γ, respectively, sampled from their priors (equation (4.11)). The γ variables represent the effect of the unoccupied auxiliary clusters, and using the equation (4.10), the prior for each is then defined as

$$\frac{\alpha/\gamma}{n-1+\alpha}. \tag{4.14}$$

The algorithm has been written in the statistical package `R 4.1.1`, which can be easily used for any data dimensionality d. Specifically, when $d = 1$ the model reduces to the univariate hierarchical infinite Gaussian mixture model (Rasmussen, 2000).

[1] A function f(x) is logarithmically concave (or log-concave) if and only if $\log f(\lambda x + (1-\lambda)y) \geq \lambda \log f(x) + (1-\lambda)\log f(y)$ for all $0 \leq \lambda \leq 1$ and for all x, y.

Algorithm 6 Collapsed Gibbs and ARS Sampling (Neal's Algorithm 8)

```
Precondition: Generate an initial state for each of the model parameters/hyper-
    parameters/indicators $\{\boldsymbol{\mu_j^{(0)}}\}_{j=1}^{K^{(0)}}, \{S_j^{(0)}\}_{j=1}^{K^{(0)}}, \boldsymbol{\mu_0}^{(0)}, S_0^{(0)}, \beta_0^{(0)}, W_0^{(0)}, \boldsymbol{c}^{(0)}$ and $\alpha^{(0)}$.
$K^{(1)} \leftarrow K^{(0)}$
for $t \leftarrow 1$ to $T$ do
    for $j \leftarrow 1$ to $K^{(t)}$ do
    draw $\boldsymbol{\mu_j}^{(t)} \sim p(\boldsymbol{\mu_j} | \boldsymbol{c}^{(t-1)}, S_j^{(t-1)}, \boldsymbol{\mu_0}^{(t-1)}, S_0^{(t-1)}, \boldsymbol{y})$
    draw $S_j^{(t)} \sim p(S_j | \boldsymbol{c}^{(t-1)}, \beta_0^{(t-1)}, W_0^{(t-1)}, \boldsymbol{\mu_j}^{(t)}, \boldsymbol{y})$
    end for
draw $\boldsymbol{\mu_0}^{(t)} \sim p(\boldsymbol{\mu_0} | S_0^{(t-1)}, \{\boldsymbol{\mu_j}^{(t)}\}_{j=1}^{K^{(t)}})$
draw $S_0^{(t)} \sim p(S_0 | \boldsymbol{\mu_0}^{(t)}, \{\boldsymbol{\mu_j}^{(t)}\}_{j=1}^{K^{(t)}})$
draw $W_0^{(t)} \sim p(W_0 | \beta_0^{(t-1)}, \{S_j^{(t)}\}_{j=1}^{K^{(t)}})$
draw $z^{(t)} \sim p(z | \{S_j^{(t)}\}_{j=1}^{K}, W_0^{(t)})$
update $\beta_0^{(t)} = z^{(t)} + d - 1$
    for $i \leftarrow 1$ to $n$ do
create $\gamma$ auxiliary variables from the base distribution with means and precision matrices:
  $\boldsymbol{\mu_\gamma} \sim \mathcal{N}_d(\boldsymbol{\mu_0}, S_0^{-1})$ and $S_\gamma \sim \mathcal{W}(\beta_0, (\beta_0 W_0)^{-1})$
        if $n_{-i,j} > 0$ then
        evaluate $n_{-i,j} \times \mathcal{N}_d(\boldsymbol{\mu_j}^{(t)}, S_j^{-1\,(t)})$ ($\propto$ to the likelihood of $j$th cluster)
        update $c_i^{(t)}$
        else    (i.e. $c_i^{(t)}$ is a singleton)
            for $\kappa \leftarrow 1$ to $\gamma$ do
        evaluate $(\alpha^{(t)}/\gamma) \times \mathcal{N}_d(\boldsymbol{\mu_\kappa}, S_\kappa^{-1})$ ($\propto$ to the likelihood of $\kappa$th auxiliary cluster)
        update $c_i^{(t)}$
            end for
        end if
    choose randomly one of the $K^{(t)} + \gamma$ clusters with the previously evaluated probabilities
        if an auxiliary cluster is chosen then
        assign label $K^{(t)} + 1$ to this new cluster
        update $K^{(t)}$
        end if
        if old cluster is empty then
        remove the empty cluster
        update $K^{(t)}$
        end if
    end for
draw $u^{(t)} \sim p(u | K^{(t)}, n)$ (ARS algorithmic steps - see algorithm 7)
update $\alpha^{(t)} = \exp(u^{(t)})$
$K^{(t+1)} \leftarrow K^{(t)}$
end for
```

Algorithm 7 Adaptive Rejection Sampling

```
    Given a univariate non-normalised probability density $p(u)$, perform the following
    initialisation, sampling and update steps:
 1: Initialise an abscissa set $T_k$, such that $p'(u_1) > 0$ and $p'(u_k) < 0$, the corresponding
    envelope and squeezing functions $g_h$ and $g_l$. This can be efficiently achieved
    by starting from an initial guess and stepping out in steps of exponentially
    increasing size.
 2: Draw $u' \sim \frac{g_h(u)}{\int g_h(u')du'}$ and $w \sim \text{Uniform}(0, 1)$
    perform the following squeezing test:
 3: if $w \leq \frac{g_l(u')}{g_h(u')}$ then
 4:     accept $u'$
 5: else
    perform the following rejection test:
 6:     if $w \leq \frac{\log p(u')}{g_h(u')}$ then
 7:         accept $u'$
 8:     else
 9:         reject $u'$
10:     end if
11: end if
12: if $u'$ was accepted at the squeezing test then
13:     go to step 2
14: else
15:     insert $u'$ into $T_k$ to obtain $T_{k+1}$
16:     update the piecewise exponential functions $g_l$ and $g_h$ accordingly
17:     return to step 2
18: end if
```

4.3.6 Covariate Dependent DP-GMM

We now turn our attention to a generalisation of the Dirichlet Process Gaussian Mixture Model (DP-GMM) in which we are flexible to cluster sampling units according to possible patterns of covariates, such as the Dirichlet Process mixtures of Generalised Linear Models (DP-GLM) specified by Hannah et al. (2011). There is a variety of formulations of DPMMs that allow covariate information, however DP-GLM of a broader GLM framework have proved better density estimation and prediction performance compared to other DP mixture regression models.

Dirichlet Process Gaussian Mixture Models with covariates (DP-GMMx). Suppose that $(\boldsymbol{x_i}, \boldsymbol{y_i})$ are the observed regression data pair, where $\boldsymbol{x_i} \in \mathcal{X} \subseteq \mathbb{R}^p$ is a p-dimensional vector of covariates, and $\boldsymbol{y_i}$ denotes the d-dimensional vector of response variables for the ith sampling unit ($\boldsymbol{y_i} \in \mathcal{Y} \subseteq \mathbb{R}^d$). This leads to models of the form

$$\begin{aligned} \boldsymbol{y_i}|\boldsymbol{x_i}, c_i, \boldsymbol{\theta} &\sim \mathcal{N}_d(\boldsymbol{\mu_i}, S_{c_i}^{-1}) \\ \boldsymbol{\mu_{ij}} &= A_j \boldsymbol{x_i^T} = \boldsymbol{r_j}^T \boldsymbol{q_j} \boldsymbol{x_i^T} \\ c_i|\boldsymbol{\pi} &\sim Discrete(\pi_1, ..., \pi_K) \\ (\boldsymbol{q_j}, \boldsymbol{r_j}, S_j) &\sim G_0 \\ \boldsymbol{\pi}|\alpha &\sim Dirichlet(\alpha/K, ..., \alpha/K), \end{aligned} \tag{4.15}$$

where $\boldsymbol{\theta_j} = \{\pi_j, A_j, S_j\}$ is the *parameter set* for the j^{th} component, and $\boldsymbol{\pi}$ denotes the vector with the *mixing proportions* (which are positive and $\sum_{j=1}^{K} \pi_j = 1$). In this GLM framework, each observation i has its mean vector $\boldsymbol{\mu_{ij}}$, which results from the linear relationship between the covariates vector x_i and the regression coefficients matrix A_j. We make the assumption that the regression coefficients consist of two sources of coefficients, $\boldsymbol{q_j} \in \mathbb{R}^p$ and $\boldsymbol{r_j} \in \mathbb{R}^d$, that clearly separate the effect of p covariates from the effect of the d-variate response, respectively. In order to make the parameters $\boldsymbol{q_j}$ and $\boldsymbol{r_j}$ identifiable, we specify Gaussian distributions with their own hyperparameters as priors. The prior distributions for each parameter are defined independently as

$$G_0(\boldsymbol{q_j, r_j}, S_j) = \mathcal{N}_p(\boldsymbol{q_j}|\boldsymbol{\mu_q}, S_q^{-1})\ \mathcal{N}_d(\boldsymbol{r_j}|\boldsymbol{\mu_r}, S_r^{-1})\ \mathcal{W}(S_j|\beta_0, (\beta_0 W_0)^{-1}), \tag{4.16}$$

where the prior of $\boldsymbol{q_j}$ is a Gaussian distribution centred at $\boldsymbol{\mu_q}$ with precision matrix S_q, the prior of $\boldsymbol{r_j}$ is a Gaussian distribution centred at $\boldsymbol{\mu_r}$ with precision matrix S_r, and the prior of S_j is a Wishart distribution with β_0 degrees of freedom and mean W_0^{-1}. Also, the parameter set $\boldsymbol{\phi} = \{\boldsymbol{\mu_q}, \boldsymbol{\mu_r}, S_q, S_r, \beta_0, W_0\}$ includes the hyperparameters which are common to all mixture components. This implies that the prior of each model parameter is conjugate to the likelihood conditional on the other. Next, we choose conjugate-style hyperpriors (second level priors) on the hyperparameters of the model to enhance its flexibility and robustness. These are very broad prior distributions in line with the model of Görür and Rasmussen (2010); Gaussians for means $\boldsymbol{\mu_q}$ and $\boldsymbol{\mu_r}$, and Wishart distributions for precisions S_r and S_q

$$\boldsymbol{\mu_q} \sim \mathcal{N}_p(\mathbf{1}, \Sigma_0), \quad S_q \sim \mathcal{W}(p, (p\Sigma_0)^{-1}), \tag{4.17}$$

$$\boldsymbol{\mu_r} \sim \mathcal{N}_d(\boldsymbol{\mu_y}, \Sigma_y), \quad S_r \sim \mathcal{W}(d, (d\Sigma_y)^{-1}), \tag{4.18}$$

where $\boldsymbol{\mu_y}$ and Σ_y are, respectively, the empirical mean and covariance matrix of the data, and Σ_0 is initially set equal to $\mathbb{I}_p$. For the hyperparameter W_0 a Wishart hyperprior is given, whereas the hyperprior for β_0 is defined indirectly by a Gamma distribution for the parameter $(\beta_0 - d + 1)^{-1}$ with shape and rate parameters $1/2$ and $d/2$, respectively. It is important to note the presence of the restriction $\beta_0 \geq d - 1$. Lastly, a Gamma hyperprior is specified for α^{-1}. Specifically, we specify

$$W_0 \sim \mathcal{W}(d, (d\Sigma_y)^{-1}), \quad (\beta_0 - d + 1)^{-1} \sim \mathcal{G}(1/2, d/2), \quad \alpha^{-1} \sim \mathcal{G}(1/2, 1/2). \tag{4.19}$$

S_j is the *precision* (inverse covariance) matrix of the jth cluster. A visual representation of the hierarchical Dirichlet process Gaussian mixture model is depicted in Figure 4.3. A similar algorithmic procedure to Algorithm 6 of the covariate-free DP-GMM is followed for inference[1].

[1] The derivation of the conditional posterior distributions are presented in more detail in Appendix A; A.9.1 and A.9.2.

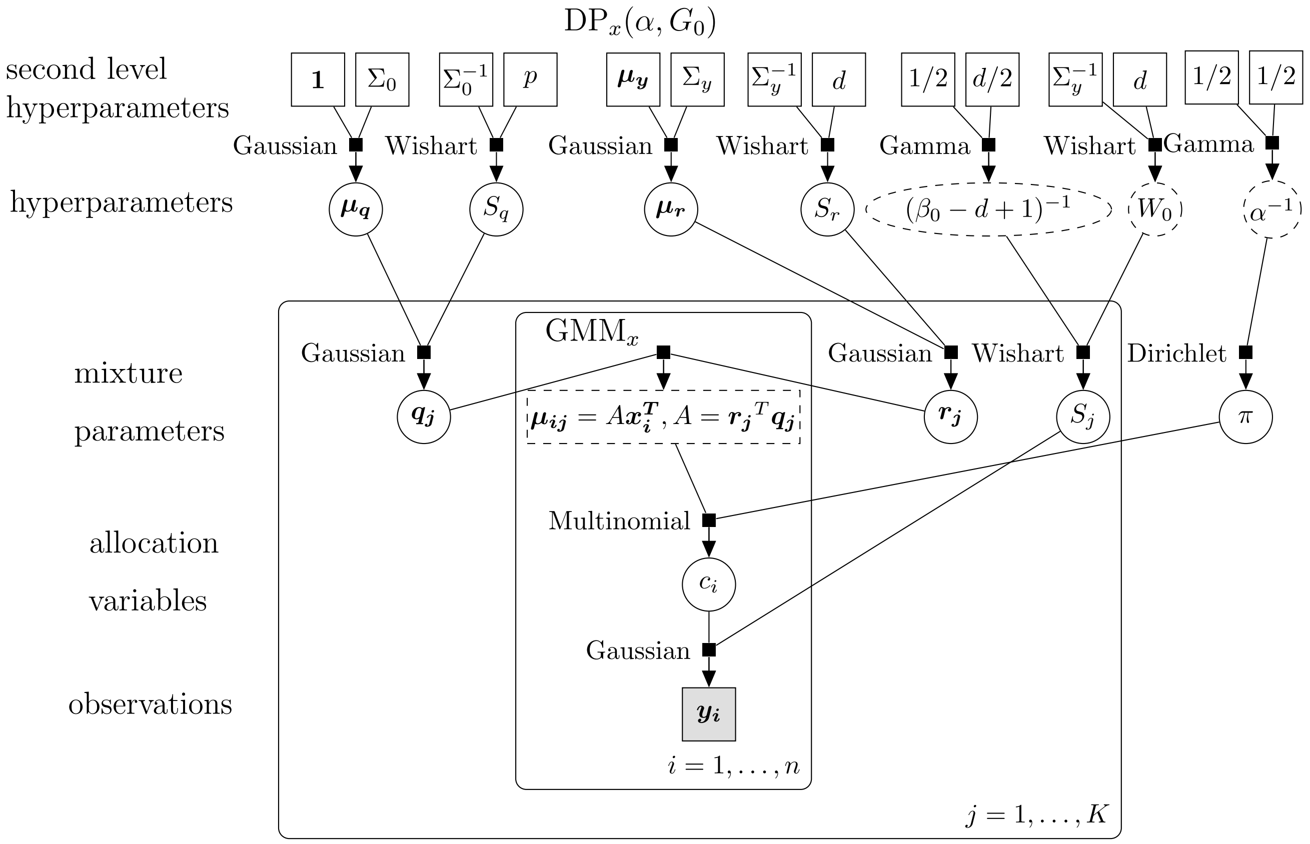


FIGURE 4.3: A graphical representation of the infinite hierarchical DP-GMMx model with covariates and conditionally conjugate priors, where there is no dependency between the component means and precisions. The dashed nodes denote the indirect prior assignment on some of the model hyperparameters, including the transformed parameters on which a prior has been assigned.

4.3.7 Predictive Distribution

In the covariate-free case the predictive distribution of a new data vector $\boldsymbol{\tilde{y}}$ given the training data $\boldsymbol{y} = (\boldsymbol{y_1}, \ldots, \boldsymbol{y_n})$ is obtained by marginalising over the parameters sampled by the Markov chain with their posterior distributions as

$$p(\boldsymbol{\tilde{y}}|\boldsymbol{y}) = \int p(\boldsymbol{\tilde{y}}|\boldsymbol{\mu}, S)p(\boldsymbol{\mu}, S|\boldsymbol{\mu_0}, S_0, \beta_0, W_0)d\boldsymbol{\mu}\, dS. \tag{4.20}$$

Mixture modelling leads to a mixture representation of the posterior predictive distribution consisting of two segments, the finite mixture of Gaussians for which observations have been assigned (occupied clusters), and the infinite mixture of

Gaussians for which there were no observations assigned (unoccupied/empty clusters) with weights $n/(n+\alpha)$ and $\alpha/(n+\alpha)$ for each segment, respectively. The integral in equation (4.20) cannot be evaluated analytically because the use of conditional conjugate priors generates a posterior of a non tractable form. Consequently, we focus on approximately calculating the predictive distribution using the parameters drawn from the T Gibbs iterations as follows

$$p(\tilde{\boldsymbol{y}}|\boldsymbol{y}) \approx \frac{1}{T}\sum_{t=1}^{T}\left[\frac{\alpha^{(t)}}{n+\alpha^{(t)}}\mathcal{N}_d(\tilde{\boldsymbol{y}}|\boldsymbol{\mu_0^{(t)}},\, S_0^{-1\,(t)}) + \sum_{j=1}^{K^{(t)}}\frac{n_j}{n+\alpha^{(t)}}\mathcal{N}_d(\tilde{\boldsymbol{y}}|\boldsymbol{\mu_j^{(t)}},\, S_j^{-1\,(t)})\right]. \tag{4.21}$$

Notice that the posterior predictive distribution is a combination of the infinite and finite mixtures of Gaussians, whose parameters are based on the prior and the data, respectively. Similarly, we compute the predictive distribution for DP-GMMx[1].

[1] See Appendix A; A.9.3.

Chapter 5

Dirichlet Process Gaussian Mixture Models: applications on prostate cancer prediction

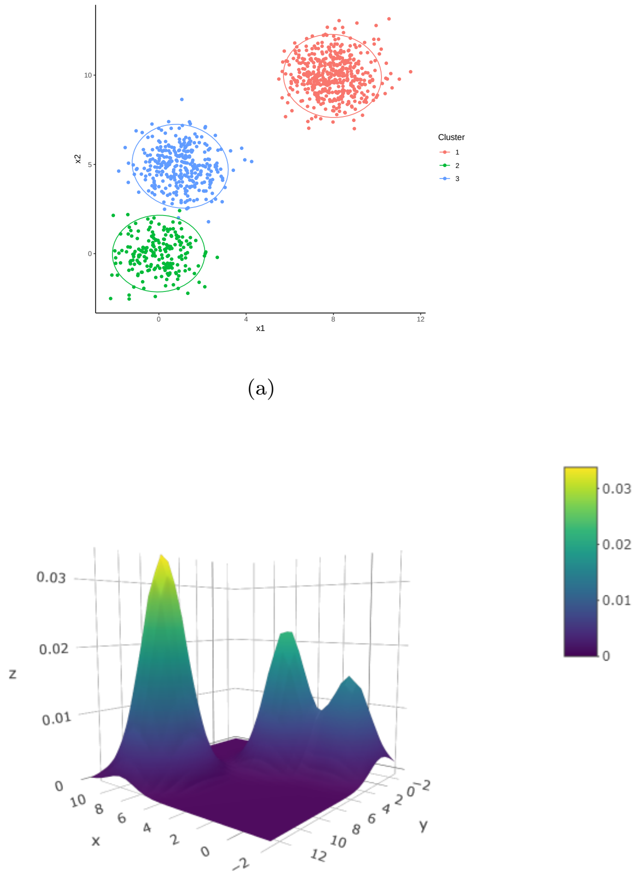
5.1 Experiments

In this section the covariate-free DP-GMM model has been employed first on simulated data to compare its clustering performance with other two clustering methods. We then use the *iris*, *crabs* and breast cancer data to evaluate the classification performance of the developed methodology.

5.1.1 Clustering Performance

Initially, a simulation study is carried out to assess the performance of DP-GMM modelling in terms of density estimation. A set of 1000 simulated bivariate data points are generated from a mixture of three Gaussians and the weights of components are $\pi_1 = 0.2$, $\pi_2 = 0.3$ and $\pi_3 = 0.5$ (see Figure 5.1). The actual sample sizes in each cluster were $n_1 = 200$, $n_2 = 320$, and $n_3 = 480$. A large number of Gibbs sweeps for DP-GMM clustering on the simulated data is initially performed, sampling the parameters and hyperparameters of the model in turn from their conditional distributions derived in previous sections. To the left of Figure 5.2, the autocorrelation for several quantities is plotted, which is dropped close to zero after the 30th iteration time of lag 1. Then, we performed 3000 iterations of which the first 100 draws were discarded as a burn-in period being based on the first Gibbs implementation in which the algorithm converges after the 100th iteration (Figure B.8). The remaining 2,900 draws are used to generate 100 independent samples from the posterior (equally spaced).

The results of the DP-GMM clustering are summarised in Table 5.1 compared to the results of other two techniques of clustering using the *Dirichletprocess* and *mclust* packages (Ross and Markwick, 2018; Scrucca et al., 2016). The main difference between the two DP methods is that in the DP-GMM model we use conditionally conjugate priors, while the *Dirichletprocess* package is based on conjugate priors. All techniques indicate three clusters with similar data allocations. On the other hand, the Adjusted Rand Index, a common metric to evaluate clustering methods, shows that all methods have similar grouping with the known true grouping (Gates and Ahn, 2017). In terms of clustering, DP mixture modelling with conditionally conjugate priors and a smaller sample proved its superiority over the corresponding model with conjugate priors and the mclust method. The DP-GMM performs well in estimating the true parameters of the model of three Gaussian mixtures with no use of any prior information. For this particular sample, the median of the concentration parameter α over the MCMC iterations is approximately 0.45, where the three occupied clusters account for $\frac{n}{n+\text{median}(\alpha)} \simeq 99.96$ % of the mass, and only the 0.04% of the mass of the predictive distribution belongs to unoccupied clusters. Figure 5.2 (b) represents the clustering from the samples of DP-GMM on the last iteration, which agrees with the clustering of mclust and Dirichletprocess (Figures B.9, B.10, B.11).



