UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS MATEMÁTICAS Y
DE LA ESPECIALIZACIÓN EN ESTADÍSTICA APLICADA

DISCRIMINANT ANALYSIS WITH GAUSSIAN GRAPHICAL TREE MODELS

TESIS
QUE PARA OPTAR POR EL GRADO DE:
DOCTOR EN CIENCIAS

PRESENTA:
GONZALO PÉREZ DE LA CRUZ

DIRECTOR DE LA TESIS

DRA. GUILLERMINA ESLAVA GÓMEZ
FACULTAD DE CIENCIAS, UNAM

MIEMBROS DEL COMITÉ TUTOR

DR. JOHAN JOZEF LODE VAN HOREBEEK
PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS MATEMÁTICAS Y
DE LA ESPECIALIZACIÓN EN ESTADÍSTICA APLICADA

DR. IGNACIO MÉNDEZ RAMÍREZ
INSTITUTO DE INVESTIGACIONES EN MATEMÁTICAS APLICADAS Y EN SISTEMAS, UNAM

MÉXICO, D. F. FEBRERO DE 2015

Discriminant analysis with Gaussian graphical tree models

Abstract

This thesis presents a study of the use of Gaussian graphical models in discriminant analysis for two populations. By using estimates of the associated concentration matrices, the plug-in allocation rules are defined. Two estimation problems are involved when considering graphical models with restrictions: parameter and structure estimation.

In the first part of the thesis, we consider the problem of parameter estimation when a linear concentration model for each of the two populations is considered. Additionally, we consider the case where some equalities between elements of the two concentration matrices are imposed. Considering Maximum Likelihood, we derive the analytical expressions needed to use the iterative partial maximization algorithm, to then solve the parameter estimation problem.

When considering these linear restrictions, the estimation of the structure is complex, and for this, we restrict the study to Gaussian graphical models with a tree graph in the second part of the thesis. In this part of the thesis the interest focusses on the structure estimation problem, since the parameter estimation is solved by using Maximum Likelihood Estimators which have explicit expressions.

We describe six methods for estimating the structure of Gaussian graphical tree models. In each method an unknown tree structure is assumed for each concentration matrix involved in the discriminant function. By finding a minimum weight spanning tree and using maximum likelihood estimation, the concentration matrices are estimated. Three of these methods have been introduced in the literature, and based on these, three others are introduced in this thesis.

The six methods take advantage of the tree structure, specifically of an efficient algorithm for finding the minimum weight spanning tree, and the existence of closed form expressions for the maximum likelihood estimator of the concentration matrices.

A numerical study is presented, where the performance of the six methods are compared when the training sample size is the same in both populations, though in this case two pair of methods are equivalent. The comparison of the four different methods is based on estimated error rates obtained from real and simulated data. Diagonal discriminant analysis is considered as a benchmark, as well as quadratic and linear discriminant analysis whenever the sample size is sufficient.

In spite of showing that none of the methods based on tree models outperforms the benchmarks in all data sets, tree models offer a simple and computationally inexpensive alternative to well established discriminant methods in high dimensional settings, where sample size is similar to, or smaller than, the number of variables.

Acknowledgements

First, I would like to express my appreciation and gratitude to my advisor, Dr. Guillermina Eslava, for her endless support and guidance. Her advice and coaching went far beyond research and I am very grateful to her for introducing me into her network of colleagues and for encouraging me to do two research stays abroad.

I would also like to thank Dr. Van Horebeek for the many discussions with him and for his suggestions. I very much enjoyed working at CIMAT during the short visits to his Department.

My sincere thanks also goes to Dr. Jin Tian and Dr. Søren Højsgaard for their generous hospitality when I visited their Departments in Iowa State University and Aalborg University, respectively. Also to Dr. Poul Svante for interesting and productive discussions.

Finally, I would like to thank my family for all their love and encouragement, and my friends for always being there with support and advice.

Contents

Introduction

Discriminant analysis considers the problem of classifying new observations into different populations. In discriminant analysis for two populations with p-variate Gaussian distributions, the optimal allocation rule is based on the log-likelihood ratio, and hence on the means and concentration matrices. In practical applications, these parameters are estimated and replaced into the allocation rule, giving what is known as the plug-in rule. Typically, the means and concentration matrices are considered with no restriction, so that their maximum likelihood estimators (MLEs) correspond to the sample means and the inverse of each sample covariance matrix.

However, the number of parameters involved in the optimal rule grows quickly as a function of the number of variables. And in current statistical applications, problems involving a large number of variables, but only a small number of observations often appear. In these cases, the use of the plug-in rule associated with the optimal rule becomes impractical and the use of allocation rules with less number of parameters is an alternative.

Alternative rules can be obtained when using some restrictions on the parameters; noticing that the restrictions can be considered either on the covariance or the concentration matrices, and on the mean vectors as well. For instance, when considering zero mean differences, as in the twins example given by Bartlett and Please (1963). Or when considering a diagonal matrix for each covariance matrix. Here, the allocation rule becomes simpler, since the number of parameters and the minimum required sample size diminish considerably.

Another example is the use of restrictions on each concentration matrix with the purpose of reflecting particular characteristics of the populations or distributions such as symmetry, marginal or conditional independences. Some models which consider these kind of restrictions are Gaussian graphical models (Lauritzen, 1996), symmetry models (Højsgaard and Lauritzen, 2008), and linear concentration models (Anderson, 1970 and Sturmfels and Uhler, 2010).

The kind of restrictions considered in the linear concentration models were introduced by Anderson (1970), and later on Sturmfels and Uhler (2010) gave this name to these models. In Chapter 3, we consider a linear concentration model independently for each of the two populations; additionally, we consider the case where some corresponding elements between the two concentration matrices are equal.

When considering these models with restrictions, two estimation problems arise: *parameter estimation*, which corresponds to estimating the concentration matrix given a specific structure, and *structure estimation or model selection*, which corresponds to identifying or estimating the structure.

Exploiting that these models are special cases of the regular exponential family, we adapt the iterative partial maximization (IPM) algorithm (Jensen et al., 1991) to solve the parameter estimation problem. That is, we derive the analytical expressions needed to use this algorithm. On the other hand, the estimation of the structure is complex. For this aspect, we list some algorithms which are found in the literature and have been developed for models with less restrictions. For instance, when the models are restricted to be Gaussian graphical models in both populations without considering equalities between corresponding elements of the two concentration matrices. However, we note that there is no single method which can be used when all the restrictions are considered.

Considering the case where the structure is unknown and with the purpose of studying the use of linear restrictions on the concentration matrices in the context of classification, we restrict the study to the case where the concentration matrices are associated with Gaussian graphical tree models. These models are Gaussian graphical models (GGMs) with a tree graph, and they are one of the simplest decomposable models for which a explicit expression exists for the MLE of the concentration matrix.

We consider six methods for structure estimation. They all use Maximum Likelihood (ML) estimation for the parameters, but use a different function to be optimized for the estimation of the tree structure. Three of these methods have been studied in the literature: Chow and Liu (1968), Friedman et al. (1997, 1998), and Tan et al. (2010). And based on these, three others are introduced in this thesis, for which the function to be optimized is the J-divergence for one of them, and the empirical log-likelihood ratio (log-ratio) for the other two. We show in Propositions 4.5.1 and 4.5.2 that their associated optimization problems are equivalent to one of finding a minimum weight spanning tree (MWST). As a result of this last property and the existence of closed form expressions for the MLEs, any of these methods offers a simple and computationally inexpensive alternative in high dimensional settings, where sample size is similar to, or smaller than, the number of variables. This also permit using simulation or cross-validation procedures to estimate error rates for the comparison of methods for structure estimation in the context of classification of observations.

We compare the classification performance of these methods using estimated error rates. These are computed for breast cancer data and for simulated data sets. The latter were generated from four specific models: an autoregressive of order 1; a moving average of order 1; a set of equal correlated variables; and a set of variables from a Gaussian distribution with a random concentration matrix.

For the numerical comparison, we consider training samples of equal size in both populations, and for this case the method given in Tan et al. (2010) and the one based on the log-ratio are equivalent, as well as the methods based on the J-divergence and the log-ratio with equal trees. We also compute error rates for diagonal quadratic and linear discriminant analysis, DQDA and DLDA respectively, and also for QDA and LDA when the sample size is large enough.

We also consider HIV data to illustrate the case when the data correspond to repeated measures and the training samples sizes are different. In this case, diagonal discriminant analysis is also considered as a benchmark.

The results of the study show that among the methods based on trees, there is no single one that outperforms the others. It is observed that there are cases where DQDA and DLDA are better than the methods based on trees, and in some cases LDA is also better. However, in most cases the performance of the methods using trees is superior, especially when the group sample size is similar to the number of variables. In this case, we conclude that using trees is a good alternative to diagonal, linear and quadratic discriminant analysis.

The present thesis is organized as follows. Chapter 1 provides an introduction to discriminant analysis, Gaussian graphical models, and models with linear restrictions on the concentration matrices as those introduced by Anderson (1970). The statistical theory of Gaussian graphical models, which is briefly described, appears in Lauritzen (1996) and the implementation in R of algorithms for these models in Højsgaard et al. (2012).

Chapter 2 gives some early examples found in the literature, where some linear structures on the concentration matrices are used in discriminant analysis or in the context of one population. In Chapter 3, we describe the linear concentration models with some equalities between corresponding elements of the two matrices. We also derive the expressions for the IPM algorithm to obtain

ML estimates and list some algorithms used for structure estimation. An example in the context of an educational testing problem is presented to illustrate the use of the algorithm to obtain the ML estimates, and the breast cancer data is introduced to illustrate the use of algorithms to estimate the structure. In the latter example, error rates are estimated using the repeated holdout method assuming a given structure, though in Chapter 4, this method is used for this data including the structure estimation.

In Chapter 4, we state the six methods for structure estimation of Gaussian graphical tree models, giving their corresponding optimization problems together with the weights needed for the associated MWST problems. We also present the results of the numerical study for the case of equal group sample size and for data on repeated measures. In Chapter 5, we give conclusions of the work and describe future work. Finally, in the appendices, the proofs of the propositions and corollary are given, also some details needed for the use of the IPM algorithm and supplementary figures.

1. Preliminaries

1.1 Discriminant Analysis

Discriminant analysis considers the problem of discriminating between different populations or groups. It has been used with at least one of two objectives: (i) to understand the discriminating power of the variables on the observations, and (ii) to classify new observations into one of the groups based on their characteristics.

In its linear form and without distributional assumptions, discriminant analysis was introduced by Fisher (1936). Welch (1939) formulated the discrimination rule based on probability distributions for two populations, Π_1 and Π_2, as follows. Assign an observation $\boldsymbol{x}$ into population Π_1 when

$$\pi_1 f_1(\boldsymbol{x}) > \pi_2 f_2(\boldsymbol{x}),$$

and otherwise to Π_2. Or equivalently, based on the log-likelihood ratio criterion, when

$$\ln \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} > \ln \frac{\pi_2}{\pi_1},$$

where $f_c(\boldsymbol{x}) = f(\boldsymbol{x}|C = c)$ is an arbitrary density or probability function and $\pi_c = P(C = c)$, $c = 1, 2$. This rule is also equivalent to the Bayes rule: Assign an observation $\boldsymbol{x}$ into Π_1 when

$$P(C = 1|\boldsymbol{x}) = \pi_1 \frac{f_1(\boldsymbol{x})}{f(\boldsymbol{x})} > P(C = 2|\boldsymbol{x}) = \pi_2 \frac{f_2(\boldsymbol{x})}{f(\boldsymbol{x})},$$

and into Π_2 otherwise.

In particular, when f_c is the density of a Gaussian distribution $N(\boldsymbol{\mu}_c, \boldsymbol{K}_c^{-1})$ with $\boldsymbol{\Sigma_c} = \boldsymbol{K}_c^{-1}$, $c = 1, 2$,

$$\begin{aligned} \ln \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} &= \tfrac{1}{2}\ln \frac{\mid \boldsymbol{K}_1 \mid}{\mid \boldsymbol{K}_2 \mid} + \tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^t \boldsymbol{K}_2 (\boldsymbol{x}-\boldsymbol{\mu}_2) - \tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^t \boldsymbol{K}_1 (\boldsymbol{x}-\boldsymbol{\mu}_1) \\ &= \boldsymbol{x}^t \boldsymbol{A} \boldsymbol{x} + \boldsymbol{b}^t \boldsymbol{x} + d, \end{aligned} \tag{1.1}$$

where $\boldsymbol{A} = \frac{1}{2}(\boldsymbol{K}_2 - \boldsymbol{K}_1)$, $\boldsymbol{b} = \boldsymbol{K}_1\boldsymbol{\mu}_1 - \boldsymbol{K}_2\boldsymbol{\mu}_2$, and $d = \frac{1}{2}\ln\frac{|\boldsymbol{K}_1|}{|\boldsymbol{K}_2|} + \frac{1}{2}(\boldsymbol{\mu}_2^t \boldsymbol{K}_2 \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1^t \boldsymbol{K}_1 \boldsymbol{\mu}_1)$. Expression (1.1) is called the quadratic discriminant function.

This rule is optimal in the sense that minimizes the error rate or probability of misclassification $P(e)$ defined as

$$P(e) = \pi_1 P(2|1) + \pi_2 P(1|2), \tag{1.2}$$

where $P(i|j)$ denotes the probability of assigning an observation from population Π_j to Π_i.

We also obtain the same rule if we consider that the joint distribution of the vector $(C, \boldsymbol{X})$ is a Conditional Gaussian distribution, CG-distribution, with density (see Lauritzen, 1996, s. 6.1.1)

$$f(c, \boldsymbol{x}) = \pi_c (2\pi)^{-p/2} \mid \boldsymbol{K}_c \mid^{1/2} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_c)^t \boldsymbol{K}_c (\boldsymbol{x}-\boldsymbol{\mu}_c)\right], \tag{1.3}$$

where $\boldsymbol{X}|C = c \sim N(\boldsymbol{\mu}_c, \boldsymbol{K}_c^{-1})$, $P(C = c) = \pi_c$, $c = 1, 2$, and $\pi_1 + \pi_2 = 1$.

In general, we note that the joint distribution can be factorized as follows

$$f(c, \boldsymbol{x}) = \begin{cases} f(\boldsymbol{x}|c)P(c), \\ P(c|\boldsymbol{x})f(\boldsymbol{x}). \end{cases}$$

When the interest focuses on estimating $f(\boldsymbol{x}|c)$, the methods are called generative. On the

other hand, when it focuses on $P(c|\boldsymbol{x})$, the methods are called discriminative.

In discriminant analysis, the interest focuses on $f(\boldsymbol{x}|c) = f_c(\boldsymbol{x})$. Hence, in a practical application, we need to estimate the unknown parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$. When the parameter estimates are inserted in $f_c(\boldsymbol{x})$, we obtain what is called the plug-in allocation rule.

ML can be used to estimate $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$ as follows. Given a sample $\boldsymbol{x}_1^{(c)}, ..., \boldsymbol{x}_{n_c}^{(c)}$ from each group c, $c = 1, 2$, the MLEs are

$$\widehat{\boldsymbol{\mu}}_c = \overline{\boldsymbol{x}}^{(c)} = \sum_{k=1}^{n_c} \boldsymbol{x}_k^{(c)}/n_c; \tag{1.4}$$

$$\widehat{\boldsymbol{K}_c} = \boldsymbol{W}_c^{-1} \quad \text{with} \quad \boldsymbol{W}_c = \sum_{k=1}^{n_c} (\boldsymbol{x}_k^{(c)} - \overline{\boldsymbol{x}}^{(c)})(\boldsymbol{x}_k^{(c)} - \overline{\boldsymbol{x}}^{(c)})^t/n_c, \;\; c = 1, 2; \tag{1.5}$$

see, for example, Anderson (2003, p. 613). The estimator $\widehat{\boldsymbol{\mu}}_c$ is unbiased; but $\widehat{\boldsymbol{\Sigma}}_c = \boldsymbol{W}_c$ is biased; and an unbiased estimator is $\frac{n_c}{n_c-1}\boldsymbol{W}_c$, $c = 1, 2$. We refer to the plug-in allocation rule obtained with the MLEs as quadratic discriminant analysis (QDA).

When $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, the quadratic term in (1.1) vanishes giving

$$\ln \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \;\; = \;\; (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \boldsymbol{K}[\boldsymbol{x} - \tfrac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)] = \boldsymbol{b}^t\boldsymbol{x} + d, \tag{1.6}$$

where $\boldsymbol{b} = \boldsymbol{K}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ and $d = -\frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t\boldsymbol{K}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$. (1.6) is called the linear discriminant function. In this case given a sample $\boldsymbol{x}_1^{(c)}, ..., \boldsymbol{x}_{n_c}^{(c)}$ on each group c, $c = 1, 2$, the MLEs are

$$\widehat{\boldsymbol{K}} = \boldsymbol{W}^{-1} \quad \text{with} \quad \boldsymbol{W} = \sum_{c=1}^{2}\sum_{k=1}^{n_c} (\boldsymbol{x}_k^{(c)} - \overline{\boldsymbol{x}}^{(c)})(\boldsymbol{x}_k^{(c)} - \overline{\boldsymbol{x}}^{(c)})^t/(n_1 + n_2), \tag{1.7}$$

and $\widehat{\boldsymbol{\mu}}_c$ as in (1.4). The plug-in allocation rule obtained with these MLEs is referred as linear

discriminant analysis (LDA).

Discriminant analysis and logistic regression are two closely related models in supervised statistical classification, however, in logistic regression interest focuses on the conditional probabilities:

$$P(C = c|\boldsymbol{X} = \boldsymbol{x}) = \frac{\pi_c f(\boldsymbol{x}|C = c)}{\pi_2 f(\boldsymbol{x}|C = 2) + \pi_1 f(\boldsymbol{x}|C = 1)}, \quad c = 1, 2.$$

In order to obtain the conditional probabilities, logistic regression models the function:

$$\ln \frac{P(C = 1|\boldsymbol{X} = \boldsymbol{x})}{P(C = 2|\boldsymbol{X} = \boldsymbol{x})} = \ln \frac{\pi_1}{\pi_2} + \ln \frac{f(\boldsymbol{x}|C = 1)}{f(\boldsymbol{x}|C = 2)}. \tag{1.8}$$

And in its linear form, logistic regression models (1.8) assuming a linear function of $\boldsymbol{x}$,

$$\ln \frac{P(C = 1|\boldsymbol{X} = \boldsymbol{x})}{P(C = 2|\boldsymbol{X} = \boldsymbol{x})} = \boldsymbol{\beta}^t \boldsymbol{x} + \beta_0. \tag{1.9}$$

Therefore

$$P(C = 1|\boldsymbol{X} = \boldsymbol{x}_i) = \frac{\exp(\boldsymbol{\beta}^t \boldsymbol{x}_i + \beta_0)}{1 + \exp(\boldsymbol{\beta}^t \boldsymbol{x}_i + \beta_0)} \quad \text{and} \quad P(C = 2|\boldsymbol{X} = \boldsymbol{x}_i) = \frac{1}{1 + \exp(\boldsymbol{\beta}^t \boldsymbol{x}_i + \beta_0)}.$$

In this case, the goal is to estimate the parameters $(\boldsymbol{\beta}, \beta_0)$ directly from the sample and not through a function of the estimated group means and concentration matrices.

In order to use ML, it is assumed that the distribution of C conditional on $\boldsymbol{X} = \boldsymbol{x}$ is Bernoulli with probabilities $P(C = c|\boldsymbol{X} = \boldsymbol{x})$. Then given a sample $\{(c, \boldsymbol{x})_1, ..., (c, \boldsymbol{x})_n\}$, the MLEs are those that maximize the log-likelihood function

$$ln(\ell(\boldsymbol{\beta}, \beta_0)) = \sum_{i=1}^{n} \delta_1(c_i) \ln(P(C = 1|\boldsymbol{X} = \boldsymbol{x}_i)) + (1 - \delta_1(c_i)) \ln(P(C = 2|\boldsymbol{X} = \boldsymbol{x}_i)) \tag{1.10}$$

where $\delta_1(c) = \begin{cases} 1 & \text{if} \quad c = 1 \\ 0 & \text{if} \quad c \neq 1 \end{cases}$.

We note that when the distribution of $\boldsymbol{X}$ conditional on $C = c$ is $N(\boldsymbol{\mu}_c, \boldsymbol{K}^{-1})$, function (1.8) is in fact linear,

$$\ln \frac{P(C = 1|\boldsymbol{X} = \boldsymbol{x})}{P(C = 2|\boldsymbol{X} = \boldsymbol{x})} = \ln \frac{\pi_1}{\pi_2} + \ln \frac{f(\boldsymbol{x}|C = 1)}{f(\boldsymbol{x}|C = 2)} = \ln \frac{\pi_1}{\pi_2} + \boldsymbol{b}^t \boldsymbol{x} + d, \tag{1.11}$$

with $\boldsymbol{b}$ and d as in (1.6). And where the parameters in (1.9) and (1.11) are related by

$$\boldsymbol{\beta} = \boldsymbol{b} = \boldsymbol{K}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \quad \text{and} \quad \beta_0 = \ln \frac{\pi_1}{\pi_2} + d = \ln \frac{\pi_1}{\pi_2} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \boldsymbol{K}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2).$$

However, the estimates for $\boldsymbol{\beta}$ and β_0 obtained by maximizing (1.10) are not the same as those obtained by plugging in the ML estimates computed using (1.4) and (1.7), see for example Ripley (1996, p. 45) and Cox and Snell (1989, p. 136).

Anderson (1975) considers the logistic regression in its quadratic form as

$$\ln \frac{P(C = 1|\boldsymbol{X} = \boldsymbol{x})}{P(C = 2|\boldsymbol{X} = \boldsymbol{x})} = \boldsymbol{x}^t \boldsymbol{\Omega} \boldsymbol{x} + \boldsymbol{\beta}^t \boldsymbol{x} + \beta_0. \tag{1.12}$$

This model, for example, is obtained when the distribution of $\boldsymbol{X}$ conditional on $C = c$ is $N(\boldsymbol{\mu}_c, \boldsymbol{K}_c^{-1})$, where $\boldsymbol{\Omega}$ corresponds to $\boldsymbol{A}$ given in (1.1).

Due to the large number of parameters in the quadratic logistic regression model, Anderson

(1975) proposes to approximate $\boldsymbol{\Omega}$ with $l_{max}\boldsymbol{v}_{max}\boldsymbol{v}_{max}^t$, where l_{max} is the largest eigenvalue and $\boldsymbol{v}_{max}$ its associated eigenvector from the spectral decomposition $\boldsymbol{\Omega} = \sum_{i=1}^{p} l_i \boldsymbol{v}_i \boldsymbol{v}_i^t$.

Therefore, the number of parameters to be estimated in (1.12) are only $2p + 1$, because the coefficients of quadratic and crossed terms are functions of p parameters: l_{max}, $v_{1_{max}}$, ..., $v_{(p-1)_{max}}$. For example, when $p = 4$

$$\boldsymbol{x}^t\boldsymbol{\Omega}\boldsymbol{x} + \boldsymbol{\beta}^t\boldsymbol{x} + \beta_0 \simeq l_{max}\sum_{k=1}^{4}\sum_{s=1}^{4} x_k x_s v_{k_{max}} v_{s_{max}} + \boldsymbol{\beta}^t\boldsymbol{x} + \beta_0. \tag{1.13}$$

where $v_{4_{max}}^2 = 1 - v_{1_{max}}^2 - v_{2_{max}}^2 - v_{3_{max}}^2$. The author also suggests to use more than one from the p terms in the spectral decomposition and mentions other approximations when dealing with binary variables instead of continuous.

In discriminant analysis, usually no assumption is made on the pattern of the mean vectors or the concentration matrices. However, we can consider particular patterns or structures on the group means and either on the covariance or the concentration matrices. In this work, we consider linear restrictions on the concentration matrices. These restrictions may be considered for two reasons: (i) to help to diminish the number of unknown parameters and (ii) to identify a model which could be more readily interpretable for a specific problem. To diminish the number of parameters is a growing need in applications where the number of variables outgrows the number of observations.

For example, when n_1 or n_2 are smaller than p, QDA cannot be used since $\boldsymbol{W}_1$ or $\boldsymbol{W}_2$ could be singular matrices. In this case, LDA is often used, even when the assumption of equal covariance matrices is not plausible, though its performance could be poor when $n_1 + n_2$ is similar to the number of variables, and it cannot be used when $n_1 + n_2 < p$.

We remark that linear restrictions on the concentration matrices also impact the logistic regres-

sion function such that the number of unknown parameters in (1.12) may diminish. For example, when $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$ are diagonal distinct matrices, (1.8) can be expressed as

$$\ln \frac{P(C=1|\boldsymbol{X}=\boldsymbol{x})}{P(C=2|\boldsymbol{X}=\boldsymbol{x})} = \boldsymbol{\alpha}^t \boldsymbol{z} + \boldsymbol{\beta}^t \boldsymbol{x} + \beta_0, \tag{1.14}$$

where $\boldsymbol{z} = (x_1^2, ..., x_p^2)$. Though, there are other linear restrictions that imply this expression.

The assumption of diagonal matrices is made very often when p is larger than n_1 and n_2. Diagonal quadratic discriminant analysis (DQDA) is based on the plug-in rule with $\widehat{\boldsymbol{D}}_c^{-1}$ instead of $\widehat{\boldsymbol{K}}_c$, where $\widehat{\boldsymbol{D}}_c = \text{diag}(\boldsymbol{W}_c)$, $c = 1, 2$; and diagonal linear discriminant analysis (DLDA) uses $\widehat{\boldsymbol{D}}_1 = \widehat{\boldsymbol{D}}_2 = \text{diag}(\boldsymbol{W})$. Bickel and Levina (2004) studied the behaviour of DLDA, using $(n_1+n_2)\boldsymbol{W}/(n_1+n_2-2)$ instead of $\boldsymbol{W}$, and under some conditions when the number of variables grows: a Mahalanobis distance at least c, where c is constant; a bound on the ratio of the largest and the smallest eigenvalue; and the mean vectors belonging to a specific compact subset of l_2. They showed that when $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ and under these conditions, the performance of DLDA is better than random guessing and in some cases is asymptotically optimal when $n = n_1 = n_2 \to \infty$, $p \to \infty$ and $(\ln p)/n \to 0$. Under the same conditions they also showed that DLDA can indeed greatly outperform LDA.

The good classification performance of DQDA and DLDA led us to consider a spectrum of possibilities spanning the range between assuming full independence and arbitrary dependence; whereas the good classification performance of LDA between assuming equal and arbitrary matrices. In the following section, we describe the models with linear restrictions on the concentration matrices that we consider.

1.2 Gaussian Graphical models

In terms of Gaussian graphical models (GGMs), a zero on the concentration matrix is equivalent to assume conditional independence between the corresponding variables. Specifically, a GGM has an undirected graph $G = (V, E)$ where $V = \{v_1, ..., v_p\}$ is a set of vertices representing the set of variables $\{x_1, ..., x_p\}$ and E a set of edges. G, also referred as dependence graph, represents conditional independence relations for the set of random variables. The model satisfies the following relations:

$$x_i \perp\!\!\!\perp x_j | \{x_1, ..., x_p\} \setminus \{x_i, x_j\} \iff k_{ij} = 0 \iff (i, j) \notin E,$$

where k_{ij} is the entry ij of the concentration matrix $\boldsymbol{K}$.

For example, we consider three particular patterned concentration matrices associated with the following graphs: a cycle C_p, a path τ and the edgeless $\overline{\kappa}$; and compare them with a non restricted concentration matrix which corresponds to the complete graph κ, as shown in Figure 1.1 for $p = 4$ variables.

The associated concentration matrices to the four graphs in Figure 1.1 are the following.

$$\boldsymbol{K}_{\kappa} = \begin{pmatrix} k_{11} & k_{12} & k_{13} & k_{14} \\ k_{12} & k_{22} & k_{23} & k_{24} \\ k_{13} & k_{23} & k_{33} & k_{34} \\ k_{14} & k_{24} & k_{34} & k_{44} \end{pmatrix}, \quad \boldsymbol{K}_{\mathrm{C}_4} = \begin{pmatrix} k_{11} & k_{12} & 0 & k_{14} \\ k_{12} & k_{22} & k_{23} & 0 \\ 0 & k_{23} & k_{33} & k_{34} \\ k_{14} & 0 & k_{34} & k_{44} \end{pmatrix},$$

$$\boldsymbol{K}_{\tau} = \begin{pmatrix} k_{11} & k_{12} & 0 & k_{14} \\ k_{12} & k_{22} & k_{23} & 0 \\ 0 & k_{23} & k_{33} & k_{34} \\ k_{14} & 0 & k_{34} & k_{44} \end{pmatrix}, \quad \boldsymbol{K}_{\overline{\kappa}} = \begin{pmatrix} k_{11} & 0 & 0 & 0 \\ 0 & k_{22} & 0 & 0 \\ 0 & 0 & k_{33} & 0 \\ 0 & 0 & 0 & k_{44} \end{pmatrix},$$

with $k_{ij} \neq 0$, $i, j = 1, ..., 4$.

Hereafter, we use the notation $\boldsymbol{K}_G$ to make the dependence on the specific graph G clear.

Considering these four graphs in the context of discriminant analysis with the same kind of graph in both populations, the number of parameters to be estimated for the quadratic and linear discriminant functions given in (1.1) and (1.6) diminishes, see Table 1.1.

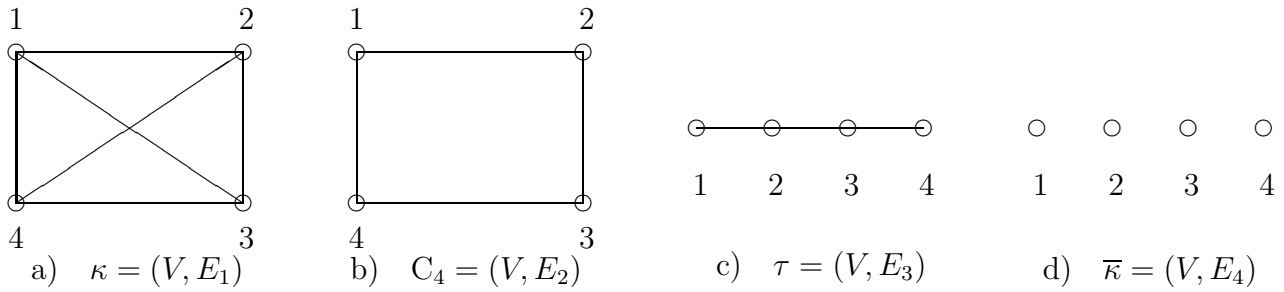


Figure 1.1: Four examples of graphs $G = (V, E_i), i = 1, 2, 3$, with $V = \{1, 2, 3, 4\}$. a) complete with $E_1 = \{(1,2), (1,3), (1,4), (2,3), (2,4), (3,4)\}$, b) a cycle with $E_2 = \{(1,2), (1,4), (2,3), (3,4)\}$, c) a tree graph called path with $E_3 = \{(1,2), (2,3), (3,4)\}$, and d) empty graph with $E_4 = \emptyset$.

Discriminant	Number of parameters for $p = 100$							
function	κ	C_4	τ	$\overline{\kappa}$	κ	C_4	τ	$\overline{\kappa}$
Quadratic	$3p + p^2$	$6p$	$6p - 2$	$4p$	10,300	600	598	400
Linear	$(p^2 + 5p)/2$	$4p$	$4p - 1$	$3p$	5,250	400	399	300

Table 1.1: Number of parameters in the quadratic and linear discriminant functions when considering concentration matrices associated with graphs of order p: κ, C_p, τ, and $\overline{\kappa}$.

We note that particular cases of discriminant analysis described in Section 1.1 are associated with known graphs: (i) LDA and QDA are associated with κ; and (ii) DLDA and DQDA with $\overline{\kappa}$. For example, for $p = 4$, the associated graphs are given in Figure 1.1a) and 1.1d), respectively.

When considering a GGM, two estimation problems exist: *parameter estimation* which corresponds to estimate $\boldsymbol{K}$ given an specific graph G, and *structure estimation or model selection*

which is to identify or estimate the dependence graph G.

ML can be used for the estimation of the concentration matrix. For an arbitrary GGM, ML estimates must be found iteratively, for example using the iterative proportional scaling (IPS) algorithm, see for example, Speed and Kiiveri (1986). Though for decomposable models, MLEs can be found in closed form (Lauritzen, 1996, and Højsgaard et al., 2012). Decomposable models are GGM with triangulated or chordal dependence graphs and their properties follow from the existence of the clique-separator factorization, see Lauritzen (1996, p. 145) or Green and Thomas (2013) for details. That is, the density function of a decomposable model associated with graph G can be factorized in terms of the set of cliques $\mathscr{C}$, from which a perfect sequence can be formed, and the set of separators $\mathscr{S}$ as

$$f_G(x_1, ..., x_p) = \frac{\prod_{\mathcal{C} \in \mathscr{C}} f(\boldsymbol{x}_\mathcal{C})}{\prod_{\mathcal{S} \in \mathscr{S}} f(\boldsymbol{x}_\mathcal{S})^{v(\mathcal{S})}}, \tag{1.15}$$

where $v(\mathcal{S})$ is the number of times that $\mathcal{S}$ occurs in a perfect sequence.

For example, the decomposable model associated with the chordal graph (Jiroušek and Přeučil, 1995) given in Figure 1.2 has a density that can be factorized as

$$f(x_1, ..., x_7) = \frac{f(x_1, x_2, x_3)\ f(x_2, x_3, x_6)\ f(x_2, x_5, x_6)\ f(x_4)\ f(x_6, x_7)}{f(x_2, x_3)\ f(x_2, x_6)\ f(x_6)},$$

where the set of *Cliques* is $\mathscr{C} = \{C_1, C_2, C_3, C_4, C_5\}$ with $C_1 = \{1, 2, 3\}$, $C_2 = \{2, 3, 6\}$, $C_3 = \{2, 5, 6\}$, $C_4 = \{4\}$ and $C_5 = \{6, 7\}$. The set of *separators* is $\mathscr{S} = \{S_1, S_2, S_3\}$ with $S_1 = \{2, 3\}$, $S_2 = \{2, 6\}$ and $S_3 = \{6\}$.

Other examples of triangulated graphs are given in Figures 1.1a), 1.1c) and 1.1d). The cycle given in 1.1b) is an example of a graph that is not chordal.

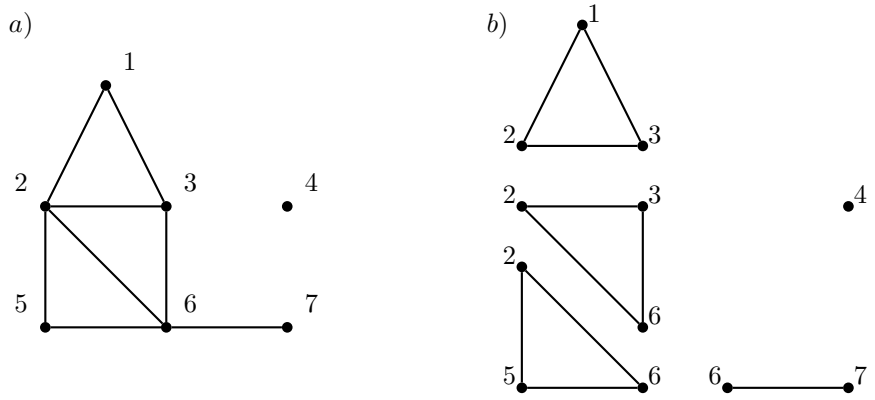


Figure 1.2: Example of a decomposable model: (a) the associated dependence graph and (b) the set of cliques: $C_1 = \{1, 2, 3\}$, $C_2 = \{2, 3, 6\}$, $C_3 = \{2, 5, 6\}$, $C_4 = \{4\}$, $C_5 = \{6, 7\}$

The MLE of the concentration matrix $\boldsymbol{K}_G$ for a decomposable model with graph G is given by

$$\widehat{\boldsymbol{K}}_G = \sum_{\mathcal{C} \in \mathscr{C}} [\boldsymbol{W}_{\mathcal{C}}^{-1}]^p - \sum_{\mathcal{S} \in \mathscr{S}} v(\mathcal{S}) [\boldsymbol{W}_{\mathcal{S}}^{-1}]^p. \tag{1.16}$$

where $\boldsymbol{W}$ is the sample covariance matrix given in (1.5) for a specific population; and for any vector of indexes $\boldsymbol{m}$, we let $\boldsymbol{W}_{\boldsymbol{m}}^{-1}$ denote the inverse of the submatrix of $\boldsymbol{W}$ formed by selecting the subset of rows and columns of $\boldsymbol{W}$ given by $\boldsymbol{m}$. We note that entry lk of $\boldsymbol{W}_{\boldsymbol{m}}^{-1}$ is associated with indexes m_l and m_k, so that we let $[\boldsymbol{W}_{\boldsymbol{m}}^{-1}]^p$ denote the $p \times p$ matrix such that $([\boldsymbol{W}_{\boldsymbol{m}}^{-1}]^p)_{m_l m_k} = (\boldsymbol{W}_{\boldsymbol{m}}^{-1})_{lk}$, $l, k = 1, 2, ..., |\boldsymbol{m}|$, and zero otherwise. For example, the matrix $[\boldsymbol{W}_{\mathcal{C}}^{-1}]^p$ associated to the clique $C_2 = \{2, 3, 6\}$ of the decomposable model with graph given in Figure 1.2a) is as follows

$$[\boldsymbol{W}_{\mathcal{C}_2}^{-1}]^7 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & a_{11} & a_{12} & 0 & 0 & a_{13} & 0 \\ 0 & a_{12} & a_{22} & 0 & 0 & a_{23} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & a_{13} & a_{23} & 0 & 0 & a_{33} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

where $\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{pmatrix} = \begin{pmatrix} w_{22} & w_{23} & w_{26} \\ w_{23} & w_{33} & w_{36} \\ w_{26} & w_{36} & w_{66} \end{pmatrix}^{-1}$ and w_{ij} is the entry ij of the matrix $\boldsymbol{W}$.

The analytical expressions for the MLEs in decomposable models are useful in high dimensional settings, where iterative methods can be computationally expensive. However, prior to the parameter estimation, the structure of G has to be specified, which is not an easy task.

1.3 Gaussian linear concentration models

We consider Gaussian models with linear structure in the concentration matrix. This structure was already suggested by Anderson (1970), where the concentration matrix can be expressed as a linear combination of given symmetric matrices as

$$\boldsymbol{K} = \boldsymbol{\Sigma}^{-1} = \sum_{h=0}^{q} \psi_h \boldsymbol{H}_h, \tag{1.17}$$

where $\boldsymbol{H}_h$, $h = 0, ..., q$, are symmetric and given matrices which are linearly independent and the values $\psi_0, \psi_1, ..., \psi_q$ are such as to make $\boldsymbol{K}$ positive definite.

These models have been recently called linear concentration models by Sturmfels and Uhler (2010). Particular cases of these patterns in the context of discriminant analysis have already been studied and applied, see for example, Bartlett and Please (1963), Penrose (1946–47), Smith (1946–47) and Dudoit and Fridlyand (2003); we describe some of them in Chapter 2.

Concentration matrices expressed as in (1.17) have taken relevance with the recent advances on GGMs, in the following subsections we consider some particular forms of expression 1.17 which correspond to models known as Coloured Gaussian graphical models (Højsgaard and Lauritzen, 2005, 2007 and 2008).

We note that the concentration matrix associated with a GGM with graph $G = (V, E)$ can be expressed as in (1.17) as follows

$$\boldsymbol{K}_G = \sum_{i=1}^{p} k_{ii}\boldsymbol{H}_{ii} + \sum_{\substack{i<j \\ (i,j)\in E}} k_{ij}\boldsymbol{H}_{ij},$$

where $\boldsymbol{H}_{ij}$ is the $p \times p$ matrix with all entries equal to zero, except for entries ij and ji which are equal to 1, $i, j = 1, ..., p$. For example, the concentration matrix associated with the complete graph in Figure 1.1 can be expressed as

$$\boldsymbol{K}_\kappa = \sum_{h=0}^{q} \psi_h \boldsymbol{H}_h, \tag{1.18}$$

where $q = 9$, $(\psi_0, \psi_1, \psi_2, \psi_3, \psi_4, \psi_5, \psi_6, \psi_7, \psi_8, \psi_9) = (k_{11}, k_{12}, k_{13}, k_{14}, k_{22}, k_{23}, k_{24}, k_{33}, k_{34}, k_{44})$, and $\boldsymbol{H}_h$ is a 4×4 matrix such that $\{\boldsymbol{H}_h\}_{ij}$ equals to 1 when $\{\boldsymbol{K}\}_{ij} = \psi_h$ and 0 otherwise, $h = 0, ..., 9$. And the one associated with $\overline{\kappa}$ is as follows

$$\boldsymbol{K}_{\overline{\kappa}} = \psi_0 \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} + \psi_1 \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} + \psi_2 \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} + \psi_3 \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{1.19}$$

where $q = 3$ and $(\psi_0, \psi_1, \psi_2, \psi_3) = (k_1, k_2, k_3, k_4)$.

1.3.1 Coloured graphical Gaussian models

Coloured graphical Gaussian models as introduced by Højsgaard and Lauritzen (2005, 2007, 2008) and further studied in Neufeld (2009) are models that impose constraints on some elements of the concentration matrix, or on the partial correlations. Two examples are RCON and RCOR models which are GGMs with some elements of the concentration and partial correlation matrix,

respectively, restricted to be equal. Another one is RCOP models which are RCON models such that the restrictions on the concentration matrix are also reflected on the partial correlation matrix.

A coloured graphical Gaussian model is associated with a coloured graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a partition of V into k subsets, $V = \{V_1, ..., V_k\}$, and $\mathcal{E}$ is a partition of E into l subsets, $E = \{E_1, ..., E_l\}$. Each subset $V_i, i = 1, ..., k$, represents a subset of variables whose partial variances are equal and whose vertices are coloured with the same colour. Each subset $E_i, i = 1, ..., l$, represents a subset of edges which are coloured with the same colour and whose corresponding partial covariances or partial correlations are equal. That is

$$V = V_1 \cup ... \cup V_k, \quad 1 \leq k \leq |V|, \qquad E = E_1 \cup ... \cup E_l, \quad 1 \leq l \leq |E|.$$

An RCON model has a coloured graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with restrictions on the elements of the concentration matrix, $\{k_{ij},\ i, j = 1, ..., p\}$, as follows.

a) $k_{ij} = 0 \iff (i, j) \notin E$.

b) $k_{ii} = k_{jj} \iff \{v_i, v_j\} \in V_r$, for some r, $r = 1, ..., k$.

c) $k_{ij} = k_{nm} \neq 0 \iff \{(i, j), (n, m)\} \in E_r$, for some r, $r = 1, ..., l$.

As an example of an RCON model, we consider one related to a coloured graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} = \{V_1, V_2\}$ and $\mathcal{E} = \{E_1, E_2\}$, where $V_1 = \{1, 3\}$, $V_2 = \{2, 4\}$, $E_1 = \{(1, 2), (2, 3), (1, 4)\}$ and $E_2 = \{(3, 4)\}$; and whose graph is the cycle of order 4 displayed in Figure 1.3.

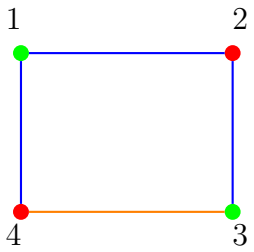


Figure 1.3: A coloured graph associated with an RCON model with $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, $\mathcal{V} = \{V_1, V_2\}$ and $\mathcal{E} = \{E_1, E_2\}$, where $V_1 = \{1, 3\}$, $V_2 = \{2, 4\}$, $E_1 = \{(1, 2), (2, 3), (1, 4)\}$ and $E_2 = \{(3, 4)\}$.

The corresponding concentration matrix is given by

$$\boldsymbol{K} = \begin{pmatrix} a & d & 0 & d \\ d & b & d & 0 \\ 0 & d & a & c \\ d & 0 & c & b \end{pmatrix} = \psi_0 \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} + \psi_1 \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} + \psi_2 \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} + \psi_3 \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \tag{1.20}$$

where $q = 3$, $(\psi_0, \psi_1, \psi_2, \psi_3) = (a, b, d, c)$, and a, b, c and d are different from zero.

In general, the concentration matrix of RCON models can be expressed as in (1.17) as follows

$$\boldsymbol{K} = \boldsymbol{\Sigma}^{-1} = \sum_{h=0}^{q} \psi_h \boldsymbol{H}_h, \tag{1.21}$$

where if $\{\boldsymbol{H}_k\}_{ij} \neq 0$, then $\{\boldsymbol{H}_l\}_{ij} = 0 \; \forall \; l \neq k$, $k, l = 0, ..., q$; and each matrix $\boldsymbol{H}_h$ corresponds to a vertex or edge colour class.