(a)

(b)

FIGURE 5.1: (a) Scatterplot of the 1000 simulated 2D data points drawn from $0.2\mathcal{N}_2\big(\mu_1 = (0,0), \Sigma_1 = \left[\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right]\big) + 0.3\mathcal{N}_2\big(\mu_2 = (1,5), \Sigma_2 = \left[\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right]\big) + 0.5\mathcal{N}_2\big(\mu_3 = (8,10), \Sigma_3 = \left[\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right]\big)$ with the 95th percentile ellipsoids for each component distribution. (b) Density estimate of the mixture of three Gaussians.

TABLE 5.1: Summary table of the clustering results from DP-GMM model in comparison to Dirichletprocess (applied on the scaled simulated data) and mclust packages.

Package	iterations	Clusters	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	ARI[e]
DP-GMM[a]	100[d]	3	0.199	0.321	0.480	0.9934
Dirichletprocess[b]	1000	3	0.201	0.319	0.480	0.9846
mclust[c]	1000	3	0.201	0.319	0.480	0.9823

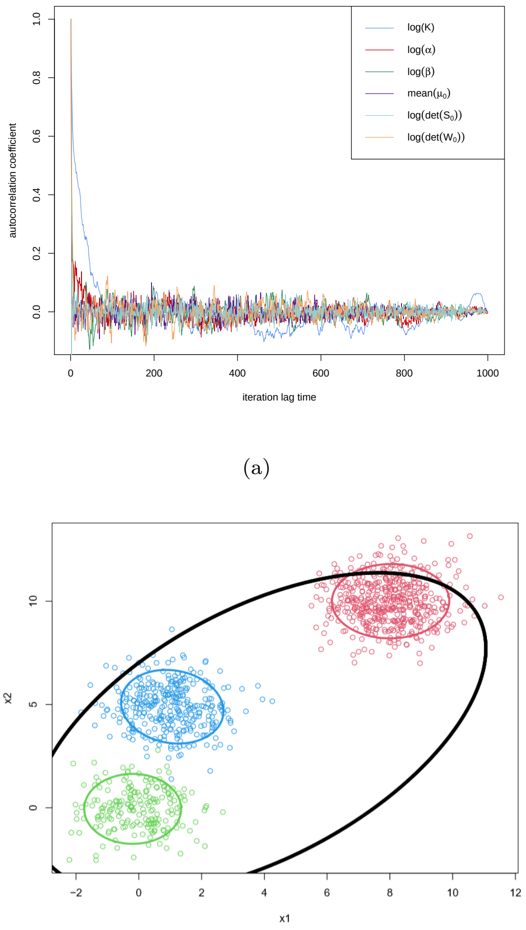
[a] The DP-GMM package uses conditionally conjugate priors.

[b] The Dirichletprocess package is based on conjugate priors.

[c] The mclust package is based on Gaussian finite mixture modelling fitted by EM algorithm.

[d] 100 independent samples have been generated after thinning.

[e] Adjusted Rand Index: $ARI = \frac{RI - E(RI)}{\max(RI) - E(RI)}$, where the raw Rand Index is given by $RI = \frac{\text{count of pairs in agreement}}{\text{total number of possible pairs}}$.



(a)

(b)

Figure 5.2: (a) Autocorrelation length for various parameters in the Markov chain, based on 10^3 iteration. Only the number of occupied clusters, K, shows a small correlation; the effective sample size is approximately 30. (b) The samples from the last (3000th) iteration of Algorithm 2 separates the data points into three distinct clusters. Coloured ellipses are centered at the posterior mean $\boldsymbol{\mu_j}$ and covariance matrix S_j^{-1} for each cluster j. The big black ellipse is the one created by the prior hyperparameters $\boldsymbol{\mu_0}$ and S_0^{-1}.

5.1.2 Classification Performance

To evaluate the classification performance of the DP-GMM, we used three publicly available datasets to classify different species of flowers and crabs, and the two classes of breast cancer (benign and malignant). First, we modelled Fisher's *iris* data as a four-dimensional example to classify 150 flowers into the three species of iris (Setosa, Versicolor and Virginica) (Fisher, 1936). The data consist of measurements in centimetres on four variables (sepal length and petal width, petal length and petal width) and 50 flowers are recorded from each species. Secondly, we fitted *crabs* data as a mixture of six-variate Gaussian distributions. The data frame contains 50 crabs from each species (Blue and Orange) and both sex, along with their five morphological measurements (Campbell and Mahon, 1974). Finally, we modelled 32 features from the *breast* data (BCa) applying the DP-GMM to distinguish between benign and malignant classes. More information about this dataset can be found here.

We start by splitting each dataset into a training set and a test set, 70%/30%, respectively. We train the conditionally conjugate DP-GMM without covariates to approximate the predictive density of each class, estimate the mixing proportions and the parameters of the model. We run 3000 Gibbs sampling iterations, where the first 1000 are removed, and then we evaluate the predictive posterior distributions for each distinct class on the test set. For example, for breast cancer classification we calculate $p(\boldsymbol{y_{test}}|M_{\text{Benign}})$ and $p(\boldsymbol{y_{test}}|M_{\text{Malignant}})$ to categorise "unlabelled" data to the most likely class; i.e. benign or malignant. Convergence and mixing for all samplers is achieved fast. Table 5.2 presents the good predictive performance of the DP-GMM classification model on each dataset. Overall accuracy for iris and crabs data classification are 0.98, 95% CI (0.88,1) and 1, 95% CI (0.93,1), respectively. Benign and malignant classes of the breast cancer dataset seemed to be slightly difficult to separate, because they might have similar densities. To achieve a better classification accuracy, we might investigate which of the 32 variables we have available are the most powerful and useful to model. For comparison reasons, we also present the classification results obtained from SVM and LDA classification methods in Table 5.2 (Cortes and Vapnik, 1995; McLachlan, 2005). Compared to the classification performance of SVM and LDA, the Dirichlet-Process Gaussian mixture model (DP-GMM) was equally accurate, achieving very high values of overall accuracy, however for breast cancer data, SVM and LDA methods achieved higher values; 0.96, 95% CI (0.92,0.98) and 0.94, 95% CI (0.89,97), respectively.

Table 5.2: Predictive performance of the classification methods DP-GMM, SVM and LDA on the iris, crabs and breast cancer datasets. "Blue" and "Malignant" were set as the positive classes in the classification process. "Dim" denotes the dimensionality of the Gaussian components.

					Balanced	Overall
Dataset	Dim	Class	Sensitivity	Specificity	Accuracy	Accuracy (95% CI)
DP-GMM						
Iris	4	Setosa	1	1	1	0.98 (0.88, 1)
		Versicolor	0.93	1	0.97	
		Virginica	1	0.97	0.98	
Crabs	6	Orange-Blue	1	1	1	1 (0.93, 1)
BCa	32	Benign-Malignant	0.57	0.93	0.75	0.72 (0.62,0.78)
SVM						
Iris	4	Setosa	1	1	1	0.93 (0.82, 0.99)
		Versicolor	0.80	1	0.90	
		Virginica	1	0.90	0.95	
Crabs	6	Orange-Blue	1	1	1	1 (0.94, 1)
BCa	32	Benign-Malignant	0.92	0.98	0.95	0.96 (0.92,0.98)
LDA						
Iris	4	Setosa	1	1	1	0.96 (0.85, 1)
		Versicolor	0.93	0.97	0.95	
		Virginica	0.93	0.97	0.95	
Crabs	6	Orange-Blue	1	1	1	1 (0.94, 1)
BCa	32	Benign-Malignant	0.86	0.98	0.92	0.94 (0.89,0.97)

5.2 Applications on Prostate Cancer Data

5.2.1 Data Description

In this section, we present the results of the proposed methodology on real prostate cancer datasets of different dimensionality for three classes of individuals; healthy (H), benign prostatic hyperplasia (BPH) and prostate cancer (PCa) patients. The first study (dataset 1) included 96 healthy males, 127 patients with BPH and 83 patients with confirmed PCa. The second study (dataset 2) included 212 healthy males, 144 patients with BPH and 140 patients with confirmed PCa. The range of age of all individuals in the dataset 2 was from 16 to 83 years. The mean age in each class was 54 for healthy men and 67 for both BPH and PCa patients. The mean serum PSA level was 4.11 ng/mL for BPH patients and 8.40 ng/mL for PCa patients. PSA values were not available for healthy men. The ultimate goal of the DP-GMM and DP-GMMx models is to find appropriate distributions to delineate the underlying distributions of the training data, and then predict the categorical class labels of new coming data.

5.2.2 Results

For both datasets, we initially split them randomly into 70% train and 30% test, while preserving relative ratios of different labels in variable "Class". Then, the conditionally conjugate DP-GMMs with and without covariates are applied to model the predictive density of each class using its training set, and estimate its mixing proportions and model parameters. Convergence and mixing of all samplers is fast for both datasets. The first 1000 draws out of 3000 iterations of the sampling process are discarded as a burn-in period. Based on the parameters drawn from the remaining Gibbs sweeps, we use Bayes' factors evaluating the predictive posterior distributions between the three models $p(\boldsymbol{y_{test}}|M_H)$, $p(\boldsymbol{y_{test}}|M_{BPH})$ and $p(\boldsymbol{y_{test}}|M_{CAP})$ in order to classify the "unlabelled" data points to the most likely model class.

The challenging part here is to achieve good classification for all classes. It is very important to detect the BPH and PCa cases, but also crucial to correctly identify healthy individuals as healthy. The DP-GMMx model applied on the dataset 2 with age as a covariate achieves the best overall predictive performance for the three groups, showing superiority compared to the other models we applied (see Table 5.3). The covariates seem to add important information into the model and make it a more powerful tool for classification.

We have also attempted to investigate whether using a subset of individuals over 60 years old could be more representative for detecting BPH and PCa patients, taking into account that PCa mostly occurs in individuals older than 60 years. However, the results did not show any improvement compared to the model with the full set included. The densities of BPH and PCa patients seem to be similar, which creates an extra level of difficulty in discriminating between these two classes (see Figure B.12).

Lastly, we applied the model on the subset with benign prostatic hyperplasia and prostate cancer patients, adding the serum prostate specific antigen as a second covariate. We also present the classification performance using two standard classification techniques, SVM and LDA, applied on the same training and test sets in Table 5.4. It is obvious that classification via DP-GMM was superior to the classification from SVM and LDA.

TABLE 5.3: Predictive performance of various DP-GMM models applied on markers only from both available datasets. In all binary classifications "PCa" was set as the positive class.

					Balanced	Overall
Dataset*	Covariates	Class	Sensitivity	Specificity	Accuracy	Accuracy (95% CI)
1		H	1	0.94	0.97	
		BPH	0.78	0.87	0.82	0.81 (0.70, 0.89)
		PCa	0.62	0.89	0.76	
2		H	0.81	0.79	0.80	
		BPH	0.72	0.83	0.76	0.69 (0.60, 0.77)
		PCa	0.46	0.90	0.69	
2	Age	H	0.81	0.96	0.88	
		BPH	0.86	0.84	0.85	0.79 (0.71, 0.86)
		PCa	0.68	0.90	0.79	
2	Age$\geq$ 60	H	0.79	0.85	0.82	
		BPH	0.74	0.84	0.79	0.73 (0.63, 0.81)
		PCa	0.64	0.90	0.77	
2**	Age	BPH-PCa	0.84	0.75	0.80	0.80 (0.68, 0.89)
2**	Age$\geq$ 60	BPH-PCa	0.85	0.76	0.80	0.80 (0.67, 0.90)

* Table C.6 represents the variables included in each dataset.

** Including PSA levels as an additional response variable.

Table 5.4: Predictive performance of SVM and LDA classification methods applied on markers only from both available datasets. In all binary classifications "PCa" was set as the positive class.

Dataset* (Method)	Additional Response Variable	Class	Sensitivity	Specificity	Balanced Accuracy	Overall Accuracy (95% CI)
1		H	0.83	0.94	0.88	
(SVM)		BPH	0.79	0.74	0.77	0.69 (0.58, 0.78)
		PCa	0.36	0.84	0.60	
1		H	0.83	0.94	0.88	
(LDA)		BPH	0.71	0.74	0.73	0.66 (0.56, 0.76)
		PCa	0.40	0.81	0.60	
2		H	0.91	0.62	0.76	
(SVM)		BPH	0.47	0.84	0.66	0.59 (0.49, 0.69)
		PCa	0.41	0.92	0.67	
2		H	0.69	0.70	0.69	
(LDA)		BPH	0.62	0.72	0.67	0.54 (0.44, 0.64)
		PCa	0.31	0.89	0.60	
2	Age	H	0.92	0.54	0.73	
(SVM)		BPH	0.61	0.87	0.74	0.63 (0.55, 0.71)
		PCa	0.20	0.98	0.59	
2	Age	H	0.89	0.48	0.69	
(LDA)		BPH	0.47	0.87	0.67	0.59 (0.50, 0.67)
		PCa	0.23	0.96	0.59	
2	Age$\geq$ 60	H	0.78	0.71	0.75	
(SVM)		BPH	0.53	0.83	0.68	0.57 (0.47, 0.67)
		PCa	0.41	0.82	0.61	
2	Age$\geq$ 60	H	0.59	0.71	0.65	
(LDA)		BPH	0.53	0.75	0.64	0.48 (0.38, 0.58)
		PCa	0.31	0.76	0.54	
2**	Age	BPH-PCa	0.68	0.76	0.72	0.72 (0.61, 0.82)
(SVM)	PSA					
2**	Age	BPH-PCa	0.55	0.82	0.68	0.68 (0.57, 0.79)
(LDA)	PSA					
2**	Age$\geq$ 60	BPH-PCa	0.52	0.90	0.84	0.71 (0.57, 0.82)
(SVM)	PSA					
2**	Age$\geq$ 60	BPH-PCa	0.45	0.93	0.69	0.69 (0.56, 0.80)
(LDA)	PSA					

* Table C.6 represents the variables included in each dataset.

** Including PSA levels as an additional response variable.

5.3 Conclusions

The infinite Dirichlet process model for a mixture of Gaussians in a Bayesian framework has been presented and extended into a more flexible model which uses covariate information. It has been shown that DP-GMMs with and without covariates achieve remarkable clustering and classification performances on multivariate datasets compared to the widely used PSA diagnostic test and other multivariate statistical approaches, such as support vector machines (SVM) and linear discriminant analysis (LDA) or approaches used by Wolf et al. (2010) and Amante et al. (2019). Gaussian mixture models of adequate mixture components work also well when applied on the raw data due to their power to approximate any smooth density with any specific non-zero amount of error (Goodfellow et al., 2016). The implementation of the DP-GMM model on biomarkers only, while accounting for the age as a covariate, and the PSA levels as an additional response variable, showed increased prediction accuracy. It is also worth noting that the model with covariates proved computationally faster, possibly because this additional information results in a more accurate density estimation with reasonable number of components, which makes the MCMC algorithm converge faster.

Chapter 6

Discussion

6.1 On Doping Detection

6.1.1 Conclusions

Doping is a global problem in both vulnerable athletic and non-athletic populations. This research work is an attempt to improve tools to stop the spread of this problem enhancing the decisions by International sports federations, while sophisticated models for doping detection have been developed. In the first part of this thesis, we extended the standard univariate model to a multivariate model for identifying anomalous values within the steroidal profile of athletes. Non-longitudinal and longitudinal data are used to train the proposed multivariate Bayesian model for the majority class of non-doped athletes. We then investigated the performance of the one-class classifier applied on a variety of univariate and multivariate models. We focused on answering the following questions:

- How prior information about the population of the majority class (or the"target" class) is specified in the model?
- How the classifier is constructed, trained and evaluated based on data from the majority class?

- What is the decision rule for classifying a sample obtained from an athlete as belonging to the majority class?
- What is the measure to minimise the probability of accepting abnormal (or outlier) samples as normal?
- How the model deals with the population- and individual-based information in order to pass from population biomarker thresholds to individual level thresholds?
- What metrics are suitable for the evaluation of the one-class classifier when limited or no outlier samples are available?
- Which model is proposed for doping control analysis?

First, we conclude that ratios appear to be more sensitive in detecting abnormal samples within the athlete's steroid profiles. Specifically, the best one-class classification performance among the univariate models achieved when using the ratios T/E (with weak informative priors and with strong informative priors from Sottas et al. (2006) modelling), A/ETIO and A5/E.

Secondly, among the T/E models, the model of Sottas et al. (2006), which is based on strongly informative priors, showed slightly better prediction compared to the T/E model where semi-informative conjugate priors were used.

Applying the multivariate Gaussian models based on all biomarkers and ratios, only biomarkers and only ratios with and without oversampling, the values of the metrics are overall higher compared to the corresponding ones from the univariate models. We also conclude that among all the applications of the MGMMs without oversampling, the five available ratios (T/E, A/T, A/Etio, A5/B5 and A5/E) were the most powerful set of variables for detecting abnormal concentration samples. However, the best overall classification performance after applying oversampling is achieved by the multivariate model where all biomarkers and ratios are included.

Furthermore, we compared the classification performance of these models with the results obtained from the applications of generalised linear mixed-effects models. Their predictive performance was not satisfactory, which might be a result of the large imbalance in the normal and abnormal distributions.

The classification and evidence evaluation procedures are implemented into an online web application, which outputs the corresponding results for a given set of measurements from an athlete's steroid profile. The BioScan App contributes a quick and easy-to-use tool for athletes' evaluation and better decision making by forensic scientists.

6.1.2 Limitations and Future Work

1. This research work used non-informative priors for all parameters of the Bayesian hierarchical model. Nevertheless, the use of more informative priors may provide a more accurate predictive distribution. For example, prior knowledge on the model parameters related to each biomarker can be very useful, as well as information on the correlation between the markers can be also included in the prior setting.

2. Another important challenge that has been raised is that of multiple statistical tests, which increases the probability of false positives. Therefore, multiple testing corrections, such as Bonferroni correction, should be also considered in order to adjust p-values derived from the tests.

3. Potential weaknesses of the one-class classifiers that we used might be hidden. The model performance could be further improved by implementing several classifiers and decision rules, or by combining them.

4. Single-component Gaussian distributions might be too restrictive to model biological markers in our attempt to capture patterns between them. More flexible modelling, such as Gaussian mixtures, could be useful in fitting high-dimensional data and, therefore, in improving the classification accuracy.

5. In summary, the classification performance of the proposed model depends not only on the scaling of the variables, preprocessing, the sample size in comparison with the dimensionality, but also on time and computation constraints, and what target rejection rate is acceptable. Furthermore, to avoid learning the noise, when the model suggests an outlier, then this observation is automatically excluded from the set of recordings that are used to compute the HPD intervals. On the other hand, leaving out samples which are assumed to be abnormal based on the 95% HPD interval criterion generates the issue of discarding 5 out of 100 false positive samples truncating the tails of the distribution over time. Therefore, all such constraints should be taken into account before selecting the best model, which is able to distinguish between normal and abnormal concentration values of EAAS of athletes.

6.2 On Prostate Cancer Prediction

6.2.1 Conclusions

Prostate cancer (PCa) is one of the most frequent cancer diagnosis made and leading cause of death worldwide after lung cancer (Litwin and Tan, 2017). The serum prostate specific antigen (PSA) is the most commonly used screening programme for prostate disease. However, its low sensitivity of detecting prostatic malignancies is a matter of concern. In this research work we focused on studying multiple urinary biomarkers and ratios to improve prognosis of prostate cancer. We are interested in identifying patterns between several metabolite biomarkers such as testosterone and androsterone in order to understand their correlation with the diagnosis of the conditions of prostate cancer and benign prostatic hyperplasia (BPH).

A machine learning model has been developed and trained by data coming from three condition groups; patients with PCa, patients with BPH, and healthy men. We demonstrated the performance of the model in terms of clustering and classification on both, simulated and real datasets. Various applications of Dirichlet Process

Gaussian mixture models with and without covariates (DP-GMM and DP-GMMx) were conducted under a Bayesian nonparametric framework in order to classify new unlabelled data points. We finally presented the classification results from the applications of the DP-GMMs including the steroids with highest significance (e.g. testosterone, epitestosterone and etiocholanolone) and the PSA levels. All models yielded high sensitivity and specificity scores, resulting significantly higher than PSA itself. Specifically, the best performance was given by the models on the markers including the PSA levels as a $(d+1)$th response variable with sensitivity and specificity of 84% and 75% (for all men), and 85% and 76% (for men ≥ 60 years old).

The DP-GMM classification model contributes to the following domains:

- The model proved a powerful tool in terms of density estimation, clustering and classification performance for both, applications on real and artificial data.
- The extension of the DP-GMM model controlling for any potential covariates enhances the model performance and provides the ability of making inference on the model parameters.
- The addition of covariates in the model showed a significant decrease of the computational cost. A possible cause for this observation is that fewer mixture components were created compared to the model without covariates which reduces the complexity of the model due to the additional information. However, this statement should be further examined.
- DP-GMMx uses two distinct sources for the effect of the p covariates and the effect of the $d-$variate response as regression coefficients. The assumption we made about the structure of these coefficients reduces significantly the model parameters which are needed to adjust for covariates.
- The DP-GMM and DP-GMMx models using urinary biomarkers were suitable for prostate cancer prognosis and showed superiority compared to the widely used PSA test, but also to other multivariate statistical methods, such as SVM and LDA.

- It is worth mentioning that the developed model is flexible to use data obtained only from a non-invasive procedure. It is also a low cost and easy to apply screening tool.

6.2.2 Limitations and Future Work

This work has some limitations based on which the DP-GMM model could be further examined and developed.

- Exploring additional covariate information as potential risk factors such as ethnicity, family history and genetic factors, dietary factors, alcohol and coffee consumption, smoking status, obesity and physical activity and medications, would potentially enable the model to capture hidden patterns and increase its power for prediction.
- It would also be useful if a variable selection technique is incorporated in the analysis, so it will take into account only the highly significant variables automatically.
- If repeated recordings from men across various time points were available, it would be a great addition for monitoring the progress of prostate cancer therapy in the context of personalised medicine. This aspect requires that the DP-GMM can be easily extended to a model with fixed and random effects terms.

Appendix A

Derivation of technical details

A.1 Likelihood function

$$\begin{aligned}
\mathcal{L}|\boldsymbol{y}(\mu,\sigma^2) &\propto \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\Big\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\Big\} \propto \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\Big\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2-2n\bar{y}\mu+n\mu^2)\Big\} \\
&\propto \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\Big\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2-2n\bar{y}^2+n\bar{y}^2+n\bar{y}^2-2n\bar{y}\mu+n\mu^2)\Big\} \\
&\propto \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\Big\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\bar{y})^2\Big\} \exp\Big\{-\frac{1}{2\sigma^2}n(\bar{y}-\mu)^2\Big\} \\
&\overset{\sigma^2\propto 1/\tau}{\propto} \tau^{n/2} \exp\Big\{-\frac{\tau}{2}\sum_{i=1}^{n}(y_i-\bar{y})^2\Big\} \exp\Big\{-\frac{\tau}{2}n(\bar{y}-\mu)^2\Big\}
\end{aligned}$$

A.2 Conjugate Joint Posterior Distribution

$$\begin{aligned}
p(\mu,\tau|\boldsymbol{y}) &\propto f(\boldsymbol{y}|\mu,1/\tau)p(\mu,\tau) \propto \mathcal{N}(\boldsymbol{y}|\mu,\tau) \times \mathcal{NG}a(\mu_0,\kappa_0,\alpha_0,\beta_0) \\
&\propto \tau^{1/2}e^{-\frac{\tau\kappa_0}{2}(\mu-\mu_0)^2}\tau^{\alpha_0-1}e^{-\beta_0\tau}\tau^{n/2}e^{-\frac{\tau}{2}\sum_{i=1}^{n}(y_i-\mu)^2} \\
&\propto \tau^{1/2}\tau^{\alpha_0+n/2-1}e^{-\beta_0\tau}e^{-\frac{\tau}{2}[\kappa_0(\mu-\mu_0)^2+\sum_{i=1}^{n}(y_i-\mu)^2]} \\
&\propto \tau^{1/2}\tau^{\alpha_0+n/2-1}e^{-\frac{\tau}{2}[\kappa_0(\mu-\mu_0)^2+\sum_{i=1}^{n}[(y_i-\bar{y})-(\mu-\bar{y})]^2]}
\end{aligned}$$

$$
\begin{aligned}
&\propto \tau^{1/2}\tau^{\alpha_0+n/2-1}e^{-\beta_0\tau}e^{-\frac{\tau}{2}[\kappa_0(\mu-\mu_0)^2+\sum_{i=1}^n(y_i-\bar{y})^2+\sum_{i=1}^n(\bar{y}-\mu)^2-2\overbrace{\sum_{i=1}^n(y_i-\bar{y})(\mu-\bar{y})}^{0}]}\\
&\propto \tau^{1/2}\tau^{\alpha_0+n/2-1}e^{-\beta_0\tau}e^{-\frac{\tau}{2}[\kappa_0(\mu-\mu_0)^2+\sum_{i=1}^n(y_i-\bar{y})^2+n(\mu-\bar{y})^2]}\\
&\propto \tau^{1/2}\tau^{\alpha_0+n/2-1}e^{-\beta_0\tau}e^{-\frac{\tau}{2}[(\kappa_0+n)(\mu-\frac{\kappa_0\mu_0+n\bar{y}}{\kappa_0+n})^2+\frac{\kappa_0 n(\bar{y}-\mu_0)^2}{\kappa_0+n}+\sum_{i=1}^n(y_i-\bar{y})^2]}.
\end{aligned}
$$

Therefore, the joint posterior distribution is a product of Gaussian and Gamma distributions as

$$
\begin{aligned}
p(\mu,\tau|\boldsymbol{y}) &\propto \tau^{1/2}e^{-\frac{\tau}{2}(\kappa_0+n)(\mu-\frac{\kappa_0\mu_0+n\bar{y}}{\kappa_0+n})^2}\\
&\quad\times \tau^{\alpha_0+n/2-1}e^{-\beta_0\tau}e^{-\frac{\tau}{2}\sum_{i=1}^n(y_i-\bar{y})^2}e^{-\frac{\tau}{2}\frac{\kappa_0 n(\bar{y}-\mu_0)^2}{\kappa_0+n}}\\
&\quad\propto \mathcal{N}(\boldsymbol{y}|\mu_n,(\kappa_n\tau)^{-1})\times\mathcal{G}a(\alpha_n,\beta_n),
\end{aligned}
$$

with mean μ_n, variance $(\kappa_n\tau)^{-1}$, shape parameter α_n, and rate parameter β_n, where

$$
\mu_n = \frac{\kappa_0\mu_0+n\bar{y}}{\kappa_0+n}
$$

$$
\kappa_n = \kappa_0+n
$$

$$
\alpha_n = \alpha_0+n/2
$$

$$
\beta_n = \beta_0+\frac{1}{2}\sum_{i=1}^n(y_i-\bar{y})^2+\frac{\kappa_0 n(\bar{y}-\mu_0)^2}{\kappa_0+n}
$$

A.3 Non-Conjugate Joint Posterior Distribution

$$
\begin{aligned}
p(\mu,\tau|\boldsymbol{y}) &\propto f(\boldsymbol{y}|\mu,\tau)p(\mu,\tau) = \mathcal{N}(\boldsymbol{y}|\mu,\tau)\times\mathcal{N}\mathcal{G}a(\mu_0,\kappa_0,\alpha_0,\beta_0)\\
&\quad\propto \tau^{\frac{2\alpha_0+n+1}{2}-1}e^{-\frac{\tau}{2}[\beta_0+(\kappa_0+n)(\mu-\frac{\kappa_0\mu_0+n\bar{y}}{\kappa_0+n})^2+\frac{\kappa_0 n(\bar{y}-\mu_0)^2}{\kappa_0+n}+\sum_{i=1}^n(y_i-\bar{y})^2]}
\end{aligned}
$$

A.4 Conditional Posterior of parameter μ

$$
\begin{aligned}
p(\mu|\tau,\boldsymbol{y}) \propto p(\mu,\tau|\boldsymbol{y}) &\propto e^{-\frac{\tau}{2}(\sum_{i=1}^{n} y_i^2 - 2\mu\sum_{i=1}^{n} y_i + n\mu^2 + \kappa_0\mu^2 - 2\kappa_0\mu_0\mu + \kappa_0\mu_0^2)} \\
&\propto e^{-\frac{\tau}{2}[(\kappa_0+n)\mu^2 - 2(n\bar{y}+\kappa_0\mu_0)\mu]} \\
&\propto e^{-\frac{\tau(\kappa_0+n)}{2}(\mu^2 - 2\mu\frac{n\bar{y}+\kappa_0\mu_0}{\kappa_0+n} + \frac{2\bar{y}+\kappa_0\mu_0}{\kappa_0+n})} \\
&\propto e^{-\frac{\tau(\kappa_0+n)}{2}(\mu - \frac{n\bar{y}+\kappa_0\mu_0}{\kappa_0+n})^2}
\end{aligned}
$$

A.5 Deriving the joint distribution of μ and σ

Let $\theta = (\mu, C)$ be a 2-dimensional vector of the population characteristics, denoting the mean and coefficient of variation with probability density function

$$
\begin{aligned}
p(\theta) = p(\mu, C) = p(\mu)p(C|\mu) = p(\mu)p(C), \\
\text{where} \quad C = \frac{\sigma}{\mu}.
\end{aligned}
$$

Also, let the transformation $g : \mathbb{R}^2 \to \mathbb{R}^2$,

$$
\theta^\top \mapsto g(\theta^\top) = g\begin{pmatrix} \mu \\ C \end{pmatrix} = \begin{pmatrix} \theta_1'(\mu, C) \\ \theta_2'(\mu, C) \end{pmatrix} = \begin{pmatrix} \mu \\ C \cdot \mu \end{pmatrix} = \begin{pmatrix} \mu \\ \sigma \end{pmatrix} = \theta'^\top,
$$

so we have $g(\mu) = \mu = \theta_1'$ and $g(C) = C \cdot \mu = \sigma = \theta_2'$ with joint range space $R_{\theta'} = (0, \infty)$. Thus,

$$
\theta^\top = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} \mu \\ C \end{pmatrix} = g^{-1}\begin{pmatrix} \theta_1' \\ \theta_2' \end{pmatrix} = \begin{pmatrix} \theta_1' \\ {}^{\theta_2'}/_{\theta_1'} \end{pmatrix}.
$$

As we are interested in obtaining the distribution of θ', we first need to compute the Jacobian term. Therefore, the joint distribution is given by

$$
p(\theta') = p(\mu, \sigma) = p(\theta) \cdot |J| = p(\mu) \cdot p(C) \cdot |J|,
$$

$$
J = \det\left(\frac{\partial g^{-1}(\theta')}{\partial \theta'}\right) = \begin{vmatrix} \frac{\partial g^{-1}(\theta_1')}{\partial \theta_1'} & \frac{\partial g^{-1}(\theta_1')}{\partial \theta_2'} \\ \frac{\partial g^{-1}(\theta_2')}{\partial \theta_1'} & \frac{\partial g^{-1}(\theta_2')}{\partial \theta_2'} \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ -\frac{\theta_2'}{\theta_1'^2} & \frac{1}{\theta_1'} \end{vmatrix} = \frac{1}{\mu}.
$$

A.6 Full Conditional Posterior Distributions for the parameters of the univariate multilevel model

Joint posterior:

$$
\begin{aligned}
p(\mu, \{\mu_j\}_{j=1}^J, \sigma_b^2, \sigma_e^2 | \boldsymbol{y}) = & \prod_{j=1}^{J} \prod_{i=1}^{n_j} \mathcal{N}(y_{ij}|\mu_j, \sigma_e^2) \times \prod_{j=1}^{J} \mathcal{N}(\mu_j|\mu, \sigma_b^2) \\
& \times \mathcal{N}(\mu|\mu_0, \sigma_0^2) \times \mathcal{IG}a(\sigma_e^2|\alpha_1, \beta_1) \times \mathcal{IG}a(\sigma_b^2|\alpha_2, \beta_2) \\
& \propto \frac{1}{(\sigma_e^2)^{n/2}} e^{-\frac{1}{2\sigma_e^2}\sum_{j=1}^J \sum_{i=1}^{n_j}(y_{ij}-\mu_j)^2} \times \prod_{j=1}^{J} \frac{1}{(\sigma_b^2)^{1/2}} e^{-\frac{1}{2\sigma_b^2}(\mu_j-\mu)^2} \times e^{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2} \\
& \times \left(\frac{1}{\sigma_e^2}\right)^{\alpha_1+1} e^{-\frac{\beta_1}{\sigma_e^2}} \times \left(\frac{1}{\sigma_b^2}\right)^{\alpha_2+1} e^{-\frac{\beta_2}{\sigma_b^2}} \\
& \propto \frac{1}{(\sigma_e^2)^{n/2}} e^{-\frac{1}{2\sigma_e^2}\sum_{j=1}^J \sum_{i=1}^{n_j}(y_{ij}-\mu_j)^2} \times \frac{1}{(\sigma_b^2)^{J/2}} e^{-\frac{1}{2\sigma_b^2}\sum_{j=1}^J(\mu_j-\mu)^2} \\
& \times e^{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2} \times \left(\frac{1}{\sigma_e^2}\right)^{\alpha_1+1} e^{-\frac{\beta_1}{\sigma_e^2}} \times \left(\frac{1}{\sigma_b^2}\right)^{\alpha_2+1} e^{-\frac{\beta_2}{\sigma_b^2}}
\end{aligned}
$$

Conditional distributions:

$$
\begin{aligned}
p(\mu|\{\mu_j\}_{j=1}^J, \sigma_b^2, \sigma_e^2, \boldsymbol{y}) & \propto e^{-\frac{1}{2\sigma_b^2}\sum_{j=1}^J(\mu_j-\mu)^2} \times e^{-\frac{1}{2\sigma_e^2}(\mu^2-2\mu\mu_0)} \\
& \propto e^{-\frac{1}{2\sigma_b^2}(-2\mu\sum_{j=1}^J \mu_j + J\mu^2)} \times e^{-\frac{1}{2\sigma_0^2}(\mu^2-2\mu\mu_0)} \\
& \propto \exp\left(\frac{2\mu\sum_{j=1}^J \mu_j - J\mu^2}{2\sigma_b^2} + \frac{2\mu\mu_0 - \mu^2}{2\sigma_0^2}\right) \\
& = \exp\left(\frac{2\mu\sum_{j=1}^J \mu_j\sigma_0^2 - J\mu^2\sigma_0^2 + 2\mu\mu_0\sigma_b^2 - \mu^2\sigma_b^2}{2\sigma_b^2\sigma_0^2}\right) \\
& = \exp\left(- \frac{J\sigma_0^2 + \sigma_b^2}{2\sigma_b^2\sigma_0^2}\mu^2 + \frac{\sum_{j=1}^J \mu_j\sigma_0^2 + \mu_0\sigma_b^2}{\sigma_b^2\sigma_0^2}\mu\right)
\end{aligned}
$$

$$
\implies \mu|\{\mu_j\}_{j=1}^J, \sigma_b^2, \sigma_e^2, \boldsymbol{y} \sim \mathcal{N}\left(\frac{\sum_{j=1}^J \mu_j\,\sigma_0^2 + \mu_0\sigma_b^2}{J\sigma_0^2 + \sigma_b^2}, \frac{\sigma_b^2\sigma_e^2}{J\sigma_0^2 + \sigma_b^2}\right)
$$

$$
\begin{aligned}
p(\mu_j|\mu,\mu_{-j},\sigma_b^2,\sigma_e^2,\boldsymbol{y}) &\propto e^{-\frac{1}{2\sigma_e^2}\sum_{i=1}^{n_j}(y_{ij}-\mu_j)^2} \times e^{-\frac{1}{2\sigma_b^2}(\mu_j^2-2\mu\mu_j)} \\
&\propto e^{-\frac{1}{2\sigma_e^2}\sum_{i=1}^{n_j}(-2\mu_j y_{ij}+\mu_j^2)} \times e^{-\frac{1}{2\sigma_b^2}(\mu_j^2-2\mu\mu_j)} \\
&= \exp\left(\frac{2\mu_j\sigma_b^2\sum_{i=1}^{n_j} y_{ij} - \sigma_b^2 n_j\mu_j^2 - \sigma_e^2\mu_j^2 + 2\mu\mu_j\sigma_e^2}{2\sigma_b^2\sigma_e^2}\right) \\
&= \exp\left(- \frac{\sigma_b^2 n_j + \sigma_e^2}{2\sigma_b^2\sigma_e^2}\mu_j^2 + \frac{\sigma_b^2\sum_{i=1}^{n_j} y_{ij} + \mu\sigma_e^2}{\sigma_b^2\sigma_e^2}\mu_j\right)
\end{aligned}
$$

$$
\implies \mu_j|\mu,\mu_{-j},\sigma_b^2,\sigma_e^2,\boldsymbol{y} \sim \mathcal{N}\left(\frac{\sum_{i=1}^{n_j} y_{ij}\sigma_b^2 + \mu\sigma_e^2}{n_j\sigma_b^2+\sigma_e^2}, \frac{\sigma_b^2\sigma_e^2}{n_j\sigma_b^2+\sigma_e^2}\right)
$$

$$
p(\sigma_e^2|\mu,\{\mu_j\}_{i=1}^J,\sigma_b^2,\boldsymbol{y}) \propto \frac{1}{(\sigma_e^2)^{n/2}} e^{-\frac{1}{2\sigma_e^2}\sum_{j=1}^J\sum_{i=1}^{n_j}(y_{ij}-\mu_j)^2} \times \left(\frac{1}{\sigma_e^2}\right)^{\alpha_1+1} e^{-\frac{\beta_1}{\sigma_e^2}}
$$

$$
= \left(\frac{1}{\sigma_e^2}\right)^{\frac{n}{2}+\alpha_1+1} e^{-\frac{1}{2\sigma_e^2}\frac{\sum_{j=1}^J\sum_{i=1}^{n_j}(y_{ij}-\mu_j)^2+2\beta_1}{2}}
$$

$$
\implies \sigma_e^2|\mu,\{\mu_j\}_{j=1}^J,\sigma_b^2,\boldsymbol{y} \sim \mathcal{IG}a\left(\text{shape} = \frac{n}{2}+\alpha_1,\ \text{scale} = \frac{\sum_{j=1}^J\sum_{i=1}^{n_j}(y_{ij}-\mu_j)^2+2\beta_1}{2}\right)
$$

$$
p(\sigma_b^2|\mu,\{\mu_j\}_{j=1}^J,\sigma_e^2,\boldsymbol{y}) \propto \frac{1}{(\sigma_b^2)^{J/2}} e^{-\frac{1}{2\sigma_b^2}\sum_{j=1}^J(\mu_j-\mu)^2} \times \left(\frac{1}{(\sigma_b^2)}\right)^{\alpha_2+1} e^{-\frac{\beta_2}{\sigma_b^2}}
$$

$$
= \left(\frac{1}{\sigma_b^2}\right)^{J/2+\alpha_2+1} e^{-\frac{1}{\sigma_b^2}\frac{\sum_{j=1}^J(\mu_j-\mu)^2}{2}+\beta_2}
$$

$$
\implies \sigma_b^2|\mu,\{\mu_j\}_{j=1}^J,\sigma_e^2,\boldsymbol{y} \sim \mathcal{IG}a\left(\text{shape} = J/2+\alpha_2,\ \text{scale} = \frac{\sum_{i=j}^J(\mu_j-\mu)^2}{2}+\beta_2\right)
$$

A.7 Full Conditional Posterior Distributions for the parameters of the multivariate multilevel model

Joint posterior:

$$\begin{aligned} p(\boldsymbol{\mu}, \{\boldsymbol{\mu_j}\}_{j=1}^J, \Omega_e, \Omega_b, \Omega_\mu|\boldsymbol{y}) \propto {} & \prod_{j=1}^J \prod_{i=1}^{n_j} \mathcal{N}_K(y_{ij}|\boldsymbol{\mu_j}, \Omega_e^{-1}) \times \prod_{j=1}^J \mathcal{N}_K(\boldsymbol{\mu_j}|\boldsymbol{\mu}, \Omega_b^{-1}) \\ & \times \mathcal{N}_K(\boldsymbol{\mu}|\boldsymbol{\mu_0}, \Omega_\mu^{-1}) \times \mathcal{W}i(\Omega_e|d_e, S_e) \\ & \times \mathcal{W}i(\Omega_b|d_b, S_b) \times \mathcal{W}i(\Omega_\mu|d_\mu, S_\mu) \end{aligned}$$