ML estimates for these models can be found using iterative algorithms, for example: the Scoring algorithm or the Iterative Partial Maximisation algorithm. These algorithms are described in Højsgaard and Lauritzen (2007) and implemented in the *gRc* package for R.

RCOP models constitute a subset of the RCON models and an example is presented in the following subsection.

1.3.2 Other examples of linear concentration models

RCON models have concentration matrices that can be expressed as in (1.17), though not any model with this kind of matrix is an RCON model. An example of a concentration matrix corresponding to a GGM that can be expressed as in (1.17), though it does not correspond to an

RCON model, is the following.

$$\boldsymbol{K} = \begin{pmatrix} a & b & 0 & 2b \\ b & a & 2b & 0 \\ 0 & 2b & a & b \\ 2b & 0 & b & a \end{pmatrix} = a \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} + b \begin{pmatrix} 0 & 1 & 0 & 2 \\ 1 & 0 & 2 & 0 \\ 0 & 2 & 0 & 1 \\ 2 & 0 & 1 & 0 \end{pmatrix}.$$

RCON models whose concentration matrix can be expressed as in (1.17) impose restrictions on the off-diagonal elements of the concentration matrix which are linear, whereas RCOR models impose non linear restrictions on the concentrations or partial covariances. An RCOR model has a coloured graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where the elements of the concentration matrix, $\{k_{ij},\ i, j = 1, ..., p\}$, are constrained as follows.

a) $k_{ij} = 0 \Longleftrightarrow (i, j) \notin E$.

b) $k_{ii} = k_{jj} \Longleftrightarrow \{v_i, v_j\} \in V_r$, for some r, $r = 1, ..., k$.

c) $\rho_{ij|V\setminus\{v_i,v_j\}} = \rho_{nm|V\setminus\{v_n,v_m\}} \neq 0 \Longleftrightarrow \{(i, j), (n, m)\} \in E_r$, for some r, $r = 1, ..., l$;

where $\rho_{ij|V\setminus\{v_i,v_j\}} = -\dfrac{k_{ij}}{\sqrt{k_{ii}}\sqrt{k_{jj}}}$.

An example of an RCOR model with coloured graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} = \{V_1, V_2\}$ and $\mathcal{E} = \{E_1, E_2, E_3\}$, where $V_1 = \{1, 2\}$, $V_2 = \{3, 4\}$, $E_1 = \{(1, 2), (3, 4)\}$, $E_2 = \{(1, 4)\}$ and $E_3 = \{(2, 3)\}$; is shown in Figure 1.4.

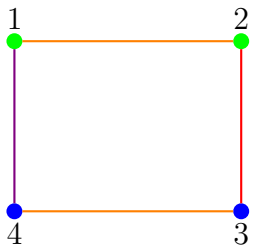


Figure 1.4: A coloured graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ associated with an RCOR model; with $\mathcal{V} = \{V_1, V_2\}$ and $\mathcal{E} = \{E_1, E_2, E_3\}$, where $V_1 = \{1, 2\}$, $V_2 = \{3, 4\}$, $E_1 = \{(1, 2), (3, 4)\}$, $E_2 = \{(1, 4)\}$ and $E_3 = \{(2, 3)\}$.

The corresponding concentration matrix to the RCOR model with graph shown in Figure 1.4 is as follows.

$$\boldsymbol{K} = \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} a & \frac{ca}{b} & 0 & d \\ \frac{ca}{b} & a & e & 0 \\ 0 & e & b & c \\ d & 0 & c & b \end{pmatrix},$$

where a, b, c, d and e represent values different from zero. We note that this concentration matrix cannot be expressed as in (1.17).

An RCOP model is a GGM that is both an RCON and an RCOR model. These models are defined using some properties of the associated graph $G = (V, E)$ as follows. Let $Aut(G)$ denote the group of automorphisms of G. Let Γ be a subgroup of $Aut(G)$, vertex orbits are defined as the class of equivalence

$$i \equiv_\Gamma j \Leftrightarrow j = \sigma(i) \quad \text{for some} \quad \sigma \in \Gamma,$$

where $i, j \in V$. Also, edge orbits are defined as the class relation

$$(i, j) \equiv_\Gamma (k, l) \Leftrightarrow (k, l) = (\sigma(i), \sigma(j)) \quad \text{for some} \quad \sigma \in \Gamma,$$

with $\{(i, j), (k, l)\} \in E$.

An RCOP model is a model whose vertex colour classes correspond to the vertex orbits for a subgroup Γ of $Aut(G)$. An the edge colour classes correspond to the edge orbits of Γ.

For example, let $G = (V, E)$ be the 4–cycle graph shown in Figure 1.1b). Then the group of automorphisms of G is $Aut(G) = \{id, (13), (24), (13)(24), (12)(34), (14)(23), (1234), (1432)\}$. Let $\Gamma = \{id, (13)(24)\}$, then $\mathcal{V} = \{V_1, V_2\}$ and $\mathcal{E} = \{E_1, E_2\}$, where $V_1 = \{1, 3\}$, $V_2 = \{2, 4\}$, $E_1 = \{(1, 2), (3, 4)\}$ and $E_2 = \{(1, 4), (2, 3)\}$. Then the coloured graph associated with the RCOP model generated considering $\Gamma = \{id, (13)(24)\}$ for the 4–cycle graph is the one shown in Figure 1.5.

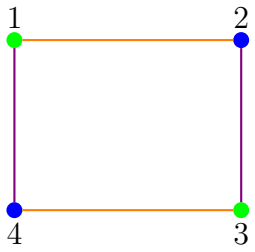


Figure 1.5: Coloured 4–cycle graph associated with the RCOP model generated by $\Gamma = \{id, (13)(24)\}$.

The concentration matrix associated with the RCOP model shown in Figure 1.5 is as follows.

$$\boldsymbol{K} = \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} a & c & 0 & d \\ c & b & d & 0 \\ 0 & d & a & c \\ d & 0 & c & b \end{pmatrix},$$

where a, b, c and d are all different from zero.

We note that the symmetries imposed by the coloured graph in an RCOP model are also reflected in the partial correlation and covariance matrices. For example, the covariance matrix associated with the RCOP model shown in Figure 1.5 is

$$\boldsymbol{\Sigma} = \frac{1}{h} \begin{pmatrix} b(ba - c^2 - d^2) & -c(ba - c^2 + d^2) & 2bcd & -(ba - d^2 + c^2)d \\ -c(ba - c^2 + d^2) & a(ba - c^2 - d^2) & -(ba - d^2 + c^2)d & 2cda \\ 2bcd & -(ba - d^2 + c^2)d & b(ba - c^2 - d^2) & -c(ba - c^2 + d^2) \\ -(ba - d^2 + c^2)d & 2cda & -c(ba - c^2 + d^2) & a(ba - c^2 - d^2) \end{pmatrix},$$

where $h = a^2b^2 - 2bac^2 - 2d^2ba + d^4 + c^4 - 2c^2d^2$.

Votaw (1948) and Wilks(1946) imposed symmetry restrictions on the covariance matrix which are also reflected in the concentration matrices. These models are examples of RCOP models.

1.3.3 Comments

For the parameter estimation in GGMs, the algorithms depend on whether the graph is decomposable. When it is, there exists an analytical expression for the MLE of the concentration matrix.

When it is not, there are some algorithms to find numerical solutions for the ML estimates, for example: the IPS algorithm (Lauritzen, 1996, p134) and another given in Hastie *et al.* (2009, p. 634). Both algorithms require that the structure of the matrix is known, and the IPS additionally requires that the cliques of the graph are given.

For the estimation of coloured graphical Gaussian models, at least two algorithms are implemented in the package *gRc* in R: Scoring algorithm and Iterative partial maximization algorithm. Both are described in Højsgaard and Lauritzen (2007).

For the estimation of linear concentration models, given that the structure of the matrix is known, Anderson (1970) described an algorithm to solve the corresponding nonlinear equations system using the Newton-Raphson algorithm.

For the case of RCON models with additional restrictions on the mean vectors, Gehrmann and Lauritzen (2012) give the necessary and sufficient condition for the estimation of the means to be independent from the estimation of the concentration matrix when using ML.

The existence of MLEs has been assumed, however, Sturmfels and Uhler (2010) and Ulher (2012) study the problem of the minimum number of observations that ensure the existence of MLEs with probability one.

In the following chapter, some early examples in discriminant analysis are described, where some linear restrictions on the concentration matrices appear.

2. Early examples of the use of Gaussian linear concentration models

In Chapter 1, we described some models with linear restrictions on the concentration matrix with the purpose of using them in the context of discriminant analysis. Instances of these models have been studied recently, and in this chapter, we describe some of the early cases considered in the literature. We also describe some restrictions on the mean vectors which can be considered in discriminant analysis: (i) means with some elements, in one or both vectors, restricted to be equal, and (ii) zero mean differences. We conclude the chapter giving two practical examples, with data taken from the described cases, to show how the estimation of the parameters can be done using tools already developed.

2.1 Unrestricted group means

With no assumption on the structure of the mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, their MLEs are independent from the corresponding to the concentration matrices $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$, and are given by the sample mean vectors as in (1.4). $\boldsymbol{K}_c$ is then estimated by maximizing the profile likelihood function, that is the likelihood function evaluated in $\widehat{\boldsymbol{\mu}}_c$, $c = 1, 2$.

An early example in the context of classification, where there is no restriction on the mean vectors but with some structure on the concentration matrices, is given by Penrose (1946–1947) using the concepts of size y shape. Also, we can find examples in the context of the classification of biological samples using gene expression data from DNA microarray experiments where naive

Bayes classification is very useful, see Dudoit and Fridlyand (2003).

2.1.1 Penrose's concepts of size and shape

The old idea of describing an object with two measurements, one of size and another of shape, has been referred and studied in various areas of research, particularly in morphometrics. In the context of Principal components, Rao (1964) gives some comments on what different authors call size and shape. Size is basically a weighted sum of p variables with positive weights; whereas shape is a weighted sum of p variables with weights summing to zero.

In the context of discriminant analysis, Penrose (1946–47) defines size as the linear function $W = X_1 + ... + X_p$ and shape as $P = k_1X_1 + ... + k_pX_p$ with $\sum_{i=1}^{p} k_i = 0$. He also assumes that on each of the two populations the covariance and concentration matrices has the same pattern:

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho & \cdots & \rho & \rho \\ \rho & 1 & \cdots & \rho & \rho \\ & & \ddots & & \\ \rho & \rho & \cdots & 1 & \rho \\ \rho & \rho & \cdots & \rho & 1 \end{pmatrix} = (1-\rho)\boldsymbol{I} + \rho \boldsymbol{z}\boldsymbol{z}^t, \qquad \boldsymbol{K} = \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} c & d & \cdots & d & d \\ d & c & \cdots & d & d \\ & & \ddots & & \\ d & d & \cdots & c & d \\ d & d & \cdots & d & c \end{pmatrix} = (c-d)\boldsymbol{I} + d\boldsymbol{z}\boldsymbol{z}^t,$$

where $\boldsymbol{z} = (1, ..., 1)$, $c = \dfrac{-(\rho(p-2)+1)}{(\rho-1)[\rho(p-1)+1]}$ and $d = \dfrac{\rho}{(\rho-1)[\rho(p-1)+1]}$. When $\boldsymbol{K} = \boldsymbol{\Sigma}^{-1}$ has this special pattern, the linear discriminant function (1.6) becomes:

$$\begin{aligned} \ln \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} = & \frac{-P\sum_{i=1}^{p}(\mu_{1i}-\mu_{2i})}{(\rho-1)p} + \frac{W\sum_{i=1}^{p}(\mu_{1i}-\mu_{2i})}{p[\rho(p-1)+1]} + \frac{1}{2}\frac{1}{\rho-1}\sum_{i=1}^{p}(\mu_{1i}-\mu_{2i})(\mu_{1i}+\mu_{2i}) \\ & -\frac{1}{2}\frac{\rho}{(\rho-1)[\rho(p-1)+1]}\sum_{i=1}^{p}(\mu_{1i}-\mu_{2i})\sum_{i=1}^{p}(\mu_{1i}+\mu_{2i}) \\ = & \frac{-\sum_{i=1}^{p}(\mu_{1i}-\mu_{2i})}{(\rho-1)p}\left[P + \frac{W(1-\rho)}{[\rho(p-1)+1]}\right] + \frac{1}{2}\frac{1}{\rho-1}\sum_{i=1}^{p}(\mu_{1i}-\mu_{2i})(\mu_{1i}+\mu_{2i}) \\ & -\frac{1}{2}\frac{\rho}{(\rho-1)[\rho(p-1)+1]}\sum_{i=1}^{p}(\mu_{1i}-\mu_{2i})\sum_{i=1}^{p}(\mu_{1i}+\mu_{2i}). \end{aligned}$$

This model considers strong assumptions on the concentration matrix, though the graph associated with this model is the complete graph κ. We note, however, that when these assumptions are satisfied, the parameters to be estimated reduce to the means and one parameter ρ.

2.1.2 Diagonal discriminant analysis

Naive Bayes classification for Gaussian class conditional densities corresponds to diagonal discriminant analysis where the covariance and concentration matrices are diagonal as follows

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & 0 & \cdots & 0 & 0 \\ 0 & \sigma_{22} & \cdots & 0 & 0 \\ & & \ddots & & \\ 0 & 0 & \cdots & \sigma_{(p-1)(p-1)} & 0 \\ 0 & 0 & \cdots & 0 & \sigma_{pp} \end{pmatrix} \quad \text{and} \quad \boldsymbol{K} = \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} k_{11} & 0 & \cdots & 0 & 0 \\ 0 & k_{22} & \cdots & 0 & 0 \\ & & \ddots & & \\ 0 & 0 & \cdots & k_{(p-1)(p-1)} & 0 \\ 0 & 0 & \cdots & 0 & k_{pp} \end{pmatrix},$$

with $\sigma_{ii} > 0$, $i = 1, ..., p$.

The fundamental assumption of naive Bayes classification is that the variables are marginally independent given the class or group. The graph for each group is a set of p dots without edges, the edgeless graph $\overline{\kappa}$.

We can see that when $\boldsymbol{K} = \boldsymbol{\Sigma}^{-1}$ has this special pattern and under the assumption of equal covariance matrices, the linear discriminant function in (1.6) becomes:

$$\ln \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} = \sum_{i=1}^{p} k_{ii} \left[(\mu_{1i} - \mu_{2i}) x_i - \tfrac{1}{2}(\mu_{1i}^2 - \mu_{2i}^2) \right].$$

On the other hand, assuming that both $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$ have this special pattern and are arbitrary, the quadratic function in (1.1) reduces to a simpler expression as follows.

$$\ln \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} = \tfrac{1}{2} \sum_{i=1}^{p} \left[(x_i - \mu_{2i})^2 k_{ii}^{(2)} - \ln(k_{ii}^{(2)}) - (x_i - \mu_{1i})^2 k_{ii}^{(1)} + \ln(k_{ii}^{(1)}) \right].$$

In the context of Bioinformatics and Genetics, where it is needed to estimate covariance matrices of large dimension with relatively few observations, this pattern in the covariance matrices is often used. For example, Dudoit and Fridlyand (2003) compared Diagonal discriminant analysis with other four models using microarrays to study the molecular variations among tumours, we describe their main results in Section 2.4 where we consider the data for the illustrative example.

2.2 Group means vectors partitioned into subsets of equal elements

It is also possible to impose restrictions on the elements of the mean vectors. Anderson (1970) gives some examples where it is possible to assume a linear structure on the mean vector, he also gives some comments on the estimation of the mean vector, covariance and concentration matrices. Recently, Gehrmann and Lauritzen (2012) studied the problem of estimation when linear constraints are imposed on the mean vector, in particular when the concentration matrix has also restrictions associated with RCON, RCOR or RCOP models.

The problem of restrictions on the means and on elements of the concentration matrix has been studied in Votaw (1948), and some examples are given in Votaw et al. (1950), though these examples are not given in the context of discriminant analysis.

2.2.1 Votaw's concept of compound symmetry

Votaw (1948) studies the case of compound symmetry in multivariate Gaussian distributions, as a generalization of the concept of symmetry given in Wilks (1946). Compound symmetry corresponds to Gaussian distributions with concentration matrices with a structure of symmetry by blocks. One of the assumed structures for the mean vectors and concentration matrices is as follows.

$$\boldsymbol{K} = \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \boldsymbol{A}_1 & \boldsymbol{D}_{12} & \boldsymbol{D}_{13} & . & . & \boldsymbol{D}_{1h} \\ \boldsymbol{D}_{12}^t & \boldsymbol{A}_2 & \boldsymbol{D}_{23} & . & . & \boldsymbol{D}_{2h} \\ \boldsymbol{D}_{13}^t & \boldsymbol{D}_{23}^t & \boldsymbol{A}_3 & . & . & \boldsymbol{D}_{3h} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \boldsymbol{D}_{1h}^t & \boldsymbol{D}_{2h}^t & \boldsymbol{D}_{3h}^t & . & . & \boldsymbol{A}_h \end{pmatrix} \quad \text{and} \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \boldsymbol{\mu}_3 \\ . \\ . \\ \boldsymbol{\mu}_h \end{pmatrix},$$

where $\boldsymbol{A}_j$ is a matrix of dimension $p_j \times p_j$ and $\boldsymbol{\mu}_j$ is a mean vector of dimension p_j with the following structures

$$\boldsymbol{A}_j = \begin{pmatrix} a_j & b_j & . & b_j \\ b_j & a_j & . & b_j \\ . & . & . & . \\ b_j & b_j & . & a_j \end{pmatrix}, \quad \boldsymbol{\mu}_j = \begin{pmatrix} \mu_j \\ \mu_j \\ . \\ \mu_j \end{pmatrix},$$

where $p_j \geq 1$, $j = 1, 2, ..., h$ and $\Sigma_{j=1}^{h} p_j = p$. $\boldsymbol{D}_{ij}$ is a matrix of dimension $p_i \times p_j$ structured as

$$\boldsymbol{D}_{ij} = \begin{pmatrix} c_{ij} & c_{ij} & . & c_{ij} \\ c_{ij} & c_{ij} & . & c_{ij} \\ . & . & . & . \\ c_{ij} & c_{ij} & . & c_{ij} \end{pmatrix}.$$

In Wilks (1946), symmetry corresponds to the case of elements instead of matrices in $\boldsymbol{\Sigma}^{-1}$, *i.e.* the case with $h = 1$.

Votaw et al. (1950) illustrate the concept of compound symmetry with an example of four measurements on each of 16 rabbits. These measurements correspond to the anterior or posterior muscle weights in both of the hind legs. Let X_1, X_2, X_3 and X_4 be the anterior left, anterior right, posterior left, and posterior right muscle, respectively. From six test of hypothesis presented in the article, one is of our interest, the hypothesis $H_0 : \{\mu_1 = \mu_2,\ \mu_3 = \mu_4,\ \sigma_1 = \sigma_2,\ \sigma_3 = \sigma_4,\ \rho_{13} = \rho_{14} = \rho_{23} = \rho_{24}\}$, which corresponds to test whether variable X_1 is interchangeable with X_2, and X_3 with X_4. The concentration matrix and mean vector for this hypothesis are as follows.

$$\boldsymbol{K} = \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} a & b & c & c \\ b & a & c & c \\ c & c & d & e \\ c & c & e & d \end{pmatrix} \quad \text{and} \quad \boldsymbol{\mu} = \begin{pmatrix} w \\ w \\ q \\ q \end{pmatrix}. \tag{2.1}$$

The example on the rabbits can be seen as a particular case of the models studied by Gehrmann and Lauritzen (2012). The coloured graph associated with the rabbits example is shown in Figure 2.1.

The model associated with the concentration matrix given in (2.1), with no assumption about the mean vectors, is an RCOP model generated by $\Gamma = \{id, (12), (34), (12)(34)\}$ with incident graph the complete of order 4.

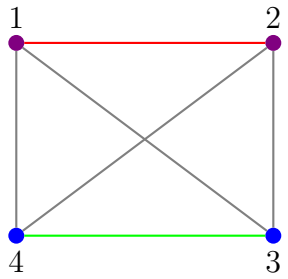


Figure 2.1: Coloured graph corresponding to the concentration matrix for the rabbits example presented in Votaw et al. (1950).

2.3 Zero mean differences

In some practical cases of discrimination, it is assumed equality of the group mean vectors, $\boldsymbol{\mu_1} = \boldsymbol{\mu_2}$. In spite of this assumption, it is still possible to discriminate among the two populations, provided that the covariance matrices are different. Bartlett and Please (1963) studied this case when equal means are assumed. They considered patterned concentration matrices that correspond to an RCOP model.

2.3.1 Bartlett's model for discriminant analysis with zero mean differences

Bartlett and Please (1963) consider the particular case of discriminant analysis with zero mean differences, $\boldsymbol{\mu_1} = \boldsymbol{\mu_2}$, and different covariance and hence concentration matrices with the following pattern:

$$\boldsymbol{\Sigma} = \begin{pmatrix} a & b & \cdots & b & b \\ b & a & \cdots & b & b \\ & & \ddots & & \\ b & b & \cdots & a & b \\ b & b & \cdots & b & a \end{pmatrix} = (a-b)\boldsymbol{I} + b\boldsymbol{z}\boldsymbol{z}^t, \quad \boldsymbol{K} = \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} c & d & \cdots & d & d \\ d & c & \cdots & d & d \\ & & \ddots & & \\ d & d & \cdots & c & d \\ d & d & \cdots & d & c \end{pmatrix} = (c-d)\boldsymbol{I} + d\boldsymbol{z}\boldsymbol{z}^t, \tag{2.2}$$

where $c = \dfrac{-[b(p-2)+a]}{(b-a)[b(p-1)+a]}$ and $d = \dfrac{b}{(b-a)[b(p-1)+a]}$.

A model with this concentration matrix is an RCON, whose dependence graph is the complete of order p and is coloured as follows, $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, with $\mathcal{V} = \{V\}$ and $\mathcal{E} = \{E\}$. For example, the graph associated with the corresponding RCON model with $p = 4$ is given in Figure 2.2.

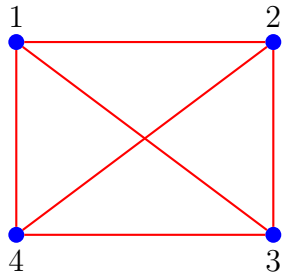


Figure 2.2: Coloured graph associated with the model considered by Bartlett and Please (1963) when $p = 4$; $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ with $\mathcal{V} = \{V\}$ and $\mathcal{E} = \{E\}$.

We observe that Penrose (1946–47) uses a special case for these matrices with $a = 1$. Bartlett and Please (1963) consider the pattern on the concentration matrix given in (2.2) with $a = 1$ and $b = \rho_1$ for group one, and with $a = \sigma^2$ and $b = \sigma^2\rho_2$ for group two. That is

$$\boldsymbol{\Sigma}_1 = (1-\rho_1)\boldsymbol{I} + \rho_1\boldsymbol{z}\boldsymbol{z}^t, \ \boldsymbol{\Sigma}_2 = \sigma^2[(1-\rho_2)\boldsymbol{I} + \rho_2\boldsymbol{z}\boldsymbol{z}^t].$$

Assuming that the covariance matrices have this pattern, their corresponding inverses also have the same pattern, as:

$$\boldsymbol{K}_1 = (c-d)\boldsymbol{I} + d\boldsymbol{z}\boldsymbol{z}^t, \ \boldsymbol{K}_2 = \frac{1}{\sigma^2}\left((e-f)\boldsymbol{I} + f\boldsymbol{z}\boldsymbol{z}^t\right), \tag{2.3}$$

where $c = \dfrac{-[\rho_1(p-2)+1]}{(\rho_1-1)[\rho_1(p-1)+1]}$, $d = \dfrac{\rho_1}{(\rho_1-1)[\rho_1(p-1)+1]}$, $e = \dfrac{-[\rho_2(p-2)+1]}{(\rho_2-1)[\rho_2(p-1)+1]}$ and $f = \dfrac{\rho_2}{(\rho_2-1)[\rho_2(p-1)+1]}$.

Assuming $\boldsymbol{\mu_1} = \boldsymbol{\mu_2} = \mathbf{0}$ and concentration matrices as in (2.3), the quadratic function in (1.1) reduces to a simpler expression as follows.

$$\ln \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} = -\frac{1}{2}\left\{\left[\frac{1}{1-\rho_1} - \frac{1}{\sigma^2(1-\rho_2)}\right] z_1 - \left[\frac{\rho_1}{1-\rho_1}\frac{1}{1+(p-1)\rho_1} - \frac{\rho_2}{\sigma^2(1-\rho_2)}\frac{1}{1+(p-1)\rho_2}\right] z_2\right\}$$
$$+\frac{1}{2}\ln\left(\frac{1+(p-1)\rho_2}{1+(p-1)\rho_1}\left(\frac{\rho_2-1}{\rho_1-1}\right)^{p-1}\left(\sigma^2\right)^p\right),$$

where $z_1 = x_1^2 + ... + x_p^2$ and $z_2 = (x_1 + ... + x_p)^2$. In Penrose (1946–47), $\sqrt{z_2}$ corresponds to the size variable.

The covariance pattern in (2.2) considered by Bartlett and Please (1963) corresponds to an RCOP model generated by $\Gamma = Aut(G)$ where G is a complete graph of order p.

2.4 Two practical examples

We consider two data sets. One has been analysed and reported by Dudoit and Fridlyand (2003), we replicate part of the analysis of this data and present the obtained results. The second data set has been presented and analysed by Votaw et al. (1950) in the context of one population.

2.4.1 Breast cancer data from Dudoit and Fridlyand (2003)

The data set deals with the supervised classification of 49 breast tumour mRNA samples, each sample or observation has 7,129 measurements or variables. Also, estrogen receptor status was measured for each tumour sample: $ER+$ (25 samples) and $ER-$ (24 samples). It is believed that different biological mechanisms are involved in the development of breast cancer depending on the

ER status of a patient, and we consider this variable as the one that identifies the two groups.

Dudoit and Fridlyand (2003) compared six different methods for the classification: knn with Euclidean distance; Diagonal linear discriminant analysis, DLDA; Diagonal quadratic discriminant analysis, DQDA; LogitBoost; Random forest; and Support vector machine, SVM. They used subsets of different number of variables: 10, 50, 100, 500, 1000, and 7129, and found that the two simple methods of DLDA and DQDA had a good performance when compared to the other more complex methods.

We downloaded the data base from *http* : *//data.cgt.duke.edu/west.php*, PNAS_paper.zip. This file contains 49 original files, one for each sample or observation. Each of the 49 observations was preprocessed, as indicated by Dudoit and Fridlyand (2003), as follows.

i) thresholding of each x_{jk} value, $j = 1, ..., 49, k = 1, ..., 7129$, with a floor of 100 and ceiling of 16,000; $100 \leq x_{jk} \leq 16,000$.
ii) base-10 logarithmic transformation of each value; $y_{jk} = \log_{10}(x_{jk})$.
iii) standardization of each of the 49 observations to have zero mean and unit variance; $z_{jk} = \dfrac{y_{jk} - \overline{y}_j}{\sqrt{\sum_{k=1}^{7129}(y_{jk} - \overline{y}_j)^2/7129}}$.

The classification errors were estimated by cross-validation in two ways: external and internal. The external is performed in three steps as described bellow.

For each observation, $j = 1, ..., 49$, we did the following.

i) With the rest of 48 observations, we selected 10 variables from 7,129 available, those with the largest t absolute value statistic for the two sample t-test of equality of means.
ii) With the 10 selected variables, the functions corresponding to DLDA and DQDA are estimated.

iii) The observation j is classified with the estimated functions.

The internal cross-validation is performed in two steps as follows. Considering the 49 observations, we selected 10 variables from 7129 available, those with the largest absolute value of the t statistic for the two sample t-test of equality of means, then for the jth observation, $j = 1, ..., 49$, we did the following.

i) With the 10 selected variables, the functions corresponding to DLDA and DQDA are estimated.

ii) The jth observation is classified with the estimated functions.

The previous cross-validation procedures were also performed using the W statistic for the Wilcoxon rank sum test, a nonparametric alternative to the two sample t-test, instead of the t statistic. The estimated classification errors by external cross-validation are displayed in Table 2.1 and the ones by internal cross-validation are displayed in Table 2.2.

		Observed							
		DLDA				DQDA			
		t-statistic		W-statistic		t-statistic		W-statistic	
		$ER-$	$ER+$	$ER-$	$ER+$	$ER-$	$ER+$	$ER-$	$ER+$
Predicted	$ER-$	20	4	21	4	20	3	20	3
	$ER+$	4	21	3	21	4	22	4	22

Table 2.1: Classification errors estimated by external cross-validation. Ten variables or genes were selected out of 7,129 using the t-statistic and the W-statistic. DLDA: Diagonal Linear Discriminant Analysis and DQDA: Diagonal Quadratic Discriminant analysis.

		Observed							
		DLDA				DQDA			
		t-statistic		W-statistic		t-statistic		W-statistic	
		$ER-$	$ER+$	$ER-$	$ER+$	$ER-$	$ER+$	$ER-$	$ER+$
Predicted	$ER-$	22	2	22	2	21	3	22	2
	$ER+$	2	23	2	23	3	22	2	23

Table 2.2: Classification errors estimated by internal cross-validation. Ten variables or genes were selected out of 7,129 using the t-statistic and the W-statistic, and the 49 observations. DLDA: Diagonal Linear Discriminant Analysis and DQDA: Diagonal Quadratic Discriminant analysis.

Figures 2.3 and 2.4 display the linear projection in two dimensions of the 49 observations in a 10–dimensional space. Ten variables or genes were selected out of 7,129 using the t-statistic and the W-statistic, respectively, and the 49 observations. The axes corresponds to the directions given by the first two principal components.

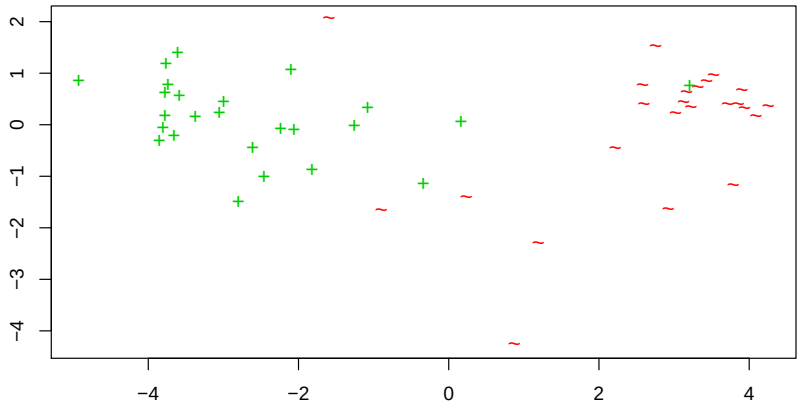


Figure 2.3: Linear projection in two dimensions of the 49 observations in a 10–dimensional space. The axes corresponds to the directions given by the first two principal components. The ten selected variables or genes using the t statistic, those with the largest absolute value of the t statistic for the two sample t-test of equality of means, and the 49 observations were: $L08044\_s\_at$, $M26311\_s\_at$, $U39840\_at$, $U41060\_at$, $U42408\_at$, $U79293\_at$, $X03635\_at$, $X17059\_s\_at$, $X58072\_at$, $X83425\_at$.

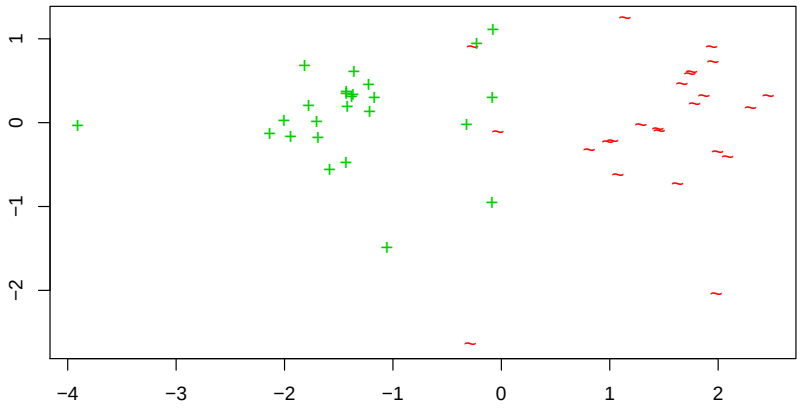


Figure 2.4: Linear projection in two dimensions of the 49 observations in a 10–dimensional space. The axes corresponds to the direction given by the first two principal components. The ten selected variables or genes using the W statistic, those with the largest absolute value of the W statistic for the two sample Wilcoxon test, and the 49 observations were: $D38521\_at$, $HG4716.HT5158\_at$, $J03827\_at$, $L17131\_rna1\_at$, $M16038\_at$, $M24485\_s\_at$, $M26062\_at$, $M26311\_s\_at$, $U42408\_at$, $X87212\_at$.

2.4.2 Rabbit example from Votaw et al. (1950)

The second example is taken from Votaw et al. (1950). There are four measurements on each of 16 rabbits corresponding to the anterior or posterior muscle weights in both of the hind legs. Let X_1, X_2, X_3 and X_4 be the measurement on the anterior left, anterior right, posterior left, and posterior right muscle, respectively. We consider the estimation for the hypothesis

$$H_0 : \{\mu_1 = \mu_2,\ \mu_3 = \mu_4,\ \sigma_1 = \sigma_2,\ \sigma_3 = \sigma_4,\ \rho_{13} = \rho_{14} = \rho_{23} = \rho_{24}\},$$

which corresponds to test whether variable X_1 is interchangeable with X_2, and X_3 with X_4. The concentration matrix and mean vector for this hypothesis are as in (2.1). The coloured graph associated with this example is shown in Figure 2.1. In Figure 2.5, we present some graphs that display the 16 observations and four variables. We observe that the assumptions given by $H_0 : \{\mu_1 = \mu_2,\ \mu_3 = \mu_4,\ \sigma_1 = \sigma_2,\ \sigma_3 = \sigma_4,\ \rho_{13} = \rho_{14} = \rho_{23} = \rho_{24}\}$ are plausible.

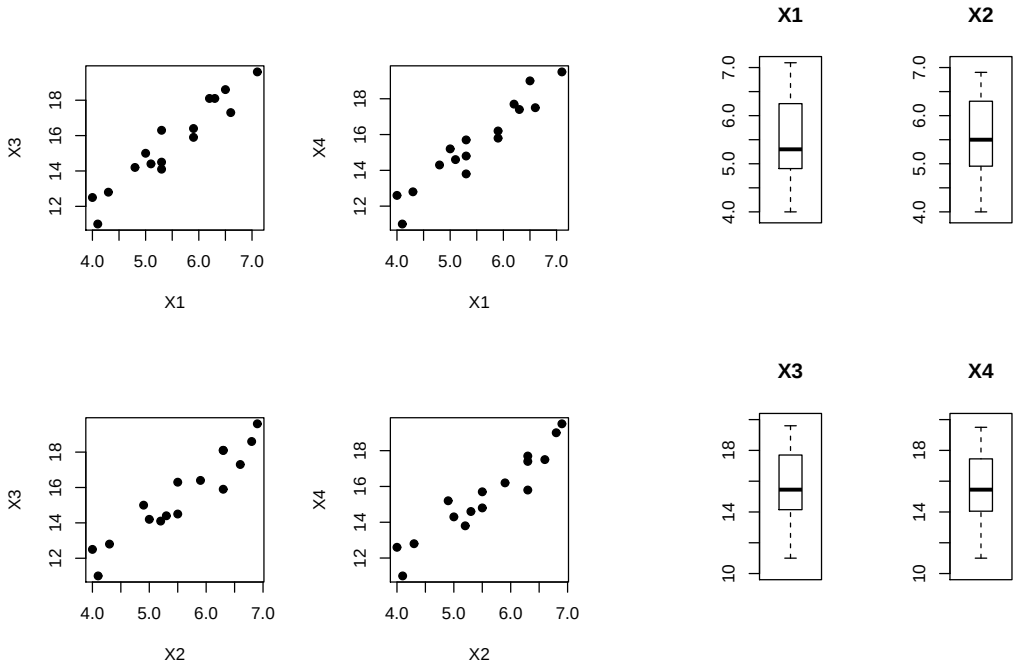


Figure 2.5: Scatter plots and Box plots for variables X1, X2, X3, and X4.

We use the gRc R-package, Højsgard and Lauritzen (2007), to get the estimates. First without considering restrictions on the means and then considering the restrictions. For the estimation with restricted means we have used the fact that the model is RCOP, and hence the estimation of the means is independent from the one of the concentration matrix, see Gehrmann and Lauritzen (2012), so we estimate the means and then use the gRc package to estimate the concentration matrix with given estimated means. The results are as follows. Considering restrictions on the concentration matrix and unrestricted means.

$$\widehat{\boldsymbol{K}} = \widehat{\boldsymbol{\Sigma}}^{-1} = \begin{pmatrix} 41.8509 & -35.0722 & -1.2683 & -1.2683 \\ -35.0722 & 41.8509 & -1.2683 & -1.2683 \\ -1.2683 & -1.2683 & 10.5230 & -9.4853 \\ -1.2683 & -1.2683 & -9.4853 & 10.5230 \end{pmatrix} \quad \text{and} \quad \widehat{\boldsymbol{\mu}} = \begin{pmatrix} 5.4812 \\ 5.5562 \\ 15.5500 \\ 15.4937 \end{pmatrix}.$$

Considering restrictions on both, the concentration matrix and the vector of means.

$$\widehat{\boldsymbol{K}} = \widehat{\boldsymbol{\Sigma}}^{-1} = \begin{pmatrix} 36.4119 & -31.6992 & -1.4431 & -1.4431 \\ -31.6992 & 36.4119 & -1.4431 & -1.4431 \\ -1.4431 & -1.4431 & 11.6011 & -10.4204 \\ -1.4431 & -1.4431 & -10.4204 & 11.6011 \end{pmatrix} \quad \text{and} \quad \widehat{\boldsymbol{\mu}} = \begin{pmatrix} 5.5187 \\ 5.5187 \\ 15.5219 \\ 15.5219 \end{pmatrix}.$$

We have compared these estimates with those calculated with an analytical expression given in Votaw et al. (1950). We did not find differences. In Appendix H, we present the R scripts used to generate the numerical results.

These examples consider some restrictions on each concentration matrix for which algorithms or closed form expressions exist to obtain ML estimates. However, other restrictions for elements within each concentration matrix or between elements of the two matrices can be considered, and the structure of these matrices in most of the cases is unknown. In the following chapters, these two aspects are studied.

3. Restrictions between elements of two concentration matrices

In this chapter, we consider that the vector $(C, \boldsymbol{X})$ follows a CG-distribution (see, for example, Lauritzen, 1996, s. 6.1.1) and that we are interested in using discriminant analysis in order to predict the membership group of a new observation $\boldsymbol{x}$. As mentioned in Chapter 1, we need to estimate the parameters $\pi_1,\ \pi_2,\ \boldsymbol{\mu}_1,\ \boldsymbol{\mu}_2,\ \boldsymbol{K}_1$ and $\boldsymbol{K}_2$.

Usually, the group mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, and covariance matrices or their inverses, are not restricted. For example, there is no assumption on the presence of marginal or conditional independences among the continuous variables.

We consider restrictions as in (1.17) and some elements of one concentration matrix equal to their corresponding ones in the other concentration matrix. We adapt the Iterative Partial Maximization (IPM) algorithm (Jensen et al., 1991) to get ML estimates, and derive the necessary expressions for this.

3.1 Equalities between elements of concentration matrices with linear restrictions

One could consider some constrains as we described in Chapters 1 and 2. Additionally, since in discriminant analysis we have two populations, we can also consider that some parameters in $\boldsymbol{K}_1$

are restricted to being equal to their corresponding in $\boldsymbol{K}_2$. Usually, the only assumption is to assume that $\boldsymbol{K}_1 = \boldsymbol{K}_2$.

If we assume that there is no relation between parameters from $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$, but there are some linear restrictions within elements of each matrix, the estimation of the parameters for each matrix can be done independently. In this case the IPM algorithm can be adapted and used for the estimation in each population. This algorithm was adapted in Højsgaard and Lauritzen (2007, 2008) for RCON models.

In the case where some parameters in $\boldsymbol{K}_1$ are equal to their corresponding in $\boldsymbol{K}_2$, the estimation cannot be performed independently. Only in some special cases the MLEs can be found analytically. For example, when the only restriction is that both concentration matrices are equal, in such case, the MLE is as in (1.7). Another example is when assuming that both concentration matrices are equal and the covariance matrices have a pattern of compound symmetry (Votaw, 1948) or circular symmetry (Olkin and Press, 1969). Some conditions for explicit MLEs are given in Szatrowsky (1978, 1979) for some specific linear restrictions on the covariance matrices in the context of one population and when assuming that both populations have the same covariance matrix with linear restrictions.

In order to express the equalities between the parameters of the two concentration matrices in a simple way, we observe that restrictions considered in (1.17) can be divided in the following three cases.

i) $q_1 = q_2 = q$ and $H_h^{(1)} = H_h^{(2)}$, $h = 0, 1, ..., q$.

ii) $q_1 = q_2 = q$ and $H_h^{(1)} \neq H_h^{(2)}$ for some $h \in \{0, 1, 2, ..., q\}$.

iii) $q_1 \neq q_2$.

We note that case i) corresponds to concentration matrices with the same structure, unlike

cases ii) and iii). Here, the IPM algorithm is adapted and implemented to compute numerical estimates in case i). These models satisfy (1.17) as follows

$$\boldsymbol{K}_c = \boldsymbol{\Sigma}_c^{-1} = \sum_{h=0}^{q} \psi_h^{(c)} \boldsymbol{H}_h, \quad c = 1, 2. \tag{3.1}$$

The type of relations between parameters of $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$ are limited to the following.

I. $\psi_h^{(1)} \neq \psi_h^{(2)},\ h = 0, 1, ..., q.$
II. $\psi_h^{(1)} = \psi_h^{(2)} = \psi_h,\ h = 0, 1, ..., q.$
III. $\psi_h^{(1)} = \psi_h^{(2)} = \psi_h$, $h = 0, 1, ..., f$, and $\psi_h^{(1)} \neq \psi_h^{(2)}$, $h = f+1, , ..., q$, $0 \leq f < q$.

Examples of cases I and II have been previously considered, for example in Lauritzen (1996), Abreu et al. (2010) and Edwards et al. (2010). Models in case I are called heterogeneous and those in II homogeneous. In the last two articles, the authors consider suitable models for applications with a large number of variables, as in those in Genetics and Bioinformatics. They restrict themselves to models whose associated graph is a forest or a tree, as we shall do in Chapter 4.

We present two instances of models that follow (3.1) and some examples for cases I, II and III.

1. Considering a saturated model on each population and $p = 3$, we can consider restrictions as follows

   (a) $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$ arbitrary matrices, that is, case I with

$$\boldsymbol{K}_1 = \psi_0^{(1)} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \psi_1^{(1)} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \psi_2^{(1)} \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} + \psi_3^{(1)} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$+ \psi_4^{(1)} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} + \psi_5^{(1)} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} \psi_0^{(1)} & \psi_1^{(1)} & \psi_2^{(1)} \\ \psi_1^{(1)} & \psi_3^{(1)} & \psi_4^{(1)} \\ \psi_2^{(1)} & \psi_4^{(1)} & \psi_5^{(1)} \end{pmatrix},$$

$$\boldsymbol{K}_2 = \psi_0^{(2)} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \psi_1^{(2)} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \psi_2^{(2)} \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} + \psi_3^{(2)} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$
$$+ \psi_4^{(2)} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} + \psi_5^{(2)} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} \psi_0^{(2)} & \psi_1^{(2)} & \psi_2^{(2)} \\ \psi_1^{(2)} & \psi_3^{(2)} & \psi_4^{(2)} \\ \psi_2^{(2)} & \psi_4^{(2)} & \psi_5^{(2)} \end{pmatrix}. \tag{3.2}$$

(b) $\boldsymbol{K}_1 = \boldsymbol{K}_2$, that is, case II with

$$\boldsymbol{K}_1 = \boldsymbol{K}_2 = \boldsymbol{K} = \psi_0 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \psi_1 \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \psi_2 \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} + \psi_3 \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$
$$+ \psi_4 \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} + \psi_5 \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} \psi_0 & \psi_1 & \psi_2 \\ \psi_1 & \psi_3 & \psi_4 \\ \psi_2 & \psi_4 & \psi_5 \end{pmatrix}. \tag{3.3}$$

(c) Parameters of the two concentration matrices corresponding to the submatrix composed of rows and columns 2 and 3 are equal. That is, case III with

$$\boldsymbol{K}_1 = \psi_0 \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \psi_1 \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} + \psi_2 \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} + \psi_3^{(1)} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$
$$+ \psi_4^{(1)} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \psi_5^{(1)} \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} \psi_3^{(1)} & \psi_4^{(1)} & \psi_5^{(1)} \\ \psi_4^{(1)} & \psi_0 & \psi_1 \\ \psi_5^{(1)} & \psi_1 & \psi_2 \end{pmatrix},$$

$$\boldsymbol{K}_2 = \psi_0 \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \psi_1 \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} + \psi_2 \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} + \psi_3^{(2)} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$
$$+ \psi_4^{(2)} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \psi_5^{(2)} \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} \psi_3^{(2)} & \psi_4^{(2)} & \psi_5^{(2)} \\ \psi_4^{(2)} & \psi_0 & \psi_1 \\ \psi_5^{(2)} & \psi_1 & \psi_2 \end{pmatrix}. \tag{3.4}$$

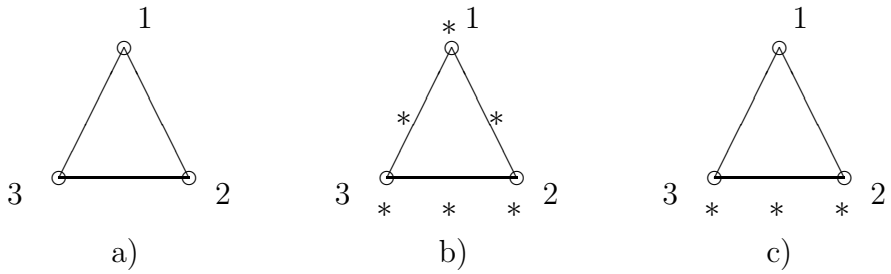


Figure 3.1: Graphs associated with models 1a, 1b and 1c, respectively. The concentration matrices associated with these graphs are (3.2), (3.3) and (3.4), respectively. Symbol $*$ on a node or an edge indicates that the corresponding parameters in both matrices are equal.