Conditional distributions:

$$\begin{aligned} p(\boldsymbol{\mu}|\{\boldsymbol{\mu_j}\}_{j=1}^J,\Omega_e, \Omega_b, \Omega_\mu, \boldsymbol{y}) &\propto \prod_{j=1}^J e^{-\frac{1}{2}(\boldsymbol{\mu_j}-\boldsymbol{\mu})^T\Omega_b(\boldsymbol{\mu_j}-\boldsymbol{\mu})} \times e^{-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu_0})^T\Omega_\mu(\boldsymbol{\mu}-\boldsymbol{\mu_0})} \\ &\propto e^{-\frac{1}{2}(-2\sum_{j=1}^J \boldsymbol{\mu}^T\Omega_b\boldsymbol{\mu_j}+\sum_{j=1}^J \boldsymbol{\mu}^T\Omega_b\boldsymbol{\mu})} \times e^{-\frac{1}{2}(\boldsymbol{\mu}^T\Omega_\mu\boldsymbol{\mu}-2\boldsymbol{\mu}^T\Omega_\mu\boldsymbol{\mu_0})} \\ &= \exp\Bigg(-\frac{1}{2}(-2\boldsymbol{\mu}^T\Omega_b J\bar{\boldsymbol{\mu}}_{\boldsymbol{j}} + J\boldsymbol{\mu}^T\Omega_b\boldsymbol{\mu})\Bigg) \times \exp\Bigg(-\frac{1}{2}(\boldsymbol{\mu}^T A_0\boldsymbol{\mu} - 2\boldsymbol{\mu}^T b_0)\Bigg) \\ &= \exp\Bigg(-\frac{1}{2}(-2\boldsymbol{\mu}^T b_1 + \boldsymbol{\mu}^T A_1\boldsymbol{\mu})\Bigg) \times \exp\Bigg(-\frac{1}{2}(\boldsymbol{\mu}^T A_0\boldsymbol{\mu} - 2\boldsymbol{\mu}^T b_0)\Bigg) \\ &= \exp\Bigg(\boldsymbol{\mu}^T(b_0+b_1) - \frac{1}{2}\boldsymbol{\mu}^T(A_0+A_1)\boldsymbol{\mu}\Bigg) = \exp\Bigg(\boldsymbol{\mu}^T b_n - \frac{1}{2}\boldsymbol{\mu}^T A_n\boldsymbol{\mu}\Bigg), \end{aligned}$$

where

$$\begin{aligned} b_0 &= \Omega_\mu\boldsymbol{\mu_0} \\ A_0 &= \Omega_\mu \\ b_1 &= J\Omega_b\bar{\boldsymbol{\mu}}_{\boldsymbol{j}} \\ \bar{\boldsymbol{\mu}}_{\boldsymbol{j}} &= (\frac{1}{J}\sum_{j=1}^J \mu_{j[1]}, ..., \frac{1}{J}\sum_{j=1}^J \mu_{j[K]}) \\ A_1 &= J\Omega_b \\ b_n = b_0 + b_1 &= \Omega_\mu\boldsymbol{\mu_0} + J\Omega_b\bar{\boldsymbol{\mu}}_{\boldsymbol{j}} \\ A_n = A_0 + A_1 &= \Omega_\mu + J\Omega_b \end{aligned}$$

$$\implies \boldsymbol{\mu}|\{\boldsymbol{\mu_j}\}_{j=1}^J, \Omega_e, \Omega_b, \Omega_\mu, \boldsymbol{y} \sim \mathcal{N}_K\Bigg(\mu_n = A_n^{-1}b_n, \Sigma_n = A_n^{-1}\Bigg),$$

where

$$
\begin{aligned}
\mu_n = A_n^{-1}b_n = (\Omega_\mu + J\Omega_b)^{-1}(\Omega_\mu\boldsymbol{\mu_0} + J\Omega_b\boldsymbol{\bar{\mu}_j})\\
\Sigma_n = A_n^{-1} = (\Omega_\mu + J\Omega_b)^{-1}
\end{aligned}
$$

$$
\begin{aligned}
p(\boldsymbol{\mu_j}|\boldsymbol{\mu},\{\boldsymbol{\mu_{-j}}\},\Omega_e,\Omega_b,\Omega_\mu,\boldsymbol{y}) &\propto \prod_{i=1}^{n_j} e^{-\frac{1}{2}(\boldsymbol{y_{ij}}-\boldsymbol{\mu_j})^T\Omega_e(\boldsymbol{y_{ij}}-\boldsymbol{\mu_j})} \times e^{-\frac{1}{2}(\boldsymbol{\mu_j}-\boldsymbol{\mu})^T\Omega_b(\boldsymbol{\mu_j}-\boldsymbol{\mu})}\\
&\propto e^{-\frac{1}{2}(-2\boldsymbol{\mu_j}^T\Omega_e\sum_{i=1}^{n_j} y_{ij}+n_j\boldsymbol{\mu_j}^T\Omega_e\boldsymbol{\mu_j})} \times e^{-\frac{1}{2}(\boldsymbol{\mu_j}^T\Omega_b\boldsymbol{\mu_j}-2\boldsymbol{\mu_j}^T\Omega_b\boldsymbol{\mu})}\\
&= \exp\bigg(-\frac{1}{2}(-2\boldsymbol{\mu_j}^T b_1^{'} + \boldsymbol{\mu_j}^T A_1^{'}\boldsymbol{\mu_j})\bigg)\\
&\times \exp\bigg(-\frac{1}{2}(\boldsymbol{\mu_j}^T A_0^{'}\boldsymbol{\mu_j} - 2\boldsymbol{\mu_j}^T b_0^{'}\bigg)\\
&= \exp\bigg(\boldsymbol{\mu_j}^T(b_0^{'}+b_1^{'}) - \frac{1}{2}\boldsymbol{\mu_j}^T(A_0^{'}+A_1^{'})\boldsymbol{\mu_j})\bigg)\\
&= \exp\bigg(\boldsymbol{\mu_j}^T b_n^{'} - \frac{1}{2}\boldsymbol{\mu_j}^T A_n^{'}\boldsymbol{\mu_j})\bigg),
\end{aligned}
$$

where

$$
\begin{aligned}
b_0^{'} = \Omega_b\boldsymbol{\mu}\\
A_0 = \Omega_b\\
b_1^{'} = \Omega_e\sum_{j=1}^{n_j} y_{ij}\\
A_1^{'} = n_j\Omega_e\\
b_n^{'} = b_0^{'} + b_1^{'} = \Omega_b\boldsymbol{\mu} + \Omega_e\sum_{j=1}^{n_j} y_{ij}\\
A_n^{'} = A_0^{'} + A_1^{'} = \Omega_b + n_j\Omega_e
\end{aligned}
$$

$$
\implies \boldsymbol{\mu_j}|\boldsymbol{\mu},\{\boldsymbol{\mu_{-j}}\},\Omega_e,\Omega_b,\Omega_\mu,\boldsymbol{y} \sim \mathcal{N}_K\bigg(\mu_n^{'} = A_n^{-1'}b_n^{'}, \Sigma_n^{'} = A_n^{-1'}\bigg)\ \forall\ j = 1,...,J,
$$

where

$$\mu_n^{'} = A_n^{-1'} b_n^{'} = (\Omega_b + n_j \Omega_e)^{-1} (\Omega_b \boldsymbol{\mu} + \Omega_e \sum_{j=1}^{n_j} y_{ij})$$

$$\Sigma_n^{'} = A_n^{-1'} = (\Omega_b + n_j \Omega_e)^{-1}$$

$$p(\Omega_e | \boldsymbol{\mu}, \{\boldsymbol{\mu_j}\}_{j=1}^{J}, \Omega_b, \Omega_\mu, \boldsymbol{y}) \propto \prod_{j=1}^{J} \prod_{i=1}^{n_j} \mathcal{N}_K(\boldsymbol{\mu_j}, \Omega_e^{-1}) \times \mathcal{W}i(d_e, S_e)$$

$$\propto |\Omega_e^{-1}|^{-\frac{n}{2}} e^{-\frac{1}{2} \sum_{j=1}^{J} \sum_{i=1}^{n_j} (\boldsymbol{y_{ij}} - \boldsymbol{\mu_j})^T \Omega_e (\boldsymbol{y_{ij}} - \boldsymbol{\mu_j})} \times |\Omega_e^{-1}|^{\frac{d_e - K - 1}{2}} e^{-\frac{tr(S_e^{-1} \Omega_e)}{2}}$$

$$\propto \left(\frac{1}{|\Omega_e|} \right)^{-\frac{n}{2}} |\Omega_e|^{\frac{d_e - K - 1}{2}} e^{-\frac{1}{2} (\sum_{j=1}^{J} \sum_{i=1}^{n_j} (\boldsymbol{y_{ij}} - \boldsymbol{\mu_j})^T \Omega_e (\boldsymbol{y_{ij}} - \boldsymbol{\mu_j}) + tr(S_e^{-1} \Omega_e))}$$

$$\propto |\Omega_e|^{\frac{(d_e + n) - K - 1}{2}} e^{-\frac{1}{2} tr(S^{-1} \Omega_e)},$$

where

$$n = \sum_{j=1}^{J} n_j$$

$$S = \sum_{j=1}^{J} \sum_{i=1}^{n_j} (\boldsymbol{y_{ij}} - \boldsymbol{\mu_j})(\boldsymbol{y_{ij}} - \boldsymbol{\mu_j})^T + S_e^{-1}$$

$$\implies \Omega_e | \boldsymbol{\mu}, \{\boldsymbol{\mu_j}\}_{j=1}^{J}, \Omega_b, \Omega_\mu, \boldsymbol{y} \sim \mathcal{W}i \ (d_e^{'} = d_e + n, S_e^{'} = S^{-1})$$

$$p(\Omega_b|\boldsymbol{\mu}, \{\boldsymbol{\mu_j}\}_{j=1}^{J}, \Omega_e, \Omega_\mu, \boldsymbol{y}) \propto \prod_{j=1}^{J} \mathcal{N}_K(\boldsymbol{\mu}, \Omega_b^{-1}) \times \mathcal{W}i(d_b, S_b)$$

$$\propto \prod_{j=1}^{J} |\Omega_b^{-1}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{\mu_j}-\boldsymbol{\mu})^T \Omega_b (\boldsymbol{\mu_j}-\boldsymbol{\mu})} \times |\Omega_b|^{\frac{d_b-K-1}{2}} e^{-\frac{tr(S_b^{-1}\Omega_b)}{2}}$$

$$\propto \left(\frac{1}{|\Omega_b|}\right)^{-\frac{J}{2}} e^{-\frac{1}{2}(\sum_{j=1}^{J}(\boldsymbol{\mu_j}-\boldsymbol{\mu})^T \Omega_b (\boldsymbol{\mu_j}-\boldsymbol{\mu})} \times |\Omega_b|^{\frac{d_b-K-1}{2}} e^{-\frac{1}{2}tr(S_b^{-1}\Omega_b)}$$

$$\propto |\Omega_b|^{\frac{(J+d_b)-K-1}{2}} e^{-\frac{1}{2}tr(S^{'}\Omega_b)},$$

where

$$S^{'} = \sum_{j=1}^{J} (\boldsymbol{\mu_j} - \boldsymbol{\mu})(\boldsymbol{\mu_j} - \boldsymbol{\mu})^T + S_b^{-1}$$

$$\implies \Omega_b|\boldsymbol{\mu}, \{\boldsymbol{\mu_j}\}_{j=1}^{J}, \Omega_e, \Omega_\mu, \boldsymbol{y} \sim \mathcal{W}i\ (d_b^{'} = J + d_b, S_b^{'} = S^{\prime -1})$$

$$p(\Omega_\mu|\boldsymbol{\mu}, \{\boldsymbol{\mu_j}\}_{j=1}^{J}, \Omega_e, \Omega_b, \boldsymbol{y}) \propto |\Omega_\mu^{-1}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu_0})^T \Omega_\mu (\boldsymbol{\mu}-\boldsymbol{\mu_0})} \times |\Omega_\mu|^{\frac{d_\mu-K-1}{2}} e^{-\frac{tr(S_\mu^{-1}\Omega_\mu)}{2}}$$

$$\propto |\Omega_\mu|^{\frac{(d_\mu+1)-K-1}{2}} e^{-\frac{1}{2}tr(S^{''}\Omega_b)},$$

where

$$S^{''} = (\boldsymbol{\mu} - \boldsymbol{\mu_0})(\boldsymbol{\mu} - \boldsymbol{\mu_0})^T + S_\mu^{-1}$$

$$\implies \Omega_\mu|\boldsymbol{\mu}, \{\boldsymbol{\mu_j}\}_{j=1}^{J}, \Omega_e, \Omega_b, \boldsymbol{y} \sim \mathcal{W}i\ (d_\mu^{'} = d_\mu + 1, S_\mu^{'} = S^{\prime\prime -1})$$

A.8 Technical details for Inference of DP-GMM

A.8.1 Conditional posterior distributions for the means $\boldsymbol{\mu_j}$s

The conditional posterior distributions for the means $\boldsymbol{\mu_j}$s are derived by the multiplication of the likelihood given the variable $\boldsymbol{c}$ of indicators from equation (4.3) by the prior for the means from equation (4.11) as follows:

$$
\begin{aligned}
p(\boldsymbol{\mu_j}|\boldsymbol{c}, S_j, \boldsymbol{\mu_0}, S_0, \boldsymbol{y}) &\propto \prod_{i=1}^{n_j} \mathcal{N}_d(\boldsymbol{y_i}|\boldsymbol{\mu_j}, S_j^{-1}) \times \mathcal{N}_d(\boldsymbol{\mu_j}|\boldsymbol{\mu_0}, S_0^{-1}) \\
&\propto \prod_{i=1}^{n_j} e^{-\frac{1}{2}(\boldsymbol{y_i}-\boldsymbol{\mu_j})^T S_j (\boldsymbol{y_i}-\boldsymbol{\mu_j})} \times e^{-\frac{1}{2}(\boldsymbol{\mu_j}-\boldsymbol{\mu_0})^T S_0 (\boldsymbol{\mu_j}-\boldsymbol{\mu_0})} \\
&\propto e^{-\frac{1}{2}(-2\boldsymbol{\mu_j}^T S_j n_j \bar{\boldsymbol{y}}_{\boldsymbol{j}} + n_j \boldsymbol{\mu_j}^T S_j \boldsymbol{\mu_j})} \times e^{-\frac{1}{2}(\boldsymbol{\mu_j}^T S_0 \boldsymbol{\mu_j} - 2\boldsymbol{\mu_j}^T S_0 \boldsymbol{\mu_0})} \\
&= \exp\Bigg(-\frac{1}{2}(-2\boldsymbol{\mu_j}^T b_1 + \boldsymbol{\mu_j}^T A_1 \boldsymbol{\mu_j})\Bigg) \times \exp\Bigg(-\frac{1}{2}(\boldsymbol{\mu_j}^T A_0 \boldsymbol{\mu_j} - 2\boldsymbol{\mu_j}^T b_0\Bigg) \\
&= \exp\Bigg(\boldsymbol{\mu_j}^T (b_0 + b_1) - \frac{1}{2}\boldsymbol{\mu_j}^T (A_0 + A_1)\boldsymbol{\mu_j})\Bigg) = \exp\Bigg(\boldsymbol{\mu_j}^T b_n - \frac{1}{2}\boldsymbol{\mu_j}^T A_n \boldsymbol{\mu_j})\Bigg),
\end{aligned}
$$

where

$$
b_0 = S_0\boldsymbol{\mu_0}, \quad A_0 = S_0, \quad b_1 = n_j \bar{\boldsymbol{y}}_{\boldsymbol{j}} S_j, \quad \bar{\boldsymbol{y}}_{\boldsymbol{j}} = (\frac{1}{n_j}\sum_{i=1}^{n_j} \boldsymbol{y_{i[1]}}, ..., \frac{1}{n_j}\sum_{i=1}^{n_j} \boldsymbol{y_{i[d]}})
$$

$$
A_1 = n_j S_j, \quad b_n = b_0 + b_1 = S_0\boldsymbol{\mu_0} + n_j \bar{\boldsymbol{y}}_{\boldsymbol{j}} S_j, \quad A_n = A_0 + A_1 = S_0 + n_j S_j
$$

$$
\implies \boldsymbol{\mu_j}|\boldsymbol{c}, S_j, \boldsymbol{\mu_0}, S_0, \boldsymbol{y} \sim \mathcal{N}_d\Bigg(\mu_n = A_n^{-1} b_n, \Sigma_n = A_n^{-1}\Bigg) \ \forall \ j = 1, ..., K,
$$

where

$$\mu_n = A_n^{-1} b_n = (S_0 + n_j S_j)^{-1}(S_0 \boldsymbol{\mu_0} + n_j \bar{\boldsymbol{y}}_{\boldsymbol{j}} S_j)$$

$$\Sigma_n = A_n^{-1} = (S_0 + n_j S_j)^{-1}$$

A.8.2 Conditional posterior distributions for precisions S_j

We obtain the conditional posteriors for precisions S_js by multiplying the likelihood conditioned on the indicators from equation (4.3) by the prior for precisions from equation (4.11) as:

$$p(S_j|\boldsymbol{c}, \boldsymbol{\mu_j}, \beta_0, W_0, \boldsymbol{y}) \propto \prod_{i=1}^{n_j} \mathcal{N}_d(\boldsymbol{y_i}|\boldsymbol{\mu_j}, S_j^{-1}) \times \mathcal{W}(S_j|\beta_0, (\beta_0 W_0)^{-1})$$

$$\propto \prod_{i=1}^{n_j} |S_j^{-1}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{y_i}-\boldsymbol{\mu_j})^T S_j (\boldsymbol{y_i}-\boldsymbol{\mu_j})} \times |S_j|^{\frac{\beta_0-d-1}{2}} e^{-\frac{tr((\beta_0 W_0)S_j)}{2}}$$

$$\propto \left(\frac{1}{|S_j|}\right)^{-\frac{n_j}{2}} e^{-\frac{1}{2}\sum_{j=1}^{n_j}(\boldsymbol{y_i}-\boldsymbol{\mu_j})^T S_j (\boldsymbol{y_i}-\boldsymbol{\mu_j})} \times |S_j|^{\frac{\beta_0-d-1}{2}} e^{-\frac{1}{2}tr((\beta_0 W_0)S_j)}$$

$$\propto |S_j|^{\frac{(n_j+\beta_0)-d-1}{2}} e^{-\frac{1}{2}tr(S' S_j)},$$

where

$$S' = \sum_{j=1}^{n_j}(\boldsymbol{y_i} - \boldsymbol{\mu_j})(\boldsymbol{y_i} - \boldsymbol{\mu_j})^T + \beta_0 W_0$$

$$\implies S_j|\boldsymbol{c}, \boldsymbol{\mu_j}, \beta_0, W_0, \boldsymbol{y} \sim \mathcal{W}(n_j + \beta_0, S'^{-1})$$

A.8.3 Conditional posterior distributions for the hyperparameters $\boldsymbol{\mu_0}$, S_0, W_0 and y

The conditional posterior distributions for the hyperparameters $\boldsymbol{\mu_0}$, S_0, W_0 and $z = \beta_0 - d + 1$ are obtained from the corresponding distributions in equation (4.11), which acts as likelihood, and the hyperpriors from equations (4.12) and (4.19) as:

$$p(\boldsymbol{\mu_0}|\{\boldsymbol{\mu_j}\}_{j=1}^{K}, S_0) \propto \prod_{j=1}^{K} \mathcal{N}_d(\boldsymbol{\mu_j}|\boldsymbol{\mu_0}, S_0^{-1}) \times \mathcal{N}_d(\boldsymbol{\mu_0}|\boldsymbol{\mu_y}, \Sigma_y)$$

$$\propto \prod_{j=1}^{K} e^{-\frac{1}{2}(\boldsymbol{\mu_j}-\boldsymbol{\mu_0})^T S_0 (\boldsymbol{\mu_j}-\boldsymbol{\mu_0})} \times e^{-\frac{1}{2}(\boldsymbol{\mu_0}-\boldsymbol{\mu_y})^T \Sigma_y^{-1}(\boldsymbol{\mu_0}-\boldsymbol{\mu_y})}$$

$$\propto e^{-\frac{1}{2}(-2\boldsymbol{\mu_0}^T S_0 \sum_{j=1}^{K}\boldsymbol{\mu_j} + K\boldsymbol{\mu_0}^T S_0 \boldsymbol{\mu_0})} \times e^{-\frac{1}{2}(\boldsymbol{\mu_0}^T \Sigma_y^{-1}\boldsymbol{\mu_0} - 2\boldsymbol{\mu_0}^T\Sigma_y^{-1}\boldsymbol{\mu_y})}$$

$$= \exp\Bigg(-\frac{1}{2}(-2\boldsymbol{\mu_0}^T b_1 + \boldsymbol{\mu_0}^T A_1 \boldsymbol{\mu_0})\Bigg) \times \exp\Bigg(-\frac{1}{2}(\boldsymbol{\mu_0}^T A_0 \boldsymbol{\mu_0} - 2\boldsymbol{\mu_0}^T b_0\Bigg)$$

$$= \exp\Bigg(\boldsymbol{\mu_0}^T(b_0 + b_1) - \frac{1}{2}\boldsymbol{\mu_0}^T(A_0 + A_1)\boldsymbol{\mu_0})\Bigg) = \exp\Bigg(\boldsymbol{\mu_0}^T b_n - \frac{1}{2}\boldsymbol{\mu_0}^T A_n \boldsymbol{\mu_0})\Bigg),$$

where

$$b_0 = \Sigma_y^{-1}\boldsymbol{\mu_y}, \quad A_0 = \Sigma_y^{-1}, \quad b_1 = K\bar{\boldsymbol{\mu}}S_0, \quad \bar{\boldsymbol{\mu}} = \big(\frac{1}{K}\sum_{j=1}^{K}\boldsymbol{\mu_{j[1]}}, ..., \frac{1}{K}\sum_{j=1}^{K}\boldsymbol{\mu_{j[d]}}\big)$$

$$A_1 = KS_0, \quad b_n = b_0 + b_1 = \Sigma_y^{-1}\boldsymbol{\mu_y} + K\bar{\boldsymbol{\mu}}S_0, \quad A_n = A_0 + A_1 = \Sigma_y^{-1} + KS_0$$

$$\implies \boldsymbol{\mu_0}|\{\boldsymbol{\mu_j}\}_{j=1}^{K}, S_0 \sim \mathcal{N}_d\Bigg(\mu_n = A_n^{-1}b_n, \Sigma_n = A_n^{-1}\Bigg),$$

where

$$\mu_n = A_n^{-1} b_n = (\Sigma_y^{-1} + KS_0)^{-1}(\Sigma_y^{-1}\boldsymbol{\mu_y} + K\bar{\boldsymbol{\mu}}S_0)$$

$$\Sigma_n = A_n^{-1} = (\Sigma_y^{-1} + KS_0)^{-1}$$

$$p(S_0|\{\boldsymbol{\mu_j}\}_{j=1}^K, \boldsymbol{\mu_0}) \propto \prod_{j=1}^K \mathcal{N}_d(\boldsymbol{\mu_j}|\boldsymbol{\mu_0}, S_0^{-1}) \times \mathcal{W}(S_0|d, (d\Sigma_y)^{-1})$$

$$\propto \prod_{j=1}^K |S_0^{-1}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{\mu_j}-\boldsymbol{\mu_0})^T S_0 (\boldsymbol{\mu_j}-\boldsymbol{\mu_0})} \times |S_0|^{\frac{d-d-1}{2}} e^{-\frac{tr(d\Sigma_y S_0)}{2}}$$

$$\propto \left(\frac{1}{|S_0|}\right)^{-\frac{K}{2}} e^{-\frac{1}{2}(\sum_{j=1}^K (\boldsymbol{\mu_j}-\boldsymbol{\mu_0})^T S_0 (\boldsymbol{\mu_j}-\boldsymbol{\mu_0})} \times |S_0|^{\frac{d-d-1}{2}} e^{-\frac{1}{2}tr(d\Sigma_y S_0)}$$

$$\propto |S_0|^{\frac{(K+d)-d-1}{2}} e^{-\frac{1}{2}tr(S^{'} S_0)},$$

where

$$S^{'} = \sum_{j=1}^K (\boldsymbol{\mu_j} - \boldsymbol{\mu_0})(\boldsymbol{\mu_j} - \boldsymbol{\mu_0})^T + d\Sigma_y$$

$$\implies S_0|\{\boldsymbol{\mu_j}\}_{j=1}^K, \mu_0 \sim \mathcal{W}\ (K + d, S'^{-1})$$

$$p(W_0|\{S_j\}_{j=1}^K, \beta_0) \propto \prod_{j=1}^{K} \mathcal{W}(S_j|\beta_0, (\beta_0 W_0)^{-1}) \times \mathcal{W}(W_0|d, (d\Sigma_y)^{-1})$$

$$\propto |\beta_0 W_0|^{\frac{K\beta_0}{2}} \prod_{j=1}^{K} e^{-\frac{tr(\beta_0 W_0 S_j)}{2}} \times |W_0|^{\frac{d-d-1}{2}} e^{-\frac{tr(d\Sigma_y W_0)}{2}}$$

$$\propto |W_0|^{\frac{K\beta_0}{2}} e^{-\frac{1}{2}\sum_{j=1}^{K} tr(\beta_0 W_0 S_j)} \times |W_0|^{\frac{d-d-1}{2}} e^{-\frac{tr(d\Sigma_y W_0)}{2}}$$

$$= |W_0|^{\frac{(K\beta_0+d)-d-1}{2}} e^{-\frac{1}{2}tr\left((\beta_0 \sum_{j=1}^{K} S_j + d\Sigma_y) W_0\right)} \propto |W_0|^{\frac{(K\beta_0+d)-d-1}{2}} e^{-\frac{1}{2}tr(S^{'} W_0)},$$

where

$$S^{'} = \beta_0 \sum_{j=1}^{K} S_j + d\Sigma_y$$

$$\implies W_0|\{S_j\}_{j=1}^K, \beta_0 \sim \mathcal{W}\ (K\beta_0 + d, S'^{-1})$$

$$p(z|\{S_j\}_{j=1}^K, W_0) \overset{(\beta_0 = z+d-1)}{\propto} \prod_{j=1}^{K} \mathcal{W}(S_j|z+d-1, \frac{W_0^{-1}}{z+d-1}) \times \mathcal{IG}(z|1/2, d/2)$$

$$\propto z^{-\frac{3}{2}} e^{-\frac{d}{2z}} |W_0|^{\frac{(z+d-1)K}{2}} \prod_{j=1}^{K} \frac{|S_j|^{\frac{z}{2}} e^{-\frac{z+d-1}{2} tr(W_0 S_j)}}{\prod_{i=0}^{d-1} \Gamma_d\big(\frac{z+i}{2}\big)},$$

where

$$\Gamma_d(\phi) = \pi^{d(d-1)/4} \prod_{i=1}^{d} \Gamma\big(\phi + \frac{i-d}{2}\big)$$

$$
\begin{aligned}
\log(p(z|\{S_j\}_{j=1}^{K}, W_0)) \propto & -\frac{3}{2}\log z - \frac{d}{2z} \\
& + \frac{(z+d-1)K}{2}\log|W_0| + \frac{Kd}{2}(z+d-1)(\log(z+d-1) - \log 2) \\
& + \frac{z}{2}\sum_{j=1}^{K}[\log|S_j| - tr(W_0 S_j)] - K\sum_{i=0}^{d-1}\log(\Gamma_d(\frac{z+i}{2}))
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \log(p(z|\{S_j\}_{j=1}^{K}, W_0))}{\partial z} \propto & -\frac{3}{2z} + \frac{d}{2z^2} + \frac{K}{2}\log|W_0| + \frac{Kd}{2}(\log(z+d-1) - \log 2) \\
& + \frac{Kd}{2} + \frac{1}{2}\sum_{j=1}^{K}[\log|S_j| - tr(W_0 S_j)] - \frac{K}{2}\frac{\partial\Big(\sum_{i=0}^{d-1}\log(\Gamma_d(\frac{z+i}{2}))\Big)}{\partial z}
\end{aligned}
$$

Note. Inverse Gamma and Wishart distributions are log-concave functions. Since the product of log-concave functions is log-concave, therefore the probability density function $p(z|\{S_j\}_{j=1}^{K}, W_0)$, which is proportional to the product of K independent Wishart distributions and an Inverse Gamma distribution, is log-concave. Hence, its logarithm, i.e. $\log(p(z|\{S_j\}_{j=1}^{K}, W_0))$ is concave and ARS method can be applied to sample from the posterior distribution for z (Gilks and Wild, 1992).

A.8.4 Prior probability of the assignments $c_{1:n}$

Using equations (4.5) and (4.6) along with the standard Dirichlet integral, we can integrate out the mixing proportions, $\boldsymbol{\pi}$, and then obtain the probability of a particular set of assignments, $c_{1:n}$, as follows:

$$
\begin{aligned}
p(c_{1:n}|\alpha, n) &= \int p(c_{1:n}|\boldsymbol{\pi})p(\boldsymbol{\pi}|\alpha)d\boldsymbol{\pi} = \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K}\int\prod_{j=1}^{K}\pi_j^{n_j+\alpha/K-1}d\pi_j = \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K}B(\boldsymbol{n}+\boldsymbol{\alpha}) \\
&= \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K}\frac{\prod_{j=1}^{K}\Gamma(n_j+\alpha/K)}{\Gamma(\sum_{j=1}^{K}(n_j+\alpha/K))} = \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)}\prod_{j=1}^{K}\frac{\Gamma(n_j+\alpha/K)}{\Gamma(\alpha/K)}.
\end{aligned}
$$

Note that $n = \sum_{j=1}^{K} n_j$, $\boldsymbol{n} = (n_1, ..., n_k)$ and $\boldsymbol{\alpha} = (\alpha/K, ..., \alpha/K)$.

A.8.5 Conditional prior of the assignment c_i

Keeping all the allocation variables fixed except for a single one, the conditional prior for this allocation given the rest is described by:

$$p(c_i = j|\boldsymbol{c_{-i}}, \alpha) = \frac{p(c_{1:(i-1)}, c_i = j)}{p(c_{1:(i-1)})} = \frac{\frac{\Gamma(\alpha)}{\Gamma(n+\alpha)}\prod_{j=1}^{K}\frac{\Gamma(n_j+\alpha/K)}{\Gamma(\alpha/K)}}{\frac{\Gamma(\alpha)}{\Gamma(n-1+\alpha)}\prod_{j=1}^{K}\frac{\Gamma(n_{-i,j}+\alpha/K)}{\Gamma(\alpha/K)}}$$

$$= \frac{\Gamma(n-1+\alpha)}{\Gamma(n+\alpha)}\prod_{j=1}^{K}\frac{\Gamma(n_j+\alpha/K)}{\Gamma(n_{-i,j}+\alpha/K)}$$

$$\overset{\Gamma(x+1)=x\Gamma(x)}{=} \frac{\Gamma(n-1+\alpha)}{\Gamma(n+\alpha)}\frac{(n_{-i,j}+\alpha/K)\Gamma(n_{-i,j}+\alpha/K)}{\Gamma(n_{-i,j}+\alpha/K)}$$

$$\overset{\Gamma(x)=(x-1)!}{=} \frac{(n+\alpha-2)!}{(n+\alpha-1)!}(n_{-i,j}+\alpha/K) = \frac{n_{-i,j}+\alpha/K}{n-1+\alpha},$$

where the subscript $-i$ denotes all indices except i, and $n_{-i,j}$ is the number of observations excluding the ith observation that are associated with component j; i.e. $n_{-i,j} = n_j - 1$.

A.8.6 Probability distributions of allocations in CRP

In the CRP context, when the number of clusters (tables) approaches the infinity ($K \to \infty$), a new customer i, who enters the restaurant, either randomly chooses to sit at an occupied table (e.g. jth table) with probability proportional to the number of customers they already sit there $n_{-i,j}$.:

$$p(c_i = j|\boldsymbol{c_{-i}}, \alpha) = \frac{n_{-i,j}}{n-1+\alpha},$$

or to sit at the first available currently empty table (e.g. a new table K + 1) with probability proportional to the concentration parameter α/K:

$$p(c_i = K + 1 | \boldsymbol{c_{-i}}, \alpha) = 1 - \frac{\sum_{j=1}^{K} n_{-i,j}}{n - 1 + \alpha} = \frac{\alpha}{n - 1 + \alpha}.$$

A.8.7 Probability distribution of the occupation numbers $n_{1:K}$

The probability of the occupation numbers $\{n_i\}_{i=1}^{K}$, conditioned on α and the number of occupied components K, is derived by equation (4.7) and denotes the likelihood function of α:

$$p(\{n_i\}_{i=1}^{K} | \alpha, K) = \alpha^K \prod_{i=1}^{n} \frac{1}{i - 1 + \alpha} = \frac{\alpha^K \Gamma(\alpha)}{\Gamma(n + \alpha)}.$$

A.8.8 Conditional posterior for the hyperparameter α

The conditional posterior distribution for the hyperparameter α is obtained from the likelihood $p(\{n_i\}_{i=1}^{K} | \alpha, K)$ and the hyperprior $p(\alpha)$ (derived by $p(\alpha^{-1})$ in 4.19) as:

$$p(\alpha | K, n) \propto p(\{n_i\}_{i=1}^{K} | \alpha, K) \times p(\alpha) = \frac{\alpha^K \Gamma(\alpha)}{\Gamma(n + \alpha)} \times \alpha^{-\frac{3}{2}} e^{-\frac{1}{2\alpha}} = \frac{\alpha^{K-\frac{3}{2}} e^{-\frac{1}{2\alpha}} \Gamma(\alpha)}{\Gamma(n + \alpha)}$$

$$\log(p(u | K, n)) \overset{(u = \log(\alpha))}{\propto} (K - \frac{3}{2})u - \frac{1}{2e^u} + \log \Gamma(e^u) - \log \Gamma(n + e^u) + u.$$

Note. Since the distribution $p(\log(\alpha) | K, n)$ is log-concave, the distribution $\log(p(\log(\alpha) | K, n))$ is concave, thus the adaptive rejection sampling technique can be applied to draw independent samples from this posterior distribution for $\log(\alpha)$ (Gilks and Wild, 1992).

A.8.9 Conditional posterior of the assignment c_i

To define the conditional posterior density of a single indicator c_i, we multiply the likelihood function given the rest indicators $\boldsymbol{c_{-i}}$ from equation (4.3) with the prior distribution from equations (4.9) and (4.10). Therefore, when $n_{-i,j} > 0$ the conditional posterior is:

$$\begin{aligned} p(c_i = j|\boldsymbol{c_{-i}}, \boldsymbol{\mu_j}, S_j, \alpha) &\propto p(c_i = j|\boldsymbol{c_{-i}}, \alpha) \times p_j(\boldsymbol{y_i}|\boldsymbol{\mu_j}, S_j, \boldsymbol{c_{-i}}) \\ &\propto \frac{n_{-i,j}}{n-1+\alpha} \times \mathcal{N}_d(\boldsymbol{y_i}|\boldsymbol{\mu_j}, S_j^{-1}) \\ &\propto \frac{n_{-i,j}}{n-1+\alpha} \times |S_j|^{\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{y_i}-\boldsymbol{\mu_j})^T S_j (\boldsymbol{y_i}-\boldsymbol{\mu_j})}, \end{aligned} \tag{A.1}$$

otherwise

$$\begin{aligned} p(c_i = K+1|\boldsymbol{c_{-i}}, \boldsymbol{\mu_0}, S_0, \beta_0, W_0, \alpha) &\propto p(c_i = K+1|\boldsymbol{c_{-i}}, \alpha) \times \int p_j(\boldsymbol{y_i}|\boldsymbol{\mu_j}, S_j) G_0(\boldsymbol{\mu_j}, S_j) d\boldsymbol{\mu_j} dS_j \\ &\propto \frac{\alpha}{n-1+\alpha} \times \int \mathcal{N}_d(\boldsymbol{y_i}|\boldsymbol{\mu_j}, S_j^{-1}) \mathcal{N}_d(\boldsymbol{\mu_j}|\boldsymbol{\mu_0}, S_0^{-1})\ \mathcal{W}(S_j|\beta_0, (\beta_0 W_0)^{-1}) d\boldsymbol{\mu_j} dS_j. \end{aligned} \tag{A.2}$$

A.9 Technical details for Inference of DP-GMMx

A.9.1 Conditional posterior distributions for the coefficients $\boldsymbol{q_j}$s and $\boldsymbol{r_j}$s

Let $\boldsymbol{x_i}$ be the $(p+1)$-dimensional vector including the covariate information, where p denotes the number of covariates, and $\boldsymbol{y_i}$ is the d-dimensional response variable for the ith unit. The conditional posterior distributions for the effect due to the p covariates $\boldsymbol{q_j}$s are derived by the multiplication of the likelihood given the variable $\boldsymbol{c}$ of indicators from equation (4.3) and the prior for $\boldsymbol{q_j}$s from

$$G_0(\boldsymbol{q_j}, \boldsymbol{r_j}, S_j) = \mathcal{N}_{(p+1)}(\boldsymbol{q_j}|\boldsymbol{\mu_q}, S_q^{-1})\ \mathcal{N}_d(\boldsymbol{r_j}|\boldsymbol{\mu_r}, S_r^{-1})\ \mathcal{W}(S_j|\beta_0, (\beta_0 W_0)^{-1}) \tag{A.3}$$

as follows:

$$p(\boldsymbol{q_j}|\boldsymbol{c}, S_j, \boldsymbol{\mu_q}, S_q, \boldsymbol{x}, \boldsymbol{y}) \propto \prod_{i=1}^{n_j} \mathcal{N}_d(\boldsymbol{y_i}|\boldsymbol{\mu_{ij}} = \boldsymbol{q_j}\boldsymbol{x_i}^T\boldsymbol{r_j}, S_j^{-1}) \times \mathcal{N}_{(p+1)}(\boldsymbol{q_j}|\boldsymbol{\mu_q}, S_q^{-1})$$

$$\overset{1}{=} \prod_{i=1}^{n_j} \mathcal{N}_d(\boldsymbol{y_i}|\boldsymbol{\lambda_i}, \Lambda_i) \times \mathcal{N}_{(p+1)}(\boldsymbol{q_j}|\boldsymbol{\xi}, \Xi) \propto \mathcal{N}_{(p+1)}(\boldsymbol{q_j}|\boldsymbol{\xi}, \Xi)$$

Initialise: $\Xi_0^{-1} = S_q$; $\boldsymbol{\xi_0} = \boldsymbol{\mu_q}$; $\boldsymbol{z_0} = \boldsymbol{\mu_q} S_q$

Iterate: $\Lambda_i = S_j^{-1} + M_i^T \Xi_{i-1} M_i$, where $M_i = \boldsymbol{x_i}^T\boldsymbol{r_j}$ and $i = 1, \ldots, n_j$

$$\boldsymbol{\lambda_i} = \boldsymbol{\xi_{i-1}} M_i$$

$$\Xi_i^{-1} = \Xi_{i-1}^{-1} + M_i S_j M_i^T$$

$$\boldsymbol{z_i} = \boldsymbol{z_{i-1}} + \boldsymbol{y_i} S_j M_i^T$$

$$\boldsymbol{\xi_i} = \boldsymbol{z_i} \Xi_i$$

Finish: $\Xi = \Xi_{n_j}$ and $\boldsymbol{\xi} = \boldsymbol{\xi_{n_j}}$

The conditional posterior distributions for the effect due to the d response variables $\boldsymbol{r_j}$s are derived by the multiplication of the likelihood given the variable $\boldsymbol{c}$ of indicators from equation (4.3) and the prior for $\boldsymbol{r_j}$s from equation (A.3) as follows:

[1] Make use of the mathematical derivations made in Hogg et al. (2020).

$$\begin{aligned}
p(\boldsymbol{r_j}|\boldsymbol{c}, S_j, \boldsymbol{\mu_r}, S_r, \boldsymbol{x}, \boldsymbol{y}) &\propto \prod_{i=1}^{n_j} \mathcal{N}_d(\boldsymbol{y_i}|\boldsymbol{\mu_{ij}} = \boldsymbol{r_j}^T \boldsymbol{q_j} \boldsymbol{x_i}^T, S_j^{-1}) \times \mathcal{N}_d(\boldsymbol{r_j}|\boldsymbol{\mu_r}, S_r^{-1}) \\
&\overset{a_i = \boldsymbol{q_j}\boldsymbol{x_i}^T}{=} \prod_{i=1}^{n_j} e^{-\frac{1}{2}(\boldsymbol{y_i} - a_i \boldsymbol{r_j})^T S_j (\boldsymbol{y_i} - a_i \boldsymbol{r_j})} \times e^{-\frac{1}{2}(\boldsymbol{r_j} - \boldsymbol{\mu_r})^T S_r (\boldsymbol{r_j} - \boldsymbol{\mu_r})} \\
&\propto e^{-\frac{1}{2}\sum_{i=1}^{n_j}(\boldsymbol{y_i} - a_i \boldsymbol{r_j})^T S_j (\boldsymbol{y_i} - a_i \boldsymbol{r_j})} \times e^{-\frac{1}{2}(\boldsymbol{r_j}^T S_r \boldsymbol{r_j} - 2\boldsymbol{r_j}^T S_r \boldsymbol{\mu_r})} \\
&= e^{-\frac{1}{2}[-2\sum_{i=1}^{n_j}(a_i \boldsymbol{r_j})^T S_j \boldsymbol{y_i}^T + \sum_{i=1}^{n_j}(a_i \boldsymbol{r_j})^T S_j (a_i \boldsymbol{r_j})]} \times e^{-\frac{1}{2}[\boldsymbol{r_j}^T A_0 \boldsymbol{r_j} - 2\boldsymbol{r_j}^T b_0]} \\
&= e^{-\frac{1}{2}[-2\sum_{i=1}^{n_j} \boldsymbol{r_j}^T a_i^T S_j \boldsymbol{y_i}^T + \sum_{i=1}^{n_j} \boldsymbol{r_j}^T a_i^T S_j a_i \boldsymbol{r_j}]} \times e^{-\frac{1}{2}[\boldsymbol{r_j}^T A_0 \boldsymbol{r_j} - 2\boldsymbol{r_j}^T b_0]} \\
&= e^{-\frac{1}{2}[-2\boldsymbol{r_j}^T \sum_{i=1}^{n_j} a_i^T S_j \boldsymbol{y_i}^T + \boldsymbol{r_j}^T \sum_{i=1}^{n_j} a_i^T S_j a_i \boldsymbol{r_j}]} \times e^{-\frac{1}{2}[\boldsymbol{r_j}^T A_0 \boldsymbol{r_j} - 2\boldsymbol{r_j}^T b_0]} \\
&= \exp\bigg(- \frac{1}{2}(-2\boldsymbol{r_j}^T b_1 + \boldsymbol{r_j}^T A_1 \boldsymbol{r_j})\bigg) \times \exp\bigg(- \frac{1}{2}[\boldsymbol{r_j}^T A_0 \boldsymbol{r_j} - 2\boldsymbol{r_j}^T b_0]\bigg) \\
&= \exp\bigg(\boldsymbol{r_j}^T (b_0 + b_1) - \frac{1}{2}\boldsymbol{r_j}^T (A_0 + A_1)\boldsymbol{r_j})\bigg) = \exp\bigg(\boldsymbol{r_j}^T b_n - \frac{1}{2}\boldsymbol{r_j}^T A_n \boldsymbol{r_j})\bigg),
\end{aligned}$$

where

$$b_0 = S_r \boldsymbol{\mu_r}, \quad A_0 = S_r, \quad b_1 = S_j \sum_{i=1}^{n_j} a_i^T \boldsymbol{y_i}^T, \quad A_1 = S_j \sum_{i=1}^{n_j} a_i^2$$

$$b_n = b_0 + b_1 = \boldsymbol{\mu_r} + S_j \sum_{i=1}^{n_j} a_i^T \boldsymbol{y_i}^T, \quad A_n = A_0 + A_1 = S_r + S_j \sum_{i=1}^{n_j} a_i^2$$

$$\implies \boldsymbol{r_j}|\boldsymbol{c}, S_j, \boldsymbol{\mu_r}, S_r, \boldsymbol{x}, \boldsymbol{y} \sim \mathcal{N}_d\bigg(\mu_n = A_n^{-1} b_n, \Sigma_n = A_n^{-1}\bigg) \ \forall\ j = 1, ..., K,$$

where

$$\mu_n = A_n^{-1} b_n = (S_r + S_j \sum_{i=1}^{n_j} a_i^2)^{-1}(\boldsymbol{\mu_r} + S_j \sum_{i=1}^{n_j} a_i^T \boldsymbol{y_i}^T)$$

$$\Sigma_n = A_n^{-1} = (S_r + S_j \sum_{i=1}^{n_j} a_i^2)^{-1}$$

A.9.2 Conditional posterior distributions for the hyperparameters $\boldsymbol{\mu_q}$, S_q, $\boldsymbol{\mu_r}$ and S_r