The graphs associated with models 1a, 1b and 1c, are shown in Figure 3.1, where symbol $*$ on a node or an edge indicates that the corresponding parameters in both concentration matrices are equal.

2. Considering an RCON model with a cycle of size 4 for each population and the restrictions: all diagonal elements are equal, and all off-diagonal elements different from zero are equal.

   (a) Case I, $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$ such that

$$\boldsymbol{K}_1 = \psi_0^{(1)} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} + \psi_1^{(1)} \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} \psi_0^{(1)} & \psi_1^{(1)} & 0 & \psi_1^{(1)} \\ \psi_1^{(1)} & \psi_0^{(1)} & \psi_1^{(1)} & 0 \\ 0 & \psi_1^{(1)} & \psi_0^{(1)} & \psi_1^{(1)} \\ \psi_1^{(1)} & 0 & \psi_1^{(1)} & \psi_0^{(1)} \end{pmatrix},$$

$$\boldsymbol{K}_2 = \psi_0^{(2)} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} + \psi_1^{(2)} \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} \psi_0^{(2)} & \psi_1^{(2)} & 0 & \psi_1^{(2)} \\ \psi_1^{(2)} & \psi_0^{(2)} & \psi_1^{(2)} & 0 \\ 0 & \psi_1^{(2)} & \psi_0^{(2)} & \psi_1^{(2)} \\ \psi_1^{(2)} & 0 & \psi_1^{(2)} & \psi_0^{(2)} \end{pmatrix}. \quad (3.5)$$

   (b) Case II, $\boldsymbol{K}_1 = \boldsymbol{K}_2$, with

$$\boldsymbol{K}_1 = \boldsymbol{K}_2 = \psi_0 \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} + \psi_1 \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} \psi_0 & \psi_1 & 0 & \psi_1 \\ \psi_1 & \psi_0 & \psi_1 & 0 \\ 0 & \psi_1 & \psi_0 & \psi_1 \\ \psi_1 & 0 & \psi_1 & \psi_0 \end{pmatrix}. \quad (3.6)$$

(c) Parameters corresponding to diagonal elements are equal. That is, case III with

$$\boldsymbol{K}_1 = \psi_0 \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} + \psi_1^{(1)} \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} \psi_0 & \psi_1^{(1)} & 0 & \psi_1^{(1)} \\ \psi_1^{(1)} & \psi_0 & \psi_1^{(1)} & 0 \\ 0 & \psi_1^{(1)} & \psi_0 & \psi_1^{(1)} \\ \psi_1^{(1)} & 0 & \psi_1^{(1)} & \psi_0 \end{pmatrix},$$

$$\boldsymbol{K}_2 = \psi_0 \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} + \psi_1^{(2)} \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} \psi_0 & \psi_1^{(2)} & 0 & \psi_1^{(2)} \\ \psi_1^{(2)} & \psi_0 & \psi_1^{(2)} & 0 \\ 0 & \psi_1^{(2)} & \psi_0 & \psi_1^{(2)} \\ \psi_1^{(2)} & 0 & \psi_1^{(2)} & \psi_0 \end{pmatrix}. \quad (3.7)$$

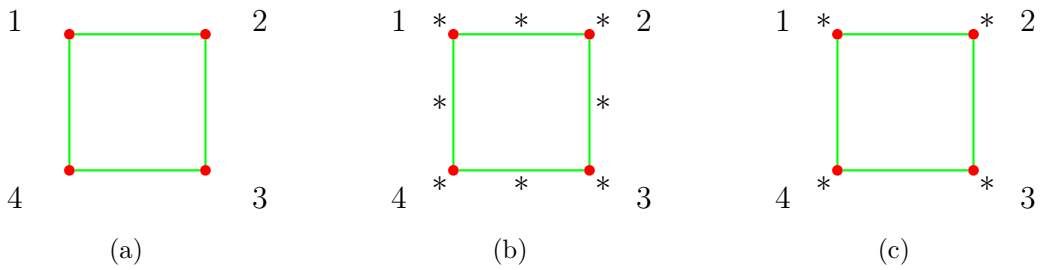


Figure 3.2: Graphs associated with models 2a, 2b and 2c, respectively. The concentration matrices associated with these graphs are (3.5), (3.6) and (3.7), respectively. Symbol $*$ on a node or an edge indicates that the corresponding parameters in both matrices are equal.

The graphs associated with models 2a, 2b and 2c, respectively, are shown in Figure 3.2, where an $*$ on a node or an edge indicates that the corresponding parameters in both concentration matrices are equal.

In this work we are not considering any assumption on the structure of the group mean vectors. Anderson (1970, 1973) considers some models where the mean vector can be expressed as a linear combination of known linearly independent vectors, that is, $\boldsymbol{\mu} = \sum_j \beta_j \boldsymbol{z}_j$. Particular instances are presented in Votaw (1948), Olkin and Press (1969) and Szatrowsky (1979). Gehrmann and Lauritzen (2012) and Gehrmann (2011) also study models where some elements of the mean vector are restricted to being equal.

In the following section, we present the algorithm to be implemented for fitting these restricted models.

3.2 An algorithm for finding ML estimates.

There are two estimation problems in the models we are considering, as those of GGMs: (1) parameter estimation (quantitative aspect) and (2) structure estimation (structural aspect). In this section, we adapt the IPM algorithm for parameter estimation, that is, we derive the analytical expressions needed to use this algorithm. For structure estimation, in Section 3.3, we comment on some algorithms already studied for finding some specific linear restrictions on the concentration matrices.

In order to use the IPM algorithm, we have to show that i) the CG-distribution with density given in (1.3) and with restrictions on the concentration matrices given by (3.1) for cases I, II and III, belongs to the regular exponential family; and ii) the parameters that we want to estimate are the canonical parameters of the CG-distribution with the assumed restrictions.

In Appendix A.1 we show i). In Appendix A.2, we use the result that the MLE, if it exists, is unique and can be obtained by equating the sufficient canonical statistics to their expectations (see Lauritzen, 1996, Theorem D.1, p. 268). We present the system of equations and observe that the unknown parameters from the concentration matrices: ψ_h, $h = 0, ..., f$, $\psi_h^{(1)}$, $h = f + 1, ..., q$, and $\psi_h^{(2)}$, $h = f + 1, ..., q$, are the canonical parameters. Results are given for case III, results for case I and II can be obtained in a similar way.

Since the CG-distribution, with density given in (1.3) and with restrictions on the concentration matrices given by (3.1) for cases I, II and III, belongs to the regular exponential family, numerical

solutions for the system of equations given in A.5 can be found using the IPM algorithm. This converges globally to a unique solution when it exists. The IPM algorithm can be described as follows.

Suppose that there are m canonical parameters θ_1, θ_2, ..., θ_m. Let $\theta_1^{(0)}$, $\theta_2^{(0)}$, ..., $\theta_m^{(0)}$ be initial values, then

1. Let $i = 1$
2. Let $j = 1$
3. The value for $\theta_j^{(i)}$ that maximizes the likelihood function is found, assuming that the rest of the parameters are known, as

$$\theta_k = \begin{cases} \theta_k^{(i)} & \text{if} \quad k = 1, 2, ..., j-1 \\ \theta_k^{(i-1)} & \text{if} \quad k = j+1, j+2, ..., m \end{cases}.$$

   The modified Newton method is applied to the derivative of the n-th root of the reciprocal likelihood function.
4. Step 3 is repeated for $j = j + 1$ if $j < m$, otherwise the following step follows.
5. Start in step 2 with $i = i + 1$ until convergence is reached.

When applying the modified Newton method in step 3, variances of the sufficient statistics are needed. These are obtained from the second derivatives of the cumulant function. The first and second derivatives of the cumulant function are given in (A.6) and (A.7), respectively, in Appendix A.2. Specifically, the iterative step 3 of the IPM algorithm for each parameter is the following.

$$\psi_{h_j} = \psi_{h_{j-1}} + \left. \frac{-\frac{1}{2n}\sum_{i=1}^{n} tr(\boldsymbol{H}_h \boldsymbol{x}_i \boldsymbol{x}_i^t) - \frac{\partial \psi(\boldsymbol{\theta})}{\partial \psi_h}}{\frac{\partial^2 \psi(\boldsymbol{\theta})}{\partial \psi_h^2} + \left\{ -\frac{1}{2n}\sum_{i=1}^{n} tr(\boldsymbol{H}_h \boldsymbol{x}_i \boldsymbol{x}_i^t) - \frac{\partial \psi(\boldsymbol{\theta})}{\partial \psi_h} \right\}^2} \right|_{\psi_h = \psi_{h_{j-1}}}, \quad h = 0, ..., f$$

$$\psi_{h_j}^{(1)} = \psi_{h_{j-1}}^{(1)} + \left. \frac{-\frac{1}{2n}\sum_{i=1}^{n} tr(\boldsymbol{H}_h \boldsymbol{x}_i \boldsymbol{x}_i^t \delta_1(c_i)) - \frac{\partial \psi(\boldsymbol{\theta})}{\partial \psi_h^{(1)}}}{\frac{\partial^2 \psi(\boldsymbol{\theta})}{\partial {\psi_h^{(1)}}^2} + \left\{ -\frac{1}{2n}\sum_{i=1}^{n} tr(\boldsymbol{H}_h \boldsymbol{x}_i \boldsymbol{x}_i^t \delta_1(c_i)) - \frac{\partial \psi(\boldsymbol{\theta})}{\partial \psi_h^{(1)}} \right\}^2} \right|_{\psi_h^{(1)} = \psi_{h_{j-1}}^{(1)}}, \quad h = f+1, ..., q$$

$$\psi_{h_j}^{(2)} = \psi_{h_{j-1}}^{(2)} + \left. \frac{-\frac{1}{2n}\sum_{i=1}^{n} tr(\boldsymbol{H}_h \boldsymbol{x}_i \boldsymbol{x}_i^t (1-\delta_1(c_i))) - \frac{\partial \psi(\boldsymbol{\theta})}{\partial \psi_h^{(2)}}}{\frac{\partial^2 \psi(\boldsymbol{\theta})}{\partial {\psi_h^{(2)}}^2} + \left\{ -\frac{1}{2n}\sum_{i=1}^{n} tr(\boldsymbol{H}_h \boldsymbol{x}_i \boldsymbol{x}_i^t (1-\delta_1(c_i))) - \frac{\partial \psi(\boldsymbol{\theta})}{\partial \psi_h^{(2)}} \right\}^2} \right|_{\psi_h^{(2)} = \psi_{h_{j-1}}^{(2)}}, \quad h = f+1, ..., q. \tag{3.8}$$

The algorithm has been implemented in C++, some comments are the following.

1. The number of matrices $\boldsymbol{H}$, q+1, has to be known and given.
2. The number of parameters that are equal in the two concentration matrices has to be given.
3. The matrices $\boldsymbol{H}$ have to be given, first those associated to the parameters that are equal in both concentration matrices followed by the rest.
4. The initial values correspond to those obtained from the diagonal of the sample covariance matrices.

We note that the algorithm is used only to find ML estimates but not for finding the graph or the structure of the concentration matrices. This is, a) zeroes, b) linear restrictions, and c) equalities between corresponding parameters of the two concentration matrices, are assumed to be known. When this structure is unknown, additional algorithms should be developed for the structure estimation, which is a complex problem.

3.2.1 Illustrative example on parameter estimation.

We consider the data on educational testing analysed in Szatrowski (1982). There are $p = 8$ variables and two groups (1-male and 2-female): X_1 and X_2 represent verbal and quantitative test

scores; (X_3, X_4, X_5) and (X_6, X_7, X_8) represent Mathematics, Science, and Social Studies scores on achievement test 1 and 2, respectively. The sample means and covariance matrices are the following.

$$\overline{\boldsymbol{x}}_1 = \begin{pmatrix} 390 \\ 460 \\ \hline 362 \\ 418 \\ 385 \\ \hline 355 \\ 410 \\ 394 \end{pmatrix}, \qquad \boldsymbol{W}_1 = \left(\begin{array}{cc|ccc|ccc} 7365 & 4845 & 6649 & 4988 & 6567 & 5142 & 3828 & 6613 \\ 4845 & 6300 & 5718 & 4435 & 4992 & 4582 & 2685 & 5440 \\ \hline 6649 & 5718 & 9199 & 5889 & 7367 & 5582 & 3910 & 6978 \\ 4988 & 4435 & 5889 & 6423 & 6065 & 4902 & 3688 & 6176 \\ 6567 & 4992 & 7367 & 6065 & 8675 & 6221 & 4632 & 7864 \\ \hline 5142 & 4582 & 5582 & 4902 & 6221 & 7880 & 4888 & 6625 \\ 3828 & 2685 & 3910 & 3688 & 4632 & 4888 & 4970 & 5275 \\ 6613 & 5440 & 6978 & 6176 & 7864 & 6625 & 5275 & 10764 \end{array}\right)$$

$$\overline{\boldsymbol{x}}_2 = \begin{pmatrix} 436 \\ 487 \\ \hline 388 \\ 440 \\ 439 \\ \hline 400 \\ 437 \\ 440 \end{pmatrix}, \qquad \boldsymbol{W}_2 = \left(\begin{array}{cc|ccc|ccc} 5974 & 2632 & 3050 & 2726 & 5471 & 2104 & 2694 & 4804 \\ 2632 & 4825 & 2872 & 1826 & 3335 & 1871 & 1642 & 2191 \\ \hline 3050 & 2872 & 3671 & 2040 & 3806 & 1843 & 1597 & 3288 \\ 2726 & 1826 & 2040 & 3437 & 4228 & 1817 & 2198 & 3815 \\ 5471 & 3335 & 3806 & 4228 & 9144 & 3010 & 3727 & 7103 \\ \hline 2104 & 1817 & 1843 & 1817 & 3010 & 3622 & 2108 & 3702 \\ 2694 & 1642 & 1597 & 2198 & 3727 & 2108 & 3825 & 4254 \\ 4804 & 2191 & 3288 & 3815 & 7103 & 3702 & 4254 & 8945 \end{array}\right).$$

Szatrowski (1982) considered different hypothesis tests about the block compound symmetry (BCS) assumption on the means and covariances matrices of each population, and equalities of means or covariance matrices of the populations given the BCS assumption. We note that the pattern in the covariance matrix assumed in the BCS coincides with the one in the concentration matrix.

The type of BCS assumed for the data in each covariance matrix and therefore in each concentration matrix is

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{A}^* & \boldsymbol{C}^* & \boldsymbol{C}^* \\ \boldsymbol{C}^{*t} & \boldsymbol{B}^* & \boldsymbol{D}^* \\ \boldsymbol{C}^{*t} & \boldsymbol{D}^* & \boldsymbol{B}^* \end{pmatrix}, \quad \boldsymbol{K} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{C} & \boldsymbol{C} \\ \boldsymbol{C}^t & \boldsymbol{B} & \boldsymbol{D} \\ \boldsymbol{C}^t & \boldsymbol{D}^t & \boldsymbol{B} \end{pmatrix}, \tag{3.9}$$

where $\boldsymbol{A}$ and $\boldsymbol{A}^*$ are 2×2 symmetric matrices, $\boldsymbol{C}$ and $\boldsymbol{C}^*$ are 2×3 matrices, and $\boldsymbol{B}$, $\boldsymbol{D}$, $\boldsymbol{B}^*$ and $\boldsymbol{D}^*$ are 3×3 symmetric matrices.

The BCS assumptions correspond to a RCON model with a coloured graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{G}$ is the complete graph κ, $\mathcal{V} = \{V_1, ..., V_5\} = \{\{v_1\}, \{v_2\}, \{v_3, v_6\}, \{v_4, v_7\}, \{v_5, v_8\}\}$ and $\mathcal{E} = \{E_1, ..., E_{16}\} = \{\{(1,2)\}, \{(1,3),(1,6)\}, \{(1,4),(1,7)\}, \{(1,5),(1,8)\}, \{(2,3),(2,6)\}, \{(2,4),(2,7)\}, \{(2,5),(2,8)\}, \{(3,4),(6,7)\}, \{(3,5),(6,8)\}, \{(4,5),(7,8)\}, \{(3,6)\}, \{(3,7),(4,6)\}, \{(3,8),(5,6)\}, \{(4,7)\}, \{(5,8)\}, \{(4,8),(5,7)\}\}$.

In this work, we consider the BCS and three types of equalities between the parameters of $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$ considered in the description of the algorithm given in this section.

a) Heterogeneous case, $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$ are patterned matrices due to the BCS assumption with arbitrary entries.

$$\boldsymbol{K}_1 = \left(\begin{array}{cc|ccc|ccc} a_{11}^{(1)} & a_{12}^{(1)} & c_{11}^{(1)} & c_{12}^{(1)} & c_{13}^{(1)} & c_{11}^{(1)} & c_{12}^{(1)} & c_{13}^{(1)} \\ a_{12}^{(1)} & a_{22}^{(1)} & c_{21}^{(1)} & c_{22}^{(1)} & c_{23}^{(1)} & c_{21}^{(1)} & c_{22}^{(1)} & c_{23}^{(1)} \\ \hline c_{11}^{(1)} & c_{21}^{(1)} & b_{11}^{(1)} & b_{12}^{(1)} & b_{13}^{(1)} & d_{11}^{(1)} & d_{12}^{(1)} & d_{13}^{(1)} \\ c_{12}^{(1)} & c_{22}^{(1)} & b_{12}^{(1)} & b_{22}^{(1)} & b_{23}^{(1)} & d_{12}^{(1)} & d_{22}^{(1)} & d_{23}^{(1)} \\ c_{13}^{(1)} & c_{23}^{(1)} & b_{13}^{(1)} & b_{23}^{(1)} & b_{33}^{(1)} & d_{13}^{(1)} & d_{23}^{(1)} & d_{33}^{(1)} \\ \hline c_{11}^{(1)} & c_{21}^{(1)} & d_{11}^{(1)} & d_{12}^{(1)} & d_{13}^{(1)} & b_{11}^{(1)} & b_{12}^{(1)} & b_{13}^{(1)} \\ c_{12}^{(1)} & c_{22}^{(1)} & d_{12}^{(1)} & d_{22}^{(1)} & d_{23}^{(1)} & b_{12}^{(1)} & b_{22}^{(1)} & b_{23}^{(1)} \\ c_{13}^{(1)} & c_{23}^{(1)} & d_{13}^{(1)} & d_{23}^{(1)} & d_{33}^{(1)} & b_{13}^{(1)} & b_{23}^{(1)} & b_{33}^{(1)} \end{array}\right) = \begin{pmatrix} \boldsymbol{A}_1 & \boldsymbol{C}_1 & \boldsymbol{C}_1 \\ \boldsymbol{C}_1^t & \boldsymbol{B}_1 & \boldsymbol{D}_1 \\ \boldsymbol{C}_1^t & \boldsymbol{D}_1 & \boldsymbol{B}_1 \end{pmatrix},$$

$$\boldsymbol{K}_2 = \left(\begin{array}{cc|ccc|ccc} a_{11}^{(2)} & a_{12}^{(2)} & c_{11}^{(2)} & c_{12}^{(2)} & c_{13}^{(2)} & c_{11}^{(2)} & c_{12}^{(2)} & c_{13}^{(2)} \\ a_{12}^{(2)} & a_{22}^{(2)} & c_{21}^{(2)} & c_{22}^{(2)} & c_{23}^{(2)} & c_{21}^{(2)} & c_{22}^{(2)} & c_{23}^{(2)} \\ \hline c_{11}^{(2)} & c_{21}^{(2)} & b_{11}^{(2)} & b_{12}^{(2)} & b_{13}^{(2)} & d_{11}^{(2)} & d_{12}^{(2)} & d_{13}^{(2)} \\ c_{12}^{(2)} & c_{22}^{(2)} & b_{12}^{(2)} & b_{22}^{(2)} & b_{23}^{(2)} & d_{12}^{(2)} & d_{22}^{(2)} & d_{23}^{(2)} \\ c_{13}^{(2)} & c_{23}^{(2)} & b_{13}^{(2)} & b_{23}^{(2)} & b_{33}^{(2)} & d_{13}^{(2)} & d_{23}^{(2)} & d_{33}^{(2)} \\ \hline c_{11}^{(2)} & c_{21}^{(2)} & d_{11}^{(2)} & d_{12}^{(2)} & d_{13}^{(2)} & b_{11}^{(2)} & b_{12}^{(2)} & b_{13}^{(2)} \\ c_{12}^{(2)} & c_{22}^{(2)} & d_{12}^{(2)} & d_{22}^{(2)} & d_{23}^{(2)} & b_{12}^{(2)} & b_{22}^{(2)} & b_{23}^{(2)} \\ c_{13}^{(2)} & c_{23}^{(2)} & d_{13}^{(2)} & d_{23}^{(2)} & d_{33}^{(2)} & b_{13}^{(2)} & b_{23}^{(2)} & b_{33}^{(2)} \end{array}\right) = \left(\begin{array}{ccc} \boldsymbol{A}_2 & \boldsymbol{C}_2 & \boldsymbol{C}_2 \\ \boldsymbol{C}_2^t & \boldsymbol{B}_2 & \boldsymbol{D}_2 \\ \boldsymbol{C}_2^t & \boldsymbol{D}_2 & \boldsymbol{B}_2 \end{array}\right).$$

b) Homogeneous case, $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$ are patterned matrices due to the BCS assumption with $\boldsymbol{K}_1 = \boldsymbol{K}_2$.

$$\boldsymbol{K}_1 = \boldsymbol{K}_2 = \left(\begin{array}{cc|ccc|ccc} a_{11} & a_{12} & c_{11} & c_{12} & c_{13} & c_{11} & c_{12} & c_{13} \\ a_{12} & a_{22} & c_{21} & c_{22} & c_{23} & c_{21} & c_{22} & c_{23} \\ \hline c_{11} & c_{21} & b_{11} & b_{12} & b_{13} & d_{11} & d_{12} & d_{13} \\ c_{12} & c_{22} & b_{12} & b_{22} & b_{23} & d_{12} & d_{22} & d_{23} \\ c_{13} & c_{23} & b_{13} & b_{23} & b_{33} & d_{13} & d_{23} & d_{33} \\ \hline c_{11} & c_{21} & d_{11} & d_{12} & d_{13} & b_{11} & b_{12} & b_{13} \\ c_{12} & c_{22} & d_{12} & d_{22} & d_{23} & b_{12} & b_{22} & b_{23} \\ c_{13} & c_{23} & d_{13} & d_{23} & d_{33} & b_{13} & b_{23} & b_{33} \end{array}\right) = \left(\begin{array}{ccc} \boldsymbol{A} & \boldsymbol{C} & \boldsymbol{C} \\ \boldsymbol{C}^t & \boldsymbol{B} & \boldsymbol{D} \\ \boldsymbol{C}^t & \boldsymbol{D} & \boldsymbol{B} \end{array}\right).$$

c) $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$ are patterned matrices due to the BCS assumption with $\boldsymbol{A}_1 = \boldsymbol{A}_2 = \boldsymbol{A}$ and $\boldsymbol{C}_1 = \boldsymbol{C}_2 = \boldsymbol{C}$.

$$\boldsymbol{K}_1 = \left(\begin{array}{cc|ccc|ccc} a_{11} & a_{12} & c_{11} & c_{12} & c_{13} & c_{11} & c_{12} & c_{13} \\ a_{12} & a_{22} & c_{21} & c_{22} & c_{23} & c_{21} & c_{22} & c_{23} \\ \hline c_{11} & c_{21} & b_{11}^{(1)} & b_{12}^{(1)} & b_{13}^{(1)} & d_{11}^{(1)} & d_{12}^{(1)} & d_{13}^{(1)} \\ c_{12} & c_{22} & b_{12}^{(1)} & b_{22}^{(1)} & b_{23}^{(1)} & d_{12}^{(1)} & d_{22}^{(1)} & d_{23}^{(1)} \\ c_{13} & c_{23} & b_{13}^{(1)} & b_{23}^{(1)} & b_{33}^{(1)} & d_{13}^{(1)} & d_{23}^{(1)} & d_{33}^{(1)} \\ \hline c_{11} & c_{21} & d_{11}^{(1)} & d_{12}^{(1)} & d_{13}^{(1)} & b_{11}^{(1)} & b_{12}^{(1)} & b_{13}^{(1)} \\ c_{12} & c_{22} & d_{12}^{(1)} & d_{22}^{(1)} & d_{23}^{(1)} & b_{12}^{(1)} & b_{22}^{(1)} & b_{23}^{(1)} \\ c_{13} & c_{23} & d_{13}^{(1)} & d_{23}^{(1)} & d_{33}^{(1)} & b_{13}^{(1)} & b_{23}^{(1)} & b_{33}^{(1)} \end{array}\right) = \left(\begin{array}{ccc} \boldsymbol{A} & \boldsymbol{C} & \boldsymbol{C} \\ \boldsymbol{C}^t & \boldsymbol{B}_1 & \boldsymbol{D}_1 \\ \boldsymbol{C}^t & \boldsymbol{D}_1 & \boldsymbol{B}_1 \end{array}\right),$$

$$\boldsymbol{K}_2 = \left(\begin{array}{cc|ccc|ccc} a_{11} & a_{12} & c_{11} & c_{12} & c_{13} & c_{11} & c_{12} & c_{13} \\ a_{12} & a_{22} & c_{21} & c_{22} & c_{23} & c_{21} & c_{22} & c_{23} \\ \hline c_{11} & c_{21} & b_{11}^{(2)} & b_{12}^{(2)} & b_{13}^{(2)} & d_{11}^{(2)} & d_{12}^{(2)} & d_{13}^{(2)} \\ c_{12} & c_{22} & b_{12}^{(2)} & b_{22}^{(2)} & b_{23}^{(2)} & d_{12}^{(2)} & d_{22}^{(2)} & d_{23}^{(2)} \\ c_{13} & c_{23} & b_{13}^{(2)} & b_{23}^{(2)} & b_{33}^{(2)} & d_{13}^{(2)} & d_{23}^{(2)} & d_{33}^{(2)} \\ \hline c_{11} & c_{21} & d_{11}^{(2)} & d_{12}^{(2)} & d_{13}^{(2)} & b_{11}^{(2)} & b_{12}^{(2)} & b_{13}^{(2)} \\ c_{12} & c_{22} & d_{12}^{(2)} & d_{22}^{(2)} & d_{23}^{(2)} & b_{12}^{(2)} & b_{22}^{(2)} & b_{23}^{(2)} \\ c_{13} & c_{23} & d_{13}^{(2)} & d_{23}^{(2)} & d_{33}^{(2)} & b_{13}^{(2)} & b_{23}^{(2)} & b_{33}^{(2)} \end{array}\right) = \left(\begin{array}{ccc} \boldsymbol{A} & \boldsymbol{C} & \boldsymbol{C} \\ \boldsymbol{C}^t & \boldsymbol{B}_2 & \boldsymbol{D}_2 \\ \boldsymbol{C}^t & \boldsymbol{D}_2 & \boldsymbol{B}_2 \end{array}\right).$$

In the three cases, the mean vectors are not restricted, then $\widehat{\boldsymbol{\mu}}_1 = \overline{\boldsymbol{x}}_1$ and $\widehat{\boldsymbol{\mu}}_2 = \overline{\boldsymbol{x}}_2$. The number of unknown parameters in the concentration matrices are 42, 21 and 33 for cases a), b) and c), respectively.

We implemented the IPM algorithm in C++ to obtain the ML estimates for the three cases. However, we note that for cases a) and b), explicit expressions for the MLEs exist (Szatrowski, 1982). The ML estimates of $\widehat{\boldsymbol{K}_1}$ and $\widehat{\boldsymbol{K}_2}$ are the following

a)

$$\widehat{\boldsymbol{K}_1} = 10^{-5} \times \left(\begin{array}{cc|ccc|ccc} 50.18 & -8.80 & -8.40 & -0.04 & -9.67 & -8.40 & -0.04 & -9.67 \\ -8.80 & 41.55 & -11.67 & 2.15 & -1.35 & -11.67 & 2.15 & -1.35 \\ \hline -8.40 & -11.67 & 42.97 & -18.87 & -7.91 & 2.00 & 0.38 & -1.99 \\ -0.04 & 2.15 & -18.87 & 55.17 & -11.64 & 0.38 & -4.64 & -7.14 \\ -9.67 & -1.35 & -7.91 & -11.64 & 44.48 & -1.99 & -7.14 & -11.26 \\ \hline -8.40 & -11.67 & 2.00 & 0.38 & -1.99 & 42.97 & -18.87 & -7.91 \\ -0.04 & 2.15 & 0.38 & -4.64 & -7.14 & -18.87 & 55.17 & -11.64 \\ -9.67 & -1.35 & -1.99 & -7.14 & -11.26 & -7.91 & -11.64 & 44.48 \end{array}\right),$$

$$\widehat{\boldsymbol{K}_2} = 10^{-5} \times \left(\begin{array}{cc|ccc|ccc} 40.19 & -7.30 & -2.38 & -1.65 & -9.70 & -2.38 & -1.65 & -9.70 \\ -7.30 & 37.72 & -15.88 & -3.61 & 4.44 & -15.88 & -3.61 & 4.44 \\ \hline -2.38 & -15.88 & 59.32 & -7.20 & -15.94 & -3.91 & 3.14 & 1.22 \\ -1.65 & -3.61 & -7.20 & 67.11 & -19.92 & 3.14 & -10.7 & -4.05 \\ -9.70 & 4.44 & -15.94 & -19.92 & 44.93 & 1.22 & -4.05 & -15.89 \\ \hline -2.38 & -15.88 & -3.91 & 3.14 & 1.22 & 59.32 & -7.2 & -15.94 \\ -1.65 & -3.61 & 3.14 & -10.7 & -4.05 & -7.2 & 67.11 & -19.92 \\ -9.70 & 4.44 & 1.22 & -4.05 & -15.89 & -15.94 & -19.92 & 44.93 \end{array}\right).$$

b)

$$\widehat{\boldsymbol{K}}_1 = \widehat{\boldsymbol{K}}_2 = 10^{-5} \times \left(\begin{array}{cc|ccc|ccc} 44.33 & -7.64 & -5.18 & -1.18 & -9.80 & -5.18 & -1.18 & -9.80 \\ -7.64 & 38.54 & -12.82 & -0.22 & 1.46 & -12.82 & -0.22 & 1.46 \\ \hline -5.18 & -12.82 & 46.05 & -15.41 & -9.22 & -2.83 & 1.71 & 0.37 \\ -1.18 & -0.22 & -15.41 & 59.20 & -14.21 & 1.71 & -7.42 & -5.84 \\ -9.80 & 1.46 & -9.22 & -14.21 & 42.79 & 0.37 & -5.84 & -14.02 \\ \hline -5.18 & -12.82 & -2.83 & 1.71 & 0.37 & 46.05 & -15.41 & -9.22 \\ -1.18 & -0.22 & 1.71 & -7.42 & -5.84 & -15.41 & 59.20 & -14.21 \\ -9.80 & 1.46 & 0.37 & -5.84 & -14.02 & -9.22 & -14.21 & 42.79 \end{array}\right).$$

c)

$$\widehat{\boldsymbol{K}}_1 = 10^{-5} \times \left(\begin{array}{cc|ccc|ccc} 44.33 & -7.64 & -5.17 & -1.17 & -9.80 & -5.17 & -1.17 & -9.80 \\ -7.64 & 38.55 & -12.83 & -0.24 & 1.47 & -12.83 & -0.24 & 1.47 \\ \hline -5.17 & -12.83 & 42.91 & -17.87 & -9.30 & 1.93 & 1.38 & -3.40 \\ -1.17 & -0.24 & -17.87 & 55.05 & -11.24 & 1.38 & -4.76 & -6.66 \\ -9.80 & 1.47 & -9.30 & -11.24 & 44.55 & -3.40 & -6.66 & -11.21 \\ \hline -5.17 & -12.83 & 1.93 & 1.38 & -3.40 & 42.91 & -17.87 & -9.30 \\ -1.17 & -0.24 & 1.38 & -4.76 & -6.66 & -17.87 & 55.05 & -11.24 \\ -9.80 & 1.47 & -3.40 & -6.66 & -11.21 & -9.30 & -11.24 & 44.55 \end{array}\right),$$

$$\widehat{\boldsymbol{K}}_2 = 10^{-5} \times \left(\begin{array}{cc|ccc|ccc} 44.33 & -7.64 & -5.17 & -1.17 & -9.80 & -5.17 & -1.17 & -9.80 \\ -7.64 & 38.55 & -12.83 & -0.24 & 1.47 & -12.83 & -0.24 & 1.47 \\ \hline -5.17 & -12.83 & 57.63 & -8.84 & -14.13 & -5.74 & 1.64 & 3.10 \\ -1.17 & -0.24 & -8.84 & 66.69 & -19.72 & 1.64 & -11.14 & -3.95 \\ -9.80 & 1.47 & -14.13 & -19.72 & 44.57 & 3.10 & -3.95 & -16.27 \\ \hline -5.17 & -12.83 & -5.74 & 1.64 & 3.10 & 57.63 & -8.84 & -14.13 \\ -1.17 & -0.24 & 1.64 & -11.14 & -3.95 & -8.84 & 66.69 & -19.72 \\ -9.80 & 1.47 & 3.10 & -3.95 & -16.27 & -14.13 & -19.72 & 44.57 \end{array}\right).$$

Three types of hypothesis test can be done with the ML estimates:

i) Homogeneous vs heterogeneous,

$$\text{Ho: } \boldsymbol{K}_1 = \boldsymbol{K}_2 = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{C} & \boldsymbol{C} \\ \boldsymbol{C}^t & \boldsymbol{B} & \boldsymbol{D} \\ \boldsymbol{C}^t & \boldsymbol{D} & \boldsymbol{B} \end{pmatrix} \text{ vs Ha: } \boldsymbol{K}_1 = \begin{pmatrix} \boldsymbol{A}_1 & \boldsymbol{C}_1 & \boldsymbol{C}_1 \\ \boldsymbol{C}_1^t & \boldsymbol{B}_1 & \boldsymbol{D}_1 \\ \boldsymbol{C}_1^t & \boldsymbol{D}_1 & \boldsymbol{B}_1 \end{pmatrix}, \; \boldsymbol{K}_2 = \begin{pmatrix} \boldsymbol{A}_2 & \boldsymbol{C}_2 & \boldsymbol{C}_2 \\ \boldsymbol{C}_2^t & \boldsymbol{B}_2 & \boldsymbol{D}_2 \\ \boldsymbol{C}_2^t & \boldsymbol{D}_2 & \boldsymbol{B}_2 \end{pmatrix}.$$

ii) Ho: $\boldsymbol{K}_1 = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{C} & \boldsymbol{C} \\ \boldsymbol{C}^t & \boldsymbol{B}_1 & \boldsymbol{D}_1 \\ \boldsymbol{C}^t & \boldsymbol{D}_1 & \boldsymbol{B}_1 \end{pmatrix}$, $\boldsymbol{K}_2 = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{C} & \boldsymbol{C} \\ \boldsymbol{C}^t & \boldsymbol{B}_2 & \boldsymbol{D}_2 \\ \boldsymbol{C}^t & \boldsymbol{D}_2 & \boldsymbol{B}_2 \end{pmatrix}$ vs Ha: $\boldsymbol{K}_1 = \begin{pmatrix} \boldsymbol{A}_1 & \boldsymbol{C}_1 & \boldsymbol{C}_1 \\ \boldsymbol{C}_1^t & \boldsymbol{B}_1 & \boldsymbol{D}_1 \\ \boldsymbol{C}_1^t & \boldsymbol{D}_1 & \boldsymbol{B}_1 \end{pmatrix}$, $\boldsymbol{K}_2 = \begin{pmatrix} \boldsymbol{A}_2 & \boldsymbol{C}_2 & \boldsymbol{C}_2 \\ \boldsymbol{C}_2^t & \boldsymbol{B}_2 & \boldsymbol{D}_2 \\ \boldsymbol{C}_2^t & \boldsymbol{D}_2 & \boldsymbol{B}_2 \end{pmatrix}$.

iii) Ho: $\boldsymbol{K}_1 = \boldsymbol{K}_2 = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{C} & \boldsymbol{C} \\ \boldsymbol{C}^t & \boldsymbol{B} & \boldsymbol{D} \\ \boldsymbol{C}^t & \boldsymbol{D} & \boldsymbol{B} \end{pmatrix}$ vs Ho: $\boldsymbol{K}_1 = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{C} & \boldsymbol{C} \\ \boldsymbol{C}^t & \boldsymbol{B}_1 & \boldsymbol{D}_1 \\ \boldsymbol{C}^t & \boldsymbol{D}_1 & \boldsymbol{B}_1 \end{pmatrix}$, $\boldsymbol{K}_2 = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{C} & \boldsymbol{C} \\ \boldsymbol{C}^t & \boldsymbol{B}_2 & \boldsymbol{D}_2 \\ \boldsymbol{C}^t & \boldsymbol{D}_2 & \boldsymbol{B}_2 \end{pmatrix}$.

Using a likelihood ratio test for each case, we have the following value for the test statistic $-2\ln\lambda$: 23.3045, 3.0011 and 20.3034 for i), ii) and iii), respectively. The degrees of freedom are 21, 9 and 12, respectively. Considering the approximation of a chi-squared distribution, the p-values are 0.328, 0.964 and 0.062.

From the test in i), the null hypothesis is not rejected when compared with the heterogeneous case, that is, it is plausible to consider

$$\boldsymbol{K}_1 = \boldsymbol{K}_2 = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{C} & \boldsymbol{C} \\ \boldsymbol{C}^t & \boldsymbol{B} & \boldsymbol{D} \\ \boldsymbol{C}^t & \boldsymbol{D} & \boldsymbol{B} \end{pmatrix}. \tag{3.10}$$

The null hypothesis is not rejected either with test in ii), when compared with the heterogeneous case, that is, it is also plausible to consider

$$\boldsymbol{K}_1 = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{C} & \boldsymbol{C} \\ \boldsymbol{C}^t & \boldsymbol{B}_1 & \boldsymbol{D}_1 \\ \boldsymbol{C}^t & \boldsymbol{D}_1 & \boldsymbol{B}_1 \end{pmatrix}, \ \boldsymbol{K}_2 = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{C} & \boldsymbol{C} \\ \boldsymbol{C}^t & \boldsymbol{B}_2 & \boldsymbol{D}_2 \\ \boldsymbol{C}^t & \boldsymbol{D}_2 & \boldsymbol{B}_2 \end{pmatrix}. \tag{3.11}$$

Finally, using test iii) to compare these two cases and using $\alpha = .10$, the homogeneous case in (3.10) is rejected against the model with concentration matrices as in (3.11).

We note that the equalities between corresponding elements of $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$ in model c) are not

reflected in the covariance matrices, see Appendix F.

3.3 Comments on some methods for structure estimation.

In this section, we list some methods and their associated algorithms that can be used for finding some specific restrictions on the concentration matrices considered in this work. There is no method that can be used to find all plausible restrictions, however, we could try using one of these methods in a first stage for finding some kind of restrictions and then using another one for finding other kind of restrictions. For example, this procedure was used in Højsgaard and Lauritzen (2008) in the context of one population and RCON models, they found the zeros or conditional independences in a first stage and then the equality of elements within the concentration matrix.

1. The method given in Abreu et al. (2010) and in Edwards et al. (2010) finds the zeroes for graphical models which consider a CG-distribution with trees or forests as associated graphs. The authors extended the approach studied in Chow and Liu (1968) in two ways: first, to find a forest optimizing a penalized likelihood criterion, for example AIC or BIC, and second, to handle data with both discrete and Gaussian variables. In the context of discriminant analysis, this method can be used in two ways: to find two independent tree structures, one for each population, for the set of variables $\{X_1, ..., X_p\}$; and to find a tree structure for the set of variables that includes the group variable, $\{X_1, ..., X_p, C\}$.

2. Tan et al. (2010) developed a method to learn tree-structured graphical models which optimizes an approximation of the log-likelihood ratio of the densities of the two populations. This method does not consider directly a CG-distribution but could handle data with both discrete and continuous variables on each population. The trees obtained on each population could be different. We will talk about this method in Chapter 4.

3. Gou et al. (2011) considered Gaussian graphical models on each population. They proposed a method that jointly estimates the graphical models corresponding to the different groups

presented in the data. They used a penalized criterion with two tuning parameters in order to identify the common zero elements across the populations whereas allowing graphs belonging to different categories to have different zero structures. This method just finds the zero structures. The algorithm that they proposed uses the glasso algorithm (Friedman et al., 2008) in one of its iterative steps.

4. Zhang and Wang (2010) also considered the case of continuous variables on each population and graphical Gaussian models. They proposed a method that jointly estimates the graphical models corresponding to two different groups, aiming to identify the common edge parameters between the two concentration matrices, while allowing differences for the zero structures. They used a penalized criterion with two tuning parameters and a block coordinate descent algorithm to solve the problem.

5. Hara and Washio (2011) also considered continuous variables on each population and a Gaussian graphical model. They considered the case where there are more than two populations or groups. The proposed method is based on a block coordinate descent optimization, where subproblems can be solved efficiently by existing algorithms. They used a penalized criterion that includes two tuning parameters in order to identify the common edge parameters across the two populations and to identify the common zero elements across the populations. This method does not allow different zero structures.

6. Danaher, Wang and Witten (2014) proposed the joint graphical lasso, which is based on a penalized log-likelihood approach where the choice of penalty depends on the characteristics of the models that are expected to be shared. They considered two convex penalty functions giving what they called fussed graphical lasso and group graphical lasso. Fussed graphical lasso encourages not only similar zero structure but also similar edge values. Group graphical lasso encourages only a shared pattern of sparsity.

7. Simon and Tibshirani (2011) also proposed a regularized model which adaptively pools

elements of the concentration matrices. The proposed objective function encourages only similar edge values, it does not consider the zero structure. They used the obtained estimates in the context of discriminant analysis.

8. Gehrmann (2011) studied the properties of the RCON models for one population. She showed that these models are complete non-distributive lattices and therefore a procedure for model selection, like the one given in Edwards and Havránek (1987), could be applied to find the equalities on the elements of the concentration matrix. However, since the number of RCON models grows super-exponentially in p, this procedure could not be convenient even for the four particular classes of RCON models that she considered. She also mentioned that a viable alternative to find the equalities could be to use a penalized criterion with three tuning parameters. The first one used to find the zeroes, the second one to find the equalities on the diagonal elements of the concentration matrix, and finally, the third one to find the equalities on the off-diagonal elements of the concentration matrix.

We note that methods described in 1 to 3 are aimed at finding only the zero structure. Methods in 4 to 6 find the zero structure and equalities of corresponding elements between two concentration matrices, whereas method in 7 only equalities of corresponding elements between two concentration matrices. Finally, method in 8 only finds equalities within the elements of one single concentration matrix.