The conditional posterior distributions for the hyperparameters $\boldsymbol{\mu_q}$, S_q, $\boldsymbol{\mu_r}$ and S_r are obtained from the corresponding distributions in equation (A.3), which acts as likelihood, and the hyperpriors $\boldsymbol{\mu_q} \sim \mathcal{N}_p(\boldsymbol{\mu_q}|\boldsymbol{\mu_0}, \Sigma_0)$, $S_q \sim \mathcal{W}(S_q|p, (p\Sigma_0)^{-1})$, $\boldsymbol{\mu_r} \sim \mathcal{N}_d(\boldsymbol{\mu_r}|\boldsymbol{\mu_y}, \Sigma_y)$ and $S_r \sim \mathcal{W}(S_r|d, (d\Sigma_y)^{-1})$, for all hyperparameters respectively, as:

$$p(\boldsymbol{\mu_q}|\{\boldsymbol{q_j}\}_{j=1}^{K}, S_q) \propto \prod_{j=1}^{K} \mathcal{N}_p(\boldsymbol{q_j}|\boldsymbol{\mu_q}, S_q^{-1}) \times \mathcal{N}_p(\boldsymbol{\mu_q}|\boldsymbol{\mu_0}, \Sigma_0)$$

$$\propto \prod_{j=1}^{K} e^{-\frac{1}{2}(\boldsymbol{q_j}-\boldsymbol{\mu_q})^T S_q(\boldsymbol{q_j}-\boldsymbol{\mu_q})} \times e^{-\frac{1}{2}(\boldsymbol{\mu_q}-\boldsymbol{\mu_0})^T \Sigma_0^{-1}(\boldsymbol{\mu_q}-\boldsymbol{\mu_0})}$$

$$\propto e^{-\frac{1}{2}(-2\boldsymbol{\mu_q}^T S_q \sum_{j=1}^{K} \boldsymbol{q_j} + K\boldsymbol{\mu_q}^T S_q \boldsymbol{\mu_q})} \times e^{-\frac{1}{2}(\boldsymbol{\mu_q}^T \Sigma_0^{-1}\boldsymbol{\mu_q} - 2\boldsymbol{\mu_q}^T \Sigma_0^{-1}\boldsymbol{\mu_0})}$$

$$= \exp\left(-\frac{1}{2}(-2\boldsymbol{\mu_q}^T b_1 + \boldsymbol{\mu_q}^T A_1 \boldsymbol{\mu_q})\right) \times \exp\left(-\frac{1}{2}(\boldsymbol{\mu_q}^T A_0 \boldsymbol{\mu_q} - 2\boldsymbol{\mu_q}^T b_0\right)$$

$$= \exp\left(\boldsymbol{\mu_q}^T(b_0 + b_1) - \frac{1}{2}\boldsymbol{\mu_q}^T(A_0 + A_1)\boldsymbol{\mu_q})\right) = \exp\left(\boldsymbol{\mu_q}^T b_n - \frac{1}{2}\boldsymbol{\mu_q}^T A_n \boldsymbol{\mu_q})\right),$$

where

$$b_0 = \Sigma_0^{-1}\boldsymbol{\mu_0}, \quad \boldsymbol{\mu_0} = \mathbf{1}_{(1\times(p+1))}, \quad \Sigma_0 = \begin{bmatrix} 100 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 100 \end{bmatrix}$$

$$A_0 = \Sigma_0^{-1}, \quad b_1 = K\bar{\boldsymbol{q}}S_q, \quad \bar{\boldsymbol{q}} = (\frac{1}{K}\sum_{j=1}^{K} \boldsymbol{q_{j[1]}}, ..., \frac{1}{K}\sum_{j=1}^{K} \boldsymbol{q_{j[p]}})$$

$$A_1 = KS_q, \quad b_n = b_0 + b_1 = \Sigma_0^{-1}\boldsymbol{\mu_0} + K\bar{\boldsymbol{q}}S_;, \quad A_n = A_0 + A_1 = \Sigma_0^{-1} + KS_q$$

$$\implies \boldsymbol{\mu_q}|\{\boldsymbol{q_j}\}_{j=1}^{K}, S_q \sim \mathcal{N}_p\Bigg(\mu_n = A_n^{-1}b_n, \Sigma_n = A_n^{-1}\Bigg),$$

where

$$\mu_n = A_n^{-1}b_n = (\Sigma_0^{-1} + KS_q)^{-1}(\Sigma_0^{-1}\boldsymbol{\mu_0} + K\bar{\boldsymbol{q}}S_q)$$

$$\Sigma_n = A_n^{-1} = (\Sigma_0^{-1} + KS_q)^{-1}$$

$$p(S_q|\{\boldsymbol{q_j}\}_{j=1}^{K}, \boldsymbol{\mu_q}) \propto \prod_{j=1}^{K}\mathcal{N}_p(\boldsymbol{q_j}|\boldsymbol{\mu_q}, S_q^{-1}) \times \mathcal{W}(S_q|p, (p\Sigma_0)^{-1})$$

$$\propto \prod_{j=1}^{K}|S_q^{-1}|^{-\frac{1}{2}}e^{-\frac{1}{2}(\boldsymbol{q_j}-\boldsymbol{\mu_q})^T S_q(\boldsymbol{q_j}-\boldsymbol{\mu_q})} \times |S_q|^{\frac{p-p-1}{2}}e^{-\frac{tr(p\Sigma_0 S_q)}{2}}$$

$$\propto \left(\frac{1}{|S_q|}\right)^{-\frac{K}{2}} e^{-\frac{1}{2}(\sum_{j=1}^{K}(\boldsymbol{q_j}-\boldsymbol{\mu_q})^T S_q(\boldsymbol{q_j}-\boldsymbol{\mu_q})} \times |S_q|^{\frac{p-p-1}{2}}e^{-\frac{1}{2}tr(p\Sigma_0 S_q)}$$

$$\propto |S_q|^{\frac{(K+p)-p-1}{2}}e^{-\frac{1}{2}tr(S^{'}S_q)},$$

where

$$S^{'} = \sum_{j=1}^{K}(\boldsymbol{q_j} - \boldsymbol{\mu_q})(\boldsymbol{q_j} - \boldsymbol{\mu_q})^T + p\Sigma_0$$

$$\implies S_q|\{\boldsymbol{q_j}\}_{j=1}^{K}, \mu_q \sim \mathcal{W}\ (K+p, S'^{-1})$$

Similarly, the conditional posteriors of μ_r and S_r are given below:

$$p(\boldsymbol{\mu_r}|\{\boldsymbol{r_j}\}_{j=1}^{K}, S_r) \propto \prod_{j=1}^{K} \mathcal{N}_d(\boldsymbol{r_j}|\boldsymbol{\mu_r}, S_r^{-1}) \times \mathcal{N}_d(\boldsymbol{\mu_r}|\boldsymbol{\mu_y}, \Sigma_y)$$

$$\propto \exp\left(\boldsymbol{\mu_r}^T(b_0 + b_1) - \frac{1}{2}\boldsymbol{\mu_r}^T(A_0 + A_1)\boldsymbol{\mu_r})\right) = \exp\left(\boldsymbol{\mu_r}^T b_n - \frac{1}{2}\boldsymbol{\mu_r}^T A_n \boldsymbol{\mu_r})\right),$$

where

$$b_0 = \Sigma_y^{-1}\boldsymbol{\mu_y}, \quad A_0 = \Sigma_y^{-1}, \quad b_1 = K\bar{\boldsymbol{r}}S_r, \quad \bar{\boldsymbol{r}} = (\frac{1}{K}\sum_{j=1}^{K}\boldsymbol{r_{j[1]}}, ..., \frac{1}{K}\sum_{j=1}^{K}\boldsymbol{r_{j[d]}})$$

$$A_1 = KS_r, \quad b_n = b_0 + b_1 = \Sigma_y^{-1}\boldsymbol{\mu_y} + K\bar{\boldsymbol{r}}S_r, \quad A_n = A_0 + A_1 = \Sigma_y^{-1} + KS_r$$

$$\Longrightarrow \boldsymbol{\mu_r}|\{\boldsymbol{r_j}\}_{j=1}^{K}, S_r \sim \mathcal{N}_d\left(\mu_n = A_n^{-1}b_n, \Sigma_n = A_n^{-1}\right),$$

where

$$\mu_n = A_n^{-1}b_n = (\Sigma_y^{-1} + KS_r)^{-1}(\Sigma_y^{-1}\boldsymbol{\mu_y} + K\bar{\boldsymbol{r}}S_r)$$

$$\Sigma_n = A_n^{-1} = (\Sigma_y^{-1} + KS_r)^{-1}$$

$$p(S_r|\{\boldsymbol{r_j}\}_{j=1}^{K}, \boldsymbol{\mu_r}) \propto \prod_{j=1}^{K} \mathcal{N}_d(\boldsymbol{r_j}|\boldsymbol{\mu_r}, S_r^{-1}) \times \mathcal{W}(S_r|d, (d\Sigma_y)^{-1}) \propto |S_r|^{\frac{(K+d)-d-1}{2}} e^{-\frac{1}{2}tr(S^{'}S_r)},$$

where

$$S^{'} = \sum_{j=1}^{K}(\boldsymbol{r_j} - \boldsymbol{\mu_r})(\boldsymbol{r_j} - \boldsymbol{\mu_r})^T + d\Sigma_y$$

$$\Longrightarrow S_r|\{\boldsymbol{r_j}\}_{j=1}^{K}, \mu_r \sim \mathcal{W}\ (K + d, S'^{-1}).$$

The rest hyperparameters maintain the same conditional posteriors of the DP-GMM without covariates.

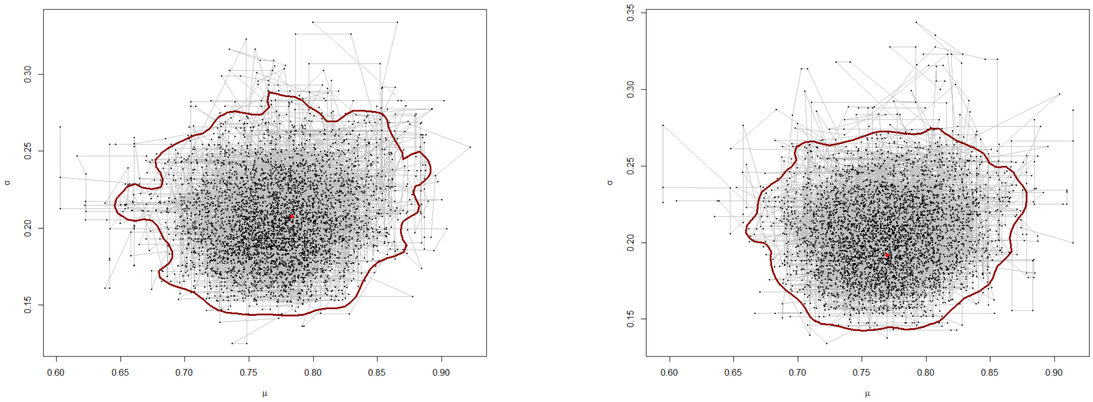
A.9.3 Predictive distribution of DP-GMMx

$$p(\tilde{\boldsymbol{y}}|\boldsymbol{y}) \approx \frac{1}{T}\sum_{t=1}^{T}\big[\frac{\alpha^{(t)}}{n+\alpha^{(t)}}\mathcal{N}_d(\tilde{\boldsymbol{y}}|\boldsymbol{\mu_r^{(t)}},\, S_r^{-1\,(t)}) + \sum_{j=1}^{K^{(t)}}\frac{n_j}{n+\alpha^{(t)}}\mathcal{N}_d(\tilde{\boldsymbol{y}}|\boldsymbol{\mu_{ij}^{(t)}},\, S_j^{-1\,(t)})\big],$$

where $\boldsymbol{\mu_{ij}^{(t)}} = \boldsymbol{q_j^{(t)}}\boldsymbol{x_i^T}\boldsymbol{r_j^{(t)}}$

Appendix B

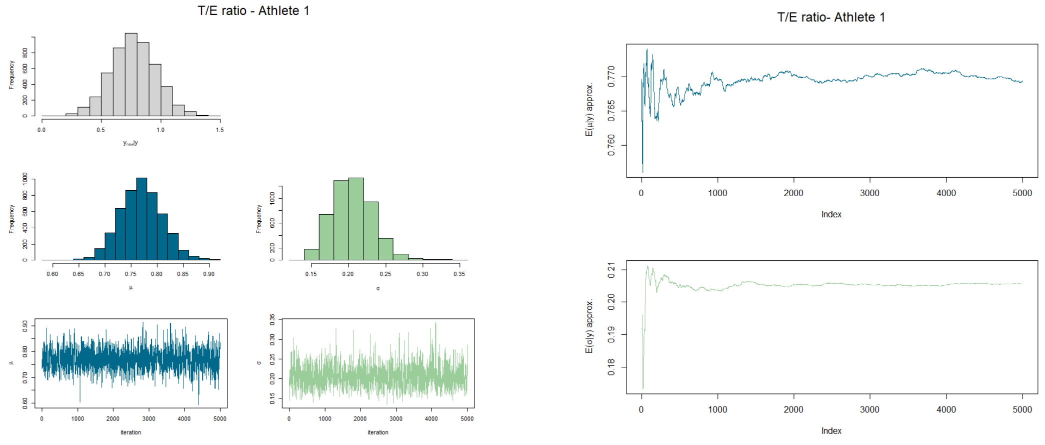
Figures



(a)

(b)

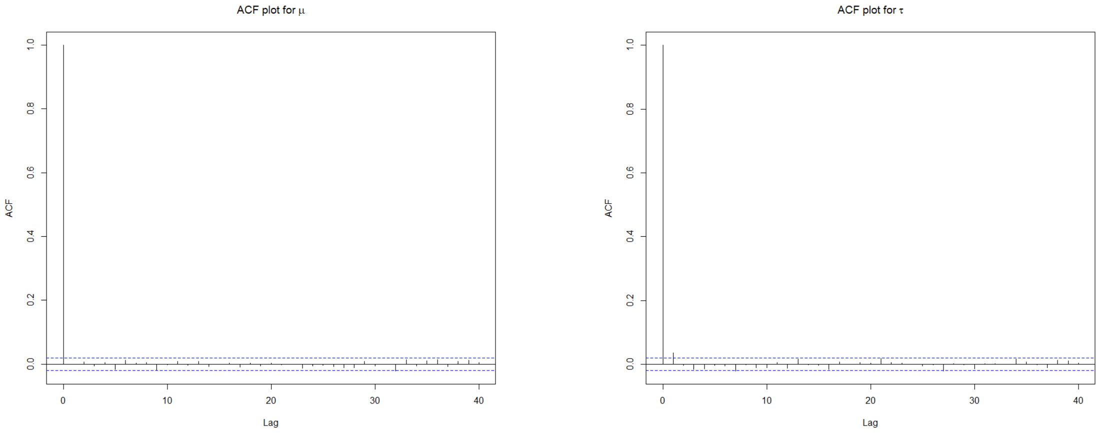
FIGURE B.1: Scatter plot of μ and σ and the 95% highest posterior region using the sampling algorithms a) MH and b) MWG and informative priors for T/E ratio of athlete 1. The red dot denotes the starting point, and a two-dimensional 95% credible region is displayed by the dark red line.



(a)

(b)

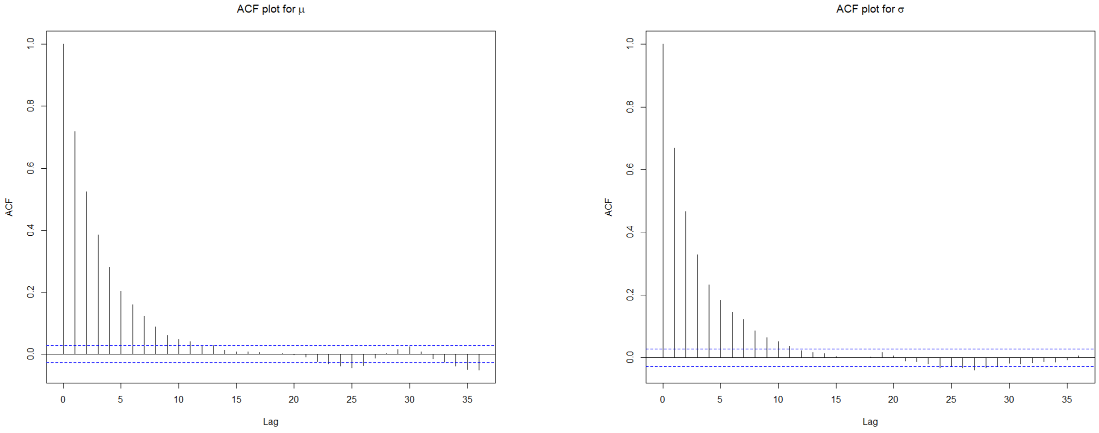
Figure B.2: (a) Histograms, traceplots and (b) running averages of the parameters μ and σ over the iterations using the MWG sampling algorithm and the informative priors from Sottas et al. (2006) for T/E of athlete 1. The posterior means for μ and σ can be approximated by the averages $E[\mu|\boldsymbol{y}] \approx \frac{1}{N}\sum_{i=1}^{N}\mu^{(i)}$ and $E[\sigma|\boldsymbol{y}] \approx \frac{1}{N}\sum_{i=1}^{N}\sigma^{(i)}$, respectively.



(a)

(b)

Figure B.3: Autocorrelation plot of μ and τ parameters - Gibbs sampler for the univariate model.



(a)

(b)

FIGURE B.4: Autocorrelation plot of μ and σ parameters - MWG sampler for the T/E model.

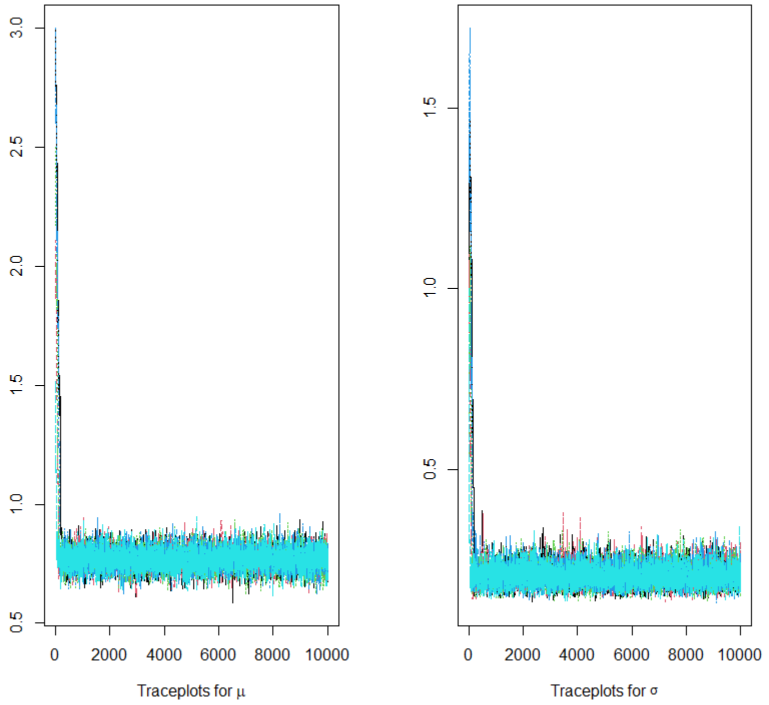
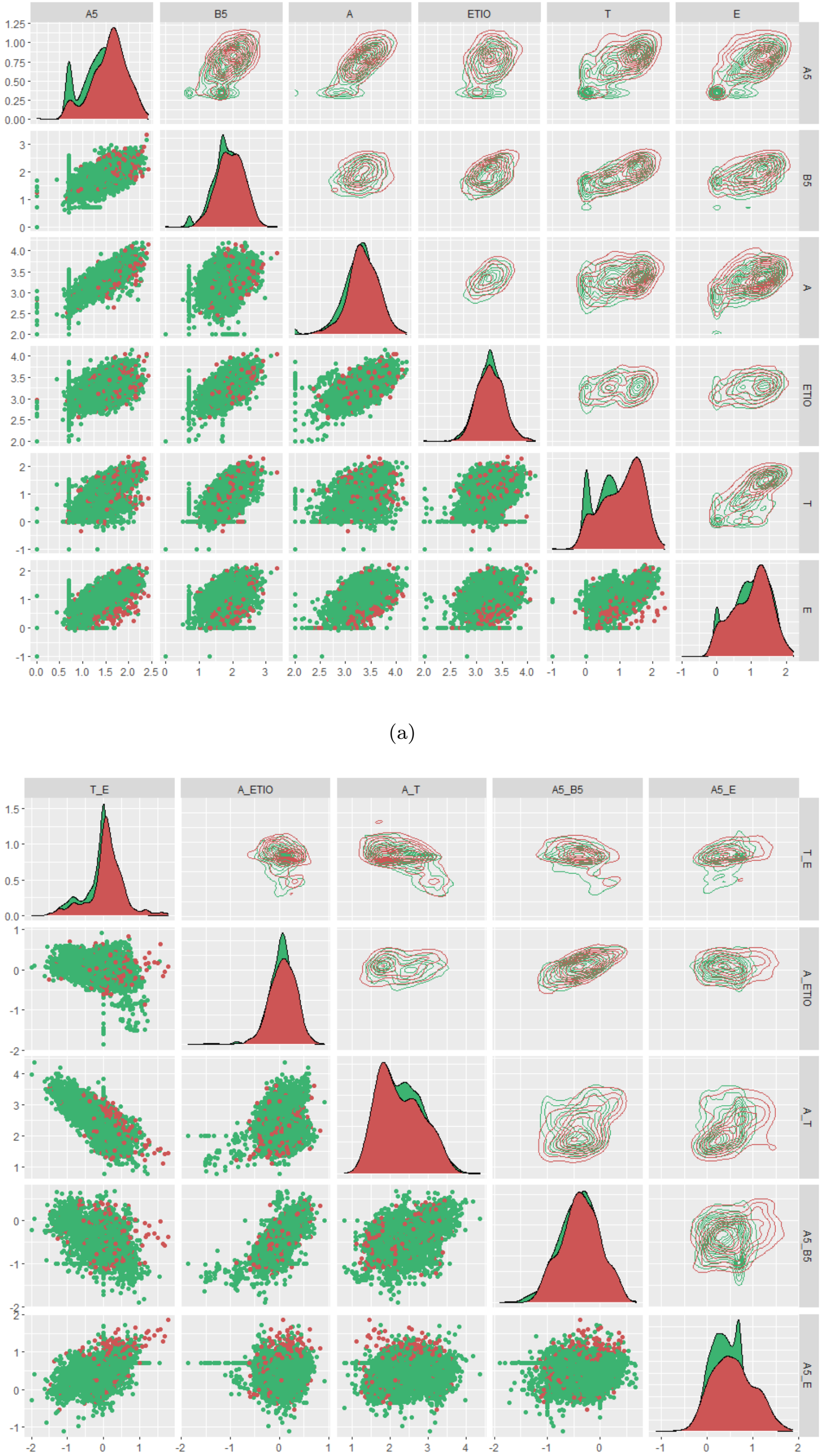


FIGURE B.5: Traceplots from five chains of Metropolis-Hastings algorithm for parameters μ and σ .



(a)

(b)

Figure B.6: Pairs plot of normal (0: pastel green) vs abnormal (1: pastel red) samples including the scatter, density and contour plots for (a) the six markers, and (b) their five ratios in the logarithmic scale.

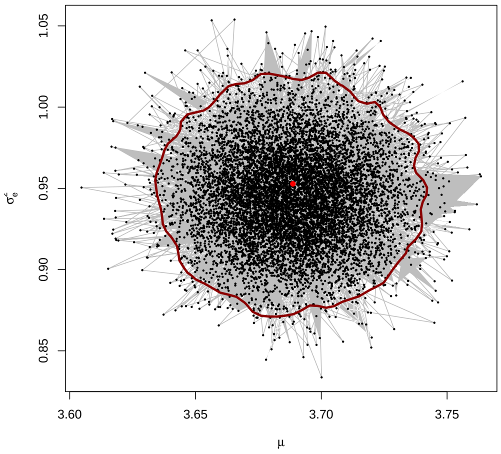
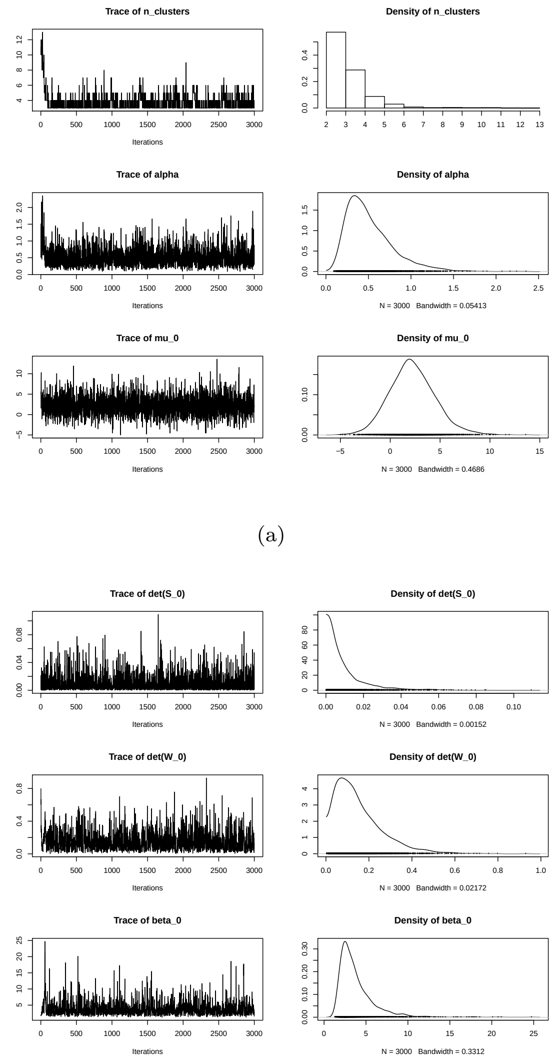


FIGURE B.7: Scatter plot of μ and σ_e^2 and the 95% highest posterior region using the Gibbs sampler and informative priors.



(a)

(b)

Figure B.8: Traceplots and density plots of a) number of clusters, concentration parameter α and hyperparameter μ_0, and b) $|S_0|$, $|W_0|$ and β_0 to verify the good mixing of Algorithm 2.

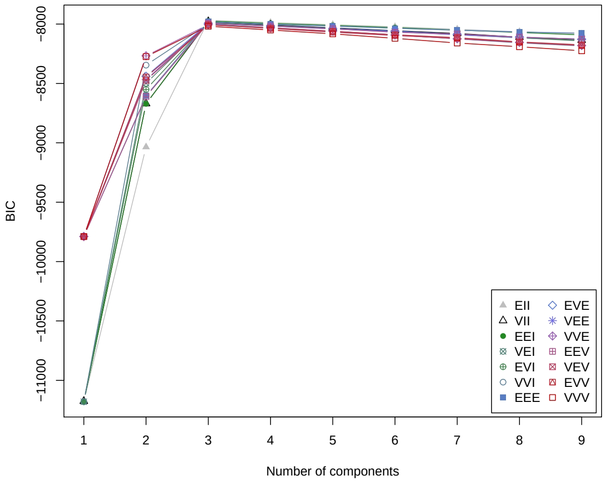
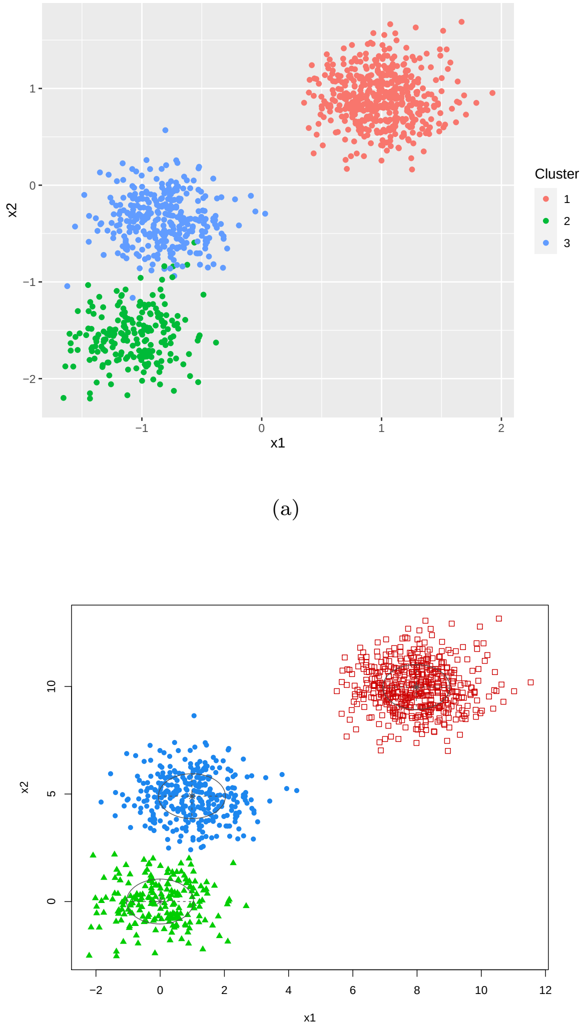


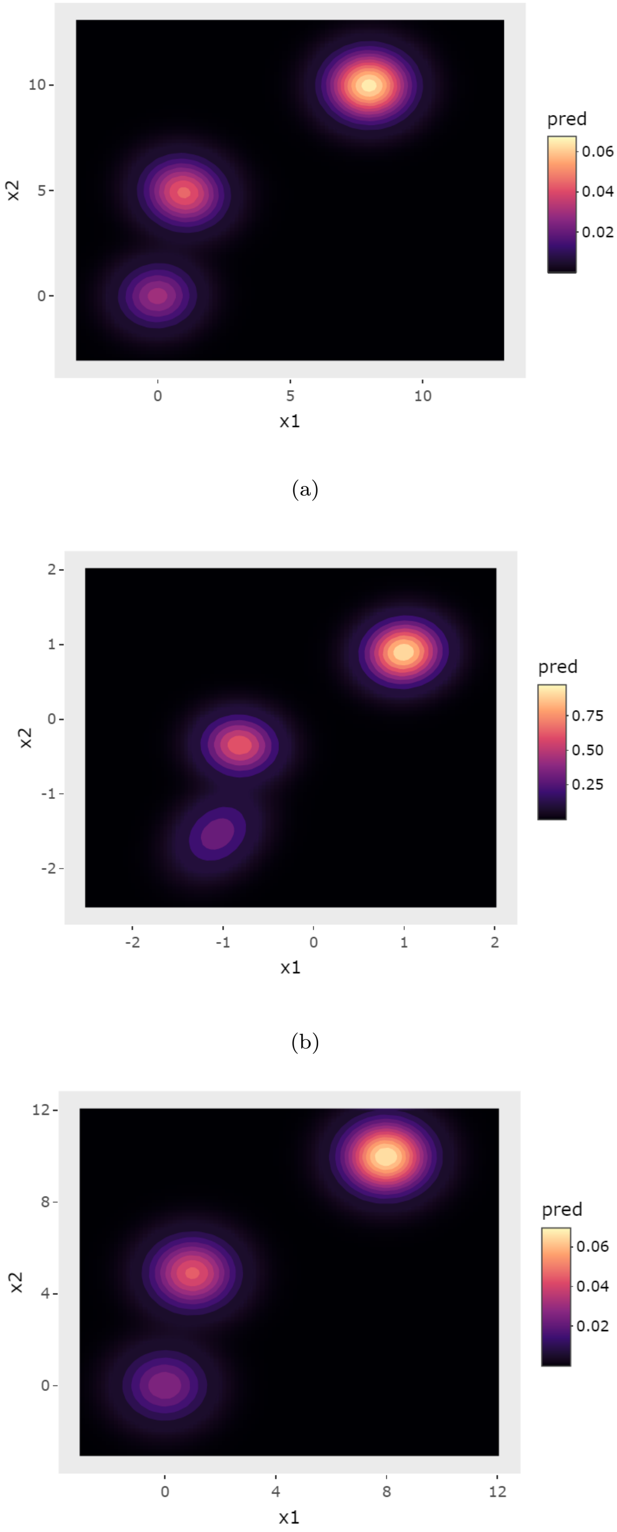
Figure B.9: The number of mixing components and the covariance parameterisation are selected using the Bayesian Information Criterion (BIC) with the mclust package. This is a plot of the BIC traces for all the models considered. There is a clear indication of a three-component mixture from all models.



(a)

(b)

Figure B.10: Scatter plots of the three mixture densities estimated with (a) Diricletprocess and (b) mclust methods on the simulated dataset.

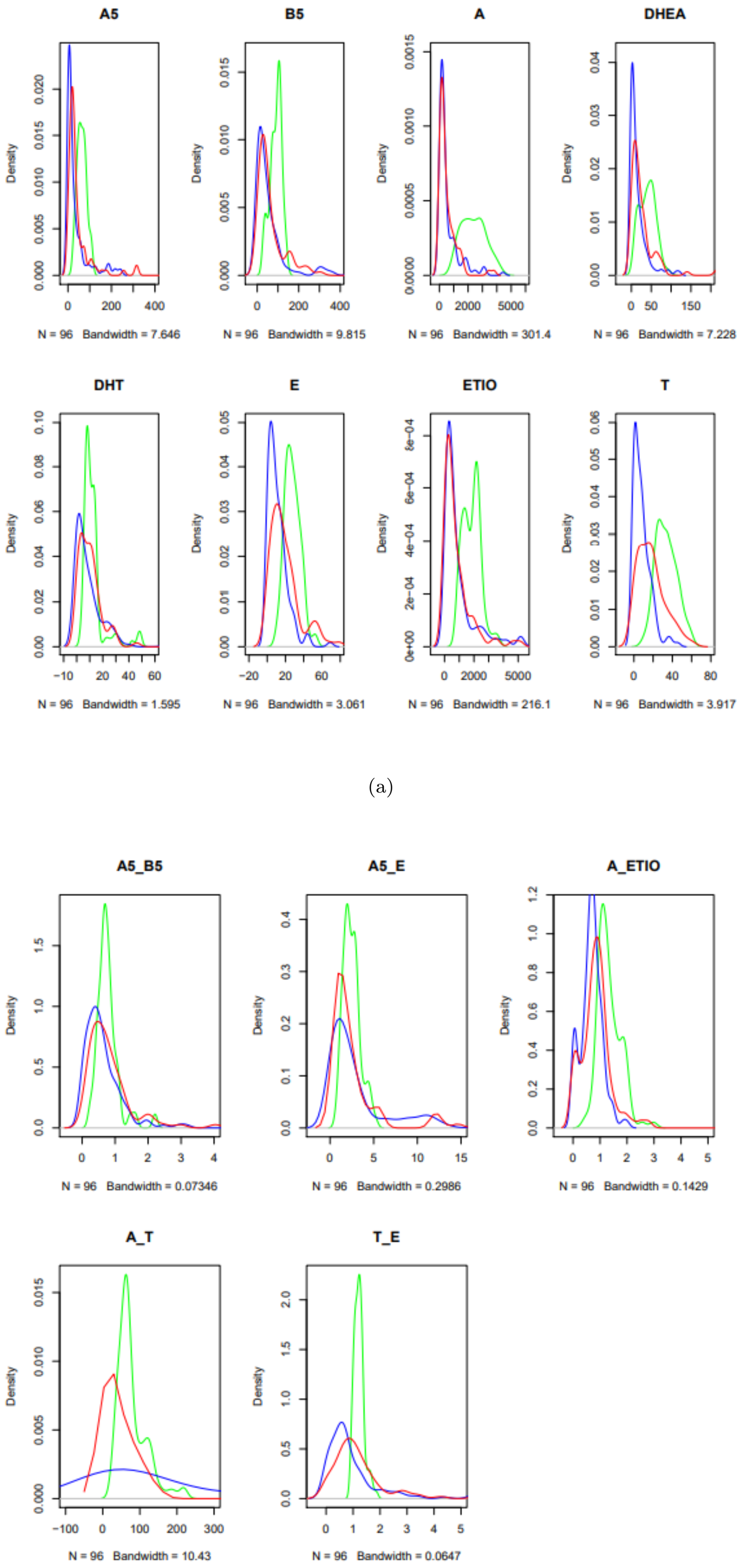


(a)

(b)

(c)

Figure B.11: Predictive densities of simulated data modelled by (a) DPGMM, (b) Dirichletprocess, and (c) mclust packages.



(a)

(b)

FIGURE B.12: Density estimates of a) biomarkers and b) ratios for each class; healthy males (green), males with BPH (blue), and males confirmed with PCa (red).

Appendix C

Tables

Table C.1: Urinary steroid metabolites quantitated in this study.

Trivial name	Abbreviation	Systematic name	Formula
5α-Adiol	A5	5α-Androstane-3α,17β-diol	$C_{19}H_{32}O_2$
5β-Adiol	B5	5β-Androstane-3α,17β-diol	$C_{19}H_{32}O_2$
Androsterone	A	3α-Hydroxy-5α-androstan-17-one	$C_{19}H_{30}O_2$
Epitestosterone	E	17α-Hydroxy-androst-4-en-3-one	$C_{19}H_{28}O_2$
Etiocholanolone	ETIO	3α-Hydroxy-5β-androstan-17-one	$C_{19}H_{30}O_2$
Testosterone	T	17β-Hydroxy-androst-4-en-3-one	$C_{19}H_{28}O_2$
5α-Adiol/5β-Adiol	A5/B5		
5α-Adiol/Epitestosterone	A5/E		
Androsterone/Etiocholanolone	A/ETIO		
Androsterone/Testosterone	A/T		
Testosterone/Epitestosterone	T/E		

Table C.2: Descriptive summaries (min, inter-quartile range; IQ1 and IQ3, mean, median and max) of the metabolites and ratios of 100 athletes with normal samples.

Target Metabolite	Min (ng/mL)	IQ1 (ng/mL)	Mean (ng/mL)	Median (ng/mL)	IQ3 (ng/mL)	Max (ng/mL)
A5	1	11	35.08	25	49	210
B5	1	34	96.25	64	120	910
A	100	1,200	2, 097	1, 800	2,600	16, 000
E	0.10	4	17.48	11	26	130
ETIO	98	1,100	1, 863	1, 700	2,400	7, 200
T	0.10	3.30	20.17	9.60	31	150
A5/B5	0.013	0.24	0.49	0.40	0.61	4.8
A5/E	0.13	1.41	4.77	2.54	4.65	160
A/ETIO	0.06	0.80	1.21	1.11	1.48	8.16
A/T	6.25	72.41	365.25	159.46	412.7	23, 000
T/E	0.012	0.75	1.35	1.0	1.6	13

Table C.3: Descriptive summaries (min, inter-quartile range; IQ1 and IQ3, mean, median and max) of the metabolites and ratios of 100 athletes with atypical samples.

Target Metabolite	Min (ng/mL)	IQ1 (ng/mL)	Mean (ng/mL)	Median (ng/mL)	IQ3 (ng/mL)	Max (ng/mL)
A5	1.0	14	36.61	28.0	50	250
B5	3.3	43	105.1	73	140	1, 400
A	100	1,200	2,318	2, 000	3,000	13, 000
E	0.10	4.50	18.53	12.0	27	160
ETIO	270	1,300	2,199	1, 900	2,700	14, 000
T	0.10	3.3	19.44	9.25	33	150
A5/B5	0.02	0.19	0.48	0.35	0.65	4.6
A5/E	0.08	1.42	3.38	2.37	4.6	53.33
A/ETIO	0.014	0.7	1.16	1.04	1.46	6.91
A/T	6.252	67.66	454.63	183.33	525.65	9, 200
T/E	0.01	0.62	1.46	1	1.64	35

Table C.4: Descriptive summaries (min, inter-quartile range; IQ1 and IQ3, mean, median and max) of the metabolites and ratios of 29 athletes with abnormal samples.

Target Metabolite	Min (ng/mL)	IQ1 (ng/mL)	Mean (ng/mL)	Median (ng/mL)	IQ3 (ng/mL)	Max (ng/mL)
A5	3.7	21	50.53	39.5	65	270
B5	4.9	24	82.69	50	97	2, 200
A	210	1,600	3, 152	2, 400	3,900	16, 000
E	1	5.90	14.41	11	19	120
ETIO	100	932.5	1, 747	1, 500	2,175	11, 000
T	0.50	1	12.51	3.50	15	220
A5/B5	0.085	0.47	0.76	1	1.42	4.2
A5/E	0.46	1.83	5.59	3.55	5.91	72.22
A/ETIO	0.50	1.17	1.92	1.69	2.53	5.08
A/T	13.18	177.55	1, 262.99	632.46	1,800	14, 828
T/E	0.03	0.15	1.58	0.34	1.3	61.11

Table C.5: Maximum EAAS values and their ratios measured in a Caucasian population consisting of 2,027 male (M) and 1,004 female (F) athletes (Van Renterghem et al., 2010) along with the available WADA's threshold limits by gender (WADA, 2018; WADA, 2021b). Plausible initial limits (IL) have been assumed for the ratios with no information regarding the population.

Target Metabolite	Max (M) (ng/mL)	Max (F) (ng/mL)	WADA TD 2021 (M) (ng/mL)	WADA TD 2021 (F) (ng/mL)
A5	652	263	250	150
B5	1,260	471		
A	20,700	17,500	10,000	10,000
E	391	51.9	200	50
ETIO	11,400	9,030	10,000	10,000
T	249	219	200	50
A/ETIO			4	4
T/E			4	4

Target Metabolite	IL (M) (ng/mL)	IL (F) (ng/mL)
A5/B5	4	4
A5/E	10	10
A/T	10000	10000

Table C.6: Variables which compose the available prostate cancer datasets.

Dataset	Variables
1	A5, B5, A, DHEA, DHT, E, ETIO, T
2	T, F, E, DHT, DHEA 7b-OH-DHEA, AND5, A5, B5, ETIO, A

* Variable names are the abbreviated names of the metabolites of the datasets as presented in Table C.1.

Bibliography

[1] Adaway, J. E., B. G. Keevil, and L. J. Owen (2015). Liquid chromatography tandem mass spectrometry in the clinical laboratory. *Annals of clinical biochemistry 52*(1), 18–38.

[2] Albini, A., A. Bruno, B. Bassani, G. D'Ambrosio, G. Pelosi, P. Consonni, L. Castellani, M. Conti, S. Cristoni, and D. M. Noonan (2018). Serum Steroid Ratio Profiles in Prostate Cancer: A New Diagnostic Tool Toward a Personalized Medicine Approach. *Frontiers in endocrinology 9*, 110.

[3] Aldous, D. (1983). Exchangeability and related topics, Ecole d'Eté de Saint-Flour XIII, Lectures Notes n 1117.

[4] Alladio, E., R. Caruso, E. Gerace, E. Amante, A. Salomone, and M. Vincenti (2016). Application of multivariate statistics to the Steroidal Module of the Athlete Biological Passport: A proof of concept study. *Analytica chimica acta 922*, 19–29.

[5] Amante, E., E. Alladio, A. Salomone, M. Vincenti, F. Marini, G. Alleva, S. De Luca, and F. Porpiglia (2018). Correlation between chronological and physiological age of males from their multivariate urinary endogenous steroid profile and prostatic carcinoma-induced deviation. *Steroids 139*, 10–17.

[6] Amante, E., C. Fiori, G. Alleva, E. Alladio, F. Marini, D. Garrou, M. Manfredi, D. Amparore, E. Checcucci, S. Pruner, et al. (2019). Prospective evaluation of urinary steroids and prostate carcinoma-induced deviation: preliminary results. *Minerva Urology and Nephrology 73*(1), 98–106.

[7] Amante, E., S. Pruner, E. Alladio, A. Salomone, M. Vincenti, and R. Bro (2019). Multivariate interpretation of the urinary steroid profile and training-induced modifications. The case study of a Marathon runner. *Drug testing and analysis 11*(10), 1556–1565.

[8] Andersen, D. W. and K. Linnet (2014). Screening for Anabolic Steroids in Urine of Forensic Cases Using Fully Automated Solid Phase Extraction and LC–MS-MS. *Journal of Analytical Toxicology 38*(9), 637–644.

[9] Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The annals of statistics*, 1152–1174.

[10] Baron, D. A., D. M. Martin, and S. A. Magd (2007). Doping in sports and its spread to at-risk populations: an international review. *World Psychiatry 6*(2), 118.

[11] Bishop, C. M. (1994). Novelty detection and neural network validation. *IEE Proceedings-Vision, Image and Signal processing 141*(4), 217–222.

[12] Blackwell, D., J. B. MacQueen, et al. (1973). Ferguson distributions via Pólya urn schemes. *The annals of statistics 1*(2), 353–355.

[13] Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association 88*(421), 9–25.