We also remark that when parameters of the mean vectors have no restrictions imposed, the canonical parameters, $\boldsymbol{\varepsilon}^{(c)} = \boldsymbol{K}_c\boldsymbol{\mu}_c$, $c = 1, 2$, have also no restrictions. However, linear restrictions on the mean vectors do not necessarily imply the same linear restrictions on the canonical parameters. For example, suppose that $\boldsymbol{\mu}_1$ and $\boldsymbol{K}_1$ have the following restrictions

$$\boldsymbol{\mu}_1 = \mu_{11}\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + \mu_{21}\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \mu_{11} \\ \mu_{21} \\ \mu_{11} \end{pmatrix},$$

$$\boldsymbol{K}_1 = \psi_1 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \psi_2 \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} + \psi_3 \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \psi_1 \begin{pmatrix} \psi_1 & \psi_3 & 0 \\ \psi_3 & \psi_2 & 0 \\ 0 & 0 & \psi_2 \end{pmatrix},$$

then $\boldsymbol{\varepsilon}^{(1)}$ does not have the same linear restriction as $\boldsymbol{\mu}^{(1)}$:

$$\boldsymbol{\varepsilon}^{(1)} = \begin{pmatrix} \psi_1\mu_{11} + \psi_3\mu_{21} \\ \psi_2\mu_{21} + \psi_3\mu_{11} \\ \psi_2\mu_{11} \end{pmatrix}.$$

This example also shows that the associated CG-distribution is not a member of the regular exponential family. Therefore, if we consider linear restrictions on the mean vectors, the associated CG-distribution does not necessarily belong to the regular exponential family and the IPM algorithm could not be used. Some linear restrictions on the mean vector, for one population, that also imply linear restrictions on the canonical parameters are studied in Gehrmann and Lauritzen (2012). Properties and a selection model procedure for a particular class of RCON models with equalities in the concentration matrices and in the mean vector are studied in Gehrmann (2011).

In the following section we present an example to illustrate the use of the algorithm developed for parameter estimation and the use of some algorithms for structure estimation, we also compare the resubstitution error rates when using the corresponding parameter estimates in the discriminant functions.

3.4 Illustrative example

We consider part of the study reported in Miller et al. (2005). The corresponding data base is available in the *gRbase* R package (Dethlefsen and Højsgaard 2005) with the name *breastcancer*. It contains information on 1,000 genes on each of 250 patients with breast cancer. These 1,000 genes had been previously selected as being the more informative under the criterion of the Wilcoxon

test. The 1,000 genes can be treated as continuous variables and have already been standardized to have mean zero and variance one.

Additionally, the data base contains a binary variable that indicates the condition of the mutation on the p53 gene. This variable takes value 1 in 58 patients whose p53 gene presents the mutation, and 0 in 192 whose p53 gene does not present the mutation. The total of 250 patients are treated as the observations, of which 58, where the binary variable is equal to one, are considered as cases, and the other 192 as controls. According to the presence or absence of the mutation the prognosis for the patient varies; in the presence of the mutation the tumours are more aggressive and more resistant.

The data were analysed by Miller et al. (2005) using three classification methods: diagonal linear discriminant analysis, k nearest neighbours, and support vector machines. They selected diagonal lineal discriminant analysis because it showed to have the highest sensitivity for detecting p53 mutants. The optimal classifier was comprised of 32 genes, the classification rates estimated using a cross validation procedure are given in Table 3.6.

The subgroup of the 58 cases and a selected subset of 150 genes, has been used to illustrate the performance of algorithms to search structure on a graph when using Coloured graphical Gaussian models in Højsgaard and Lauritzen (2008). Edwards et al. (2010) used the 250 observations, 1,000 genes and the group variable to illustrate the performance of an algorithm to search a graphical model that best fits the data, restricting the search to models that follow a CG-distribution and whose graph is a tree or a forest. In this thesis, we consider the full data set, 250 observations, 1,000 continuous variables and one binary variable.

3.4.1 Exploratory analysis

To get an idea of the linear structure of the data, we compute the linear correlations among pairs of variables, the 10 largest in absolute value are shown in Table 3.1, where we can see the presence of correlations close to one.

Computing the eigenvalues for the sample covariance matrix, taking subsets of 250 variables at the time due to the sample size of 250 observations, we observe that a large proportion of them are smaller than 1, and the largest and the smallest for each subset are: ($\lambda_1 : 108.27$, $\lambda_{250} : 9.08e{-}17$), ($\lambda_1 : 96.03$, $\lambda_{250} : 5.00e-324$), ($\lambda_1 : 103.53$, $\lambda_{250} : 1.35e-16$) and ($\lambda_1 : 97.02$, $\lambda_{250} : 1.38e-16$), respectively. The first 20 largest eigenvalues sorted in descending order for the subset of variables 1 to 250 are presented in Figure 3.3a. The frequency of eigenvalues that are less than 0.01 for this subset of variables is presented in Figure 3.3b. We observe that the values of the eigenvalues decrease very quickly and that there is a large proportion very close to zero. A similar behaviour is observed with the other three subsets of 250 variables.

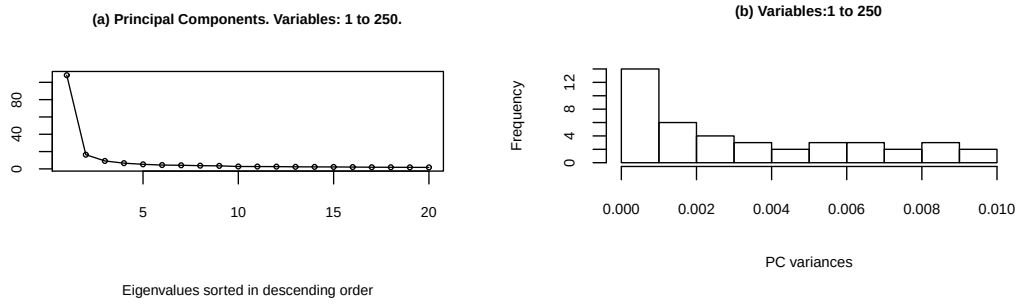


Figure 3.3: (a) The first 20 largest eigenvalues sorted in descending order for the set of variables 1 to 250. (b) Frequency of Principal Components which have a variance less than 0.01 for the set of variables 1 to 250.

X_i	468	364	243	441	243	672	319	298	304	399
X_j	490	368	437	506	298	1000	320	437	366	506
$\lvert\widehat{\rho}_{x_i x_j}\rvert$	0.996	0.982	0.979	0.978	0.978	0.978	0.975	0.970	0.970	0.967

Table 3.1: The ten largest linear correlations between pairs of variables, $|\widehat{\rho}_{x_i x_j}|$

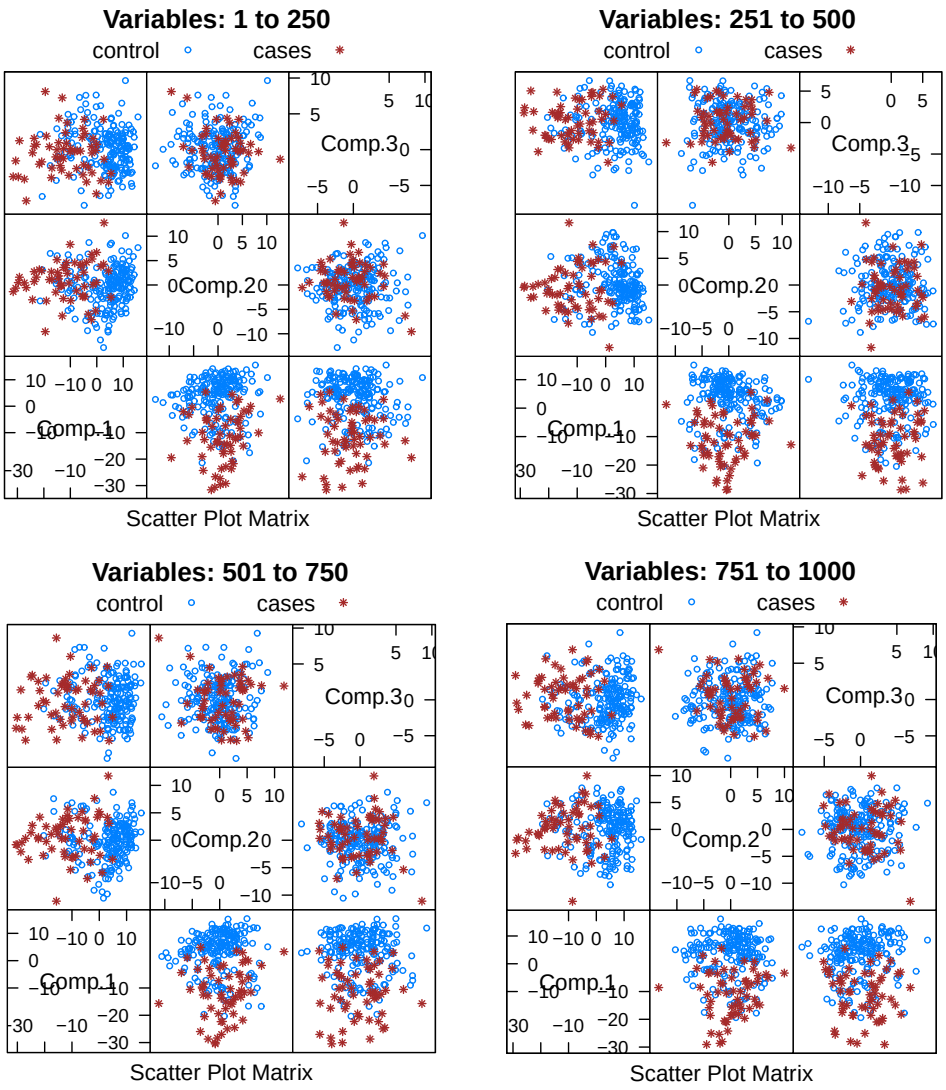


Figure 3.4: Scatterplot of the 250 observations on three Principal Components whose eigenvalues are the largest

Additionally, for each of the four matrices of dimension 250×250, the projection of the 250 points on the planes generated by the first three eigenvectors, those corresponding to the Principal components, are presented in Figure 3.4. The projection of the data set into planes suggests that it may exist some linear projections where the points can be distinguished by group, the 58 cases in red and the 192 controls in blue. In fact the projection onto the first eigenvector or principal component separates a large part of each group from the other.

3.4.2 Selection of variables and classification rates

In order to diminish the dimensionality of the space where the 250 points lie, we could: a) construct composite or derived variables, like the principal components and take a subset of them, those associated to the largest eigenvalues, or b) select a subset of variables. We do the latter. The selection of variables is done in two ways.

1. We consider a CG-distribution, $f(c, x_1, ..., x_{1000})$, whose associated graph is a tree, and use this tree as a guide to select the variables. We select those corresponding to nodes in the tree that are closest to the variable C.

2. We select a subset of variables directly from each subgroup of 100 variables. We use the relation between the linear discriminant function and the least squares linear regression $Y = \beta_0 + \beta_1 X_1 + ... + \beta_{100} X_{100} + \epsilon$. Technically, the parameters in a linear discriminant function subject to a proportionality factor can be estimated by using a least squares linear regression of Y on $X_1, ..., X_p$, taking

$$Y = \begin{cases} \frac{n}{n_1} & \text{if} \quad C = 1 \\ -\frac{n}{n_2} & \text{if} \quad C = 2, \end{cases}$$

   see e.g., Hastie *et al.* (2009, p. 135) or Fisher (1936). We then search for the structure of the graph using the set of selected variables.

We note that in case 1, we first learn the structure of the graph, then select a subset of variables, and finally estimate the parameters of the model to compute the conditional probabilities of each group. Whereas in case 2, we first select a subset of variables using least squares linear regression, then learn the structure of the graph, and finally estimate the parameters to get the conditional probabilities.

3.4.3 Results

Case 1. When considering a CG-distribution $f(c, x_1, ..., x_{1000})$ with a tree as associated graph, we used the *gRapHD* R package (Abreu et al., 2010) with the function $MinForest$ and assuming the homogeneous case, i.e., considering that the conditional distribution $f(x_1, ..., x_{1000}|c)$ has the same concentration matrix for each value of c. The $MinForest$ function searches for the best CG-distribution whose graph is a tree or a forest.

Figure 3.5 displays the tree associated to the CG-distribution that best adjust to the data set. This is the same tree as the one found and presented by Edwards et al. (2010).

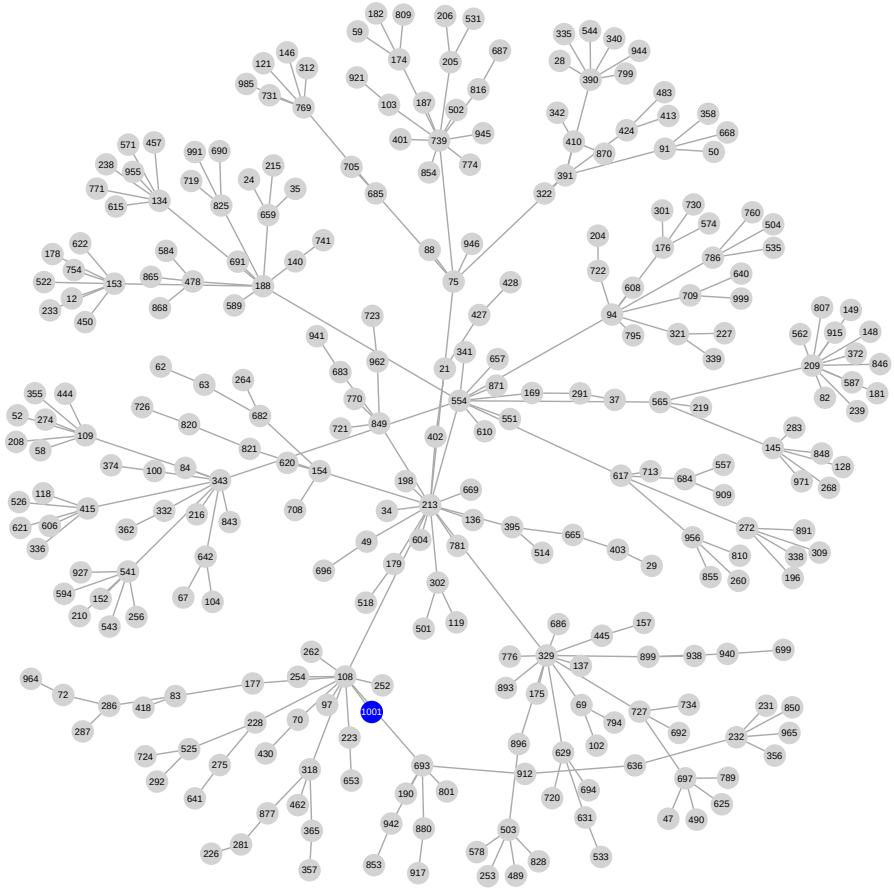


Figure 3.5: Minimal LR tree for the breast cancer data. The class variable C=1001 is shown as a blue circle.

The graph shows that the binary variable C (displayed as number 1001 and in blue) appears connected only with variable X_{108} (the path between X_{1001} and X_{693} is a path that crosses X_{108} and connects X_{693} with X_{108}). This means that variable C is conditionally independent from the other 999 variables given variable X_{108}, $(\boldsymbol{X} \setminus X_{108}) \perp\!\!\!\perp C \,|\, X_{108}$. Similar conditional independences can be read off from the graph when considering nodes in a neighbourhood of X_{108}.

When reading off conditional independences from the tree, one may think that variables whose nodes are within a neighbourhood of the grouping variable C should have some discriminative power. We selected the subset of 41 variables, those at distance less than or equal to 3 from C. Then we estimate the discriminant function and compute the conditional probabilities for three cases: saturated concentration matrices, diagonal, and those associated with a tree obtained from the original tree in Figure 3.5 when removing the class variable.

We estimate the error rates using the repeated holdout method (Kim, 2009) assuming the set of 41 variables and a fixed graph structure. In this method, approximately three fourths of the observations (144 controls and 44 cases) were randomly selected for training and one fourth for testing. This was repeated 200 times and the error rates were then estimated by the average of the 200 percentages of misclassified observations.

The discriminant analysis was done considering two cases, equal and different concentration matrices corresponding to the homogeneous and heterogeneous model, respectively. The tree graph associated with the tree model is shown in Figure 3.6. The estimated error rates are presented in Table 3.2. In Table G.1 in Appendix G, the resubstitution error rates are presented, those computed using the $n = 250$ observations to fit the model and use this to predict the class for each of these $n = 250$ observations.

Additionally, we fit a linear logistic regression with the 41 continuous variables. The estimated

error rates are presented in Table 3.2.

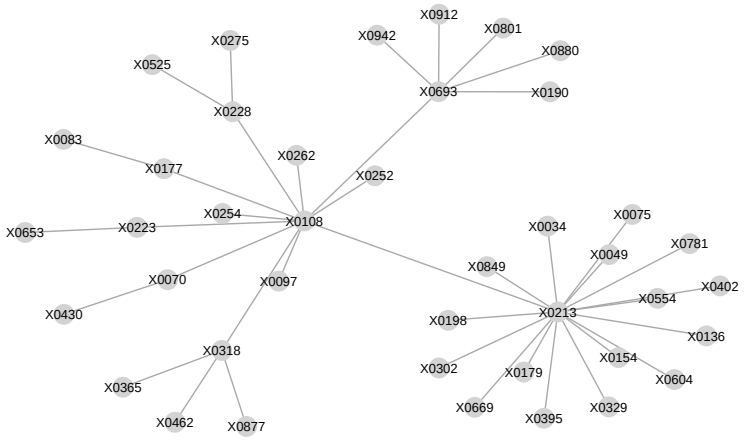


Figure 3.6: Tree graph τ associated with the tree model with $p = 41$ selected variables.

	Sample size		Discriminant Analysis						Logistic Reg
			Heterogeneous $\boldsymbol{K}_1, \boldsymbol{K}_2$			Homogeneous $\boldsymbol{K}_1 = \boldsymbol{K}_2$			
	n_t	n_v	$\boldsymbol{K}_\kappa$	$\boldsymbol{K}_{\overline{\kappa}}$	$\boldsymbol{K}_\tau$	$\boldsymbol{K}_\kappa$	$\boldsymbol{K}_{\overline{\kappa}}$	$\boldsymbol{K}_\tau$	$\boldsymbol{K}_\kappa$
Controls	144	48	0.13	16.58	14.42	10.49	16.47	10.34	14.26
Cases	44	14	99.43	35.25	37.54	52.89	30.36	38.89	54.71
Global	188	62	22.55	20.80	19.64	20.06	19.60	**16.79**	23.40
# of param			1,804	164	244	943	123	163	42

Table 3.2: Error rates computed using repeated holdout method with 200 random samples and considering 41 variables, those selected as the neighbours at distance less than 4 to variable class $C = 1001$: 108, 70, 97, 177, 213, 223, 228, 252, 254, 262, 318, 693, 430, 83, 34, 49, 75, 136, 154, 179, 198, 302, 329, 395, 402, 554, 604, 669, 781, 849, 653, 275, 525, 365, 462, 877, 190, 801, 880, 912, 942. # of param = Number of estimated parameters including the corresponding to the mean vectors.

We observe that when using the saturated models, the classification performance is poor, especially in the group of Cases. Considering the global error rates, the best performance is obtained when using a tree structure and the homogeneous case.

We also observe that the resubstitution error rates when using the empty graph or the tree graph are similar to those obtained using the holdout method. For the saturated cases, the resub-

stitution error rates are very optimistic, for example, the saturated model in the heterogeneous case has a resubstitution error rate in the group of Cases of 8.62% while the error rate estimated using the holdout method is 99.43%.

Case 2. In this case, we first select a subset of variables and then search for the structure of the graph. Using the 250 observations, we used a stepwise method for selection of the variables based on the t statistic associated with the test $H_0 : \beta_j = 0$, where β_j is one of the coefficients in the least squares linear regression. This test is equivalent to the one about comparing the Mahalanobis distance between two distributions, one with k variables and the other with $k-1$, see for example Seber (1984, p. 338) or Bodnar and Okhrin (2011). Due to the limited number of observations, the stepwise method was performed on each of ten subsets of variables, $\{X_1, ..., X_{100}\}, ..., \{X_{901}, ..., X_{1000}\}$. In this way 55 variables were selected, see Table 3.3.

$1-100$	$101-200$	$201-300$	$301-400$	$401-500$	$501-600$	$601-700$	$701-800$	$801-900$	$901-1000$
69	108	209	318	415	508	693	754	807	967
48	120	287	347	430	525	654	775	802	987
83	132	277	326	414	591	656	724	803	915
3	160	207	377	418	594	603	712	885	921
79	138	279	384	466	567	663	701	812	968
			328			652			923
			374						
			363						

Table 3.3: Subsets of selected variables when using a stepwise method in a linear regression of Y on each of ten subsets of X variables.

Using the 55 selected variables we estimated a CG-distribution assuming four different dependence graphs:

i) Complete κ.

ii) Isolated nodes, i.e. with no edges, $\overline{\kappa}$.

iii) A forest f, which is an acyclic undirected graph.

iv) A decomposable G.

And for each graph, two models were used: homogeneous and heterogeneous, i.e., considering equal or different concentration matrices on the two groups.

The structure on the concentration matrix associated with the forest graph f was found using function $MinForest$ of $gRapHD$ R package to find a tree, with the 55 continuous variables and the group variable C. Once the tree graph is found, we considered the forest which is the subgraph obtained when discarding variable C.

The structure on the concentration matrix associated with a decomposable graph G is obtained as follows. First a tree is found considering the 55 continuous variables and the group variable C. Then, starting with this tree and using function *stepw* in $gRapHD$ R package, a decomposable graph is found. Finally, the decomposable graph G is the subgraph obtained when discarding variable C.

As before, we estimate error rates using the repeated holdout method considering the graph and variables as fixed in each of the 200 iterations. In Figure 3.7 the associated graphs for the forest and decomposable models are displayed. In Table 3.4, estimated error rates are presented for each model. The error rates for the saturated model in the heterogeneous case were not computed since the sample size was not enough. However, these errors are larger than the ones computed for the other models. Error rates obtained when using a linear logistic regression on the 55 are also shown.

From the classification results we observe that the saturated model in the homogeneous case is the one with the best global performance, however this is the one with the worst performance in the Cases group. Models in the homogeneous case assuming a forest or a decomposable graph have the best general performance. We observe that the error rates with these 55 variables are

smaller than those presented in Table 3.2 obtained with a set of 41 variables.

	Sample size		Discriminant Analysis								Logistic Reg
			Heterogeneous $\boldsymbol{K}_1, \boldsymbol{K}_2$				Homogeneous $\boldsymbol{K}_1 = \boldsymbol{K}_2$				
	n_t	n_v	$\boldsymbol{K}_\kappa$	$\boldsymbol{K}_{\overline{\kappa}}$	$\boldsymbol{K}_f$	$\boldsymbol{K}_G$	$\boldsymbol{K}_\kappa$	$\boldsymbol{K}_{\overline{\kappa}}$	$\boldsymbol{K}_f$	$\boldsymbol{K}_G$	$\boldsymbol{K}_\kappa$
Controls	144	48		11.61	9.90	10.29	5.61	11.40	9.20	9.40	9.55
Cases	44	14		17.75	24.43	20.89	24.89	18.14	17.61	18.00	24.11
Global	188	62		13.00	13.18	12.69	**9.97**	12.92	11.10	11.34	12.84
# of param			3,190	220	320	354	1,650	165	270	287	56

Table 3.4: Error rates computed using repeated holdout method with 200 random samples and considering 55 variables: 3, 48, 69, 79, 83, 108, 120, 132, 138, 160, 207, 209, 277, 279, 287, 318, 326, 328, 347, 363, 374, 377, 384, 414, 415, 418, 430, 466, 508, 525, 567, 591, 594, 603, 652, 654, 656, 663, 693, 701, 712, 724, 754, 775, 802, 803, 807, 812, 885, 915, 921, 923, 967, 968, 987.
# of param = Number of estimated parameters including the corresponding to the mean vectors.

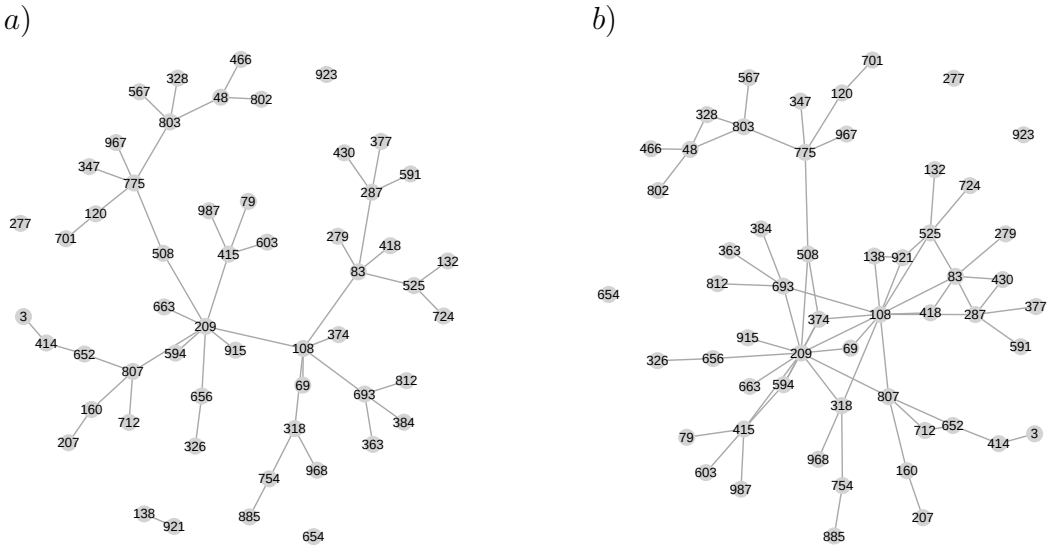


Figure 3.7: Graphs associated with CG-distribution models with 55 continuous variables. a) a forest f and b) a decomposable G

As an extra exercise, we tried a stepwise selection method on the 55 variables and got the following 15 variables: 3, 79, 132, 328, 347, 374, 415, 525, 567, 591, 654, 885, 915, 923, 987. Using this subset we got the corresponding estimated error rates shown in Table 3.5. The corresponding graphs associated with the forest and decomposable models are displayed in Figure 3.8.

	Sample size		Discriminant Analysis								Logistic Reg
			Heterogeneous $\boldsymbol{K}_1, \boldsymbol{K}_2$				Homogeneous $\boldsymbol{K}_1 = \boldsymbol{K}_2$				
	n_t	n_v	$\boldsymbol{K}_\kappa$	$\boldsymbol{K}_{\overline{\kappa}}$	$\boldsymbol{K}_f$	$\boldsymbol{K}_G$	$\boldsymbol{K}_\kappa$	$\boldsymbol{K}_{\overline{\kappa}}$	$\boldsymbol{K}_f$	$\boldsymbol{K}_G$	$\boldsymbol{K}_\kappa$
Controls	144	48	5.35	7.66	8.94	7.49	3.41	7.03	5.36	3.76	5.11
Cases	44	14	26.64	22.68	22.57	24.57	19.54	22.07	21.29	19.14	17.14
Global	188	62	10.16	11.05	12.02	11.35	**7.05**	10.43	8.96	7.23	7.83
# of param			270	60	74	92	150	45	52	61	16

Table 3.5: Error rates computed using repeated holdout method with 200 random samples and considering 15 variables: 3, 79, 132, 328, 347, 374, 415, 525, 567, 591, 654, 885, 915, 923, 987.
# of param = Number of estimated parameters including the corresponding to the mean vectors.

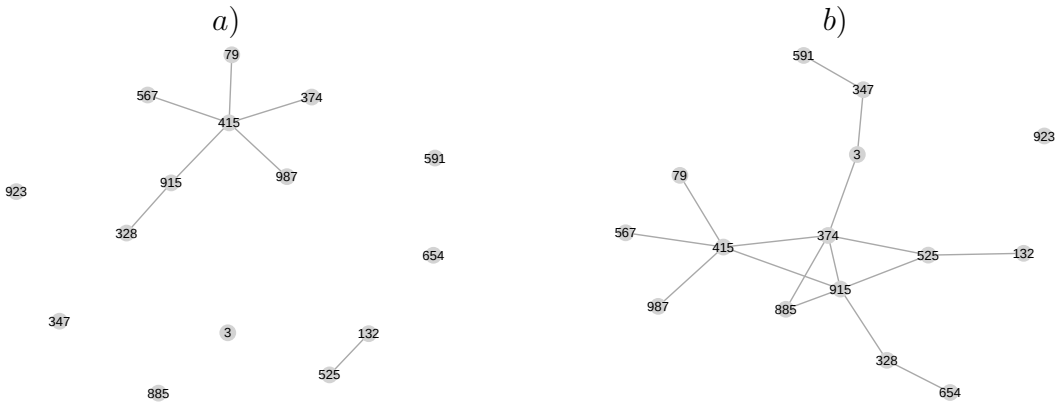


Figure 3.8: Graphs associated with CG-distribution models with 15 continuous variables. a) a forest f and b) a decomposable G

When taking 15 variables out of the 55, the global error rates are smaller, however, error rates for the Cases group are larger in some instances. The saturated model in the homogeneous case has the best performance.

Finally, we notice that when selecting a set of variables using some properties of discriminant analysis before learning the structure of the graph, we got the best classification rates.

Miller et al. (2005) reported the classification rates, estimated using a cross validation proce-

dure, given in Table 3.6. We observe that the performance of any of the models with 15 variables is better than, or similar to, the one associated with the optimal model in Miller et al. (2005).

	Predicted		
Observed	Controls	Cases	%
Controls	166	26	86.5
Cases	12	46	79.3
			84.8
# of param			96

Table 3.6: Classification rates reported in Miller et al. (2005) considering a diagonal linear discriminant analysis with $p = 32$ variables and estimated using a cross validation procedure.
# of param-Number of estimated parameters.

3.5 Comments

In the context of discriminant analysis we are considering the conditional Gaussian distribution as the joint distribution for the vector of p features and the group variable. Considering only two groups or classes, we consider linear restrictions on each of the two concentration matrices as those introduced by Anderson (1970). We have adapted the IPM algorithm to obtain estimates, only for the particular case of linear restrictions as in (3.1) and when some parameters in one concentration matrix are equal to their corresponding ones in the second.

The algorithm have been implemented in C++ and works well for small values of p, say $p \leq 30$ or when the number of matrices $\boldsymbol{H}_h$ is similar to p with $p <= 1000$. For larger values it is time consuming, therefore its implementation has to be optimized or alternative algorithms should be considered.

We remark that the algorithm is used only to obtain the parameter estimates of the concentrations matrices, but not their structure, i.e. we assume that a) the zeroes, b) the linear restrictions, and c) the equalities between corresponding parameters of the two matrices are given. An algo-

rithm for searching the structure has to be developed, or adapted from the existing ones. Some algorithms or methods have already been proposed for particular cases, for example, the ones listed in Section 3.3.

We note that linear restrictions on the concentration matrices can also be considered in the logistic regression function, and with this the number of parameters in (1.12) diminishes as we mentioned in Chapter 1, equation 1.14. In the illustrative example with $p = 1,000$ variables and 250 observations, the use of diagonal discriminant analysis and the use of a tree structure are alternatives even when considering the set of $p = 1,000$ variables, whereas linear logistic regression needs at least 1,001 observations to be used.

Another issue is variable selection for classification in high dimensionality. We have used the relation between linear discriminant analysis and linear regression to select some variables in the practical example, but we could have used the relation between discriminant analysis and logistic regression. Witten and Tibshirani (2011) and Clemmensen et al. (2011) present methods for performing linear discriminant analysis with a sparseness criterion imposed, such that, classification and variable selection are performed simultaneously, though they are not interested in the structure of the concentration matrices, that is, they estimate directly the parameters associated with the features in the linear discriminant function.

4. Some allocation rules based on trees in discriminant analysis

In this chapter, we describe and compare six methods for learning Graphical Gaussian tree models in the context of discriminant analysis for two Gaussian populations. In each method an unknown tree structure is assumed for each concentration matrix involved in the discriminant function. By finding a minimum weight spanning tree (MWST) and using maximum likelihood (ML) estimation, the concentration matrices are estimated and used in the plug-in allocation rules.

Three of these methods have been introduced in the literature: Chow and Liu (1968), Friedman et al. (1997, 1998), and Tan et al. (2010). And based on these, three others are introduced in this work, for which the function to be optimized is the J-divergence for one of them, and the empirical log-likelihood ratio (log-ratio) for the two others. We show in Propositions 4.5.1 and 4.5.2, and corollary 4.5.3, that the optimization problems of the proposed methods are equivalent to a problem of finding a MWST.

All methods take advantage of the tree structure, specifically of an efficient algorithm for finding the MWST, and the existence of analytical expressions for the Maximum Likelihood Estimators (MLEs) of the concentration matrices.

We present a numerical study where the performance of the six methods is compared when the group training sample size is the same in both populations; and for this case the method given in

Tan et al. (2010) and the one based on the log-ratio are equivalent, as well as the methods based on the J-divergence and the log-ratio with equal trees. The comparison of the different methods is based on the estimated error rates of the corresponding rules, obtained from real and simulated data. Diagonal discriminant analysis is considered as a benchmark, as well as quadratic and linear discriminant analysis but only whenever the sample size is sufficient.

We also consider HIV data to illustrate the case when the data correspond to repeated measures and the training samples sizes are different. In this case, the performance of the six methods is compared, and diagonal discriminant analysis is also considered as a benchmark.

In spite of showing that none of the methods based on tree models outperforms the benchmarks in all data sets, any of these methods offers a simple and computationally inexpensive alternative to well established discriminant methods in high dimensional settings, where sample size is similar to, or smaller than, the number of variables.

4.1 Related work

The idea of using trees for pattern recognition was introduced by Chow and Liu (1966). They proposed a procedure for finding two independent tree structures, one on each population, for a set of binary variables. Later on, Chow and Liu (1968) proposed the use of trees to approximate the distribution of a set of discrete variables, and these approximating distributions were then used to build allocation rules. The problem of estimating a single tree structure was expressed as an optimization problem, where the Kullback-Leibler divergence is optimized over the set of all possible trees with as many nodes as variables. They showed that the optimization problem is equivalent to the one of finding the ML tree model, and also that it can be formulated as the one of finding a MWST. To find the MWST, a set of weights are calculated, one for each pair of variables, and then the use of an efficient algorithm for finding a MWST for the weighted complete graph is used.

Based on Chow and Liu's idea, other classification methods have been developed, all of which are based on the MWST problem. Tree-augmented naive Bayesian network is one of them introduced in Friedman et al. (1997) for discrete variables, and in Friedman et al. (1998) for the Gaussian case. When using this network the joint likelihood is maximized assuming the same tree in both populations. Another method based on a mixture of trees is given in Meilă and Jordan (2000), for both discrete and continuous variables, where Chow and Liu's method is a special case when the mixture is based on a single tree.

More recently, Tan et al. (2010) proposed a method based on optimizing an approximation of the $J-$divergence for either discrete or continuous variables. They also considered the log-ratio measure, but only for discrete variables, in which case, showed that the problems of optimizing the approximated J-divergence and the log-ratio are equivalent. They did not consider the log-ratio for the Gaussian case, which we study in two versions: with two arbitrary and with two equal trees for both populations. The idea of using the J-divergence is also studied, it is based on their work, but avoiding the use of empirical distributions which may not be well defined for the multivariate case when the number of variables is larger than the sample size. Therefore, the proposed method optimizes the J-divergence between two tree multivariate Gaussian distributions which are well defined. To obtain the equivalence between the associated optimization problem of this method and the one of finding the MWST, the same tree in both populations is assumed.

Chow and Liu's idea has also been used in other contexts, for example, in Edwards et al. (2010) it is used to approximate single distributions based on the AIC and BIC.

Before stating the six methods, we describe some properties of GGMs with tree structure and the problem of finding the MWST.

4.2 Graphical Gaussian models with tree structure

A tree graph $\tau = (V, E)$ is an acyclic undirected graph where all vertices are connected. For example, three different tree graphs with six vertices and different edge set are shown in Figure 4.1.

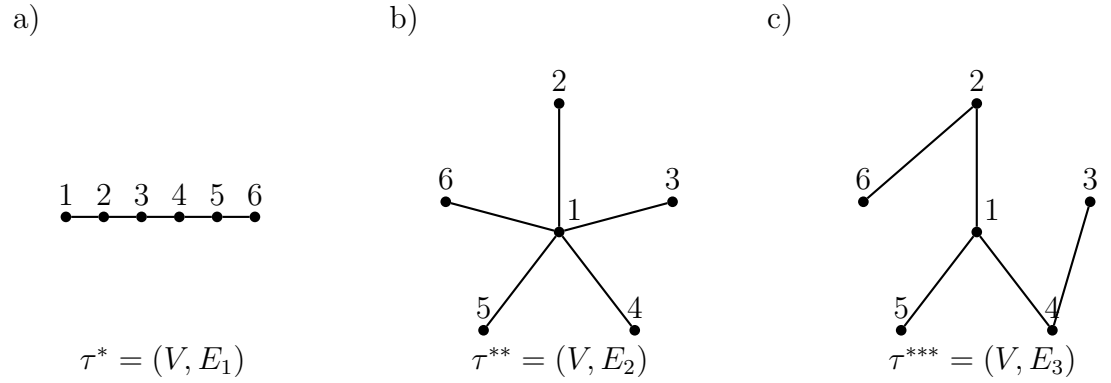


Figure 4.1: Three examples of tree graphs $\tau = (V, E_i)$, $i = 1, 2, 3$, with $V = \{1, 2, 3, 4, 5, 6\}$ and a) a path with $E_1 = \{(1, 2), (2, 3), (3, 4), (4, 5), (5, 6)\}$, b) a star with $E_2 = \{(1, 2), (1, 3), (1, 4), (1, 5), (1, 6)\}$, and c) another tree with $E_3 = \{(1, 2), (1, 4), (1, 5), (2, 6), (3, 4)\}$.

Let T_p denote the set of all the tree graphs with p vertices. When considering a graphical model with a tree $\tau = (V, E_\tau) \in T_p$ as associated graph, the distribution of the p variables can be factorized into factors which depend on the marginal distributions of one or two variables as follows

$$f_\tau(x_1, ..., x_p) = \prod_{i=1}^{p} f(x_i) \prod_{\substack{i<j \\ (i,j)\in E_\tau}} \frac{f(x_i, x_j)}{f(x_i)\ f(x_j)} \,. \tag{4.1}$$

This property follows from properties 4.2.1 to 4.2.3 described below and makes the problems of parameter and structure estimation solvable in an efficient way. Graphical models with tree graphs are particular instances of decomposable models described in Chapter 1, which also have useful properties for parameter estimation. In particular, when considering a GGM with a tree graph $\tau = (V, E_\tau)$, the following properties follow, some of them can be found in Lauritzen (2011).

4.2.1 It is a decomposable model. That is, the density function can be factorized in terms of

the set of cliques $\mathscr{C}$ and the set of separators $\mathscr{S}$ as in (1.16)

$$f_\tau(x_1, ..., x_p) = \frac{\prod_{\mathcal{C}\in\mathscr{C}} f(\boldsymbol{x}_\mathcal{C})}{\prod_{\mathcal{S}\in\mathscr{S}} f(\boldsymbol{x}_\mathcal{S})^{v(\mathcal{S})}}.$$

4.2.2 Its set of cliques $\mathscr{C}$ consists of all the edges in the tree.

4.2.3 Its set of separators $\mathscr{S}$ consists of the set of nodes such that each node is in $\mathscr{S}$ as many times as one less than its degree.

4.2.4 The clique-separator factorization is based on univariate and bivariate distributions and can be expressed as in (4.1).

4.2.5 $\boldsymbol{K}_\tau$ can be decomposed as

$$\boldsymbol{K}_\tau = \sum_{\mathcal{C}\in\mathscr{C}} [\boldsymbol{K}_\mathcal{C}]^p - \sum_{\mathcal{S}\in\mathscr{S}} v(\mathcal{S})[\boldsymbol{K}_\mathcal{S}]^p = \sum_{\substack{i<j \\ (i,j)\in E_\tau}} \left([\boldsymbol{K}_{(i,j)}]^p - [\boldsymbol{K}_{(i)}]^p - [\boldsymbol{K}_{(j)}]^p\right) + \sum_{j=1}^{p} \left([\boldsymbol{K}_{(j)}]^p\right).$$

4.2.6 The MLE of $\boldsymbol{\Sigma}_\tau$, $\widehat{\boldsymbol{\Sigma}}_\tau$, given a sample $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ has the following property

$$\widehat{\boldsymbol{\Sigma}}_\tau(\tau) = \boldsymbol{W}(\tau),$$

where $\boldsymbol{W} = \sum_{j=1}^n (\boldsymbol{x}_j - \overline{\boldsymbol{x}})(\boldsymbol{x}_j - \overline{\boldsymbol{x}})^t/n$, and for any square matrix $\boldsymbol{A}$, $\boldsymbol{A}(\tau)$ is the square matrix such that

$$(\boldsymbol{A}(\tau))_{ij} = \begin{cases} (\boldsymbol{A})_{ij} & \text{if } i = j \text{ or } (i,j) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

4.2.7 For any square matrix $\boldsymbol{A}$,

$$tr(\boldsymbol{K}_\tau \boldsymbol{A}) = tr(\boldsymbol{K}_\tau \boldsymbol{A}(\tau)).$$

Moreover, if the MLE of $\boldsymbol{K}_\tau$, $\widehat{\boldsymbol{K}_\tau}$, exists, then $tr(\widehat{\boldsymbol{K}_\tau}\boldsymbol{A}) = tr(\widehat{\boldsymbol{K}_\tau}\boldsymbol{A}(\tau))$.

4.2.8 $\widehat{\boldsymbol{K}}_\tau$ can be expressed as

$$\widehat{\boldsymbol{K}}_\tau = \sum_{\mathcal{C}\in\mathscr{C}} [\boldsymbol{W}_{\mathcal{C}}^{-1}]^p - \sum_{\mathcal{S}\in\mathscr{S}} v(\mathcal{S})[\boldsymbol{W}_{\mathcal{S}}^{-1}]^p = \sum_{\substack{i<j \\ (i,j)\in E_\tau}} \left([\boldsymbol{W}_{(i,j)}^{-1}]^p - [\boldsymbol{W}_{(i)}^{-1}]^p - [\boldsymbol{W}_{(j)}^{-1}]^p\right) + \sum_{j=1}^{p} \left([\boldsymbol{W}_{(j)}^{-1}]^p\right). \tag{4.2}$$

4.2.9 $\widehat{f}_\tau(x_1,...,x_p) = \prod_{i=1}^p \widehat{f}(x_i) \prod_{\substack{i<j \\ (i,j)\in E_\tau}} \dfrac{\widehat{f}(x_i,x_j)}{\widehat{f}(x_i)\ \widehat{f}(x_j)}$, where $\widehat{f}_\tau$ is the density of $N(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{K}}_\tau)$.

4.2.10

$$\ln \frac{\widehat{f}(x_i,x_j)}{\widehat{f}(x_i)\ \widehat{f}(x_j)} = -\tfrac{1}{2}\ln(1-\widehat{\rho}_{ij}^2) - \frac{\widehat{\rho}_{ij}^2}{2(1-\widehat{\rho}_{ij}^2)} \left\{ \frac{(x_i-\overline{x}_i)^2}{w_{ii}} + \frac{(x_j-\overline{x}_j)^2}{w_{jj}} - \frac{2(x_i-\overline{x}_i)(x_j-\overline{x}_j)}{w_{ij}} \right\},$$

where $\widehat{\rho}_{ij} = w_{ij}/\sqrt{w_{ii}w_{jj}}$.

Properties 4.2.6 and 4.2.7 are also true for any dependence graph G. We note that the existence of the MLE in equation 4.2 is assured for sample size larger than the maximal clique size, in this case $n > 2$, see for example, Proposition 7 in Frydenberg and Lauritzen (1989) or Proposition 5.9 in Lauritzen (1996). We also note that the number of unknown parameters in the concentration matrix is reduced from $p(p+1)/2$, which corresponds to the graph κ, to $2p-1$.

We consider tree models with trees $\tau_1 = (V, E_{\tau_1})$ and $\tau_2 = (V, E_{\tau_2})$. Hereafter, we use the notation $f_{1_{\tau_1}}$, $f_{2_{\tau_2}}$, $\boldsymbol{K}_{1_{\tau_1}}$ and $\boldsymbol{K}_{2_{\tau_2}}$, to clarify the dependence on the specific population and on the specific tree.