[14] Brooks, R.-V., G. Jeremiah, W. A. Webb, and M. Wheeler (1979). Detection of anabolic steroid administration to athletes. In *Hormonal Steroids: Proceedings of the Fifth International Congress on Hormonal Steroids*, pp. 913–917. Elsevier.

[15] Brown, L. D. and E. Greenshtein (2009). Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *The Annals of Statistics*, 1685–1704.

[16] Campbell, N. and R. Mahon (1974). A multivariate study of variation in two species of rock crab of the genus Leptograpsus. *Australian Journal of Zoology 22*(3), 417–425.

[17] Carruba, G. (2007). Estrogen and prostate cancer: An eclipsed truth in an androgen-dominated scenario. *Journal of cellular biochemistry 102*(4), 899–911.

[18] Casella, G. and E. I. George (1992). Explaining the Gibbs sampler. *The American Statistician 46*(3), 167–174.

[19] Catlin, D. H., C. K. Hatton, and S. H. Starcevic (1997). Issues in detecting abuse of xenobiotic anabolic steroids and testosterone by analysis of athletes' urine. *Clinical Chemistry 43*(7), 1280–1288.

[20] Chan, A. O., N. F. Taylor, S. Tiu, and C. Shek (2008). Reference intervals of urinary steroid metabolites using gas chromatography–mass spectrometry in Chinese adults. *Steroids 73*(8), 828–837.

[21] Chib, S. and E. Greenberg (1995). Understanding the metropolis-hastings algorithm. *The american statistician 49*(4), 327–335.

[22] Conover, W. (1971). One-sample" Kolmogorov" test/Two-sample" Smirnov" test. *Practical nonparametric statistics*, 295–301.

[23] Cortes, C. and V. Vapnik (1995). Support-vector networks. *Machine learning 20*(3), 273–297.

[24] Davis, K. (27 June 2018 (accessed December 10, 2018)). *Anabolic steroids: What you should know*. https://www.medicalnewstoday.com/articles/246373.php.

[25] DeGroot, M. H. (2005). *Optimal statistical decisions*, Volume 82. John Wiley & Sons.

[26] Dehennin, L. and A. M. Matsumoto (1993). Long-term administration of testosterone enanthate to normal men: alterations of the urinary profile of androgen metabolites potentially useful for detection of testosterone misuse in sport. *The Journal of steroid biochemistry and molecular biology 44*(2), 179–189.

[27] Dellaportas, P. and I. Papageorgiou (2006). Multivariate mixtures of normals with unknown number of components. *Statistics and Computing 16*(1), 57–68.

[28] Dobson, A. J. and A. G. Barnett (2018). *An introduction to generalized linear models*. Chapman and Hall/CRC.

[29] Donike, M., K. Barwald, K. Klostermann, W. Schänzer, and J. Zimmermann (1983). Nachweis von exogenem Testosteron [Detection of exogenous testosterone].

[30] Escobar, M. D. and M. West (1995). Bayesian density estimation and inference using mixtures. *Journal of the american statistical association 90*(430), 577–588.

[31] Etzioni, R. and E. Feuer (2008). Studies of prostate-cancer mortality: caution advised. *The Lancet Oncology 9*(5), 407–409.

[32] Fenton, J. J., M. S. Weyrich, S. Durbin, Y. Liu, H. Bang, and J. Melnikow (2018). Prostate-specific antigen–based screening for prostate cancer: evidence report and systematic review for the US Preventive Services Task Force. *Jama 319*(18), 1914–1931.

[33] Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The annals of statistics*, 209–230.

[34] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics 7*(2), 179–188.

[35] Gates, A. J. and Y.-Y. Ahn (2017). The impact of random models on clustering similarity. *arXiv preprint arXiv:1701.06508*.

[36] Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian data analysis*. CRC press.

[37] Gelman, A. and D. B. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical science 7*(4), 457–472.

[38] Geman, S. and D. Geman (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence* (6), 721–741.

[39] Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments. *Bayesian statistics 4*, 641–649.

[40] Geyer, C. J. (1992). Practical Markov Chain Monte Carlo. *Statistical science*, 473–483.

[41] Geyer, C. J. (2011). Introduction to Markov Chain Monte Carlo. *Handbook of markov chain monte carlo 20116022*, 45.

[42] Gilks, W. R. and P. Wild (1992). Adaptive rejection sampling for Gibbs sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 41*(2), 337–348.

[43] Glavin, F. G. and M. G. Madden (2009). Analysis of the effect of unexpected outliers in the classification of spectroscopy data. In *Irish Conference on Artificial Intelligence and Cognitive Science*, pp. 124–133. Springer.

[44] Goodfellow, M., C. Rummel, E. Abela, M. P. Richardson, K. Schindler, and J. R. Terry (2016). Estimation of brain network ictogenicity predicts outcome from epilepsy surgery. *Scientific reports 6*(1), 1–13.

[45] Görür, D. and C. E. Rasmussen (2010). Dirichlet Process Gaussian Mixture Models: Choice of the base distribution. *Journal of Computer Science and Technology 25*(4), 653–664.

[46] Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika 82*(4), 711–732.

[47] Greenshtein, E. and J. Park (2009). Application of Non Parametric Empirical Bayes Estimation to High Dimensional Classification. *Journal of Machine Learning Research 10*(7).

[48] Hannah, L. A., D. M. Blei, and W. B. Powell (2011). Dirichlet Process Mixtures of Generalized Linear Models. *Journal of Machine Learning Research 12*(6).

[49] Harris, E. K. (1974). Effects of intra-and interindividual variation on the appropriate use of normal ranges. *Clinical Chemistry 20*(12), 1535–1542.

[50] Hogg, D. W., A. M. Price-Whelan, and B. Leistedt (2020). Data Analysis Recipes: Products of multivariate Gaussians in Bayesian inferences. *arXiv preprint arXiv:2005.14199*.

[51] Horning, E. (1968). Gas phase analytical methods for the study of steroid hormones and their metabolites. In *Gas Phase Chromatography of Steroids*, pp. 1–71. Springer.

[52] Huberty, C. J. (1984). Issues in the use and interpretation of discriminant analysis. *Psychological bulletin 95*(1).

[53] Huberty, C. J. (1994). *Applied discriminant analysis.* Number 519.535 HUB. CIMMYT. New York: Wiley.

[54] Human Metabolome Database: 5a-Adiol. (2018 (accessed November 28, 2018)). *Showing metabocard for Dihydroandrosterone.* http://www.hmdb.ca/metabolites/HMDB0000554.

[55] Human Metabolome Database: DHEA. (2018 (accessed November 28, 2018)). *Showing metabocard for Dehydroepiandrosterone.* http://www.hmdb.ca/metabolites/HMDB0000077.

[56] Human Metabolome Database: Etiocholanediol. (2018 (accessed November 28, 2018)). *Showing metabocard for Etiocholanediol.* http://www.hmdb.ca/metabolites/HMDB0000551.

[57] Human Metabolome Database: Etiocholanolone. (2018 (accessed November 28, 2018)). *Showing metabocard for Etiocholanolone.* http://www.hmdb.ca/metabolites/HMDB0000490.

[58] Human Metabolome Database: Testosterone. (2018 (accessed December 10, 2018)). *Showing metabocard for Etiocholanediol.* http://www.hmdb.ca/metabolites/HMDB0000234.

[59] International Amateur Athletic Federation. (1928 (accessed December 5, 2018)). *Handbook of the International Amateur Athlete Federation 1927–1928, "Section 13: Doping".* http://www.iaaf.org/news/news/a-piece-of-anti-doping-history-iaaf-handbook.

[60] Jara, A. (2017). Theory and computations for the Dirichlet process and related models: an overview. *International Journal of Approximate Reasoning 81*, 128–146.

[61] Jarque, C. M. and A. K. Bera (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics letters 6*(3), 255–259.

[62] Johnson, R. A., D. W. Wichern, et al. (2002). *Applied multivariate statistical analysis*, Volume 5. Prentice hall Upper Saddle River, NJ.

[63] Kanayama, G. and H. G. Pope Jr (2018). History and epidemiology of anabolic androgens in athletes and non-athletes. *Molecular and Cellular Endocrinology 464*, 4–13.

[64] Kass, R. E., B. P. Carlin, A. Gelman, and R. M. Neal (1998). Markov chain Monte Carlo in practice: a roundtable discussion. *The American Statistician 52*(2), 93–100.

[65] Kelly, R. S., M. G. Vander Heiden, E. Giovannucci, and L. A. Mucci (2016). Metabolomic biomarkers of prostate cancer: prediction, diagnosis, progression, prognosis, and recurrence. *Cancer Epidemiology and Prevention Biomarkers 25*(6), 887–906.

[66] Khan, S. S. and M. G. Madden (2014). One-class classification: taxonomy of study and review of techniques. *The Knowledge Engineering Review 29*(3), 345–374.

[67] Kheirandish, P. and F. Chinegwundoh (2011). Ethnic differences in prostate cancer. *British journal of cancer 105*(4), 481–485.

[68] Kicman, A. T., S. B. Coutts, C. J. Walker, and D. A. Cowan (1995). Proposed confirmatory procedure for detecting 5 alpha-dihydrotestosterone doping in male athletes. *Clinical chemistry 41*(11), 1617–1627.

[69] Kolonel, L. N., A. M. Nomura, and R. V. Cooney (1999). Dietary fat and prostate cancer: current status. *Journal of the National Cancer Institute 91*(5), 414–428.

[70] Kotsiantis, S., D. Kanellopoulos, P. Pintelas, et al. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering 30*(1), 25–36.

[71] Krone, N., B. A. Hughes, G. G. Lavery, P. M. Stewart, W. Arlt, and C. H. Shackleton (2010). Gas chromatography/mass spectrometry (GC/MS) remains a pre-eminent discovery tool in clinical steroid investigations even in the era of fast liquid chromatography tandem mass spectrometry (LC/MS/MS). *The Journal of steroid biochemistry and molecular biology 121*(3-5), 496–504.

[72] Kuuranne, T., M. Saugy, and N. Baume (2014). Confounding factors and genetic polymorphism in the evaluation of individual steroid profiling. *Br J Sports Med 48*(10), 848–855.

[73] Li, W. K. (2003). *Diagnostic checks in time series.* CRC Press.

[74] Litwin, M. S. and H.-J. Tan (2017). The diagnosis and treatment of prostate cancer: a review. *Jama 317*(24), 2532–2542.

[75] Lunardon, N., G. Menardi, and N. Torelli (2014). ROSE: A Package for Binary Imbalanced Learning. *R journal 6*(1).

[76] Maiworm, R. and W. Langthaler (1992). Influence of Androstenol and Androsterone on the Evalulation of Men of Varying Attractiveness Levels. In *Chemical Signals in Vertebrates 6*, pp. 575–579. Springer.

[77] Mareck, U., H. Geyer, G. Opfermann, M. Thevis, and W. Schänzer (2008). Factors influencing the steroid profile in doping control analysis. *Journal of Mass Spectrometry 43*(7), 877–891.

[78] Mazzeo, F. and A. Ascione (2013). Anabolic androgenic steroids and doping in sport. *Sports Medicine Journal/Medicina Sportivâ 9*(1).

[79] McLachlan, G. J. (2005). *Discriminant analysis and statistical pattern recognition.* John Wiley & Sons.

[80] Mechergui, Y. B., A. B. Jemaa, C. Mezigh, B. Fraile, N. B. Rais, R. Paniagua, M. Royuela, and R. Oueslati (2009). The profile of prostate epithelial cytokines and its impact on sera prostate specific antigen levels. *Inflammation 32*(3), 202–210.

[81] Meigs, J. B., M. J. Barry, J. E. Oesterling, and S. J. Jacobsen (1996). Interpreting results of prostate-specific antigen testing for early detection of prostate cancer. *Journal of general internal medicine 11*(9), 505–512.

[82] Minter, T. (1975). Single-class classification. In *LARS Symposia*, pp. 54.

[83] Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics 9*(2), 249–265.

[84] Nguyen, G. H., A. Bouzerdoum, and S. L. Phung (2009). Learning pattern classification tasks with imbalanced data sets. *Pattern recognition*, 193–208.

[85] Parr, M. K. and W. Schänzer (2010). Detection of the misuse of steroids in doping control. *The Journal of steroid biochemistry and molecular biology 121*(3-5), 528–537.

[86] Pinheiro, J. and D. Bates (2000). *Mixed-effects models in S and S-PLUS.* Springer.

[87] Piper, T., H. Geyer, N. Haenelt, F. Huelsemann, W. Schaenzer, and M. Thevis (2021). Current Insights into the Steroidal Module of the Athlete Biological Passport. *International Journal of Sports Medicine*.

[88] Pottgiesser, T. and Y. O. Schumacher (2012). Biomarker monitoring in sports doping control. *Bioanalysis 4*(10), 1245–1253.

[89] Raftery, A. E. and S. Lewis (1991). How many iterations in the Gibbs sampler? Technical report, Washington University Seattle Department of Statistics.

[90] Raftery, A. E. and S. M. Lewis (1992). [Practical Markov Chain Monte Carlo]: Comment: One Long Run with Diagnostics: Implementation Strategies for Markov Chain Monte Carlo. *Statistical science 7*(4), 493–497.

[91] Rasmussen, C. E. (2000). The infinite Gaussian mixture model. In *Advances in neural information processing systems*, pp. 554–560.

[92] Rauth, S. (1994). *Referenzbereiche von urinären Steroidkonzentrationen und Steroidquotienten: ein Beitrag zur Interpretation des Steroidprofils in der Routinedopinganalytik*. Ph. D. thesis, Deutsche Sporthochschule Köln.

[93] Rawla, P. (2019). Epidemiology of Prostate Cancer. *World journal of oncology 10*(2), 63.

[94] Richardson, S. and P. J. Green (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology) 59*(4), 731–792.

[95] Ross, G. J. and D. Markwick (2018). dirichletprocess: An R Package for Fitting Complex Bayesian Nonparametric Models.

[96] Saudan, C., N. Baume, N. Robinson, L. Avois, P. Mangin, and M. Saugy (2006). Testosterone and doping control. *British journal of sports medicine 40 Suppl 1(Suppl 1)*, i21–4.

[97] Schapire, R. E. (2013). Explaining adaboost. In *Empirical inference*, pp. 37–52. Springer.

[98] Schenk, J. M., A. R. Kristal, K. B. Arnold, C. M. Tangen, M. L. Neuhouser, D. W. Lin, E. White, and I. M. Thompson (2011). Association of symptomatic benign prostatic hyperplasia and prostate cancer: results from the prostate cancer prevention trial. *American journal of epidemiology 173*(12), 1419–1428.

[99] Schulze, J. J., J. Lundmark, M. Garle, I. Skilving, L. Ekström, and A. Rane (2008). Doping Test Results Dependent on Genotype of Uridine Diphospho-Glucuronosyl Transferase 2B17, the Major Enzyme for Testosterone Glucuronidation. *The Journal of Clinical Endocrinology & Metabolism 93*(7), 2500–2506.

[100] Scrucca, L., M. Fop, T. B. Murphy, and A. E. Raftery (2016). mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R journal 8*(1), 289.

[101] Sethuraman, J. and R. C. Tiwari (1982). Convergence of Dirichlet measures and the interpretation of their parameter. In *Statistical decision theory and related topics III*, pp. 305–315. Elsevier.

[102] Snyder, A. C. and A. C. Hackney (2013). The endocrine system in overtraining. In *Endocrinology of Physical Activity and Sport*, pp. 523–534. Springer.

[103] Society for Endocrinology: Testosterone. (2018 (accessed November 28, 2018)). *You and your Hormones: an education resource from the Socety for Endocrinology.* http://www.yourhormones.info/hormones/testosterone/.

[104] Sottas, P.-E., N. Baume, C. Saudan, C. Schweizer, M. Kamber, and M. Saugy (2006). Bayesian detection of abnormal values in longitudinal biomarkers with an application to T/E ratio. *Biostatistics 8*(2), 285–296.

[105] Sottas, P.-E., N. Robinson, S. Giraud, F. Taroni, M. Kamber, P. Mangin, and M. Saugy (2006). Statistical classification of abnormal blood profiles in athletes. *The International Journal of Biostatistics 2*(1).

[106] Sottas, P.-E., N. Robinson, O. Rabin, and M. Saugy (2011). The athlete biological passport. *Clinical chemistry 57*(7), 969–976.

[107] Sottas, P.-E., C. Saudan, C. Schweizer, N. Baume, P. Mangin, and M. Saugy (2008). From population-to subject-based limits of T/E ratio to detect testosterone abuse in elite sports. *Forensic science international 174*(2-3), 166–172.

[108] Sottas, P.-E., M. Saugy, and C. Saudan (2010). Endogenous steroid profiling in the athlete biological passport. *Endocrinology and Metabolism Clinics 39*(1), 59–73.

[109] Strahm, E., P.-E. Sottas, C. Schweizer, M. Saugy, J. Dvorak, and C. Saudan (2009). Steroid profiles of professional soccer players: an international comparative study. *British journal of sports medicine*.

[110] Ulrich, R., H. G. Pope, L. Cléret, A. Petróczi, T. Nepusz, J. Schaffer, G. Kanayama, R. D. Comstock, and P. Simon (2018, Jan). Doping in Two Elite Athletics Competitions Assessed by Randomized-Response Surveys. *Sports Medicine 48*(1), 211–219.

[111] USADA: Armstrong L. (2012 (accessed December 5, 2018)). *Statement From USADA CEO Travis T. Tygart Regarding The U.S. Postal Service Pro Cycling Team Doping Conspiracy.* http://cyclinginvestigation.usada.org/.

[112] USADA: Side-Effects. (2018 (accessed December 6, 2018)). *Effects of Performance-Enhancing Drugs.* https://www.usada.org/substances/effects-of-performance-enhancing-drugs/.

[113] Valverde-Albacete, F. J., J. Carrillo-de Albornoz, and C. Peláez-Moreno (2013). A proposal for new evaluation metrics and result visualization technique for sentiment analysis tasks. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 41–52. Springer.

[114] Van Renterghem, P., P.-E. Sottas, M. Saugy, and P. Van Eenoo (2013). Statistical discrimination of steroid profiles in doping control with support vector machines. *Analytica chimica acta 768*, 41–48.

[115] Van Renterghem, P., P. Van Eenoo, H. Geyer, W. Schänzer, and F. T. Delbeke (2010). Reference ranges for urinary concentrations and ratios of endogenous steroids, which can be used as markers for steroid misuse, in a Caucasian population of athletes. *Steroids 75*(2), 154–163.

[116] Van Renterghem, P., P. van Eenoo, W. V. Thuyne, H. Geyer, W. Schaenzer, and F. T. Delbeke (2008). Validation of an extended method for the detection of the misuse of endogenous steroids in sports, including new hydroxylated metabolites. *Journal of chromatography. B, Analytical technologies in the biomedical and life sciences 876 2*, 225–35.

[117] Vernec, A. R. (2014). The Athlete Biological Passport: an integral element of innovative strategies in antidoping. *British Journal of Sports Medicine 48*(10), 817–819.

[118] WADA (2018 (accessed November 2, 2018)b). *Prohibited List, January 2018*. https://www.wada-ama.org/sites/default/files/prohibited_list_2018_en.pdf.

[119] WADA (2018 (accessed November 6, 2018)a). *Endogenous Anabolic Androgenic Steroids: Measurement and Reporting*. https://www.wada-ama.org/sites/default/files/resources/files/td2018eaas_final_eng.pdf.

[120] WADA, W. A.-D. A. (2018). WADA Technical Document – TD2018EAAS: Endogenous Anabolic Androgenic Steroids - Measurement and Reporting .

[121] WADA, W. A.-D. A. (2019). Anti-Doping Testing Figures Samples Analyzed and Reported by Accredited Laboratories in ADAMS.

[122] WADA, W. A.-D. A. (2021a). Technical Document - TD2021EAAS: Measurement and Reporting of Endogenous Anabolic Androgenic Steroid (EAAS) Markers of the Urinary Steroid Profile.

[123] WADA, W. A.-D. A. (2021b). WADA Technical Document – TD2021APMU.

[124] WADA, W. A.-D. A. (2021c). WADA Technical Document – TD2021IRMS.

[125] WADA: HGH test. (2018 (accessed December 11, 2018)). *Human Growth Hormone (HGH) testing*. https://www.wada-ama.org/en/questions-answers/human-growth-hormone-hgh-testing.

[126] Wheeler, M. J. (2013). *Hormone Assays in Biological Fluids* (2 ed.). New York: Springer.

[127] Wilkes, E. H., G. Rumsby, and G. M. Woodward (2018). Using machine learning to aid the interpretation of urine steroid profiles. *Clinical chemistry 64* (11), 1586–1595.

[128] Wolf, A. M., R. C. Wender, R. B. Etzioni, I. M. Thompson, A. V. D'Amico, R. J. Volk, D. D. Brooks, C. Dash, I. Guessous, K. Andrews, et al. (2010). American Cancer Society guideline for the early detection of prostate cancer: update 2010. *CA: a cancer journal for clinicians 60* (2), 70–98.

[129] Wu, H., T. Liu, C. Ma, R. Xue, C. Deng, H. Zeng, and X. Shen (2011). GC/MS-based metabolomic approach to validate the role of urinary sarcosine and target biomarkers for human prostate cancer by microwave-assisted derivatization. *Analytical and bioanalytical chemistry 401*(2), 635–646.

[130] Wudy, S., G. Schuler, A. Sánchez-Guijo, and M. Hartmann (2018). The art of measuring steroids: principles and practice of current hormonal steroid analysis. *The Journal of steroid biochemistry and molecular biology 179*, 88–103.

[131] Yeap, B. B. (2014). Sex steroids and cardiovascular disease. *Asian journal of andrology 16*(2), 239.

[132] Zhang, Q., Z. Chen, S. Chen, Y. Xu, and H. Deng (2017). Intraindividual stability of cortisol and cortisone and the ratio of cortisol to cortisone in saliva, urine and hair. *Steroids 118*, 61–67.