We note that when considering arbitrary covariance matrices or, equivalently, concentration matrices $\boldsymbol{K}_{1_{\tau_1}}$ and $\boldsymbol{K}_{2_{\tau_2}}$, two cases can be considered: (i) τ_1 and τ_2 are arbitrary trees and (ii) $\tau_1 = \tau_2 = \tau$. Additionally, we can also make the assumption of equal covariance matrices which implies $\tau_1 = \tau_2 = \tau$, so that $\boldsymbol{K}_{1_{\tau_1}} = \boldsymbol{K}_{2_{\tau_2}} = \boldsymbol{K}_\tau$. For example, when considering the tree τ^* given in Figure 4.1a) in both populations, the concentration matrices could be the same:

$$\boldsymbol{K}_{1_{\tau^*}} = \boldsymbol{K}_{2_{\tau^*}} = \begin{pmatrix} k_{11} & k_{12} & 0 & 0 & 0 & 0 \\ k_{12} & k_{22} & k_{23} & 0 & 0 & 0 \\ 0 & k_{23} & k_{33} & k_{34} & 0 & 0 \\ 0 & 0 & k_{34} & k_{44} & k_{45} & 0 \\ 0 & 0 & 0 & k_{45} & k_{55} & k_{56} \\ 0 & 0 & 0 & 0 & k_{56} & k_{66} \end{pmatrix},$$

or different:

$$\boldsymbol{K}_{c_{\tau^*}} = \begin{pmatrix} k_{11}^{(c)} & k_{12}^{(c)} & 0 & 0 & 0 & 0 \\ k_{12}^{(c)} & k_{22}^{(c)} & k_{23}^{(c)} & 0 & 0 & 0 \\ 0 & k_{23}^{(c)} & k_{33}^{(c)} & k_{34}^{(c)} & 0 & 0 \\ 0 & 0 & k_{34}^{(c)} & k_{44}^{(c)} & k_{45}^{(c)} & 0 \\ 0 & 0 & 0 & k_{45}^{(c)} & k_{55}^{(c)} & k_{56}^{(c)} \\ 0 & 0 & 0 & 0 & k_{56}^{(c)} & k_{66}^{(c)} \end{pmatrix}, \quad c = 1, 2.$$

4.3 Kullback and Leibler divergence

The KL-divergence and the J-divergence (Kullback and Leibler, 1951) are two measures that have been used to define some of the optimization problems for the tree structure estimation. Let $I(f_1, f_2)$ and $J(f_1, f_2)$ denote the KL-divergence and the J-divergence between two functions f_1 and f_2. When f_1 and f_2 are the densities of $N(\boldsymbol{\mu}_1, \boldsymbol{K}_1^{-1})$ and $N(\boldsymbol{\mu}_2, \boldsymbol{K}_2^{-1})$, respectively, the expressions are

$$\begin{aligned} I(f_1, f_2) &= \int f_1(\boldsymbol{x}) \ln \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} d\boldsymbol{x} \\ &= \tfrac{1}{2} \ln \frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|} + \frac{1}{2} tr\big(\boldsymbol{\Sigma}_1(\boldsymbol{K}_2 - \boldsymbol{K}_1)\big) + \frac{1}{2} tr\big(\boldsymbol{K}_2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t\big), \end{aligned} \tag{4.3}$$

$$\begin{aligned} J(f_1, f_2) &= I(f_1, f_2) + I(f_2, f_1) = \int \Big(f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})\Big) \ln \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} d\boldsymbol{x} \\ &= \frac{1}{2} tr\big((\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)(\boldsymbol{K}_2 - \boldsymbol{K}_1)\big) + \frac{1}{2} tr\big((\boldsymbol{K}_1 + \boldsymbol{K}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t\big). \end{aligned} \tag{4.4}$$

For example, if we consider GGMs with associated graph $\overline{\kappa}$ in each population, then $\boldsymbol{K}_{1_{\overline{\kappa}}}$ and $\boldsymbol{K}_{2_{\overline{\kappa}}}$ are diagonal matrices and

$$J(f_{1_{\overline{\kappa}}}, f_{2_{\overline{\kappa}}}) = \frac{1}{2} \sum_{j=1}^{p} \left(\frac{\sigma_{jj}^{(1)}}{\sigma_{jj}^{(2)}} + \frac{\sigma_{jj}^{(2)}}{\sigma_{jj}^{(1)}} + \frac{(\mu_j^{(1)} - \mu_j^{(2)})^2}{\sigma_{jj}^{(1)}} + \frac{(\mu_j^{(1)} - \mu_j^{(2)})^2}{\sigma_{jj}^{(2)}} \right) - p. \tag{4.5}$$

When the tree graph τ is a path and the covariance matrices have the following entries $\sigma_{ij}^{(c)} = \rho_c^{|i-j|}/(1-\rho_c^2)$, $i, j = 1, ..., p$, $c = 1, 2$; the J-divergence is as follows

$$\begin{aligned} J(f_{1_\tau}, f_{2_\tau}) &= \tfrac{1}{2(1-\rho_1^2)} \left(2 - 2(p-1)\rho_1\rho_2 + (p-2)(1+\rho_2^2)\right) + \tfrac{1}{2(1-\rho_2^2)} \left(2 - 2(p-1)\rho_1\rho_2 + (p-2)(1+\rho_1^2)\right) \\ &\quad + \tfrac{1}{2}(2 + \rho_1^2 + \rho_2^2) \sum_{j=2}^{p-1} (\mu_j^{(1)} - \mu_j^{(2)})^2 - (\rho_1 + \rho_2) \sum_{j=1}^{p-1} (\mu_j^{(1)} - \mu_j^{(2)})(\mu_{j+1}^{(1)} - \mu_{j+1}^{(2)}) \\ &\quad + (\mu_1^{(1)} - \mu_1^{(2)})^2 + (\mu_p^{(1)} - \mu_p^{(2)})^2 - p. \end{aligned} \tag{4.6}$$

In particular, the covariance and concentration matrices for each population, with this pattern and associated with the path τ^* given in Figure 4.1a), are

$$\boldsymbol{\Sigma}_{c_{\tau^*}} = \frac{1}{1-\rho_c^2} \begin{pmatrix} 1 & \rho_c & \rho_c^2 & \rho_c^3 & \rho_c^4 & \rho_c^5 \\ \rho_c & 1 & \rho_c & \rho_c^2 & \rho_c^3 & \rho_c^4 \\ \rho_c^2 & \rho_c & 1 & \rho_c & \rho_c^2 & \rho_c^3 \\ \rho_c^3 & \rho_c^2 & \rho_c & 1 & \rho_c & \rho_c^2 \\ \rho_c^4 & \rho_c^3 & \rho_c^2 & \rho_c & 1 & \rho_c \\ \rho_c^5 & \rho_c^4 & \rho_c^3 & \rho_c^2 & \rho_c & 1 \end{pmatrix}, \; \boldsymbol{K}_{c_{\tau^*}} = \begin{pmatrix} 1 & -\rho_c & 0 & 0 & 0 & 0 \\ -\rho_c & 1+\rho_c^2 & -\rho_c & 0 & 0 & 0 \\ 0 & -\rho_c & 1+\rho_c^2 & -\rho_c & 0 & 0 \\ 0 & 0 & -\rho_c & 1+\rho_c^2 & -\rho_c & 0 \\ 0 & 0 & 0 & -\rho_c & 1+\rho_c^2 & -\rho_c \\ 0 & 0 & 0 & 0 & -\rho_c & 1 \end{pmatrix},$$

$$\begin{aligned}\text{and} \quad J(f_{1_{\tau^*}}, f_{2_{\tau^*}}) &= \tfrac{1}{2(1-\rho_1^2)}(2-10\rho_1\rho_2+4(1+\rho_2^2)) + \tfrac{1}{2(1-\rho_2^2)}(2-10\rho_1\rho_2+4(1+\rho_1^2)) \\ &+\tfrac{1}{2}(2+\rho_1^2+\rho_2^2)\sum_{j=2}^{5}(\mu_j^{(1)}-\mu_j^{(2)})^2 - (\rho_1+\rho_2)\sum_{j=1}^{5}(\mu_j^{(1)}-\mu_j^{(2)})(\mu_{j+1}^{(1)}-\mu_{j+1}^{(2)}) \\ &+(\mu_1^{(1)}-\mu_1^{(2)})^2 + (\mu_6^{(1)}-\mu_6^{(2)})^2 - 6. \end{aligned} \tag{4.7}$$

4.4 Minimum weight spanning tree

Graham and Hell (1985) in their work 'On the history of the minimum spanning tree problem' mention that Borůvka (1926) seems to be the first work where the problem of finding a MWST was formulated. This problem has been used in various types of applications, for example, in communications and transportation networks. It also offers a method of solution to other problems where it applies less directly, for example, Chow and Liu (1968) showed that the problem of tree structure estimation when using the KL-divergence can be formulated as a problem of finding a MWST.

Given a connected and undirected graph $G = (V, E)$ with p nodes, a spanning tree of G is a subgraph that is a tree. Let $T(G) = \{\tau | \tau = (V, E_\tau)$ with $E_\tau \subseteq E$ and $\tau \in T_p\}$ denote the set of all spanning trees of G, then $T(G) \subseteq T_p$ and every connected graph contains at least one spanning tree, that is, $|T(G)| \geq 1$. For example, for the complete graph κ with p nodes, $T(\kappa) = T_p$ and $|T(\kappa)| = p^{p-2}$, the last equality is called the Cayley Tree Formula. In Figure 4.2, the complete graph κ with three nodes is shown together with its three spanning trees.

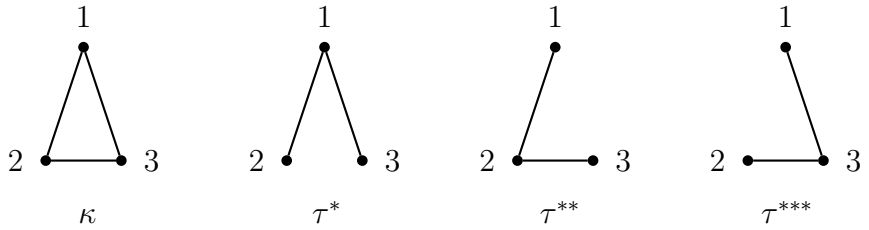


Figure 4.2: The complete graph κ with three nodes, and all its spanning trees: τ^*, τ^{**} and τ^{***}.

If a real number $\lambda(i,j)$, called weight or cost, is assigned to each edge $(i,j) \in E$, the total weight $\lambda(G)$ of G is defined as

$$\lambda(G) = \sum_{(i,j)\in E} \lambda(i,j).$$

The weight of each spanning tree of G is the sum of the weights associated with its edges. A MWST of G is any spanning tree whose weight is minimum among all the spanning trees of G. If all the weights of G are different, then the MWST is unique. The problem of finding a MWST is called minimum spanning tree problem and can be expressed as follows.

Find τ^* such that

$$\tau^* = \underset{\tau\in T(G)}{\operatorname{argmin}} \sum_{(i,j)\in E_\tau} \lambda(i,j)\ , \tag{4.8}$$

where $T(G) = \{\tau | \tau = (V, E_\tau)$ with $E_\tau \subseteq E$ and $\tau \in T_p\}$.

Among the algorithms used to solve the minimum spanning tree problem, there are two algorithms commonly used: Kruskal's (Kruskal, 1956) and Prim's algorithm (Prim, 1957). The former can also be used to find a Minimum Weight Spanning Forest (MWSF) with t edges, $t < p$, during its iterations, and in the end a MWST is found, where a forest is an acyclic undirected graph. On the other hand, in each iteration of the Prim's algorithm there is always a component which is a tree, and the other components are isolated nodes; at the end of the algorithm there is just one component which is a MWST. In Appendix B, we describe Kruskal's algorithm which is the one we have used to obtain the numerical results presented in this chapter.

In the following section, we describe the optimization problem associated with each of the six methods. Each of these problems can be formulated as one of finding a MWST. We notice that the graph G associated with each problem is the complete graph κ, so that $T(G) = T(\kappa) = T_p$.

4.5 Tree based allocation rules

The methods based on trees described in this section have the two estimation problems associated with GGMs: parameter and structure estimation. Once both problems are solved, the plug in allocation rule is specified.

The parameter estimation problem for each method is solved using the MLEs for the means and for the concentration matrices. Each concentration matrix depends on a specific tree τ_c, $c = 1, 2$. In the heterogeneous case, MLEs of the concentration matrices are $\widehat{\boldsymbol{K}}_{1_{\tau_1}}$ and $\widehat{\boldsymbol{K}}_{2_{\tau_2}}$ with

$$\widehat{\boldsymbol{K}}_{c_{\tau_c}} = \sum_{\substack{i<j \\ (i,j)\in E_{\tau_c}}} \left([\boldsymbol{W}^{-1}_{c_{(i,j)}}]^p - [\boldsymbol{W}^{-1}_{c_{(i)}}]^p - [\boldsymbol{W}^{-1}_{c_{(j)}}]^p \right) + \sum_{j=1}^{p} [\boldsymbol{W}^{-1}_{c_{(j)}}]^p, \tag{4.9}$$

and $\widehat{\boldsymbol{\mu}}_c$ and $\boldsymbol{W}_c$ as in (1.4) and (1.5), respectively. In the homogeneous case, the rule is based on (1.6) with $\widehat{\boldsymbol{\mu}}_c$ as in (1.4), and $\widehat{\boldsymbol{K}}_\tau$ as in (4.9) with $\boldsymbol{W}$ as in (1.7). For instance, when considering the path given in Figure 1.1c) and $\boldsymbol{W}$ as the sample covariance matrix, the MLE of $\boldsymbol{K}_\tau$ is

$$\begin{aligned}
\widehat{\boldsymbol{K}}_\tau \;=\; & \begin{pmatrix} \begin{pmatrix} w_{11} & w_{12} \\ w_{12} & w_{22} \end{pmatrix}^{-1} & \begin{matrix} 0 & 0 \\ 0 & 0 \end{matrix} \\ \begin{matrix} 0 & 0 \\ 0 & 0 \end{matrix} & \begin{matrix} 0 & 0 \\ 0 & 0 \end{matrix} \end{pmatrix} - \begin{pmatrix} w_{11}^{-1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} - \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & w_{22}^{-1} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \\
& + \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \begin{pmatrix} w_{22} & w_{23} \\ w_{23} & w_{33} \end{pmatrix}^{-1} & & 0 \\ 0 & & & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} - \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & w_{22}^{-1} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} - \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & w_{33}^{-1} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \\
& + \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \begin{pmatrix} w_{33} & w_{34} \\ w_{34} & w_{44} \end{pmatrix}^{-1} & \\ 0 & 0 & & \end{pmatrix} - \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & w_{33}^{-1} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} - \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & w_{44}^{-1} \end{pmatrix}
\end{aligned}$$

$$
+\begin{pmatrix} w_{11}^{-1} & 0 & 0 & 0 \\ 0 & w_{22}^{-1} & 0 & 0 \\ 0 & 0 & w_{33}^{-1} & 0 \\ 0 & 0 & 0 & w_{44}^{-1} \end{pmatrix}
$$

$$
= \begin{pmatrix} \begin{pmatrix} w_{11} & w_{12} \\ w_{12} & w_{22} \end{pmatrix}^{-1} & 0\ 0 \\ 0\ 0 & 0\ 0 \\ 0\ 0 & 0\ 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \begin{pmatrix} w_{22} & w_{23} \\ w_{23} & w_{33} \end{pmatrix}^{-1} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0\ 0 & 0 & 0 \\ 0\ 0 & 0 & 0 \\ 0\ 0 & \begin{pmatrix} w_{33} & w_{34} \\ w_{34} & w_{44} \end{pmatrix}^{-1} \end{pmatrix}
$$

$$
- \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & w_{22}^{-1} & 0 & 0 \\ 0 & 0 & w_{33}^{-1} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.
$$

On the other hand, the estimation of the tree structure depends on a a specific optimization problem for each method. Each optimization problem can be formulated as one of finding a MWST with a specific set of weights, $\{\lambda(i,j),\ i \neq j;\ i,j = 1,...,p\}$, for expression (4.8).

In the following we state the optimization problem, and the specific set of weights for the corresponding MWST problem, for each of the six methods.

4.5.1 Existing methods

We give the optimization problem corresponding to each of the three methods that have already been studied and implemented in the literature.

Let $\widehat{f}_{c_{\tau_c}}$ denote the density of $N(\widehat{\boldsymbol{\mu}}_c, \widehat{\boldsymbol{K}}_{c_{\tau_c}}^{-1})$ and $\widetilde{f}_c$ the density of the empirical distribution $N(\overline{\boldsymbol{x}}^{(c)}, \boldsymbol{W}_c)$, $c = 1, 2$. Notice that $\widehat{\boldsymbol{\mu}}_c = \overline{\boldsymbol{x}}^{(c)}$, but $\widehat{\boldsymbol{K}}_{c_{\tau_c}} \neq \boldsymbol{W}_c^{-1}$.

i) Chow and Liu (1968). Find two independent trees τ_1^* and τ_2^* such that

$$\tau_c^* = \underset{\tau_c \in T_p}{\operatorname{argmin}}\, I(\widetilde{f}_c, \widehat{f}_{c_{\tau_c}}) = \underset{\tau_c \in T_p}{\operatorname{argmin}} \int \widetilde{f}_c(\boldsymbol{x}) \ln \frac{\widetilde{f}_c(\boldsymbol{x})}{\widehat{f}_{c_{\tau_c}}(\boldsymbol{x})} d\boldsymbol{x}, \quad c = 1, 2, \tag{4.10}$$

where $T_p = \{\tau | \tau \text{ is a tree with } p \text{ nodes}\}$. The weights $\lambda^{(c)}(i,j),\ i,j = 1, ...p$, for the MWST problem associated with τ_c^* are given by

$$\lambda^{(c)}(i,j) \;=\; \frac{1}{2} \ln\left(1 - \hat{\rho}_{ij}^{(c)\,2}\right), \quad c = 1, 2, \tag{4.11}$$

where $\widehat{\rho}_{ij}^{(c)} = w_{ij}^{(c)} \left(w_{ii}^{(c)} w_{jj}^{(c)}\right)^{-1/2}$ and $w_{ij}^{(c)}$ is the entry ij of the matrix $\boldsymbol{W}_c$ given in (1.5).

ii) Friedman et al. (1997, 1998). Find a single tree τ^* such that

$$\begin{aligned} \tau^* &= \underset{\tau \in T_p}{\operatorname{argmin}} \sum_{c=1}^{2} \left(\widetilde{\pi}_c I(\widetilde{f}_c, \widehat{f}_{c_\tau}) + \widetilde{\pi}_c \ln\left(\frac{\widetilde{\pi}_c}{\widehat{\pi}_c}\right)\right) \\ &= \underset{\tau \in T_p}{\operatorname{argmax}} \left\{ \widehat{\pi}_1 \sum_{l=1}^{n_1} \ln \widehat{f}_{1_\tau}(\boldsymbol{x}_l) + \widehat{\pi}_2 \sum_{l=n_1+1}^{n_1+n_2} \ln \widehat{f}_{2_\tau}(\boldsymbol{x}_l) \right\}, \end{aligned} \tag{4.12}$$

where observations from Π_1 have indexes from 1 to n_1, and from Π_2 from n_1+1 to n_1+n_2. Here $\widetilde{\pi}_c = \widehat{\pi}_c = n_c/(n_1+n_2)$. The weights $\lambda(i,j),\ i,j = 1, ...p$, for the MWST problem associated with τ^* are

$$\lambda(i,j) \;=\; \frac{n_1}{2(n_1+n_2)} \ln\left(1 - \hat{\rho}_{ij}^{(1)\,2}\right) + \frac{n_2}{2(n_1+n_2)} \ln\left(1 - \hat{\rho}_{ij}^{(2)\,2}\right). \tag{4.13}$$

iii) Tan et al. (2010). Find two trees (τ_1^*, τ_2^*) such that

$$\begin{aligned} (\tau_1^*, \tau_2^*) &= \underset{\tau_1, \tau_2 \in T_p}{\operatorname{argmax}} \left\{ -I(\widetilde{f}_1, \widehat{f}_{1_{\tau_1}}) + I(\widetilde{f}_2, \widehat{f}_{1_{\tau_1}}) - I(\widetilde{f}_2, \widehat{f}_{2_{\tau_2}}) + I(\widetilde{f}_1, \widehat{f}_{2_{\tau_2}}) \right\} \\ &= \underset{\tau_1, \tau_2 \in T_p}{\operatorname{argmax}} \int \left(\widetilde{f}_1(\boldsymbol{x}) - \widetilde{f}_2(\boldsymbol{x})\right) \ln \frac{\widehat{f}_{1_{\tau_1}}(\boldsymbol{x})}{\widehat{f}_{2_{\tau_2}}(\boldsymbol{x})} d\boldsymbol{x}. \end{aligned} \tag{4.14}$$

The optimization problem in (4.14) is equivalent to two independent problems of finding MWSTs τ_1^* and τ_2^*, with weights $\lambda^{(c)}(i,j)$, $i,j = 1,...,p$, given by the following expression for $c = 1, 2$.

$$\begin{aligned}\lambda^{(c)}(i,j) &= \frac{-\hat{\rho}_{ij}^{(c)\,2}}{2\left(1-\hat{\rho}_{ij}^{(c)\,2}\right)}\left\{\frac{\left(\overline{x}_j^{(3-c)}-\overline{x}_j^{(c)}\right)^2}{w_{jj}^{(c)}}+\frac{\left(\overline{x}_i^{(3-c)}-\overline{x}_i^{(c)}\right)^2}{w_{ii}^{(c)}}+\frac{w_{ii}^{(3-c)}}{w_{ii}^{(c)}}\right. \\ &\left.+\frac{w_{jj}^{(3-c)}}{w_{jj}^{(c)}}-\frac{2w_{ij}^{(3-c)}}{w_{ij}^{(c)}}-\frac{2\left(\overline{x}_i^{(3-c)}-\overline{x}_i^{(c)}\right)\left(\overline{x}_j^{(3-c)}-\overline{x}_j^{(c)}\right)}{w_{ij}^{(c)}}\right\}.\end{aligned} \tag{4.15}$$

We denote these three methods as C-L, Fried and Tan, respectively.

4.5.2 Proposed methods

Propositions 4.5.1 and 4.5.2, and corollary 4.5.3, give the optimization problem associated with each of the three proposed methods, and state the equivalence between each optimization problem and one of finding a MWST. The first method, J-div, searches for the tree for which the two estimated tree distributions differ the most according to the J-divergence. To get the equivalence between its optimization problem and the MWST problem, the same tree in both populations, $\tau_1 = \tau_2 = \tau$, is assumed. The second, based on the log-ratio, is considered in Tan et al. (2010) for discrete variables, since in this case this is equivalent to Tan method based on (4.14), where the Gaussian case was mentioned, but not studied. And the third, referred as Log-ratio-equal, is also based on the log-ratio but considering equal trees.

The equivalence between each optimization problem in (4.10), (4.12) and (4.14) and one of finding a MWST has also been proven by Chow and Liu (1968), Friedman et al. (1997, 1998) and Tan et al. (2010), respectively.

Proposition 4.5.1 *(**J-div**) Let $\widehat{f}_{1_\tau}$ and $\widehat{f}_{2_\tau}$ be the densities of p-variate distributions $N(\widehat{\boldsymbol{\mu}}_1, \widehat{\boldsymbol{K}}_{1_\tau}^{-1})$*

and $N(\widehat{\boldsymbol{\mu}}_2, \widehat{\boldsymbol{K}}_{2_\tau}^{-1})$*, respectively. The problem of finding* τ^* *such that*

$$\tau^* = \underset{\tau \in T_p}{\operatorname{argmax}}\, J(\widehat{f}_{1_\tau}, \widehat{f}_{2_\tau}) = \underset{\tau \in T_p}{\operatorname{argmax}} \int \left(\widehat{f}_{1_\tau}(\boldsymbol{x}) - \widehat{f}_{2_\tau}(\boldsymbol{x})\right) \ln \frac{\widehat{f}_{1_\tau}(\boldsymbol{x})}{\widehat{f}_{2_\tau}(\boldsymbol{x})} d\boldsymbol{x}, \tag{4.16}$$

where $T_p = \{\tau | \tau$ *is a tree with* p *nodes*$\}$*, is equivalent to a problem of finding a MWST for the complete graph* κ *with weights* $\lambda(i,j),\ i,j = 1,...,p$*, given by*

$$\lambda(i,j) \;\; = \;\; \lambda^{(1)}(i,j) + \lambda^{(2)}(i,j), \tag{4.17}$$

where $\lambda^{(c)}(i,j)$ *is defined in equation 4.15 for* $c = 1, 2$*.*

Proposition 4.5.2 ***(Log-ratio)*** *Let* $\widehat{f}_{1_{\tau_1}}$ *and* $\widehat{f}_{2_{\tau_2}}$ *be the densities of* p*-variate distributions* $N(\widehat{\boldsymbol{\mu}}_1, \widehat{\boldsymbol{K}}_{1_{\tau_1}}^{-1})$ *and* $N(\widehat{\boldsymbol{\mu}}_2, \widehat{\boldsymbol{K}}_{2_{\tau_2}}^{-1})$*, respectively. The problem of finding* (τ_1^*, τ_2^*) *such that*

$$(\tau_1^*, \tau_2^*) = \underset{\tau_1, \tau_2 \in T_p}{\operatorname{argmax}} \left\{ \sum_{l=1}^{n_1} \ln \frac{\widehat{f}_{1_{\tau_1}}(\boldsymbol{x}_l)}{\widehat{f}_{2_{\tau_2}}(\boldsymbol{x}_l)} + \sum_{l=n_1+1}^{n_1+n_2} \ln \frac{\widehat{f}_{2_{\tau_2}}(\boldsymbol{x}_l)}{\widehat{f}_{1_{\tau_1}}(\boldsymbol{x}_l)} \right\}, \tag{4.18}$$

where $T_p = \{\tau | \tau$ *is a tree with* p *nodes*$\}$*, is equivalent to the problem of finding two independent MWSTs,* τ_1^* *and* τ_2^**, for the complete graph* κ *with weights* $\lambda^{(c)}(i,j),\ i,j = 1,...,p$*, given by*

$$\begin{aligned} \lambda^{(c)}(i,j) \;\; = \;\; & \frac{(n_c - n_{3-c})}{2} \ln\left(1 - \hat{\rho}_{ij}^{(c)\,2}\right) - \frac{n_b\, \hat{\rho}_{ij}^{(c)\,2}}{2\left(1 - \hat{\rho}_{ij}^{(c)\,2}\right)} \left\{ \frac{w_{ii}^{(3-c)}}{w_{ii}^{(c)}} + \frac{w_{jj}^{(3-c)}}{w_{jj}^{(c)}} - \frac{2w_{ij}^{(3-c)}}{w_{ij}^{(c)}} \right. \\ & \left. + \frac{\left(\overline{x}_i^{(3-c)} - \overline{x}_i^{(c)}\right)^2}{w_{ii}^{(c)}} + \frac{\left(\overline{x}_j^{(3-c)} - \overline{x}_j^{(c)}\right)^2}{w_{jj}^{(c)}} - \frac{2\left(\overline{x}_i^{(3-c)} - \overline{x}_i^{(c)}\right)\left(\overline{x}_j^{(3-c)} - \overline{x}_j^{(c)}\right)}{w_{ij}^{(c)}} \right\}, \end{aligned} \tag{4.19}$$

where $\widehat{\rho}_{ij}^{(c)} = w_{ij}^{(c)} \left(w_{ii}^{(c)} w_{jj}^{(c)}\right)^{-1/2}$*,* $c = 1, 2$*.*

An immediate consequence is the following

Corollary 4.5.3 ***(Log-ratio_equal)*** *When $\tau_1 = \tau_2 = \tau$ in Proposition 4.5.2, the optimization problem is reduced to the problem of finding τ^* such that*

$$\tau^* = \underset{\tau \in T_p}{\operatorname{argmax}} \left\{ \sum_{l=1}^{n_1} \ln \frac{\widehat{f}_{1_\tau}(\boldsymbol{x}_l)}{\widehat{f}_{2_\tau}(\boldsymbol{x}_l)} + \sum_{l=n_1+1}^{n_1+n_2} \ln \frac{\widehat{f}_{2_\tau}(\boldsymbol{x}_l)}{\widehat{f}_{1_\tau}(\boldsymbol{x}_l)} \right\}, \tag{4.20}$$

where $T_p = \{\tau | \tau$ is a tree with p nodes$\}$. This problem is equivalent to a problem of finding a MWST for the complete graph κ with weights $\lambda(i,j)$, $i,j = 1,...,p$, given by

$$\lambda(i,j) \;\; = \;\; \lambda^{(1)}(i,j) + \lambda^{(2)}(i,j), \tag{4.21}$$

where $\lambda^{(c)}(i,j)$ is defined in equation 4.19 for $c = 1, 2$.

The proofs of Propositions 1 and 2, and Corollary 1, are given in Appendix B.

Remark 1 We note that when $n_1 = n_2 = n$: (i) (4.15) is proportional to (4.19), and then Tan method is equivalent to Log-ratio; and (ii) (4.17) is proportional to (4.21), and then method J-div is equivalent to Log-ratio_equal.

Remark 2 Also that when $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ is assumed, $\boldsymbol{W}$, as in (1.7), is used to compute the weights for each method instead of $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$, then: (i) C-L and Fried are equivalent; and (ii) Tan, Log-ratio, J-div and Log-ratio_equal are also equivalent.

Finally, we note that the measure optimized in (4.14) under the assumption $\tau_1 = \tau_2 = \tau$ is equal to the one in (4.16), see Proposition 4.5.4 below. This is not true in general, where the measures are different as well as the solution of the corresponding optimization problems, though their solutions agree in some cases.

Proposition 4.5.4 *Let $\widehat{f}_{c_\tau}$ denote the density of $N(\widehat{\boldsymbol{\mu}}_c, \widehat{\boldsymbol{K}}_{c_\tau}^{-1})$ and $\widetilde{f}_c$ the density of the empirical distribution $N(\overline{\boldsymbol{x}}^{(c)}, \boldsymbol{W}_c)$, $c = 1, 2$, then*

$$J(\widehat{f}_{1_\tau}, \widehat{f}_{2_\tau}) = \widehat{J}(\widehat{f}_{1_\tau}, \widehat{f}_{2_\tau}; \widetilde{f}_1, \widetilde{f}_2),$$

where $\widehat{J}(\widehat{f}_{1_\tau}, \widehat{f}_{2_\tau}; \widetilde{f}_1, \widetilde{f}_2) = \int \left(\widetilde{f}_1(\boldsymbol{x}) - \widetilde{f}_2(\boldsymbol{x})\right) \ln \frac{\widehat{f}_{1_\tau}(\boldsymbol{x})}{\widehat{f}_{2_\tau}(\boldsymbol{x})} d\boldsymbol{x}.$

The proof of this proposition is also given in Appendix B. We note that the following optimization problem without the restriction of $\tau_1 = \tau_2$ can be considered:

Find (τ_1^*, τ_2^*) such that

$$(\tau_1^*, \tau_2^*) = \underset{\tau_1, \tau_2 \in T_p}{\operatorname{argmax}} J(\widehat{f}_{1_{\tau_1}}, \widehat{f}_{2_{\tau_2}}), \tag{4.22}$$

where $T_p = \{\tau | \tau \text{ is a tree with } p \text{ nodes}\}$.

However, as we have mentioned before, in this case the optimization problem is not equivalent to a problem of finding a MWST. In this case, other algorithms for combinatorial optimization problems could be used, though this could be computationally expensive.

In the following section, we illustrate the empirical performance of the six methods described in this section using some real and simulated data. For the numerical experiment in Section 4.6.1 and the breast cancer data in Section 4.6.2, we restrict the study to the case $n_1 = n_2$. The example on HIV data in Section 4.6.2 considers different group samples sizes.

4.6 Numerical studies

4.6.1 Simulated data

Structures in the covariance matrix

We consider four types of structures for the concentration matrix, each type is associated with one of the following models: autoregressive of order 1, AR(1); moving average of order 1, MA(1); a set of variables equally correlated, ECM; a model associated with a random graph generated

as a power law network, RAND. The first three were used in Bickel and Levina (2004) when comparing LDA and DLDA. The fourth one, has been used in Danaher et al. (2014) in the context of estimation of concentration matrices.

i) AR(1). For this model, the covariance matrix $\mathbf{\Sigma}$ has entries

$$\sigma_{ij} = \rho^{|i-j|}, \;\; i,j = 1,...,p, \;\; 0 < |\rho| < 1. \tag{4.23}$$

Its associated dependence graph is a tree τ called *path*. For example, when $p = 4$, the graph is as given in Figure 4.1c), and the covariance and concentration matrices are

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}, \quad \boldsymbol{K} = \frac{1}{1-\rho^2} \begin{pmatrix} 1 & -\rho & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 \\ 0 & -\rho & 1+\rho^2 & -\rho \\ 0 & 0 & -\rho & 1 \end{pmatrix}. \tag{4.24}$$

ii) MA(1). In this case the covariance matrix $\mathbf{\Sigma}$ has entries

$$\sigma_{ij} = \rho^{|i-j|} I(|i-j| \leq 1), \;\; i,j = 1,...,p, \;\; 0 < |\rho| < 0.5. \tag{4.25}$$

The dependence graph corresponds to a complete graph $G = \kappa$, though the associated concentration matrix with no zeroes depends on a single parameter ρ. For example, when $p = 4$,

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & \rho & 0 & 0 \\ \rho & 1 & \rho & 0 \\ 0 & \rho & 1 & \rho \\ 0 & 0 & \rho & 1 \end{pmatrix}, \boldsymbol{K} = \frac{1}{a} \begin{pmatrix} -2\rho^2+1 & \rho^3-\rho & \rho^2 & -\rho^3 \\ \rho^3-\rho & -\rho^2+1 & -\rho & \rho^2 \\ \rho^2 & -\rho & -\rho^2+1 & \rho^3-\rho \\ -\rho^3 & \rho^2 & \rho^3-\rho & -2\rho^2+1 \end{pmatrix}, \tag{4.26}$$

where $a = \rho^4 - 3\rho^2 + 1$, and the corresponding dependence graph as given in Figure 4.1a).

iii) ECM. For this model where all p variables are equally correlated, the covariance and concentration matrices have the same kind of pattern:

$$\boldsymbol{\Sigma} = (1-\rho)\boldsymbol{I} + \rho \boldsymbol{z}\boldsymbol{z}^t = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}, \boldsymbol{K} = (c-d)\boldsymbol{I} + d\boldsymbol{z}\boldsymbol{z}^t = \begin{pmatrix} c & d & \cdots & d \\ d & c & \cdots & d \\ \vdots & \vdots & \ddots & \vdots \\ d & d & \cdots & c \end{pmatrix}, \quad (4.27)$$

where $\boldsymbol{z} = (1, ..., 1)$, $c = \dfrac{-(\rho(p-2)+1)}{(\rho-1)[\rho(p-1)+1]}$ and $d = \dfrac{\rho}{(\rho-1)[\rho(p-1)+1]}$. Its associated graph is also a complete graph κ, and both matrices depend on a single parameter ρ.

iv) RAND. In this case, the concentration matrix is random, where (i) the number of zeroes and their position are determined by a random graph generated as a power law network (PLN) with power parameter $\alpha = 2.3$ and (ii) the non zeroes are random values from a uniform distribution. More details are given in Appendix E. Two random graphs, G_1 and G_2, generated to be used in the simulation study are given in Figure 4.3.

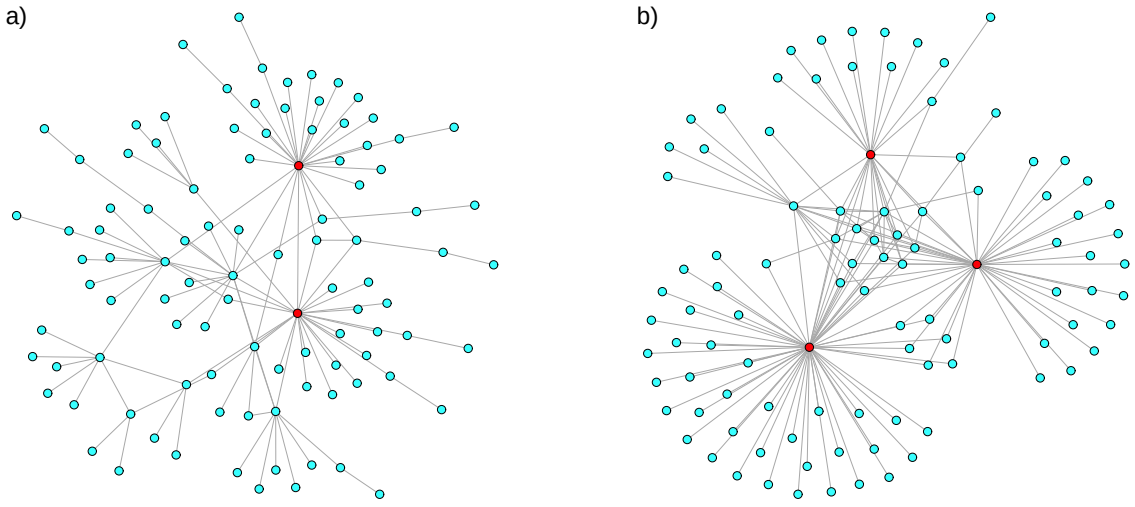


Figure 4.3: Random graphs. Two graphs generated randomly as two PLN with power parameter $\alpha = 2.3$ and $p = 100$. The nodes with a degree larger than 20 are coloured in red. a) graph G_1 with 114 edges and b) graph G_2 with 156 edges. These graphs will be used in the simulation study.

We consider both the cases, when specifying the distributions of the two populations from which the data is generated, of common and different covariance matrices. The details are the following.

Common covariance matrix

Four types of common covariance matrices and three different values for the mean vectors are considered and described below.

$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ is associated with one of the four models: AR(1) with $\rho = 0.9$; MA(1) with $\rho = 0.45$; ECM with $\rho = 0.9$; and RAND with associated graph G_1 given in Figure 4.3a).

$(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ take values: (i) $a(\mathbf{0}, \boldsymbol{v}_{max})$, (ii) $b(\mathbf{0}, \boldsymbol{v}_{min})$, and (iii) $(\mathbf{0}, \boldsymbol{u}_{rand})$; where $\boldsymbol{v}_{max}$ and $\boldsymbol{v}_{min}$ are the eigenvectors associated with the largest and smallest eigenvalue of $\boldsymbol{\Sigma}$, respectively, and $\boldsymbol{u}_{rand}$ is a vector where each element is a random number from a $U(0, t)$. For each of the three mean directions and each model, the values of a, b and t are specified to satisfy an asymptotic error rate value for LDA, given in (1.2), of around 10%, or equivalently, to satisfy $J(f_1, f_2) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \approx 6.57$. See Table D.1 in Appendix D for details.

The error rate of the optimal allocation rule (or equivalently, the asymptotic error rate of LDA) given in (1.2) with $\pi_1 = \pi_2$ is

$$P(e) = P(1|2) = P(2|1) = \Phi\Big(-\tfrac{1}{2}\sqrt{J(f_1, f_2)}\Big) = \Phi\Big(-\tfrac{1}{2}\sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}\Big)\,. \tag{4.28}$$

The asymptotic error rate of DLDA when $\pi_1 = \pi_2 = 1/2$, that is, for the allocation rule: $\boldsymbol{x}$ is assigned to Π_1 when $L(\boldsymbol{x}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{D}^{-1}) > 0$, and otherwise to Π_2, is given by

$$P(e) = \Phi\left(\frac{-A/2}{\sqrt{B}}\right) = \Phi\left(\frac{-(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \boldsymbol{D}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)/2}{\sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \boldsymbol{D}^{-1}\boldsymbol{\Sigma}\boldsymbol{D}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}}\right), \tag{4.29}$$

where $A = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \boldsymbol{D}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, $B = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \boldsymbol{D}^{-1}\boldsymbol{\Sigma}\boldsymbol{D}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, $\boldsymbol{D} = \text{diag}(\boldsymbol{\Sigma})$ and

$$L(\boldsymbol{x}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{D}^{-1}) = \boldsymbol{x}^t \boldsymbol{D}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \tfrac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^t \boldsymbol{D}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2). \tag{4.30}$$

The asymptotic error rates for LDA and DLDA for any $\boldsymbol{\Sigma}$, in particular for the four types of $\boldsymbol{\Sigma}$ used in the simulation study, are equal when $\pi_1 = \pi_2$, $n_1 = n_2 = n \to \infty$ and the mean values are $(\boldsymbol{0}, a\boldsymbol{v}_{max})$ or $(\boldsymbol{0}, b\boldsymbol{v}_{min})$. To verify this, we write $\boldsymbol{\Sigma}$ and $\boldsymbol{K}$ as

$$\boldsymbol{\Sigma} = \sum_{i=1}^{p} l_i \boldsymbol{v}_i \boldsymbol{v}_i^t \quad \text{and} \quad \boldsymbol{K} = \sum_{i=1}^{p} l_i^{-1} \boldsymbol{v}_i \boldsymbol{v}_i^t,$$

where $\{\boldsymbol{v}_1, ..., \boldsymbol{v}_p\}$ is the set of orthonormal eigenvectors of $\boldsymbol{\Sigma}$ and l_i is the eigenvalue associated with $\boldsymbol{v}_i$, $i = 1, ..., p$. Then, if $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = (\boldsymbol{0}, a\boldsymbol{v}_j)$ with $\boldsymbol{v}_j$ an eigenvector of $\boldsymbol{\Sigma}$, $a > 0$, $\pi_1 = \pi_2 = 1/2$, and $\boldsymbol{\Sigma}$ with all diagonal elements equal to one, then the probability of misclassification in (4.28) is given by

$$P(e) = \Phi\left(-\tfrac{1}{2}\sqrt{J(f_1, f_2)}\right) = \Phi\left(-\tfrac{1}{2}\sqrt{a\boldsymbol{v}_j^t \left(\textstyle\sum_{i=1}^{p} l_i^{-1} \boldsymbol{v}_i \boldsymbol{v}_i^t\right) a\boldsymbol{v}_j}\right) = \Phi\left(\frac{-a/2}{\sqrt{l_j}}\right),$$

and (4.29) is

$$P(e) = \Phi\left(\frac{-A/2}{\sqrt{B}}\right) = \Phi\left(\frac{-a^2/2}{\sqrt{a\boldsymbol{v}_j^t \left(\sum_{i=1}^{p} l_i \boldsymbol{v}_i \boldsymbol{v}_i^t\right) a\boldsymbol{v}_j}}\right) = \Phi\left(\frac{-a/2}{\sqrt{l_j}}\right).$$

For the case where the mean vector is $(\boldsymbol{0}, \boldsymbol{u}_{rand})$, the asymptotic error rates for LDA and DLDA are not equal anymore. For example, the asymptotic error rates, for the AR(1) and ECM models, are larger than 40% for DLDA and around 10% for LDA.

Different covariance matrices

In this case $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, and in all four cases $\boldsymbol{\Sigma}_1$ is the same random matrix associated with a RAND model with graph G_2 given in Figure 4.3b). $\boldsymbol{\Sigma}_2$ in each of the four cases is associated with one of the following models: AR(1) with $\rho = 0.3$; MA(1) with $\rho = 0.275$; ECM with $\rho = 0.2$; and RAND with associated graph G_1 given in Figure 4.3a).

For this case, we use only one direction for the mean values, $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = (\mathbf{0}, \boldsymbol{u}_{rand})$, where each element of $\boldsymbol{u}_{rand}$ is a random number from $U(0, t)$. The t value is such that the asymptotic global error rate for QDA is around 5%. For each model, the value of t was specified using a sample of $40,000$ observations in each population and the allocation rule in (1.1) to classify them. These asymptotic error rates are referred as optimal. See Table D.2 in Appendix D for details.

Estimation of the error rates

Estimated error rates were obtained using a Monte Carlo study, where the error rates were calculated as the average over 400 proportions of misclassified observations. Each proportion was obtained by the following procedure.

1. Generate n_1 and n_2 independent training observations from p-variate distributions $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, respectively, with $p = 100$ and $n_1 = n_2 = n \in \{10, 50, 100, 200, 1000, 50000, 200000\}$.

2. For each method, the tree structure and the parameters are estimated. Both the homogeneous and the heterogeneous case are considered, that is, with and without the assumption $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, respectively.

3. The plug-in allocation rule is then used to classify $n_{1v} = 100$ and $n_{2v} = 100$ new independent observations generated from $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, respectively.

Technical note

All the numerical results were computed using C++; Kruskal's algorithm was implemented to find the MWST for each of the six methods.

Results

Common covariance matrix

Estimated error rates calculated considering concentration matrices associated with models AR(1) and ECM are presented in Figures 4.4 and 4.5. Corresponding values for MA(1) and RAND models, presented in Appendix E, show a similar behaviour to the one of the AR(1) model. In Figures 4.4 and 4.5, the estimated error rates for each method are presented for both the heterogeneous and homogeneous case. The population or asymptotic error rate for LDA, $\Phi\left(-\frac{1}{2}\sqrt{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)}\right)$, is also included as a reference with the name *Optimal*. The AR(1) model has a path-structured graph, and the estimated error rates when assuming its tree structure are also included in Figure 4.4 with the name *TrueTree*. We observe the following.

1. The performance of tree allocation rules for the four models depends on the mean vectors' direction.

   (a) When $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = (\mathbf{0}, a\boldsymbol{v}_{max})$, tree rules are worse than DQDA or DLDA. However, they are better than LDA or QDA, except for ECM model.

   (b) When $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = (\mathbf{0}, b\boldsymbol{v}_{min})$, tree rules are better than DLDA in the homogeneous case, except for ECM when $n \geq 50,000$. In the heterogeneous case: (i) C-L, Fried and J-div are better than DQDA except for ECM when $n \geq 50,000$, though in this case Fried is similar to DQDA; and (ii) Tan is similar to, or worse than, DQDA for AR(1) and MA(1) models, and also for ECM and RAND when $n \geq 50,000$. In general, DQDA and DLDA have a bad performance when $n \leq 1,000$.

(c) When $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = (\mathbf{0}, \boldsymbol{u}_{rand})$, tree methods are better than DQDA and DLDA for AR(1) and ECM, and are similar for MA(1) and RAND.

2. In general, when $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = (\mathbf{0}, b\boldsymbol{v}_{min})$, DLDA and DQDA converge to their asymptotic error rate very slowly. For example, for AR(1) model, they have high estimated error rates even when group training sample size is $n_1 = n_2 = 50,000$, whereas C-L and Fried are close to their asymptotic error rates for $n_1 = n_2 = 200$.

3. The performance of DLDA with $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = (\mathbf{0}, b\boldsymbol{v}_{min})$, where its asymptotic error rate is the same to the one for LDA, shows that for finite sample sizes, as on this case for $n_1 = n_2 \approx 50,000$, it is not always true that a plug-in allocation rule with less parameters has a better performance than LDA. In this case, methods based on trees are a good alternative to DLDA.

4. Tan and J-div methods in none of the cases are the best methods.

5. C-L and Fried have a good performance in almost all cases. Their performance is better than, or similar to, the one of the other methods, except for ECM.

6. TrueTree method, when $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = (\mathbf{0}, a\boldsymbol{v}_{max})$, shows an example where the use of a diagonal matrix, or another matrix associated with trees, see Tan and J-div, gives a better classification performance than the use of the true tree structure for small sample sizes, in this example for $n < 200$.

We also observe the following which was expected since the design of the numerical experiment.

1. The methods in the homogeneous case have a better performance, or in the worst cases similar, when compared with their corresponding heterogeneous case counterparts.

2. The performance of QDA and LDA does not change when modifying the model used for $\boldsymbol{\Sigma}$. It does not change either for the different directions of the mean vectors. It only changes

when modifying the group training sample size $n_1 = n_2 = n$. We note that the data were generated to satisfy $J(f_1, f_2) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \approx 6.57$.

3. For the ECM model, C-L and Fried methods, which only depend on the correlation parameter ρ, can select any of the p^{p-2} possible trees since all of them have equal total weight. The choice of the tree depends on numerical differences when estimating the parameters. On the other hand, Tan and J-div methods, which also depend on the mean differences, select specific trees, but they also have a poor performance.

4. For the AR(1) model using its true tree structure, TrueTree method illustrates the two sources of estimation error: structure and parameter estimation. In this case, the estimation error is due only to the parameter estimation. We observe that the estimated error rates are very close to those corresponding to C-L and Fried, though these two consider both parameter and structure estimation.

Different covariance matrices

For each case, in population one we use a random matrix, and in two a specific matrix, that is: (RAND, AR(1)), (RAND, MA(1)), (RAND, ECM) and (RAND, RAND). Numerical results corresponding to models (RAND, ECM) and (RAND, RAND), for the heterogeneous and the homogeneous case, are shown in Figures 4.6 and 4.7. The population or asymptotic error rates for QDA are also included as *Optimal*. In the case of different concentration matrices, $P(1|2) \neq P(2|1)$, and so the estimated error rates for each group are also included in the figures. Corresponding values for (RAND, AR(1)) and (RAND, MA(1)) presented in Appendix E, show a similar behaviour to the one for (RAND, RAND). We observe the following.

1. For (RAND, AR(1)), (RAND, MA(1)) and (RAND, RAND), tree based methods have a better performance in the heterogeneous case than in the homogeneous, where their error rates are larger than 20%. These are also better than DQDA and QDA for $n_1 = n_2 > 10$.

2. For (RAND, ECM), tree methods in the heterogeneous case are very similar to DQDA, and in the homogeneous case are very similar to DLDA. All tree methods, as well as the diagonal, have a good performance for group 1 and a bad one for group 2. In spite of this, tree and diagonal methods are a good alternative to QDA and LDA for small group training sample sizes, in this example, for $n \leq 120$.

3. In the four cases, C-L and Fried have similar estimated global error rates, though slightly different group error rates for $n \in \{60, 120, 200\}$.

4. In general, global error rates are nearly the same for tree methods.

4.6.2 Real Data

Breast cancer data

We consider the data described in Section 3.4 and study the performance of the plug-in allocation rules considering subsets of p variables, $p \in \{15, 50, 100, 250, 500, 1000\}$. For $p = 15$, the variables, the ones described in 3.4.2 for Case 2, were selected using stepwise for least squares linear regression, in such a way that they gave low observed proportion of misclassified observations. These 15 variables are contained in all the subsets with $p > 15$. We note that the set of 15 variables were selected in a way that could favour LDA. As we restrict the numerical study to the case where the training sample is equal for both populations, we randomly selected 58 of the 192 observations from the control group.

The estimated error rates are computed using the repeated holdout method (Kim, 2009). For each group, approximately three fourths of the observations (44 cases and 44 controls) were randomly selected for training and one fourth for testing. This was repeated 200 times. The classification rates were then estimated by the average of the 200 percentages of misclassified observations. The estimated group and global error rates, for each method, are shown in Figure

4.8, in which the following is observed.

1. The performance of all methods is better, or similar, in the homogeneous case than in the heterogeneous. For example, global error rates are larger than 20% in the heterogeneous case for $p = 100$, whereas those in the homogeneous case are less than 20%.

2. Tree methods in the heterogeneous and homogeneous case are better than, or similar to, DQDA and DLDA, respectively, when comparing the global error rates.

3. When considering the global estimated error rates and the homogeneous case with $p \geq 50$, C-L is better than, or similar to, the other methods.

4. The performance of tree methods is different for each group.

   a) In the heterogeneous case, C-L and Fried have a better performance in the group of cases than in the group of controls, whereas Tan and J-div have a similar one in both groups.

   b) In the homogeneous case, Tan method has a different performance for each group, while C-L has almost the same performance.

Ilustrative example in the context of repeated measures

We consider a data set with a biomarker indicative for HIV therapy resistance. The goal is to classify patients either as resistant or non-resistant to HIV therapy based on longitudinal viral load profiles. The data were presented and studied in May and DeGruttola (2007) in the context of nonparametric tests. In Kohlmann et al. (2009), a subset of 85 out 356 patients is analysed in the context of classification.

The viral load was measured by the amount of HIV RNA at six occasions: at baseline, after 2, 4 8, 16 and 24 weeks. The original data set consists of 356 patients, though the data are complete for only 85 patients, 59 non-resistant and 26 resistant. We consider the set of 85 patients and

log10-transformed values of HIV RNA. The box plots of the data are presented in Figure 4.9, the scatter plots and correlations between the RNA level at two periods of time are also given.

The estimated error rates are computed using the repeated holdout method (Kim, 2009). For each group, approximately two thirds of the observations (40 non-resistant and 18 resistant) were randomly selected for training and one third for testing. This was repeated 200 times. The classification rates were then estimated by the average of the 200 percentages of misclassified observations. The estimated group and global error rates, for each method, are shown in Figure 4.10. The error rates estimated assuming the structure of a path are also included with the name *Path* in both the heterogeneous and homogeneous cases. The following is observed.

1. Global estimated error rates are similar for all methods and in both the heterogeneous and homogeneous cases. DQDA has the lowest global error rate, followed by J-div and Log-ratio_equal. These three methods consider the same structure for both concentration matrices though arbitrary values for the parameters.

2. The performance of the methods is very similar in the non-resistant group, however, Log-ratio method has the worst performance.

3. DQDA, J-div and Log-ratio_equal methods in the heterogeneous case and DLDA and Tan in the homogeneous have a better performance in the resistant group compared with the other methods. The rest of the methods have error rates larger than 50%.

4. The common assumption of a path in the context of repeated measures does not help for classification in this case. The performance of the method obtained assuming a path is very similar than the one of QDA or LDA.

5. In general, five methods have the best performance: DQDA, J-div and Log-ratio_equal in the heterogeneous case and DLDA and Tan in the homogeneous. These methods do not

estimate the real structure associated with the matrices, showing another example where a simple structure, as the one of a diagonal matrix, gives a better classification performance.

4.7 Conclusions

We have described six different methods for estimating the tree structure of graphical Gaussian models in the context of discriminant analysis. We have also compared their performance considering two populations and equal group sample size. These methods consider a tree structure on the concentration matrices and take advantage of (i) the factorization of the distribution in terms of bivariate and univariate densities, and (ii) the solution provided by an efficient algorithm to find a MWST.

The MLE of the concentration matrices has an analytical expression for decomposable models other than trees, however, the assumption of a tree structure makes it possible to find an exact solution to the optimization problem of finding the structure. Without this assumption, finding a solution could be computationally expensive for a large number of variables.

The numerical experiments show that tree methods are a good alternative to the usual QDA, LDA, DQDA and DLDA.

DQDA and DLDA require a group sample size of at least two observations for the existence of the MLEs, this makes them useful in high dimensional settings, however, as it was shown in Figures 4.4b) and 4.5b), they can converge very slowly to their asymptotic error rate or, as it was shown in Figures 4.4c) and 4.5c), they have a poor performance when they have a high asymptotic error rate. Tree methods are an alternative in all these cases and only need a group sample size of three observations for the existence of a solution.

The results also show that among the methods based on trees there is no single one that outperforms the others. Any of the tree methods can be used without an expensive computational cost in a high dimensional setting. A way to select a method based on trees, among all of them, in a practical application, is using cross-validation to estimate their error rates. These rates together with the corresponding to diagonal discriminant analysis can be compared and then used to select the best one.

gRapHD R Package can be used for the estimation of the two trees used in C-L method. For the other tree methods, *igraph* R package (Csardi and Nepusz, 2006) includes an efficient algorithm to find a MWST, which can be used using the specific weights given in this chapter for each method. We have implemented the six methods in C++ using Kruskal's algorithm.

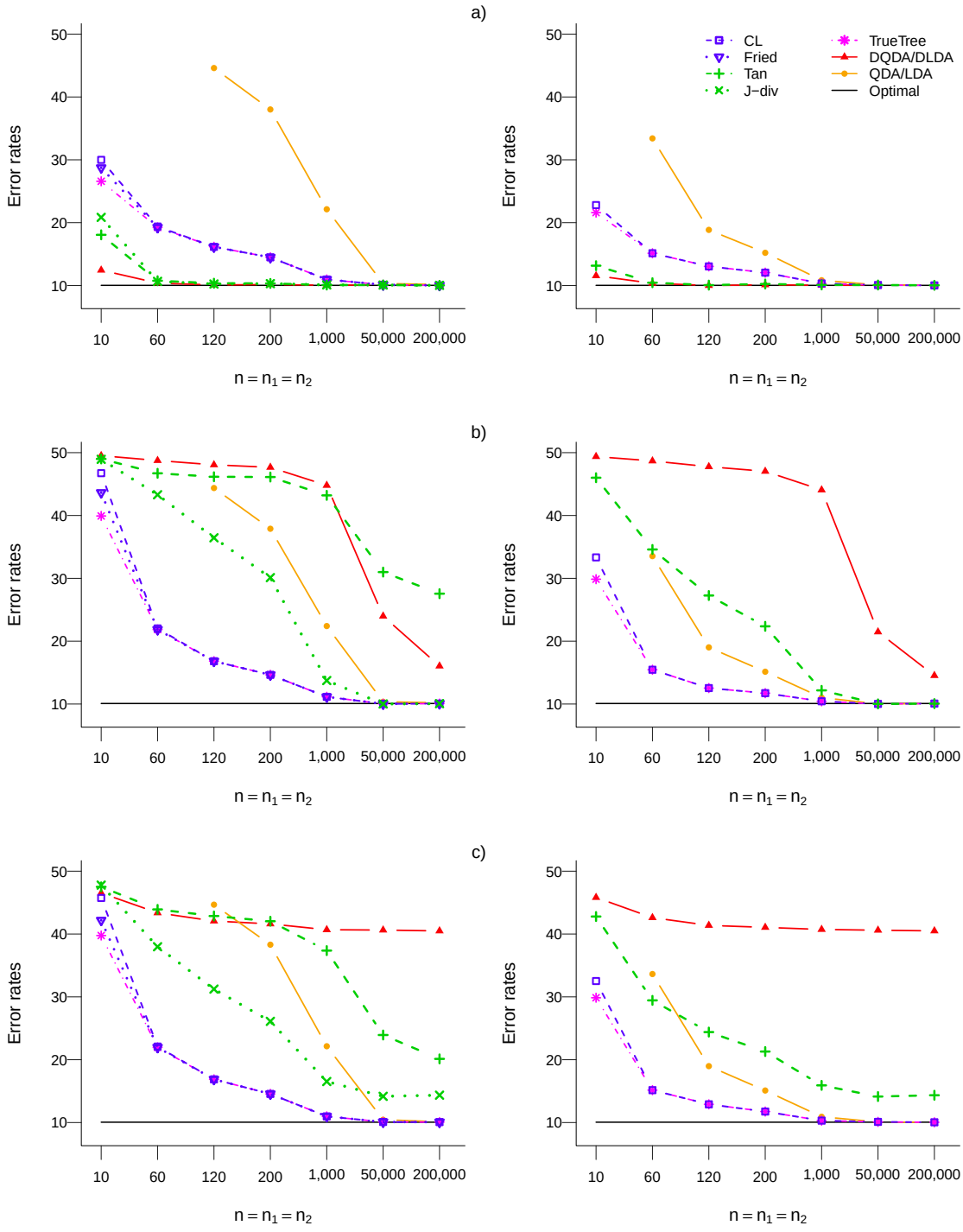


Figure 4.4: AR(1). Estimated error rates with $p = 100$ variables and $n_1 = n_2 = n$ training samples on each group. On the left the heterogeneous cases and on the right the homogeneous.
$\boldsymbol{\Sigma}$ with entries $\sigma_{ij} = 0.9^{|i-j|}$, $i, j = 1, ..., p$, and
a) $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = (\mathbf{0}, a\boldsymbol{v}_{max})$, b) $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = (\mathbf{0}, b\boldsymbol{v}_{min})$, and c) $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = (\mathbf{0}, \boldsymbol{u}_{rand})$,
where $\boldsymbol{v}_{max}$ is the largest and $\boldsymbol{v}_{min}$ the smallest eigenvalue of $\boldsymbol{\Sigma}$, and $\boldsymbol{u}_{rand}$ is a vector with random numbers from $U(0, t)$. a, b and t are constants such that the asymptotic error rate value for LDA is around 10%

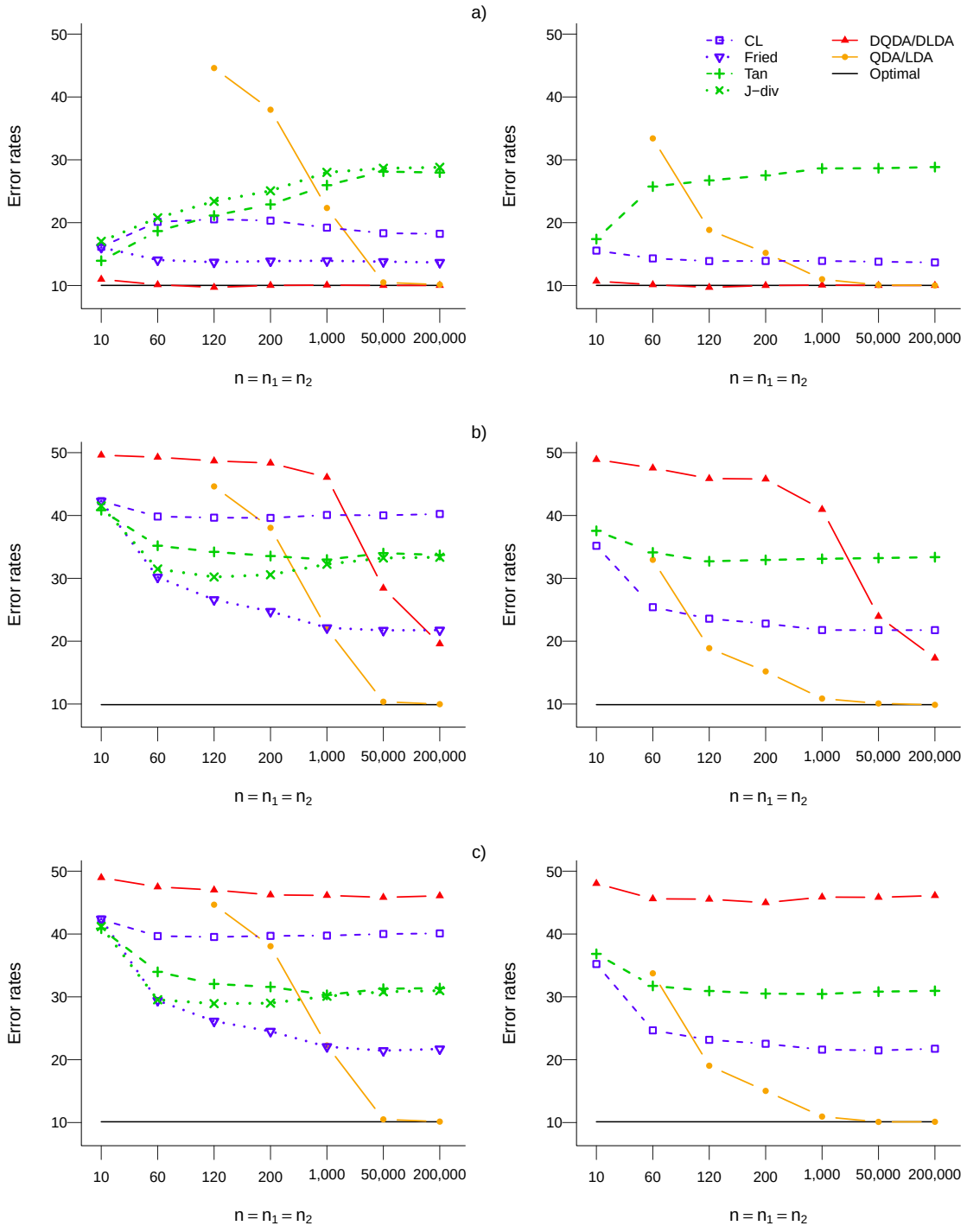


Figure 4.5: ECM. Estimated error rates with $p = 100$ variables and $n_1 = n_2 = n$ training samples on each group. On the left the heterogeneous cases and on the right the homogeneous.
$\boldsymbol{\Sigma} = (1 - 0.9)\boldsymbol{I} + 0.9\boldsymbol{z}\boldsymbol{z}^t$, where $\boldsymbol{z} = (1, ..., 1)$, and
a) $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = (\boldsymbol{0}, a\boldsymbol{v}_{max})$, b) $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = (\boldsymbol{0}, b\boldsymbol{v}_{min})$, and c) $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = (\boldsymbol{0}, \boldsymbol{u}_{rand})$,
where $\boldsymbol{v}_{max}$ is the largest and $\boldsymbol{v}_{min}$ the smallest eigenvalue of $\boldsymbol{\Sigma}$, and $\boldsymbol{u}_{rand}$ is a vector with random numbers from $U(0, t)$. a, b and t are constants such that the asymptotic error rate value for LDA is around 10%

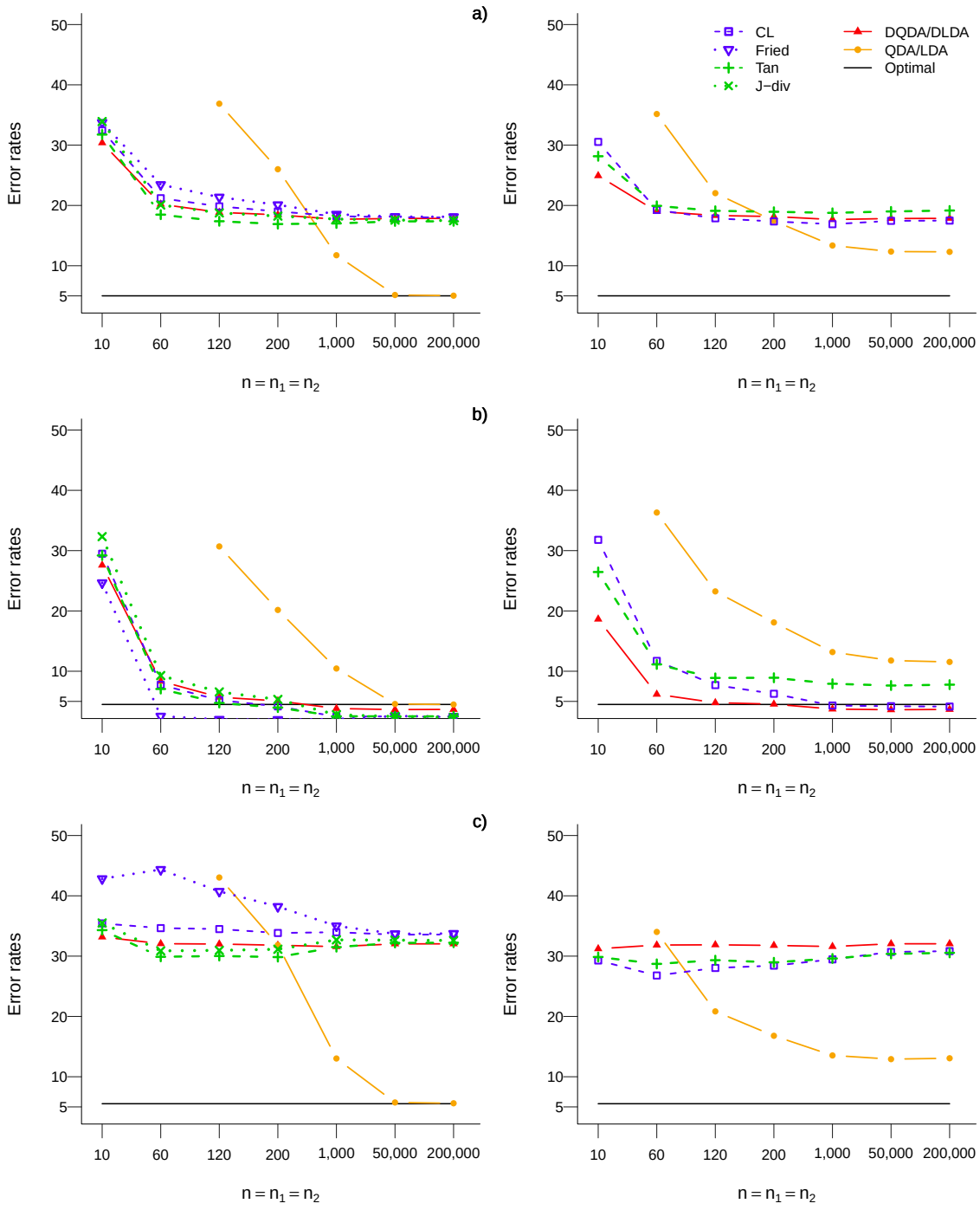


Figure 4.6: (RAND, ECM). Estimated error rates with $p = 100$ variables and $n_1 = n_2 = n$ training samples. On the left the heterogeneous cases and on the right the homogeneous. a) Global, b) Group 1 and c) Group 2.

$\mathbf{\Sigma}_1$ is associated with a RAND model with graph G_2 given in Figure 4.3b), $\mathbf{\Sigma}_2$ has entries $\sigma_{ij}^{(2)} = 0.2I(i \neq j) + 1I(i = j)$, $i, j = 1, ..., p$, and $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = (\mathbf{0}, \boldsymbol{u}_{rand})$, where $\boldsymbol{u}_{rand}$ is a vector with random numbers from $U(0, t)$ and t is such that the asymptotic error rate value for QDA is around 5%

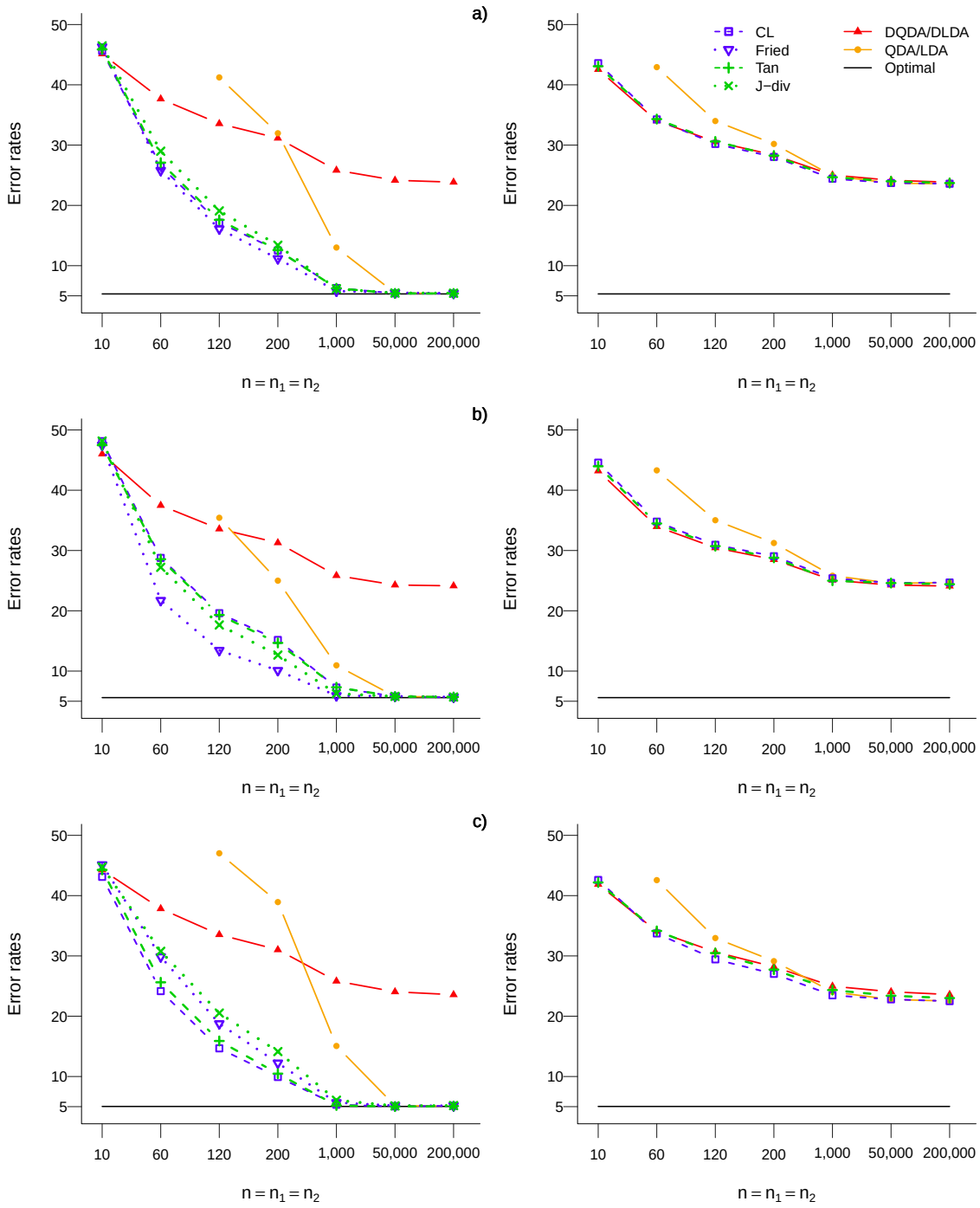


Figure 4.7: (RAND, RAND). Estimated error rates with $p = 100$ variables and $n_1 = n_2 = n$ training samples. On the left the heterogeneous cases and on the right the homogeneous. a) Global, b) Group 1 and c) Group 2.
$\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are associated with a RAND model with graph G_2 and G_1 given in Figure 4.3b) and 4.3a), respectively, and $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = (\mathbf{0}, \boldsymbol{u}_{rand})$, where $\boldsymbol{u}_{rand}$ is a vector with random numbers from $U(0, t)$ and t is such that the asymptotic error rate value for QDA is around 5%

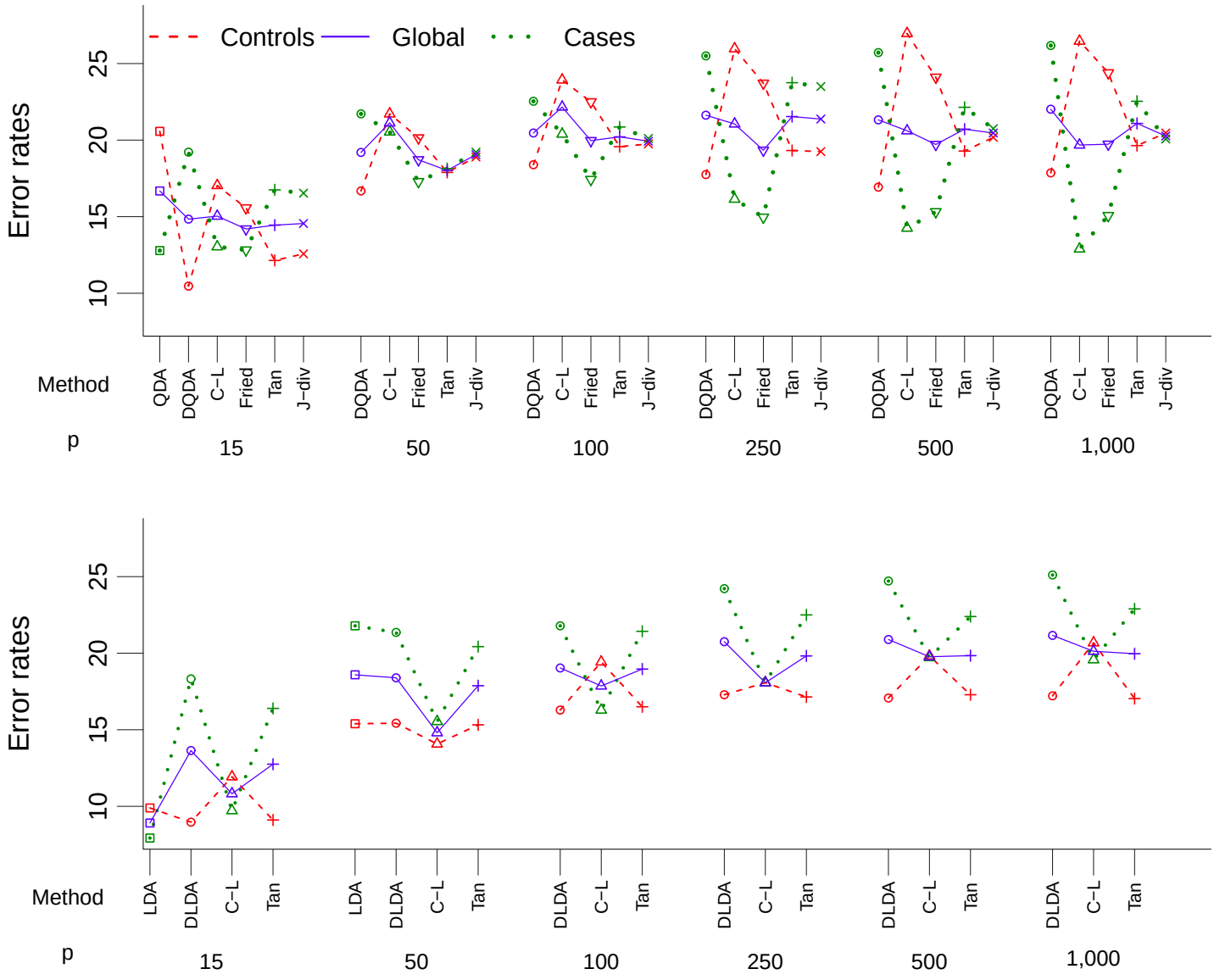


Figure 4.8: Estimated error rates computed using the repeated holdout method with 200 random samples for Breast cancer data. On the top the heterogeneous cases and on the bottom the homogeneous. The original data set has 250 observations, 58 cases and 192 controls, but we randomly selected 58 controls to satisfy $n_1 = n_2$. Training: $n_{cases} = n_{controls} = 44$. Testing: $n_{cases} = n_{controls} = 18$.

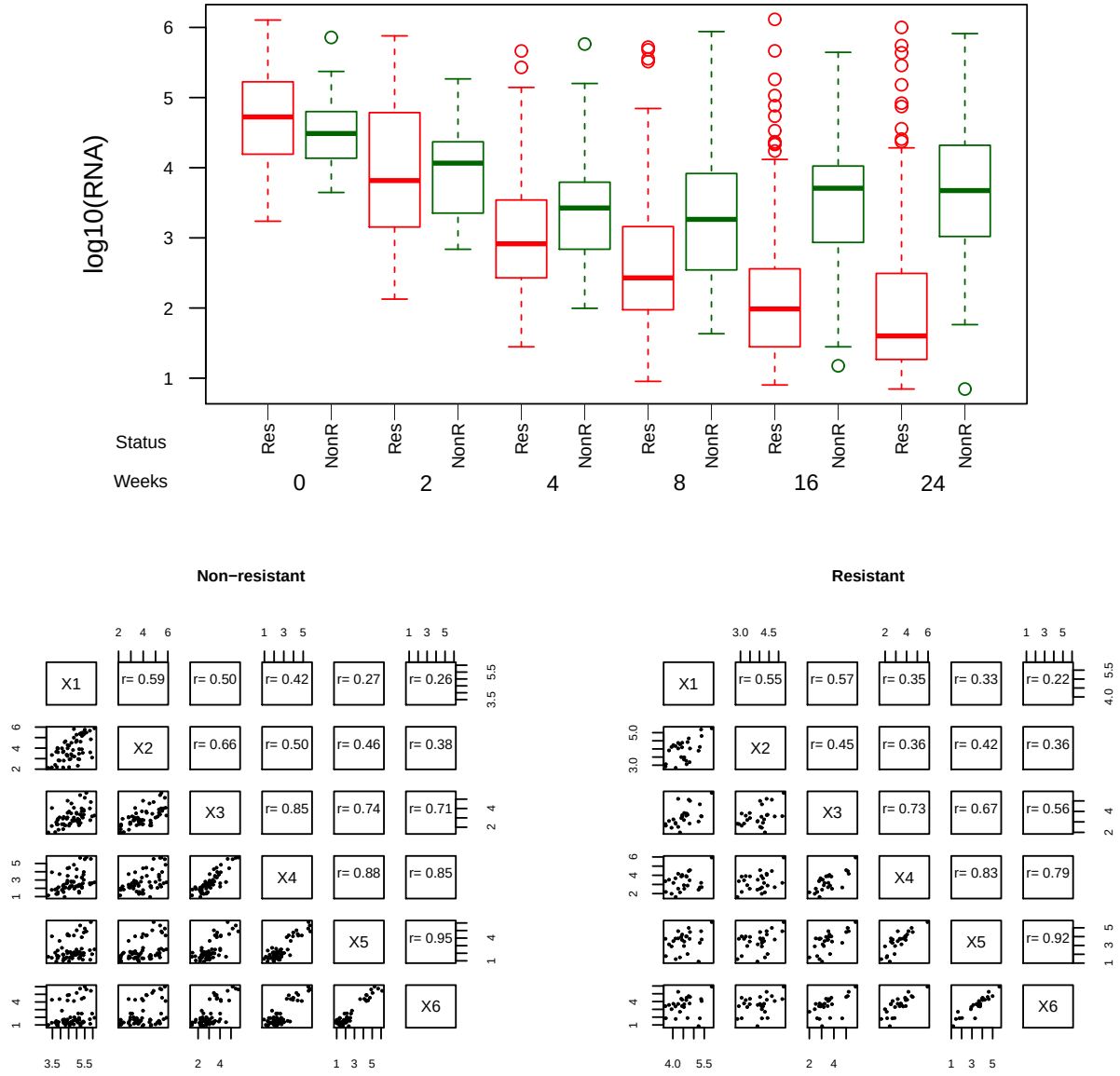


Figure 4.9: Box and scatter plots for Log10 transformed amount of HIV RNA at six occasions. At baseline, after 2, 4 8, 16 and 24 weeks (X1, X2, X3, X4, X5 and X6). Data set has 85 patients, 59 non-resistant and 26 resistant.

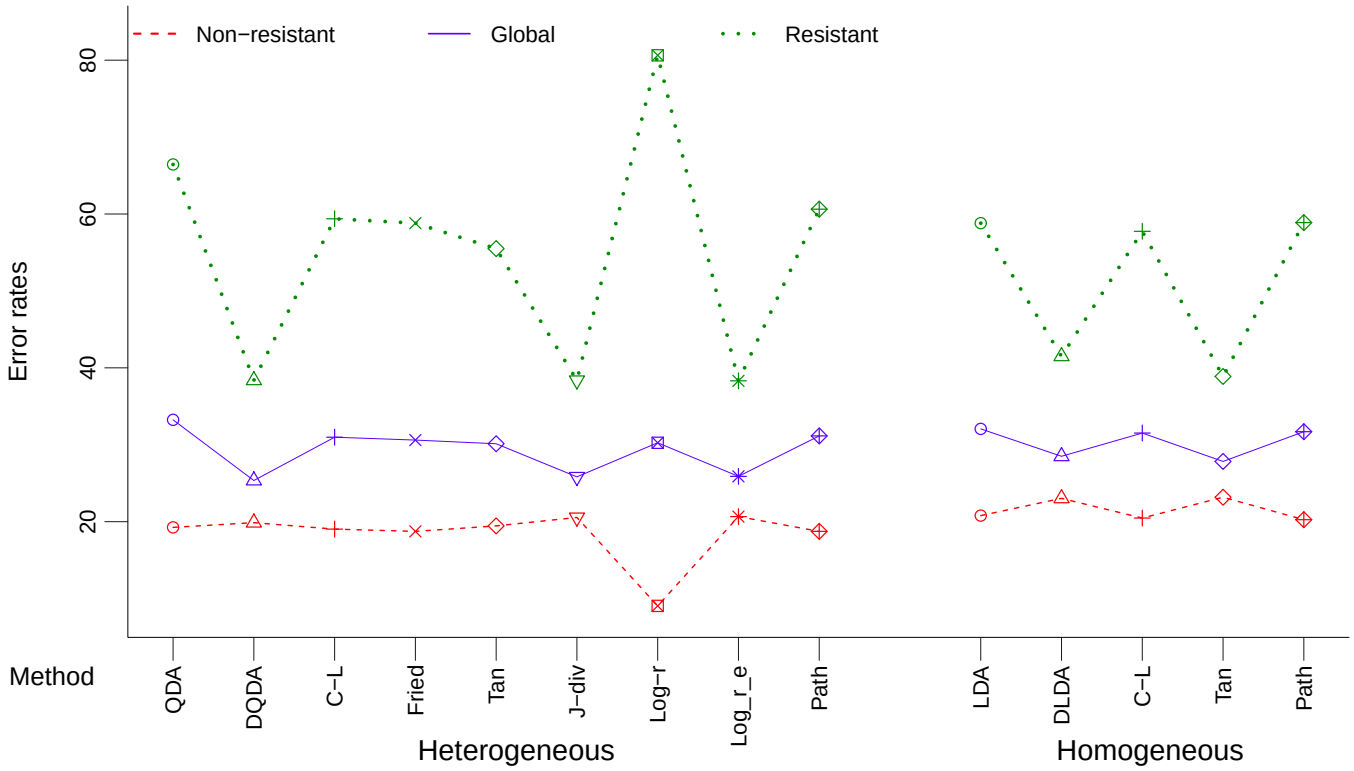


Figure 4.10: Estimated error rates computed using the repeated holdout method with 200 random samples for HIV data. $n = 85$ patients, $n_1 = 59$ non-resistant and $n_2 = 26$ resistant. Training: $n_1 = 40$ and $n_2 = 18$. Testing: $n_1 = 19$ and $n_2 = 8$.

5. Conclusions and Future Work

5.1 Conclusions

We have considered the use of patterned matrices in discriminant analysis for two Gaussian populations. In a first stage, we considered linear restrictions on the concentration matrices, as those introduced by Anderson (1970). Additionally, we considered the case where some corresponding parameters between the two matrices are equal.

When considering these restrictions, two estimation problems emerge: parameter and structure estimation. For parameter estimation, we considered ML estimation and adapted the iterative partial maximization (IPM) algorithm (Jensen et al., 1991) to obtain estimates.

The IPM algorithm was implemented in C++ and used to obtain the numerical results presented in Chapter 3. In order to use this algorithm, the structure of the concentration matrices must be known in advance. However, the estimation of the structure when considering all the restrictions is complex and requires the development of efficient algorithms. For this reason, we restricted the study to graphical tree models.

Considering Gaussian graphical tree models, we studied six methods for estimating the structure in the context of discriminant analysis. They all use ML estimation for the parameters, but use a different function to be optimized for the estimation of the tree structure. Four of these methods optimize a measure of divergence between the two Gaussian populations.

Three of these methods had been introduced in the literature, and based on these, three others were introduced in this thesis. For these three methods, we proved the equivalence of their corresponding optimization problems with one of finding a MWST. This equivalence makes the structure estimation solvable in an efficient way even if the number of variables is large.

We compared the performance of these methods in the context of discriminant analysis for equal group sample size, using real and simulated data. Diagonal discriminant analysis was considered as a benchmark, as well as quadratic and linear discriminant analysis whenever the sample size was sufficient. We observed the following.

Tree based methods were a good alternative to the usual QDA, LDA, DQDA and DLDA; especially in some of the examples presented where: (i) QDA and LDA could not be used or had a poor performance because of the training sample size, (ii) DQDA and DLDA converged very slowly to their asymptotic error rate, and (iii) DQDA and DLDA had high asymptotic error rates. The results also showed that among the methods based on trees there was no single one that outperformed the others. Even though none of the methods based on the tree models outperformed the benchmarks in all data sets, we conclude that any of the six tree methods is a simple and computationally inexpensive alternative to well established discriminant methods in high dimensional settings, where sample size is similar to, or smaller than, the number of variables.

We note the following aspects about the methods based on trees.

- They take advantage of the tree structure, specifically of (i) the existence of analytical expressions for the MLEs, and (ii) the solution provided by an efficient algorithm to find a MWST.
- The assumption of a tree structure makes it possible to find an exact and inexpensive solution

to the optimization problem of finding the structure. Without this assumption, finding a solution could be computationally expensive for a large number of variables.

- C-L and Fried methods try to estimate the zero structure associated with the true graphical models. The log-likelihood is optimized in these methods, assuming two independent and arbitrary trees, or the same tree in both populations. On the other hand, the other four methods do not necessarily focus on the zero structure. In these methods, a measure of divergence between the two distributions is optimized. This means that we cannot use the estimates of the concentration matrices to interpret conditional independences when using these four methods. In general, all the six methods were developed for the problem of classifying new observations, though Chow and Liu's idea is also used for approximating the distribution of a single population.

- In any of the six methods, we can modify the tree structure estimation such that we do not consider the space of p^{p-2} trees. We can use the subset of trees which do not have some edges; these removed edges could be selected in advance when some conditional independences are known, though the last only makes sense in C-L and Fried methods.

5.2 Future Work

- We can examine the use of decomposable models in discriminant analysis. The MLE of the concentration matrix has an analytical expression for these models. However, the structure estimation is not solvable using algorithms for the MWST problem.

  A forward selection procedure could be developed and implemented for this purpose using any of the measures associated with the six methods.

  The main problem in a forward selection method is the identification of the edges that can

be added such that the new graph remains decomposable, and some weights need to be updated at each iteration. Two algorithms to identify the edges are given in Thomas and Green (2009) and Deshpande et al. (2001). In the latter, Chow and Liu's idea has been considered for decomposable models. Abreu et al. (2010) has implemented a forward algorithm for single populations in the gRapHD R Package.

As starting point, the gRapHD R Package can be used for the structure estimation for each matrix when considering C-L method with decomposable models. For the other 5 methods, specific algorithms need to be implemented.

We note that when searching for a decomposable model, the maximal clique size should be considered since the sample size needed for the existence of the MLE depends on this value.

- Considering tree models, other measures of divergence between two populations could be optimized for structure estimation. However, their associated optimization problem could not necessarily be equivalent to a MWST problem. Therefore, finding a solution would require the use of other existing algorithms of combinatorial optimization or the development of specific ones.

  For example, when using the Binomial deviance loss function (see Hastie et al., 2009, p. 346), the optimization problem is not equivalent to one of finding the MWST. We have used this function and a forward selection procedure with the breast cancer data analysed in Chapter 4. We considered sets of variables of size $p \in \{15, 50, 100\}$ to estimate the error rates. We observed that the estimated error rates of the associated rule were better than those corresponding to any of the six methods, and similar to those for LDA and QDA. These results show that this function could be a good alternative, though for a high number

of variables, finding a solution is time-consuming and could be computationally unfeasible.

- In general, a penalized log-likelihood approach could be used to include, besides the zero structure, equalities of elements within a concentration matrix or between elements of the two matrices. However, finding a solution requires the use of algorithms for convex optimization (Boyd and Vandenberghe, 2004), for example, those implemented in the package cvx in Matlab; or the development of specific algorithms.

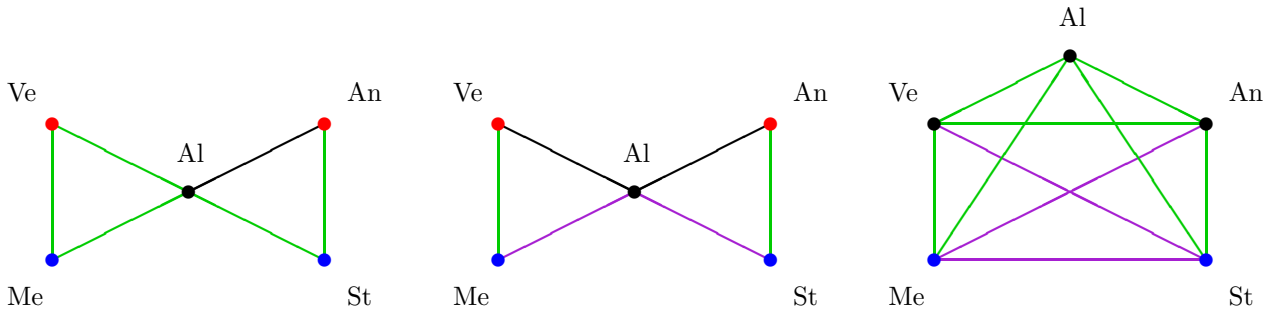
We did an exploratory experiment with a small data set of five variables and 88 observations. We considered the problem of finding the zero structure and equalities of the elements of the concentration matrix for one population. The observations correspond to examination marks of 88 students in five subjects Algebra, Analysis, Mechanics, Statistics and Vectors (Mardia et al., 1979). With this data set, Hϕjsgaard and Lauritzen (2008) adjusted the RCON model represented by the graph in Figure 5.1a); Gehrmann (2011) adjusted the one represented by the graph in Figure 5.1b); and we adjusted the one represented by the graph in Figure 5.1c) using the package cvx in Matlab to solve the optimization problem in (5.1).

$$\max_{\boldsymbol{K}>0} \left\{ (\log|\boldsymbol{K}| - \operatorname{tr}(\boldsymbol{S}\boldsymbol{K})) - \lambda_1\|\boldsymbol{K}\|_1 - \lambda_2 \sum_{i\neq j} |k_{ii} - k_{jj}| - \lambda_3 \sum_{\substack{i\neq j, k\neq l,\\ (i,j)\neq(k,l)}} |k_{ij} - k_{kl}| \right\}. \quad (5.1)$$

The solution for this example was easily found for specific values given to the tuning parameters λ_1, λ_2 and λ_3. However, when we tried with $p = 20$ variables, we realized that the development of a specific algorithm is needed. This is because cross-validation is used twice, once to determine the values of the tuning parameters and once for the estimation of the error rates. Danaher et al. (2014) and Hoefling (2010) may be useful for the development of an algorithm for structure estimation using a penalized log-likelihood approach. Recently,

Gao and Massam (2014) have considered a penalize approach to look for the equalities in RCON and RCOR models.

In any case, it is necessary to carry out numerical and theoretical studies on the performance of the different allocation rules obtained with the selected structures. In the numerical case, it is important that both parameter and structure estimation could be efficiently solved in order to use simulation procedures or a cross-validation approach.



a) $BIC = 2587.404$. # of par = 5.

b) $BIC = 2591.468$. # of par = 7.

c) $BIC = 2596.142$. # of par = 6.

Figure 5.1: RCON models fitted for Mathematics Marks Data: a) in Hϕjsgaard and Lauritzen (2008), b) in Gehrmann (2011), and c) using the penalized log-likelihood function in (5.1).

References

Abreu, G., Edwards, D. and Labouriau, R. (2010). High-Dimensional Graphical Model Search with the gRapHD R Package. *Journal of Statistical Software*, 37, 1–18.

Anderson, J. A. (1975). Quadratic logistic discrimination. *Biometrika*, 62, 149–154.

Anderson, T. W. (1970). Estimation of covariance matrices which are linear combinations or whose inverses are linear combinations of given matrices. In R. C. Bose, I. M. Chakravarti, P. C. Mahalanobis, C. R. Rao and K. J. C. Smith (Eds.), *Essays in Probability and Statistics*, Univ. North Carolina Press, Chapel Hill, 1–24.

Anderson, T. W. (1973). Asymptotically efficient estimation of covariance matrices with linear structure. *The Annals of Statistics*, 1, 135–141.

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. New York: John Wiley.

Bartlett, M. S. and Please, N. M. (1963). Discrimination in the case of zero mean differences. *Biometrika*, 50, 17–21.

Bickel, P and Levina, E. (2004). Some theory for Fisher's linear discriminant function, naive Bayes, and some alternatives when there are many more variables than observations. *Bernoulli*, 10, 989–1010.

Bodnar, T. and Okhrin, Y. (2011). On the Product of Inverse Wishart and Normal Distributions with Applications to Discriminant Analysis and Portfolio Theory. *Scandinavian Journal of Statistics*, 38,

311–331.

Borůvka, O. (1926). O jistém problému minimálním (About a certain minimal problem). *Práce mor. přírodověd. spol. v Brně III*, 3, 37–58.

Boyd S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Chow C. and Liu, C. (1966). An approach to structure adaptation in pattern recognition. *Systems Science and Cybernetics, IEEE Transactions.*, 2, 73–80.

Chow, C. and Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *Information Theory, IEEE Transactions*, 14, 462–467.

Clemmensen, L., Hastie, T., Wiiten, D. and Ersbøll, B. (2011). Sparse discriminant analysis. *Technometrics*, 53, 406-413.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695. *http : //igraph.sourceforge.net/*

Cox, D.R. and Snell, E.J. (1989). *Analysis of binary data*. 2nd ed. London: Chapman and Hall.

Danaher, P., Wang, P. and Witten, D. (2014). The joint graphical lasso for inverse covariance estimation across multiple clases. *Journal of the Royal Statistical Society: Series B.*, 76, 373–397.

Dethlefsen, C. and Højsgaard, S. (2005). A Common Platform for Graphical Models in R: The gRbase Package. *Journal of Statistical Software*. 14, 1–12.

Deshpande, A., Garofalakis, M. and Jordan, M.I. (2001). Efficient Stepwise Selection in Decomposable

Models. *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, UAI'01.* 128–135.

Dudoit, S. and Fridlyand, J. (2003). Classification in microarray experiments. In T. P. Speed (ed), *Statistical Analysis of Gene Expression Microarray Data*. Chapman & Hall/CRC, 3, 93–158.

Edwards, D., Abreu, G. and Labouriau, R. (2010). Selecting High-Dimensional Mixed Graphical Models Using Minimal AIC or BIC Forests. *BMC Bioinformatics*, 11.

Edwards, D. and Havránek, T. (1987). A fast model selection procedure for large families of models. *Journal of the American Statistical Association*, 82, 205–213.

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7, 179–188.

Friedman, N., Geiger, D. and Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning*, 29, 131–163.

Friedman, N., Goldszmidt, M. and Lee, T. (1998). Bayesian Network Classification with Continuous Attributes: Getting the Best of Both Discretization and Parametric Fitting. *In Proceedings of the International Conference on Machine Learning (ICML)*, 179–187.

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9, 432–441.

Frydenberg, M. and Lauritzen, S.L. (1989). Decomposition of Maximum Likelihood in Mixed Graphical Interaction Models. *Biometrika*, 76, 539–555.

Gao, X. and Massam H. (2014). Estimation of symmetry-constrained Gaussian graphical models: application to clustered dense networks. *Journal of Computational and Graphical Statistics.*

Gehrmann, H. (2011). Lattices of Graphical Gaussian Models with Symmetries. *Symmetry*, 3, 653–679.

Gehrmann, H. and Lauritzen, S. L. (2012). Estimation of Means in Graphical Gaussian Models with Symmetries. *The Annals of Statistics*, 40, 1061–1073.

Graham, R. L. and Hell, P. (1985). On the history of the minimum spanning tree problem. *Annals of the History of Computing*, 7, 43–57.

Green, P. and Thomas, A. (2013). Sampling decomposable graphs using a Markov chain on junction trees. *Biometrika*, 100, 91–110.

Guo, J., Levina, E., Michailidis, G. and Zhu, Ji. (2011). Joint estimation of multiple graphical models. *Biometrika*, 98, 1–15.

Hara, S. and Washio, T. (2011). Common Substructure Learning of Multiple Graphical Gaussian Models. In D. Gunopulos, T. Hofmann, D. Malerba and M. Vazirgiannis (Eds.), *Machine Learning and Knowledge Discovery in Databases*, Springer-Verlag Berlin Heidelberg, 1–16.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction.* 2nd Ed. Springer.

Hoefling, H. (2010). A path algorithm for the fused lasso signal approximator. *Journal of Computational and Graphical Statistics*, 19, 984–1006.

Højsgaard, S. and Lauritzen, S. L. (2005). Restricted concentration models-graphical Gaussian models

with concentration parameters restricted to being equal. In R. G. Cowell and Z. Ghahramani (Eds.), *Preceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, Hillsdale, NJ, 159–187.

Højsgaard, S. and Lauritzen, S. L. (2007). Inference in Graphical Gaussian Models with Edge and Vertex Symmetries with the gRc package for R. *Journal of Statistical Software*, 23.

Højsgaard, S. and Lauritzen, S. L. (2008). Graphical Gaussian models with edge and vertex symmetries. *Journal of Royal Statistical Society*, Series B, 70, 1005–1027.

Højsgaard, S., Lauritzen, S. L. and Edwards, D. (2012). *Graphical Models with R*. Springer.

Jensen, S. T., Johansen, S. and Lauritzen, S. L. (1991). Globally convergent algorithm for maximizing likelihood function. *Biometrika*, 78, 867–877.

Jiroušek, R. and Přeučil, S. (1995). On the effective implementation of the iterative proportional fitting procedure. *Computational Statistics & Data Analysis*, 19, 177–189.

Kim, J. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis*, 53, 3735–3745.

Kohlmann, M., Held, L. and Grunert V.P. (2009). Classification of therapy resistance based on longitudinal biomarker profiles. *Biometrical Journal*, 51, 610–626.

Kruskal, J. B. (1956). On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. *Proceedings of the American Mathematical Society*, 7, 48–50.

Kullback, S. and Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical*

Statistics, 22, 79–86.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press.

Lauritzen, S. L. *Elements of Graphical Models. Lectures from the XXXVIth International Probability Summer School in Saint-Flour, France, 2006.* Unpublished manuscript, electronic version, 2011.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press.

May, S. and DeGruttola, V. (2007). Nonparametric Tests for Two-Group Comparisons of Dependent Observations Obtained at Varying Time Points. *Biometrics*, 63, 194–200.

Meilă, M. and Jordan, M. (2000). Learning with Mixtures of Trees. *Journal of Machine Learning Research*, 1, 1–48.

Miller, L. D., Smeds, J., George, J., Vega, V., Vergara, L., Pawitan, Y., Hall, P., Klaar, S., Liu, E. and Bergh, J. (2005). An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciences*, 102, 13550–13555.

Neufeld, H. (2009). *Graphical Gaussian Models with Symmetries*. University of Oxford, A dissertation submitted for transfer to doctoral student status.

Olkin, I. and Press, S. J. (1969). Testing and Estimation for a Circular Stationary Model. *The Annals of Mathematical Statistics*, 40, 1358–1373.

Penrose, L. S. (1946-47). Some notes on discrimination. *Annals of Eugenics*, Lond., 13, 228–237.

Prim, R.C. (1957). Shortest connection networks and some generalizations. *Bell System Technology Journal*, 36, 1389–1401.

Rao, C. R. (1964). The Use and Interpretation of Principal Component Analysis in Applied Research. *The Indian Journal of Statistics*, Series A, 26, 329–358.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*, Cambridge: Cambridge University Press.

Seber (1984). *Multivariate Observations*, Wiley.

Simon, N. and Tibshirani, R. (2011). Discriminant Analysis with Adaptively Pooled Covariance. arXiv:1111.1687v2 [stat.ML].

Smith, C. A. B. (1946-47). Some examples of discrimination. *Annals of Eugenics*, Lond., 13, 272–82.

Speed, T. P. and Kiiveri, H. T. (1986). Gaussian Markov Distributions over Finite Graphs. *Annals of Statistics*, 14, 138–150.

Sturmfels, B and Uhler, C. (2010). Multivariate Gaussians, semidefinite matrix completion and convex algebraic geometry. *Annals of the Institute of Statistical Mathematics*, 62, 603–638.

Szatrowski, T.H. (1978). Explicit solutions, one iteration convergence and averaging in the multivariate normal estimation problem for patterned means and covariances. *Annals of the Institute of Statistical Mathematics*, 30, 81–88.

Szatrowski, T.H. (1982). Testing and Estimation in the Block Compound Symmetry Problem. *Journal of Educational Statistics*, 7, 3–18.

Tan, V., Sanghavi, V., Fisher, J., and Willsky, A. (2010). Learning Graphical Models for Hypothesis Testing and Classification. *IEEE Transactions on signal processing*, 58, 5481–5495.

Thomas, A. and Green, P. (2009). Enumerating the decomposable neighbours of a decomposable graph under a simple perturbation scheme. *Computational Statistics & Data Analysis*, 53, 1232–1238.

Uhler, C. (2012). Geometry of maximum likelihood estimation in Gaussian graphical models. *The Annals of Statistics*, 40, 238–261.

Votaw, D. (1948). Testing Compound Symmetry in a Normal Multivariate Distribution. *The Annals of Mathematical Statistics*, 19, 447–473.

Votaw, D., Kimball, A. W. and Rafferty, J. A. (1950). Compound Symmetry Tests in the Multivariate Analysis of Medical Experiments. *Biometrics*, 6, 259–281.

Welch, B. (1939). Note on Discriminant Functions. *Biometrika*, 31, 218–220.

Wilks, S. S. (1946). Sample criteria for testing equality of means, equality of variances, and equality of covariances in a normal multivariate distribution. *The Annals of Mathematical Statistics*, 17, 257–281.

Witten, D. and Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B.*, 73, 753–772.

Zhang, B. and Wang, Y. (2010). Learning structural changes of Gaussian graphical models in controlled experiments. In: *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence.*

A. CG-distribution with linear concentration matrices

A.1 CG-distribution as a member of the regular exponential family

We observe that (1.3), a Conditional Gaussian density (see Lauritzen, 1996, p. 158), can be expressed as follows

$$f(c,\boldsymbol{x}) = \exp\left[\delta_1(c)\left\{\ln\frac{\pi_1}{1-\pi_1} + \frac{1}{2}\ln\frac{\mid \boldsymbol{K}_1 \mid}{\mid \boldsymbol{K}_2 \mid} - \frac{1}{2}tr\left(\boldsymbol{K}_1\boldsymbol{\mu}_1\boldsymbol{\mu}_1^t - \boldsymbol{K}_2\boldsymbol{\mu}_2\boldsymbol{\mu}_2^t\right)\right\}\right.$$

$$\left. -\frac{1}{2}tr\left(\boldsymbol{K}_1\boldsymbol{x}\boldsymbol{x}^t\delta_1(c)\right) + tr\left(\boldsymbol{K}_1\boldsymbol{\mu}_1\boldsymbol{x}^t\delta_1(c)\right) - \frac{1}{2}tr\left(\boldsymbol{K}_2\boldsymbol{x}\boldsymbol{x}^t(1-\delta_1(c))\right) + tr\left(\boldsymbol{K}_2\boldsymbol{\mu}_2\boldsymbol{x}^t(1-\delta_1(c))\right)\right]$$

$$\exp\left[\ln(1-\pi_1) - \frac{p}{2}\ln(2\pi) + \frac{1}{2}\ln\mid \boldsymbol{K}_2 \mid - \frac{1}{2}tr\left(\boldsymbol{K}_2\boldsymbol{\mu}_2\boldsymbol{\mu}_2^t\right)\right] \tag{A.1}$$

where $\delta_1(c)$ is as in (1.10).

Taking equation 3.1 with restriction III into (A.1), one gets

$$f(c,\boldsymbol{x}) = \exp\left[\delta_1(c)\theta_1 - \frac{1}{2}\sum_{h=0}^{f}\psi_h\left(tr\left(\boldsymbol{H}_h\boldsymbol{x}\boldsymbol{x}^t\delta_1(c)\right) + tr\left(\boldsymbol{H}_h\boldsymbol{x}\boldsymbol{x}^t(1-\delta_1(c))\right)\right) - \frac{1}{2}\sum_{h=f+1}^{g}\psi_h^{(1)}tr\left(\boldsymbol{H}_h\boldsymbol{x}\boldsymbol{x}^t\delta_1(c)\right)\right.$$

$$+\sum_{i=1}^{p}\varepsilon_i^{(1)}x_i\delta_1(c)-\frac{1}{2}\sum_{h=f+1}^{g}\psi_h^{(2)}tr\left(\boldsymbol{H}_h\boldsymbol{x}\boldsymbol{x}^t(1-\delta_1(c))\right)+\sum_{i=1}^{p}\varepsilon_i^{(2)}x_i(1-\delta_1(c))\Bigg]\exp[-\psi(\boldsymbol{\theta})] \tag{A.2}$$

where $\psi(\boldsymbol{\theta})$, the cumulant function of $f(c,\boldsymbol{x})$, is equal to

$$\begin{aligned}\psi(\boldsymbol{\theta})=&\ln\left(1+\exp\left(\theta_1-\frac{1}{2}\ln\frac{\mid\boldsymbol{K}_1\mid}{\mid\boldsymbol{K}_2\mid}+\frac{1}{2}tr\left(\boldsymbol{K}_1\boldsymbol{\mu}_1\boldsymbol{\mu}_1^t-\boldsymbol{K}_2\boldsymbol{\mu}_2\boldsymbol{\mu}_2^t\right)\right)\right)\\&+\frac{p}{2}\ln(2\pi)-\frac{1}{2}\ln\mid\boldsymbol{K}_2\mid+\frac{1}{2}tr\left(\boldsymbol{K}_2\boldsymbol{\mu}_2\boldsymbol{\mu}_2^t\right).\end{aligned} \tag{A.3}$$

And the vector of canonical parameters is

$$\boldsymbol{\theta}=(\theta_1,\psi_0,...,\psi_f,\psi_{f+1}^{(1)},...,\psi_q^{(1)},\psi_{f+1}^{(2)},...,\psi_q^{(2)},\varepsilon_1^{(1)},...,\varepsilon_p^{(1)},\varepsilon_1^{(2)},...,\varepsilon_p^{(2)}),$$

where $\boldsymbol{\varepsilon}^{(1)}=(\varepsilon_1^{(1)},...,\varepsilon_p^{(1)})=\boldsymbol{K}_1\boldsymbol{\mu}_1$, $\boldsymbol{\varepsilon}^{(2)}=(\varepsilon_1^{(2)},...,\varepsilon_p^{(2)})=\boldsymbol{K}_2\boldsymbol{\mu}_2$ and

$$\theta_1=\ln\frac{\pi_1}{1-\pi_1}+\frac{1}{2}\ln\frac{\mid\boldsymbol{K}_1\mid}{\mid\boldsymbol{K}_2\mid}-\frac{1}{2}tr\left(\boldsymbol{K}_1\boldsymbol{\mu}_1\boldsymbol{\mu}_1^t-\boldsymbol{K}_2\boldsymbol{\mu}_2\boldsymbol{\mu}_2^t\right).$$

In this case, $\varepsilon_i^{(1)}$'s and $\varepsilon_p^{(2)}$'s have no restrictions since the vector means are not restricted, and hence the CG-distribution with restrictions in III belongs to the regular exponential family and (A.2) is given in his minimal expression with canonical parameter vector $\boldsymbol{\theta}$ and vector of sufficient statistics

$$\begin{aligned}\boldsymbol{t}(c,\boldsymbol{x}_1,...,\boldsymbol{x}_n)=\Bigg(&n_1=\sum_{i=1}^{n}\delta_1(c_i),\left\{-\frac{1}{2}\sum_{i=1}^{n}tr\left(\boldsymbol{H}_h\boldsymbol{x}_i\boldsymbol{x}_i^t\right)\right\}_{h=0}^{f},\left\{-\frac{1}{2}\sum_{i=1}^{n}tr\left(\boldsymbol{H}_h\boldsymbol{x}_i\boldsymbol{x}_i^t\delta_1(c_i)\right)\right\}_{h=f+1}^{q},\\&\left\{-\frac{1}{2}\sum_{i=1}^{n}tr\left(\boldsymbol{H}_h\boldsymbol{x}_i\boldsymbol{x}_i^t(1-\delta_1(c_i))\right)\right\}_{h=f+1}^{q},\left\{\sum_{i=1}^{n}x_{il}\delta_1(c_i)\right\}_{l=1}^{p},\left\{\sum_{i=1}^{n}x_{il}(1-\delta_1(c_i))\right\}_{l=1}^{p}\Bigg)\end{aligned} \tag{A.4}$$

A.2 Maximum likelihood estimation

The maximum likelihood estimator of $\boldsymbol{\theta}$, if exists, it is unique and can be obtained by equating the sufficient canonical statistics to their expectations as follows (see Lauritzen, 1996, Theorem D.1, p. 268)

$$
\begin{aligned}
n\frac{\partial\psi(\boldsymbol{\theta})}{\partial\theta_1} &= \sum_{i=1}^{n}\delta_1(c_i) = n_1 \\
n\frac{\partial\psi(\boldsymbol{\theta})}{\partial\varepsilon_l^{(1)}} &= \sum_{i=1}^{n}x_{il}\delta_1(c_i), \quad l=1,...,p \\
n\frac{\partial\psi(\boldsymbol{\theta})}{\partial\varepsilon_l^{(2)}} &= \sum_{i=1}^{n}x_{il}(1-\delta_1(c_i)), \quad l=1,...,p \\
n\frac{\partial\psi(\boldsymbol{\theta})}{\partial\psi_h} &= -\frac{1}{2}\sum_{i=1}^{n}tr\left(\boldsymbol{H}_h\boldsymbol{x}_i\boldsymbol{x}_i^t\right), \quad h=0,...,f \\
n\frac{\partial\psi(\boldsymbol{\theta})}{\partial\psi_h^{(1)}} &= -\frac{1}{2}\sum_{i=1}^{n}tr\left(\boldsymbol{H}_h\boldsymbol{x}_i\boldsymbol{x}_i^t\delta_1(c_i)\right), \quad h=f+1,...,q \\
n\frac{\partial\psi(\boldsymbol{\theta})}{\partial\psi_h^{(2)}} &= -\frac{1}{2}\sum_{i=1}^{n}tr\left(\boldsymbol{H}_h\boldsymbol{x}_i\boldsymbol{x}_i^t(1-\delta_1(c_i))\right), \quad h=f+1,...,q.
\end{aligned}
\tag{A.5}
$$

Using the results:

1. $\dfrac{\partial\boldsymbol{x}^t\boldsymbol{A}\boldsymbol{x}}{\partial\boldsymbol{A}} = \boldsymbol{x}\boldsymbol{x}^t$,
2. $\dfrac{\partial\boldsymbol{x}^t\boldsymbol{A}\boldsymbol{x}}{\partial\boldsymbol{x}} = 2\boldsymbol{A}\boldsymbol{x}$ when $\boldsymbol{A}$ is a symmetric matrix,
3. $\dfrac{\partial\mid\boldsymbol{A}\mid}{\partial x} = \mid\boldsymbol{A}\mid tr\left(\boldsymbol{A}^{-1}\dfrac{\partial\boldsymbol{A}}{\partial x}\right)$,
4. and $\dfrac{\partial\boldsymbol{A}^{-1}}{\partial x} = -\boldsymbol{A}^{-1}\dfrac{\partial\boldsymbol{A}}{\partial x}\boldsymbol{A}^{-1}$

one gets that the expectations correspond to the following expressions.

$$
\begin{aligned}
n\frac{\partial\psi(\boldsymbol{\theta})}{\partial\theta_1} &= n\frac{\exp\left(\theta_1 - \frac{1}{2}\ln\frac{|\boldsymbol{K}_1|}{|\boldsymbol{K}_2|} + \frac{1}{2}tr\left(\boldsymbol{K}_1\boldsymbol{\mu}_1\boldsymbol{\mu}_1^t - \boldsymbol{K}_2\boldsymbol{\mu}_2\boldsymbol{\mu}_2^t\right)\right)}{1+\exp\left(\theta_1 - \frac{1}{2}\ln\frac{|\boldsymbol{K}_1|}{|\boldsymbol{K}_2|} + \frac{1}{2}tr\left(\boldsymbol{K}_1\boldsymbol{\mu}_1\boldsymbol{\mu}_1^t - \boldsymbol{K}_2\boldsymbol{\mu}_2\boldsymbol{\mu}_2^t\right)\right)} = n\pi_1 \\
n\frac{\partial\psi(\boldsymbol{\theta})}{\partial\varepsilon_l^{(1)}} &= n\pi_1 tr\left(\boldsymbol{K}_1^{-1}\varepsilon^{(1)}\boldsymbol{e}_l^t\right) = n\pi_1\mu_{1_l}, \quad l=1,...,p \\
n\frac{\partial\psi(\boldsymbol{\theta})}{\partial\varepsilon_l^{(2)}} &= n(1-\pi_1) tr\left(\boldsymbol{K}_2^{-1}\varepsilon^{(2)}\boldsymbol{e}_l^t\right) = n(1-\pi_1)\mu_{2_l}, \quad l=1,...,p \\
n\frac{\partial\psi(\boldsymbol{\theta})}{\partial\psi_h} &= n\pi_1\left(-\frac{1}{2}tr(\boldsymbol{K}_1^{-1}\boldsymbol{H}_h) - \frac{1}{2}tr\left(\boldsymbol{\mu}_1\boldsymbol{\mu}_1^t\boldsymbol{H}_h\right)\right) \\
&\quad + n(1-\pi_1)\left(-\frac{1}{2}tr(\boldsymbol{K}_2^{-1}\boldsymbol{H}_h) - \frac{1}{2}tr\left(\boldsymbol{\mu}_2\boldsymbol{\mu}_2^t\boldsymbol{H}_h\right)\right), \quad h=0,...,f \qquad \text{(A.6)} \\
n\frac{\partial\psi(\boldsymbol{\theta})}{\partial\psi_h^{(1)}} &= n\pi_1\left(-\frac{1}{2}tr(\boldsymbol{K}_1^{-1}\boldsymbol{H}_h) - \frac{1}{2}tr\left(\boldsymbol{\mu}_1\boldsymbol{\mu}_1^t\boldsymbol{H}_h\right)\right), \quad h=f+1,...,q \\
n\frac{\partial\psi(\boldsymbol{\theta})}{\partial\psi_h^{(2)}} &= n(1-\pi_1)\left(-\frac{1}{2}tr(\boldsymbol{K}_2^{-1}\boldsymbol{H}_h) - \frac{1}{2}tr\left(\boldsymbol{\mu}_2\boldsymbol{\mu}_2^t\boldsymbol{H}_h\right)\right), \quad h=f+1,...,q
\end{aligned}
$$

where vector $\boldsymbol{e}_l$ has zero in all entries but in entry l has 1.

Using (A.5) and (A.6), we observe that π_1, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ can be estimated with an analytical expression: $\widehat{\pi}_1 = \frac{n_1}{n}$, $\widehat{\boldsymbol{\mu}}_1 = \overline{\boldsymbol{x}}_1$ and $\widehat{\boldsymbol{\mu}}_2 = \overline{\boldsymbol{x}}_2$; and these parameters do not depend on the canonical parameters. Additionally, the equations for the canonical parameters ψ_h, $h=0,...,f$, $\psi_h^{(1)}$, $h=f+1,...,q$, and $\psi_h^{(2)}$, $h=f+1,...,q$, depend only on the canonical parameters θ_1, $\boldsymbol{\varepsilon}^{(1)}$ and $\boldsymbol{\varepsilon}^{(2)}$ through π_1, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. Therefore, it is possible to estimate $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$ in a first stage, and afterwards to estimate $\boldsymbol{\varepsilon}^{(1)}$, $\boldsymbol{\varepsilon}^{(2)}$ and θ_1.

Variances of the sufficient statistics are obtained from the second derivatives of the cumulant function, as follows.

$$
\begin{aligned}
n\frac{\partial^2\psi(\boldsymbol{\theta})}{\partial\psi_h^2} &= n\pi_1(1+\pi_1)\left(-\frac{1}{2}tr(\boldsymbol{K}_1^{-1}\boldsymbol{H}_h)-\frac{1}{2}tr\left(\boldsymbol{\mu}_1\boldsymbol{\mu}_1^t\boldsymbol{H}_h\right)+\frac{1}{2}tr(\boldsymbol{K}_2^{-1}\boldsymbol{H}_h)+\frac{1}{2}tr\left(\boldsymbol{\mu}_2\boldsymbol{\mu}_2^t\boldsymbol{H}_h\right)\right)^2 \\
&\quad + n\pi_1(\frac{1}{2}tr(\boldsymbol{H}_h\boldsymbol{K}_1^{-1}\boldsymbol{H}_h\boldsymbol{K}_1^{-1})+tr(\boldsymbol{\mu}_1^t\boldsymbol{H}_h\boldsymbol{K}_1^{-1}\boldsymbol{H}_h\boldsymbol{\mu}_1)), \\
&\quad + n(1-\pi_1)(\frac{1}{2}tr(\boldsymbol{H}_h\boldsymbol{K}_2^{-1}\boldsymbol{H}_h\boldsymbol{K}_2^{-1})+tr(\boldsymbol{\mu}_2^t\boldsymbol{H}_h\boldsymbol{K}_2^{-1}\boldsymbol{H}_h\boldsymbol{\mu}_2)), \quad h=0,...,f \\
n\frac{\partial^2\psi(\boldsymbol{\theta})}{\partial{\psi_h^{(1)}}^2} &= n\pi_1(1+\pi_1)\left(-\frac{1}{2}tr(\boldsymbol{K}_1^{-1}\boldsymbol{H}_h)-\frac{1}{2}tr\left(\boldsymbol{\mu}_1\boldsymbol{\mu}_1^t\boldsymbol{H}_h\right)\right)^2 \\
&\quad + n\pi_1(\frac{1}{2}tr(\boldsymbol{H}_h\boldsymbol{K}_1^{-1}\boldsymbol{H}_h\boldsymbol{K}_1^{-1})+tr(\boldsymbol{\mu}_1^t\boldsymbol{H}_h\boldsymbol{K}_1^{-1}\boldsymbol{H}_h\boldsymbol{\mu}_1)), \quad h=f+1,...,q \\
n\frac{\partial^2\psi(\boldsymbol{\theta})}{\partial{\psi_h^{(2)}}^2} &= n\pi_1(1+\pi_1)\left(-\frac{1}{2}tr(\boldsymbol{K}_2^{-1}\boldsymbol{H}_h)-\frac{1}{2}tr\left(\boldsymbol{\mu}_2\boldsymbol{\mu}_2^t\boldsymbol{H}_h\right)\right)^2 \\
&\quad + n(1-\pi_1)(\frac{1}{2}tr(\boldsymbol{H}_h\boldsymbol{K}_2^{-1}\boldsymbol{H}_h\boldsymbol{K}_2^{-1})+tr(\boldsymbol{\mu}_2^t\boldsymbol{H}_h\boldsymbol{K}_2^{-1}\boldsymbol{H}_h\boldsymbol{\mu}_2)), \quad h=f+1,...,q.
\end{aligned}
\tag{A.7}
$$

Expressions in (A.6) and (A.7) are used in the IPM algorithm described in Section 3.2.

B. Kruskal's algorithm for the MWST problem

Kruskal's algorithm (Kruskal, 1956) is as follows . For a connected and weighted graph $G = (V, E)$

1. Sort all weights in increasing order.
2. Select the edge with the minimum weight among the edges not selected if it does not form a cycle with the so far selected edges, otherwise discard this edge.
3. Do step 2 until all nodes are connected.

For example, the complete graph with four nodes and weights given in Figure B.1 has 16 spanning trees and total weight $\lambda(G) = 19$. A MWST can be found using Kruskal's algorithm as in Figure B.2 considering that the edges are sorted as: $(1, 3)$, $(1, 2)$, $(2, 3)$, $(1, 4)$, $(2, 4)$ and $(3, 4)$. A red dotted edge means that this edge is discarded since it forms a cycle and a blue dashed edge means that this edge is selected.

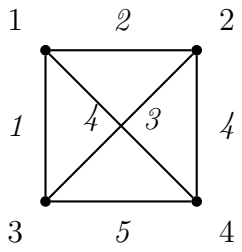


Figure B.1: A weighted complete graph G with four nodes and total weight $\lambda(G) = 19$.

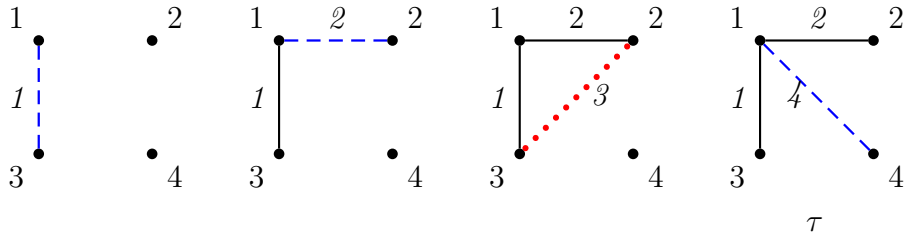


Figure B.2: Kruskal's algorithm applied to the graph in Figure B.1. A red dotted edge means that this edge is discarded since it forms a cycle. A blue dashed edge means that this edge is selected. The weight of the MWST τ is $\lambda(\tau) = 7$.

Another MWST for the graph in Figure B.1 can be obtained using Kruskal's algorithm considering the following order: $(1,3)$, $(1,2)$, $(2,3)$, $(2,4)$, $(1,4)$ and $(3,4)$. The MWST τ^* obtained is the one presented in Figure B.3 with weight $\lambda(\tau^*) = 7$.

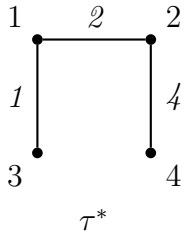


Figure B.3: Another MWST, τ^*, for the graph in Figure B.1. The weight of this spanning tree is also $\lambda(\tau^*) = 7$.

C. Proofs of propositions

Proof of proposition 4.5.1 We note that for a given tree $\tau = (V, E_\tau)$ with p nodes

$$
\begin{aligned}
J(\widehat{f}_{1_\tau}, \widehat{f}_{2_\tau}) &= \frac{1}{2}\Big[tr(\widehat{\boldsymbol{\Sigma}}_{1_\tau}\widehat{\boldsymbol{K}}_{2_\tau}) + tr(\widehat{\boldsymbol{\Sigma}}_{2_\tau}\widehat{\boldsymbol{K}}_{1_\tau}) + tr(\widehat{\boldsymbol{K}}_{1_\tau}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^t) \\
&\quad + tr(\widehat{\boldsymbol{K}}_{2_\tau}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^t)\Big] - p \qquad \text{equation 4.4} \\
&= \frac{1}{2}\Big[tr(\widehat{\boldsymbol{K}}_{2_\tau}\widehat{\boldsymbol{\Sigma}}_{1_\tau}(\tau)) + tr(\widehat{\boldsymbol{K}}_{1_\tau}\widehat{\boldsymbol{\Sigma}}_{2_\tau}(\tau)) + tr(\widehat{\boldsymbol{K}}_{1_\tau}\boldsymbol{D} + \widehat{\boldsymbol{K}}_{2_\tau}\boldsymbol{D})\Big] - p \quad \text{prop. 4.2.7} \\
&= \frac{1}{2}\Big[tr(\widehat{\boldsymbol{K}}_{2_\tau}\boldsymbol{W}_1(\tau)) + tr(\widehat{\boldsymbol{K}}_{1_\tau}\boldsymbol{W}_2(\tau)) + tr(\widehat{\boldsymbol{K}}_{1_\tau}\boldsymbol{D} + \widehat{\boldsymbol{K}}_{2_\tau}\boldsymbol{D})\Big] - p \quad \text{prop. 4.2.6} \\
&= \tfrac{1}{2}\sum_{\substack{i<j \\ (i,j)\in E_\tau}} \Big[tr([\boldsymbol{W}_1{}^{-1}_{(i,j)}]^p\boldsymbol{D}) - tr([\boldsymbol{W}_1{}^{-1}_{(i)}]^p\boldsymbol{D}) - tr([\boldsymbol{W}_1{}^{-1}_{(j)}]^p\boldsymbol{D}) \\
&\quad + tr([\boldsymbol{W}_2{}^{-1}_{(i,j)}]^p\boldsymbol{D}) - tr([\boldsymbol{W}_2{}^{-1}_{(i)}]^p\boldsymbol{D}) - tr([\boldsymbol{W}_2{}^{-1}_{(j)}]^p\boldsymbol{D}) \\
&\quad + tr([\boldsymbol{W}_2{}^{-1}_{(i,j)}]^p\boldsymbol{W}_1) - tr([\boldsymbol{W}_2{}^{-1}_{(i)}]^p\boldsymbol{W}_1) - tr([\boldsymbol{W}_2{}^{-1}_{(j)}]^p\boldsymbol{W}_1) \\
&\quad + tr([\boldsymbol{W}_1{}^{-1}_{(i,j)}]^p\boldsymbol{W}_2) - tr([\boldsymbol{W}_1{}^{-1}_{(i)}]^p\boldsymbol{W}_2) - tr([\boldsymbol{W}_1{}^{-1}_{(j)}]^p\boldsymbol{W}_2)\Big] \\
&\quad + \tfrac{1}{2}\textstyle\sum_{j=1}^p tr\left([\boldsymbol{W}_2{}^{-1}_{(j)}]^p\boldsymbol{W}_1\right) + \tfrac{1}{2}\sum_{j=1}^p tr\left([\boldsymbol{W}_1{}^{-1}_{(j)}]^p\boldsymbol{W}_2\right) \\
&\quad + \tfrac{1}{2}\textstyle\sum_{j=1}^p tr\left([\boldsymbol{W}_1{}^{-1}_{(j)}]^p\boldsymbol{D}\right) + \tfrac{1}{2}\sum_{j=1}^p tr\left([\boldsymbol{W}_2{}^{-1}_{(j)}]^p\boldsymbol{D}\right) - p \qquad \text{prop. 4.2.8} \\
&= -\sum_{\substack{i<j \\ (i,j)\in E_\tau}} \lambda(i,j) + C \;=\; -\lambda(\tau) + C,
\end{aligned}
$$

where $\lambda(\tau) = \sum_{\substack{i<j \\ (i,j)\in E_\tau}} \lambda(i,j)$ is the total weight of τ, $\boldsymbol{D} = (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^t$, $\widehat{\boldsymbol{\Sigma}}_{c_\tau} = \widehat{\boldsymbol{K}}^{-1}_{c_\tau}$, $c = 1, 2$, C is a constant, and

$$
\begin{aligned}
\lambda(i,j) \;=\; & -\tfrac{1}{2}\Bigg[tr\left(\begin{pmatrix} w_{ii}^{(2)} & w_{ij}^{(2)} \\ w_{ij}^{(2)} & w_{jj}^{(2)} \end{pmatrix}^{-1} \begin{pmatrix} w_{ii}^{(1)} & w_{ij}^{(1)} \\ w_{ij}^{(1)} & w_{jj}^{(1)} \end{pmatrix}\right) + tr\left(\begin{pmatrix} w_{ii}^{(1)} & w_{ij}^{(1)} \\ w_{ij}^{(1)} & w_{jj}^{(1)} \end{pmatrix}^{-1} \begin{pmatrix} w_{ii}^{(2)} & w_{ij}^{(2)} \\ w_{ij}^{(2)} & w_{jj}^{(2)} \end{pmatrix}\right) \\
& + tr\left(\begin{pmatrix} w_{ii}^{(1)} & w_{ij}^{(1)} \\ w_{ij}^{(1)} & w_{jj}^{(1)} \end{pmatrix}^{-1} \begin{pmatrix} (\overline{x}_i^{(1)} - \overline{x}_i^{(2)})^2 & (\overline{x}_i^{(1)} - \overline{x}_i^{(2)})(\overline{x}_j^{(1)} - \overline{x}_j^{(2)}) \\ (\overline{x}_i^{(1)} - \overline{x}_i^{(2)})(\overline{x}_j^{(1)} - \overline{x}_j^{(2)}) & (\overline{x}_j^{(1)} - \overline{x}_j^{(2)})^2 \end{pmatrix}\right) \\
& + tr\left(\begin{pmatrix} w_{ii}^{(2)} & w_{ij}^{(2)} \\ w_{ij}^{(2)} & w_{jj}^{(2)} \end{pmatrix}^{-1} \begin{pmatrix} (\overline{x}_i^{(1)} - \overline{x}_i^{(2)})^2 & (\overline{x}_i^{(1)} - \overline{x}_i^{(2)})(\overline{x}_j^{(1)} - \overline{x}_j^{(2)}) \\ (\overline{x}_i^{(1)} - \overline{x}_i^{(2)})(\overline{x}_j^{(1)} - \overline{x}_j^{(2)}) & (\overline{x}_j^{(1)} - \overline{x}_j^{(2)})^2 \end{pmatrix}\right) \\
& - \frac{w_{ii}^{(1)}}{w_{ii}^{(2)}} - \frac{w_{jj}^{(1)}}{w_{jj}^{(2)}} - \frac{w_{ii}^{(2)}}{w_{ii}^{(1)}} - \frac{w_{jj}^{(2)}}{w_{jj}^{(1)}} - \frac{\left(\overline{x}_i^{(1)} - \overline{x}_i^{(2)}\right)^2}{w_{ii}^{(1)}} - \frac{\left(\overline{x}_j^{(1)} - \overline{x}_j^{(2)}\right)^2}{w_{jj}^{(1)}} - \frac{\left(\overline{x}_i^{(1)} - \overline{x}_i^{(2)}\right)^2}{w_{ii}^{(2)}} \\
& - \frac{\left(\overline{x}_j^{(1)} - \overline{x}_j^{(2)}\right)^2}{w_{jj}^{(2)}} \Bigg].
\end{aligned}
\tag{C.1}
$$

Since $-\lambda(\tau)$ is the only term of $J(\widehat{f}_{1_\tau}, \widehat{f}_{2_\tau})$ that varies depending on a given tree τ, the problem of maximizing $J(\widehat{f}_{1_\tau}, \widehat{f}_{2_\tau})$ over T_p in equation 4.16 is equivalent to the problem of finding a MWST for the complete graph with p nodes and weights defined in equation C.1 for each edge (i,j). We noted that weights in (C.1) are equal to those in (4.17) . ∎

Proof of proposition 4.5.2 The problem in (4.18) can be expressed as finding τ_1^* and τ_2^* such that

$$
\tau_1^* = \underset{\tau_1 \in T_p}{\operatorname{argmax}} \left\{ \sum_{l=1}^{n_1} \ln \widehat{f}_{1_{\tau_1}}(\boldsymbol{x}_l) - \sum_{l=n_1+1}^{n_1+n_2} \ln \widehat{f}_{1_{\tau_1}}(\boldsymbol{x}_l) \right\} \tag{C.2}
$$

$$
\tau_2^* = \underset{\tau_2 \in T_p}{\operatorname{argmax}} \left\{ \sum_{l=n_1+1}^{n_1+n_2} \ln \widehat{f}_{2_{\tau_2}}(\boldsymbol{x}_l) - \sum_{l=1}^{n_1} \ln \widehat{f}_{2_{\tau_2}}(\boldsymbol{x}_l) \right\}. \tag{C.3}
$$

Considering the problem for τ_1^* in (C.2), we note that for a given tree $\tau = (V, E_\tau)$ with p nodes

$$\begin{aligned}
\sum_{l=1}^{n_1} \ln \widehat{f}_{1_\tau}(\boldsymbol{x}_l) \quad - \quad & \sum_{l=n_1+1}^{n_1+n_2} \ln \widehat{f}_{1_\tau}(\boldsymbol{x}_l) \\
&= \sum_{i=1}^{p} \sum_{l=1}^{n_1} \ln \widehat{f}_1(x_{i_l}) + \sum_{\substack{i<j \\ (i,j)\in E_\tau}} \sum_{l=1}^{n_1} \ln \frac{\widehat{f}_1(x_{i_l}, x_{j_l})}{\widehat{f}_1(x_{i_l})\ \widehat{f}_1(x_{j_l})} \\
&\quad -\sum_{i=1}^{p} \sum_{l=n_1+1}^{n_1+n_2} \ln \widehat{f}_1(x_{i_l}) - \sum_{\substack{i<j \\ (i,j)\in E_\tau}} \sum_{l=n_1+1}^{n_1+n_2} \ln \frac{\widehat{f}_1(x_{i_l}, x_{j_l})}{\widehat{f}_1(x_{i_l})\ \widehat{f}_1(x_{j_l})} \qquad \text{prop. } 4.2.9 \\
&= \sum_{i=1}^{p} \left(\sum_{l=1}^{n_1} \ln \widehat{f}_1(x_{i_l}) - \sum_{l=n_1+1}^{n_1+n_2} \ln \widehat{f}_1(x_{i_l}) \right) - \sum_{\substack{i<j \\ (i,j)\in E_\tau}} \lambda(i,j),
\end{aligned}$$

where

$$\lambda(i,j) = \sum_{l=n_1+1}^{n_1+n_2} \ln \frac{\widehat{f}_1(x_{i_l}, x_{j_l})}{\widehat{f}_1(x_{i_l})\ \widehat{f}_1(x_{j_l})} - \sum_{l=1}^{n_1} \ln \frac{\widehat{f}_1(x_{i_l}, x_{j_l})}{\widehat{f}_1(x_{i_l})\ \widehat{f}_1(x_{j_l})}\ . \tag{C.4}$$

Since $-\sum_{\substack{i<j \\ (i,j)\in E_\tau}} \lambda(i,j)$ is the only term that varies depending on a given tree τ, the problem of maximizing $\sum_{l=1}^{n_1} \ln \widehat{f}_{1_\tau}(\boldsymbol{x}_l) - \sum_{l=n_1+1}^{n_1+n_2} \ln \widehat{f}_{1_\tau}(\boldsymbol{x}_l)$ over T_p is equivalent to the problem of finding a MWST for the complete graph with p nodes and weights defined in equation C.4 for each edge (i,j). Using property 4.2.10 in (C.4) we can obtain the weights given in (4.19) for $c = 1$. A similar procedure can be done for the problem for τ_2^* in (C.3). ∎

Proof of corollary 4.5.3 We note that for a given tree $\tau = (V, E_\tau)$ with p nodes

$$\begin{aligned}
\sum_{l=1}^{n_1} \ln \frac{\widehat{f}_{1_\tau}(\boldsymbol{x}_l)}{\widehat{f}_{2_\tau}(\boldsymbol{x}_l)} \quad & + \sum_{l=n_1+1}^{n_1+n_2} \ln \frac{\widehat{f}_{2_\tau}(\boldsymbol{x}_l)}{\widehat{f}_{1_\tau}(\boldsymbol{x}_l)} \\
&= \left\{ \sum_{l=1}^{n_1} \ln \widehat{f}_{1_\tau}(\boldsymbol{x}_l) - \sum_{l=n_1+1}^{n_1+n_2} \ln \widehat{f}_{1_\tau}(\boldsymbol{x}_l) \right\} + \left\{ \sum_{l=n_1+1}^{n_1+n_2} \ln \widehat{f}_{2_\tau}(\boldsymbol{x}_l) - \sum_{l=1}^{n_1} \ln \widehat{f}_{2_\tau}(\boldsymbol{x}_l) \right\}.
\end{aligned}$$

The rest can be obtained using simultaneously the two procedures given in the proof of Proposition

4.5.2 for (C.2) and (C.3). ∎

Proof of proposition 4.5.4 For a given tree graph $\tau = (V, E_\tau)$ with p nodes, we have

$$\widehat{f}_{c_\tau}(\boldsymbol{x}) = \prod_{i=1}^{p} \widehat{f}_c(x_i) \prod_{\substack{i<j \\ (i,j)\in E_\tau}} \frac{\widehat{f}_c(x_i, x_j)}{\widehat{f}_c(x_i)\ \widehat{f}_c(x_j)} \tag{C.5}$$

by property 4.2.9 for $c = 1, 2$. Now,

$$\begin{aligned} J(\widehat{f}_{1_\tau}, \widehat{f}_{2_\tau}) &= \int \left(\widehat{f}_{1_\tau}(\boldsymbol{x}) - \widehat{f}_{2_\tau}(\boldsymbol{x})\right) \ln \frac{\widehat{f}_{1_\tau}(\boldsymbol{x})}{\widehat{f}_{2_\tau}(\boldsymbol{x})} d\boldsymbol{x} \\ &= \int \left(\widehat{f}_{1_\tau}(\boldsymbol{x}) - \widehat{f}_{2_\tau}(\boldsymbol{x})\right) \textstyle\sum_{i=1}^{p} \left\{\ln \widehat{f}_1(x_i) - \ln \widehat{f}_2(x_i)\right\} d\boldsymbol{x} \\ &\quad + \int \left(\widehat{f}_{1_\tau}(\boldsymbol{x}) - \widehat{f}_{2_\tau}(\boldsymbol{x})\right) \textstyle\sum_{\substack{i<j \\ (i,j)\in E_\tau}} \left\{\ln \frac{\widehat{f}_1(x_i, x_j)}{\widehat{f}_1(x_i)\ \widehat{f}_1(x_j)} - \ln \frac{\widehat{f}_2(x_i, x_j)}{\widehat{f}_2(x_i)\ \widehat{f}_2(x_j)}\right\} d\boldsymbol{x}. \end{aligned}$$

Using property 4.2.6 , this can be written as

$$\begin{aligned} J(\widehat{f}_{1_\tau}, \widehat{f}_{2_\tau}) &= \sum_{i=1}^{p} \int \left(\widehat{f}_{1_\tau}(\boldsymbol{x}) - \widehat{f}_{2_\tau}(\boldsymbol{x})\right) \left\{\ln \widehat{f}_1(x_i) - \ln \widehat{f}_2(x_i)\right\} d\boldsymbol{x} \\ &\quad + \sum_{\substack{i<j \\ (i,j)\in E_\tau}} \int \left(\widehat{f}_{1_\tau}(\boldsymbol{x}) - \widehat{f}_{2_\tau}(\boldsymbol{x})\right) \left\{\ln \frac{\widehat{f}_1(x_i, x_j)}{\widehat{f}_1(x_i)\ \widehat{f}_1(x_j)} - \ln \frac{\widehat{f}_2(x_i, x_j)}{\widehat{f}_2(x_i)\ \widehat{f}_2(x_j)}\right\} d\boldsymbol{x} \\ &= \sum_{i=1}^{p} \int \left(\widehat{f}_1(x_i) - \widehat{f}_2(x_i)\right) \left\{\ln \widehat{f}_1(x_i) - \ln \widehat{f}_2(x_i)\right\} dx_i \\ &\quad + \sum_{\substack{i<j \\ (i,j)\in E_\tau}} \int \left(\widehat{f}_1(x_i, x_j) - \widehat{f}_2(x_i, x_j)\right) \left\{\ln \frac{\widehat{f}_1(x_i, x_j)}{\widehat{f}_1(x_i)\ \widehat{f}_1(x_j)} - \ln \frac{\widehat{f}_2(x_i, x_j)}{\widehat{f}_2(x_i)\ \widehat{f}_2(x_j)}\right\} d(x_i, x_j) \end{aligned}$$

$$= \sum_{i=1}^{p} \int \left(\widetilde{f}_1(x_i) - \widetilde{f}_2(x_i) \right) \left\{ \ln \widehat{f}_1(x_i) - \ln \widehat{f}_2(x_i) \right\} dx_i$$

$$+ \sum_{\substack{i<j \\ (i,j)\in E_\tau}} \int \left(\widetilde{f}_1(x_i, x_j) - \widetilde{f}_2(x_i, x_j) \right) \left\{ \ln \frac{\widehat{f}_1(x_i, x_j)}{\widehat{f}_1(x_i)\ \widehat{f}_1(x_j)} - \ln \frac{\widehat{f}_2(x_i, x_j)}{\widehat{f}_2(x_i)\ \widehat{f}_2(x_j)} \right\} d(x_i, x_j).$$

Finally, we obtain the result

$$J(\widehat{f}_{1_\tau}, \widehat{f}_{2_\tau}) = \sum_{i=1}^{p} \int \left(\widetilde{f}_1(\boldsymbol{x}) - \widetilde{f}_2(\boldsymbol{x}) \right) \left\{ \ln \widehat{f}_1(x_i) - \ln \widehat{f}_2(x_i) \right\} d\boldsymbol{x}$$

$$+ \sum_{\substack{i<j \\ (i,j)\in E_\tau}} \int \left(\widetilde{f}_1(\boldsymbol{x}) - \widetilde{f}_2(\boldsymbol{x}) \right) \left\{ \ln \frac{\widehat{f}_1(x_i, x_j)}{\widehat{f}_1(x_i)\ \widehat{f}_1(x_j)} - \ln \frac{\widehat{f}_2(x_i, x_j)}{\widehat{f}_2(x_i)\ \widehat{f}_2(x_j)} \right\} d\boldsymbol{x}$$

$$= \widehat{J}(\widehat{f}_{1_\tau}, \widehat{f}_{2_\tau}; \widetilde{f}_1, \widetilde{f}_2).$$

∎

D. Design of the simulation experiments

Model	Graph	Mean values	Concentration	Asymp. error rates	
		$\boldsymbol{\mu_1},\ \boldsymbol{\mu_2}$	matrix $\boldsymbol{K}$	LDA	$DLDA$
$AR(1)$	tree τ	$\mathbf{0}$, $a\boldsymbol{v}_{max}$	$\boldsymbol{K_\tau}$	10.00	10.00
$\rho = .9$		$\mathbf{0}$, $b\boldsymbol{v}_{min}$		10.00	10.00
		$\mathbf{0}$, $\boldsymbol{u}_{rand}$		10.00	40.63
$MA(1)$	complete κ	$\mathbf{0}$, $a\boldsymbol{v}_{max}$	$\boldsymbol{K_\kappa}$	10.00	10.00
$\rho = .45$		$\mathbf{0}$, $b\boldsymbol{v}_{min}$		10.00	10.00
		$\mathbf{0}$, $\boldsymbol{u}_{rand}$		10.00	14.92
ECM	complete κ	$\mathbf{0}$, $a\boldsymbol{v}_{max}$	$\boldsymbol{K_\kappa}$	10.00	10.00
$\rho = .9$		$\mathbf{0}$, $b\boldsymbol{v}_{min}$		9.99	9.99
		$\mathbf{0}$, $\boldsymbol{u}_{rand}$		10.00	45.94
$RAND$	random G_1	$\mathbf{0}$, $a\boldsymbol{v}_{max}$	$\boldsymbol{K_{G_1}}$	10.00	10.00
		$\mathbf{0}$, $b\boldsymbol{v}_{min}$		10.00	10.00
		$\mathbf{0}$, $\boldsymbol{u}_{rand}$		10.00	15.32

Table D.1: Design of the Numerical simulation. Independent samples are generated from two corresponding densities of $N(\boldsymbol{\mu}_1, \boldsymbol{K})$ and $N(\boldsymbol{\mu}_2, \boldsymbol{K})$.
$\boldsymbol{v}_{max}$ is the largest and $\boldsymbol{v}_{min}$ the smallest eigenvalue of $\boldsymbol{\Sigma} = \boldsymbol{K}^{-1}$, and $\boldsymbol{u}_{rand}$ is a vector with random numbers from $U(0, t)$. a, b and t are constants such that the asymptotic error rate value for LDA is around 10%

Model	Graphs	Mean values	Concentration	Asymp. error rates for QDA		
		$\boldsymbol{\mu_1}, \boldsymbol{\mu_2}$	matrices $\boldsymbol{K_1}$, $\boldsymbol{K_2}$	Group 1	Group 2	Global
$\big(RAND, AR(1)\big)$ $\rho = .3$	(G_2, τ)	$\mathbf{0}$, $\boldsymbol{u}_{rand}$	$\boldsymbol{K}_{G_2}$, $\boldsymbol{K}_{\tau}$	4.90	4.76	4.83
$\big(RAND, MA(1)\big)$ $\rho = .275$	(G_2, κ)	$\mathbf{0}$, $\boldsymbol{u}_{rand}$	$\boldsymbol{K}_{G_2}$, $\boldsymbol{K}_{\kappa}$	5.33	4.86	5.10
$\big(RAND, ECM\big)$ $\rho = .2$	(G_2, κ)	$\mathbf{0}$, $\boldsymbol{u}_{rand}$	$\boldsymbol{K}_{G_2}$, $\boldsymbol{K}_{\kappa}$	4.49	5.53	5.01
$\big(RAND, RAND\big)$	(G_2, G_1)	$\mathbf{0}$, $\boldsymbol{u}_{rand}$	$\boldsymbol{K}_{G_2}$, $\boldsymbol{K}_{G_1}$	5.60	5.03	5.32

Table D.2: Design of the Numerical simulation. Independent samples are generated from two corresponding densities of $N(\boldsymbol{\mu}_1, \boldsymbol{K}_1)$ and $N(\boldsymbol{\mu}_2, \boldsymbol{K}_2)$.
$\boldsymbol{u}_{rand}$ is a vector with random numbers from $U(0, t)$, where t is a constant such that the asymptotic global error rate value for QDA is around 5%

E. Supplementary figures and details

a) Generation of the random concentration matrix associated with a PLN

Let p_k be the fraction of vertices in the graph that have degree k. A PLN is a graph with $p_k \propto k^{-\alpha}$, where α is the power parameter. We consider $\alpha = 2.3$ and simulate the two networks with graphs presented in Figure 4.3. Given a specific network with graph $G = (V, E)$, we use the following procedure to specify the associated covariance matrix. Let $\boldsymbol{A}$ be a matrix with entries

$$a_{ij} = \begin{cases} u_{ij} & \text{if } (i,j) \in E, \\ 0 & \text{if } (i,j) \notin E, \end{cases}$$

where u_{ij} is a random number from a uniform distribution $U(D)$. Then the diagonal elements of $\boldsymbol{A}$ are defined such that the final matrix is a diagonally dominant matrix, i.e., $a_{ii} = R \times \sum_{j \neq i} |a_{ij}|,\ i = 1, ..., p$, where $R > 1$. The covariance matrix $\boldsymbol{\Sigma}$ is then determined by $\sigma_{ij} = a^{ij}/\sqrt{a^{ii}a^{jj}}$, where a^{ij} is the entry ij of $\boldsymbol{A}^{-1}$.

For the numerical study, the RAND model associated with a PLN with graph given in Figure 4.3a) uses $D = (-1, -0.5) \cup (0.5, 1)$ and $R = 1.01$, whereas the one with graph given in Figure 4.3b) uses $D = (-0.5, 0.5)$ and $R = 1.01$.

b) Supplementary Figures

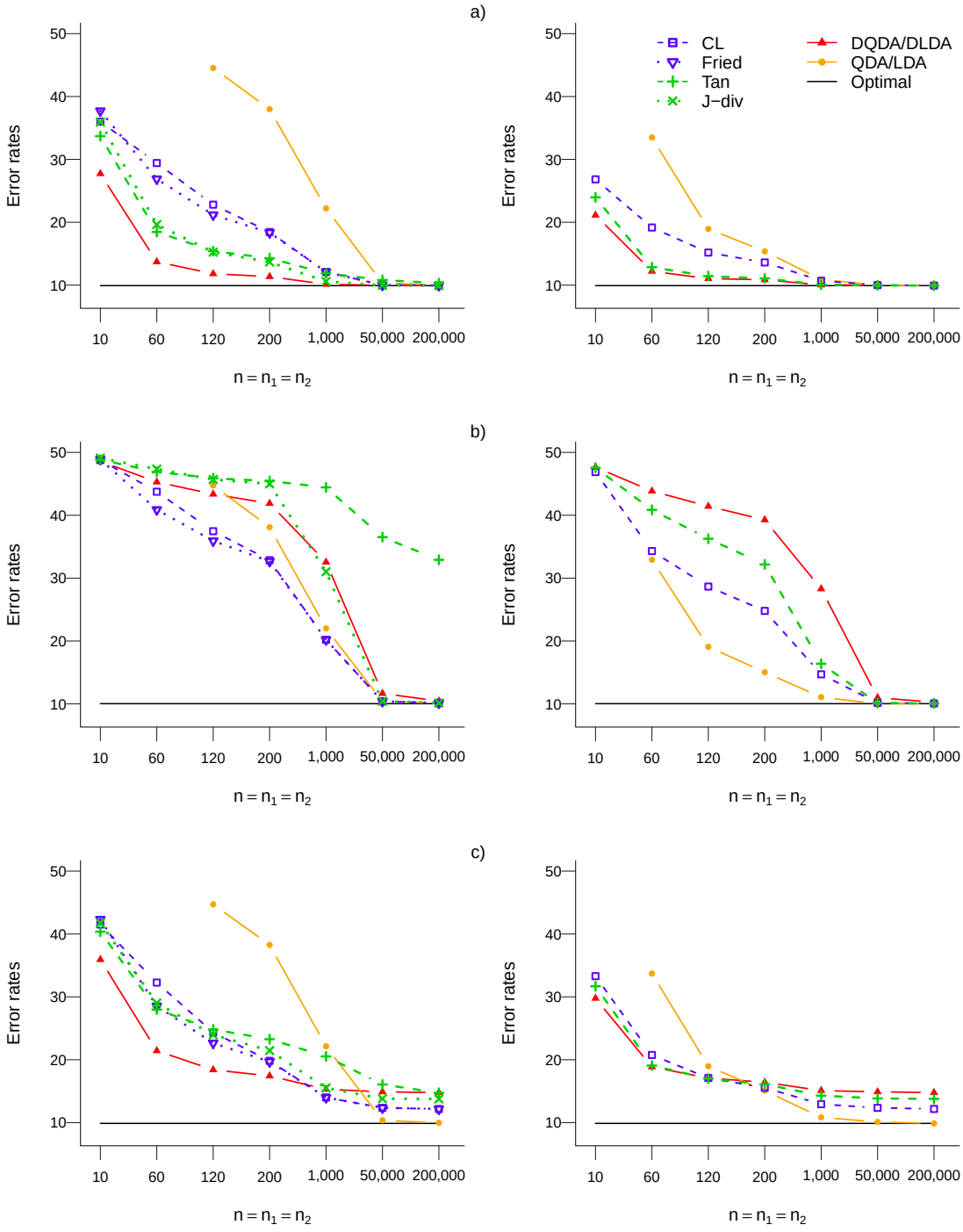


Figure E.1: MA(1). Estimated error rates with $p = 100$ variables and $n_1 = n_2 = n$ training samples on each group. On the left the heterogeneous cases and on the right the homogeneous.
$\boldsymbol{\Sigma}$ with entries $\sigma_{ij} = 0.45^{|i-j|} I(|i-j| \leq 1),\ i, j = 1, ..., p$, and
a) $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = (\mathbf{0}, a\boldsymbol{v}_{max})$, b) $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = (\mathbf{0}, b\boldsymbol{v}_{min})$, and c) $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = (\mathbf{0}, \boldsymbol{u}_{rand})$,
where $\boldsymbol{v}_{max}$ is the largest and $\boldsymbol{v}_{min}$ the smallest eigenvalue of $\boldsymbol{\Sigma}$, and $\boldsymbol{u}_{rand}$ is a vector with random numbers from $U(0, t)$. a, b and t are constants such that the asymptotic error rate value for LDA is around 10%

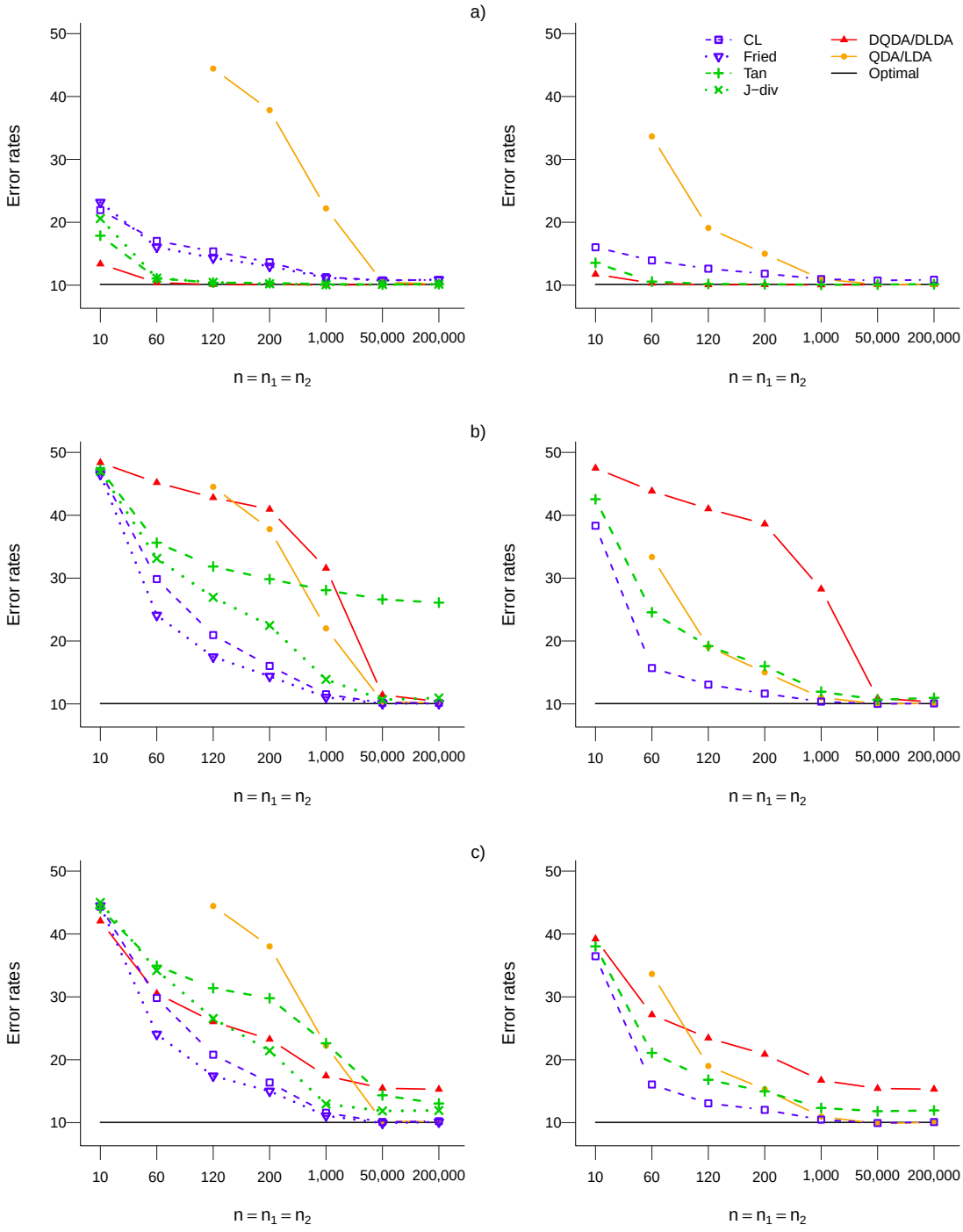


Figure E.2: RAND. Estimated error rates with $p = 100$ variables and $n_1 = n_2 = n$ training samples on each group. On the left the heterogeneous cases and on the right the homogeneous.
$\boldsymbol{\Sigma}$ is associated with a RAND model with graph G_1 given in Figure 4.3a), and
a) $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = (\mathbf{0}, a\boldsymbol{v}_{max})$, b) $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = (\mathbf{0}, b\boldsymbol{v}_{min})$, and c) $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = (\mathbf{0}, \boldsymbol{u}_{rand})$,
where $\boldsymbol{v}_{max}$ is the largest and $\boldsymbol{v}_{min}$ the smallest eigenvalue of $\boldsymbol{\Sigma}$, and $\boldsymbol{u}_{rand}$ is a vector with random numbers from $U(0, t)$. a, b and t are constants such that the asymptotic error rate value for LDA is around 10%

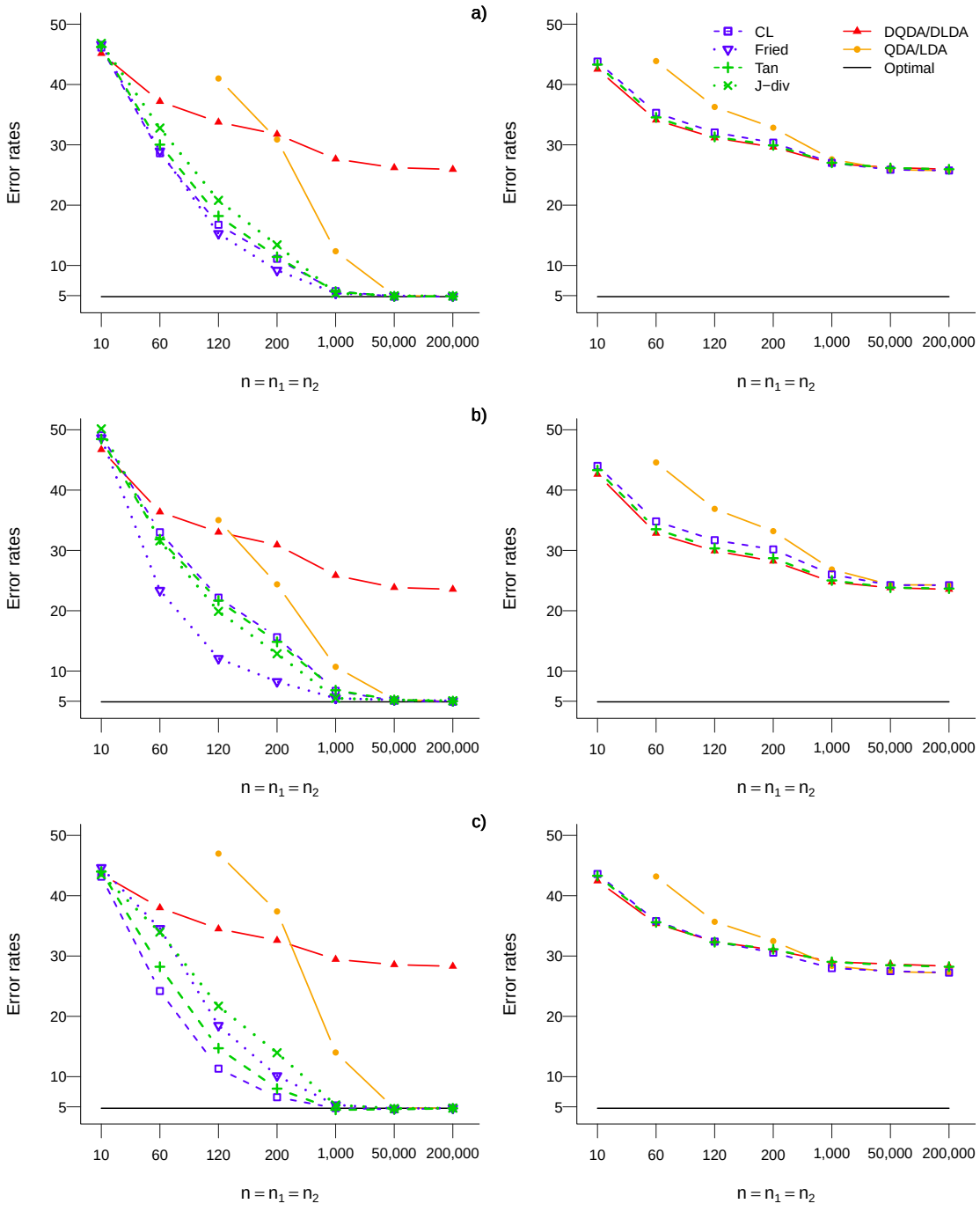


Figure E.3: (RAND, AR(1)). Estimated error rates with $p = 100$ variables and $n_1 = n_2 = n$ training samples. On the left the heterogeneous cases and on the right the homogeneous. a) Global, b) Group 1 and c) Group 2.

$\boldsymbol{\Sigma}_1$ is associated with a RAND model with graph G_2 given in Figure 4.3b), $\boldsymbol{\Sigma}_2$ has entries $\sigma_{ij}^{(2)} = 0.3^{|i-j|}$, $i, j = 1, ..., p$, and $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = (\mathbf{0}, \boldsymbol{u}_{rand})$, where $\boldsymbol{u}_{rand}$ is a vector with random numbers from $U(0, t)$ and t is such that the asymptotic error rate value for QDA is around 5%

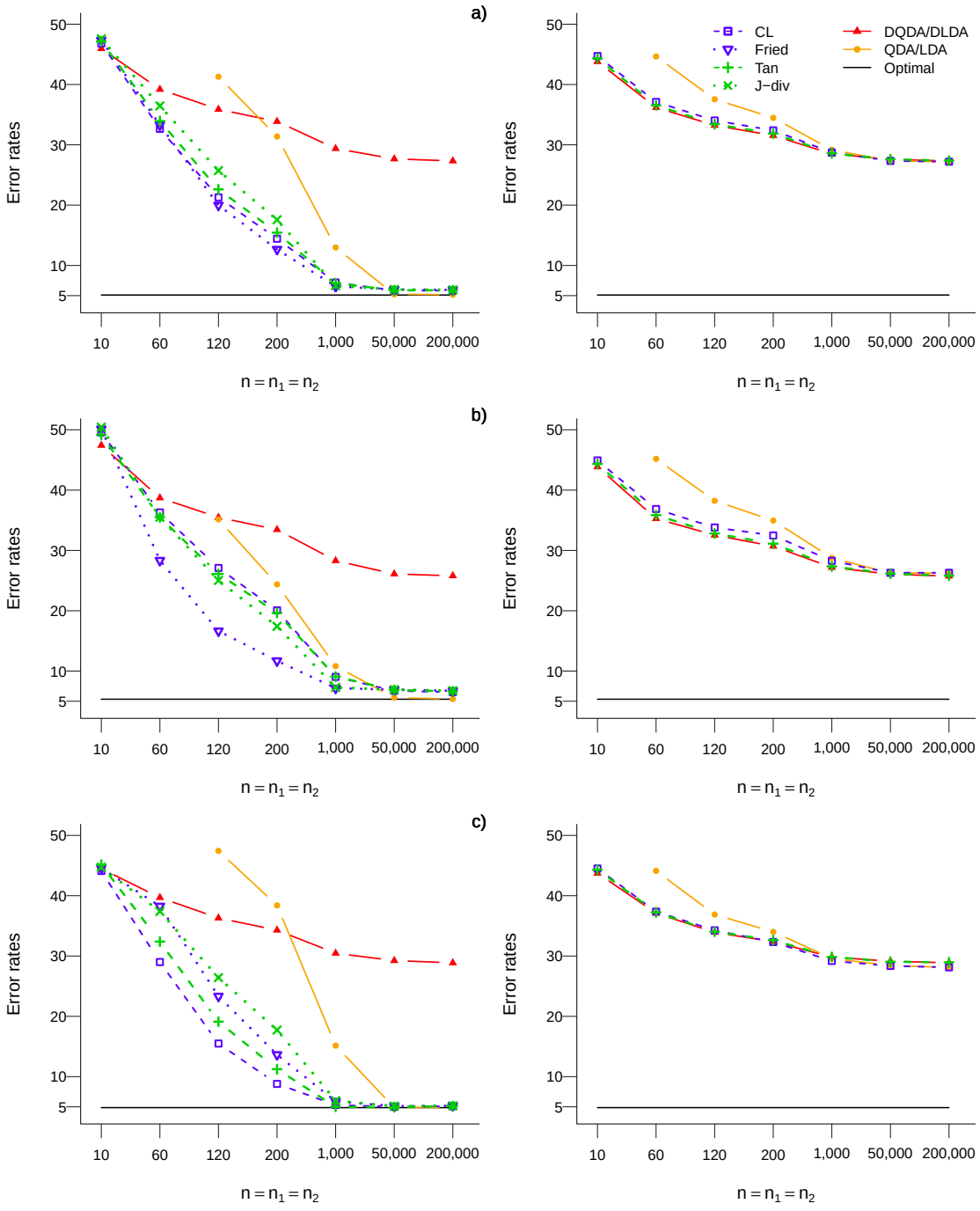


Figure E.4: (RAND, MA(1)). Estimated error rates with $p = 100$ variables and $n_1 = n_2 = n$ training samples. On the left the heterogeneous cases and on the right the homogeneous. a) Global, b) Group 1 and c) Group 2.

$\boldsymbol{\Sigma}_1$ is associated with a RAND model with graph G_2 given in Figure 4.3b), $\boldsymbol{\Sigma}_2$ has entries $\sigma_{ij}^{(2)} = 0.275^{|i-j|} I(|i-j| \leq 1)$, $i, j = 1, ..., p$, and $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = (\mathbf{0}, \boldsymbol{u}_{rand})$, where $\boldsymbol{u}_{rand}$ is a vector with random numbers from $U(0, t)$ and t is such that the asymptotic error rate value for QDA is around 5%

F. Estimated covariance matrices in the educational testing example

The ML estimates of $\widehat{\boldsymbol{\Sigma}}_1$ and $\widehat{\boldsymbol{\Sigma}}_2$ for the three types of equalities between corresponding elements of the two concentration matrices considered in Section 3.2.1, are the following

a)

$$\widehat{\boldsymbol{\Sigma}}_1 = \left(\begin{array}{cc|ccc|ccc} 7365 & 4846 & 5897 & 4406 & 6589 & 5897 & 4406 & 6589 \\ 4846 & 6299 & 5152 & 3559 & 5214 & 5152 & 3559 & 5214 \\ \hline 5897 & 5152 & 8540 & 5387 & 6994 & 5581 & 4405 & 6601 \\ 4406 & 3559 & 5387 & 5696 & 5672 & 4405 & 3688 & 5405 \\ 6589 & 5214 & 6994 & 5672 & 9720 & 6601 & 5405 & 7863 \\ \hline 5897 & 5152 & 5581 & 4405 & 6601 & 8540 & 5387 & 6994 \\ 4406 & 3559 & 4405 & 3688 & 5405 & 5387 & 5696 & 5672 \\ 6589 & 5214 & 6601 & 5405 & 7863 & 6994 & 5672 & 9720 \end{array}\right),$$

$$\widehat{\boldsymbol{\Sigma}}_2 = \left(\begin{array}{cc|ccc|ccc} 5974 & 2633 & 2575 & 2711 & 5138 & 2575 & 2711 & 5138 \\ 2633 & 4826 & 2358 & 1731 & 2766 & 2358 & 1731 & 2766 \\ \hline 2575 & 2358 & 3647 & 2072 & 3754 & 1842 & 1709 & 3150 \\ 2711 & 1731 & 2072 & 3630 & 4244 & 1709 & 2199 & 3768 \\ 5138 & 2766 & 3754 & 4244 & 9043 & 3150 & 3768 & 7104 \\ \hline 2575 & 2358 & 1842 & 1709 & 3150 & 3647 & 2072 & 3754 \\ 2711 & 1731 & 1709 & 2199 & 3768 & 2072 & 3630 & 4244 \\ 5138 & 2766 & 3150 & 3768 & 7104 & 3754 & 4244 & 9043 \end{array}\right).$$

b)

$$\widehat{\mathbf{\Sigma}}_1 = \widehat{\mathbf{\Sigma}}_2 = \left(\begin{array}{cc|ccc|cc|cc} 6669 & 3738 & 4236 & 3559 & 5864 & 4236 & 3559 & 5864 \\ 3738 & 5563 & 3754 & 2647 & 3990 & 3754 & 2647 & 3990 \\ \hline 4236 & 3754 & 6093 & 3731 & 5375 & 3713 & 3057 & 4874 \\ 3559 & 2647 & 3731 & 4664 & 4955 & 3057 & 2943 & 4588 \\ 5864 & 3990 & 5375 & 4955 & 9382 & 4874 & 4588 & 7483 \\ \hline 4236 & 3754 & 3713 & 3057 & 4874 & 6093 & 3731 & 5375 \\ 3559 & 2647 & 3057 & 2943 & 4588 & 3731 & 4664 & 4955 \\ 5864 & 3990 & 4874 & 4588 & 7483 & 5375 & 4955 & 9382 \end{array}\right).$$

c)

$$\widehat{\mathbf{\Sigma}}_1 = \left(\begin{array}{cc|ccc|cc|cc} 7659 & 4954 & 5843 & 4491 & 6670 & 5843 & 4491 & 6670 \\ 4954 & 6818 & 5400 & 3784 & 5245 & 5400 & 3784 & 5245 \\ \hline 5843 & 5400 & 8540 & 5387 & 6994 & 5581 & 4404 & 6601 \\ 4491 & 3784 & 5387 & 5697 & 5672 & 4404 & 3688 & 5403 \\ 6670 & 5245 & 6994 & 5672 & 9720 & 6601 & 5403 & 7864 \\ \hline 5843 & 5400 & 5581 & 4404 & 6601 & 8540 & 5387 & 6994 \\ 4491 & 3784 & 4404 & 3688 & 5403 & 5387 & 5697 & 5672 \\ 6670 & 5245 & 6601 & 5403 & 7864 & 6994 & 5672 & 9720 \end{array}\right),$$

$$\widehat{\mathbf{\Sigma}}_2 = \left(\begin{array}{cc|ccc|cc|cc} 5680 & 2523 & 2630 & 2627 & 5058 & 2630 & 2627 & 5058 \\ 2523 & 4307 & 2108 & 1510 & 2734 & 2108 & 1510 & 2734 \\ \hline 2630 & 2108 & 3646 & 2074 & 3756 & 1842 & 1709 & 3150 \\ 2627 & 1510 & 2074 & 3630 & 4244 & 1709 & 2199 & 3769 \\ 5058 & 2734 & 3756 & 4244 & 9043 & 3150 & 3769 & 7104 \\ \hline 2630 & 2108 & 1842 & 1709 & 3150 & 3646 & 2074 & 3756 \\ 2627 & 1510 & 1709 & 2199 & 3769 & 2074 & 3630 & 4244 \\ 5058 & 2734 & 3150 & 3769 & 7104 & 3756 & 4244 & 9043 \end{array}\right).$$

G. Resubstitution error rates in the Breast Cancer example

		Predicted											
		Discriminant Analysis									Logistic Classification		
		$\boldsymbol{K}_{\kappa}$			$\boldsymbol{K}_{\overline{\kappa}}$			$\boldsymbol{K}_{\tau}$			$\boldsymbol{K}_{\kappa}$		
Observed		Co	Ca	%	Co	Ca	%	Co	Ca	%	Co	Ca	%
$\boldsymbol{K}_1, \boldsymbol{K}_2$	Co	191	1	0.52	161	31	16.15	169	23	11.98			
	Ca	5	53	8.62	20	38	34.48	21	37	36.21			
	G			2.40			20.40			17.60			
# of param				1,804			164			244			
$\boldsymbol{K}_1 = \boldsymbol{K}_2$	Co	181	11	5.73	161	31	16.15	177	15	7.81	182	10	5.21
	Ca	21	37	36.21	17	41	29.31	20	38	34.48	16	42	27.59
	G			12.80			19.20			14.00			10.40
# of param				943			123			163			42

Table G.1: Resubstitution error rates considering 41 variables, those selected as the neighbours at distance less than 4 to variable class $C = 1001$: 108, 70, 97, 177, 213, 223, 228, 252, 254, 262, 318, 693, 430, 83, 34, 49, 75, 136, 154, 179, 198, 302, 329, 395, 402, 554, 604, 669, 781, 849, 653, 275, 525, 365, 462, 877, 190, 801, 880, 912, 942.
Co =Controls, Ca =Cases, G =Global. # of param = Number of estimated parameters including the corresponding to the mean vectors.

		Predicted														
		Discriminant Analysis												Logistic Regression		
		$\boldsymbol{K}_{\kappa}$			$\boldsymbol{K}_{\overline{\kappa}}$			$\boldsymbol{K}_{f}$			$\boldsymbol{K}_{G}$			$\boldsymbol{K}_{\kappa}$		
Observed		Co	Ca	%	Co	Ca	%	Co	Ca	%	Co	Ca	%			
$\boldsymbol{K}_1, \boldsymbol{K}_2$	Co	192	0	0.00	170	22	11.46	176	16	8.33	177	15	7.81			
	Ca	0	58	0.00	10	48	17.24	9	49	15.52	8	50	13.79			
	G			0.00			12.80			10.00			9.20			
# of param				3,190			220			320			354			
$\boldsymbol{K}_1 = \boldsymbol{K}_2$	Co	191	1	0.52	170	22	11.46	178	14	7.29	177	15	7.81	192	0	0.00
	Ca	4	54	6.90	10	48	17.24	7	51	12.07	6	52	10.34	0	58	0.00
	G			2.00			12.80			8.40			8.40			0.00
# of param				1,650			165			270			287			56

Table G.2: Resubstitution error rates considering models with 55 variables: 3, 48, 69, 79, 83, 108, 120, 132, 138, 160, 207, 209, 277, 279, 287, 318, 326, 328, 347, 363, 374, 377, 384, 414, 415, 418, 430, 466, 508, 525, 567, 591, 594, 603, 652, 654, 656, 663, 693, 701, 712, 724, 754, 775, 802, 803, 807, 812, 885, 915, 921, 923, 967, 968, 987.
Co =Controls, Ca =Cases, G =Global. # of param = Number of estimated parameters including the corresponding to the mean vectors.

		Predicted														
		Discriminant Analysis												Logistic Regression		
		$\boldsymbol{K}_{\kappa}$			$\boldsymbol{K}_{\overline{\kappa}}$			$\boldsymbol{K}_{f}$			$\boldsymbol{K}_{G}$			$\boldsymbol{K}_{\kappa}$		
Observed		Co	Ca	%	Co	Ca	%	Co	Ca	%	Co	Ca	%			
$\boldsymbol{K}_1, \boldsymbol{K}_2$	Co	186	6	3.13	180	12	6.25	177	15	7.81	181	11	5.73			
	Ca	7	51	12.07	12	46	20.69	12	46	20.69	12	46	20.69			
	G			5.20			9.60			10.80			9.20			
# of param				270			60			74			92			
$\boldsymbol{K}_1 = \boldsymbol{K}_2$	Co	188	4	2.08	179	13	6.77	184	8	4.17	187	5	2.60	188	4	2.08
	Ca	8	50	13.79	11	47	18.97	11	47	18.97	10	48	17.24	4	54	6.90
	G			4.80			9.60			7.60			6.00			3.20
# of param				150			45			52			61			16

Table G.3: Resubstitution error rates considering models with 15 variables: 3, 79, 132, 328, 347, 374, 415, 525, 567, 591, 654, 885, 915, 923, 987.
Co =Controls, Ca =Cases, G =Global. # of param = Number of estimated parameters including the corresponding to the mean vectors.

H. R scripts used to obtain the numerical results in Chapter 2

H.1 Script for the rabbits example.

```
rm(list = ls())
library(gRc)
X=matrix(c(
 5.0,  4.8,  4.3,  5.1,  4.1,  4.0,  7.1,  5.9,  5.3,  5.3,  5.3,  5.9,  6.5,  6.3,  6.6,  6.2,
 4.9,  5.0,  4.3,  5.3,  4.1,  4.0,  6.9,  6.3,  5.2,  5.5,  5.5,  5.9,  6.8,  6.3,  6.6,  6.3,
15.0, 14.2, 12.8, 14.4, 11.0, 12.5, 19.6, 15.9, 14.1, 14.5, 16.3, 16.4, 18.6, 18.1, 17.3, 18.1,
15.2, 14.3, 12.8, 14.6, 11.0, 12.6, 19.5, 15.8, 13.8, 14.8, 15.7, 16.2, 19.0, 17.4, 17.5, 17.7
), nrow=16, ncol=4)
X=as.data.frame(X)
names(X)[1:4] <- c("X1", "X2", "X3", "X4")
vcc = list(~X1+X2,~X3+X4)
ecc = list(~X1:X2, ~X1:X3+X1:X4+X2:X3+X2:X4, ~X3:X4)
####H(vc)
m1 <- rcox( vcc=vcc, ecc=ecc, data=X, method='matching')
summary(m1)
summary(m1, "K")
summary(m1, "KC")
K=summary(m1, "K")$K
solve(K)*15
mean(X)

###H(mvc)
media1=mean(c(X$X1, X$X2))
media2=mean(c(X$X3, X$X4))
X2=X
X2$X1=X$X1-media1
X2$X2=X$X2-media1
X2$X3=X$X3-media2
X2$X4=X$X4-media2
n=nrow(X2)
p=ncol(X2)
S=matrix(0,p,p)
for(i in 1:n){
```

```
S=S+t(X2[i,])%*%as.matrix(X2[i,])
}
S=S/n
m2 <- rcox( vcc=vcc, ecc=ecc, S=S, n=n, method='matching')
summary(m2)
summary(m2, "K")
summary(m2, "KC")
K2=summary(m2, "K")$K
(K2)*(n)/(n-1)
solve(K2)*(n)
media1
media2
```

H.2 Scripts for breast cancer data set.

Leave-one-out cross-validation

```
rm(list = ls())
#library(gRc)
library(doBy)
DatosBreastCancer0 <-
  read.table("C:/Users/TOSHIBA/Documents/GONZALO/Doctorado/Datos_breastCancer_West2001_Speed2003
  /Base4.csv", header=TRUE, sep=",", na.strings="NA", dec=".", strip.white=TRUE)
DatosBreastCancerE=DatosBreastCancer0[,1:3]
numren=nrow(DatosBreastCancer0)
DatosBreastCancerE$DQDAG10W=0
DatosBreastCancerE$DLDAG10W=0
DatosBreastCancerE$DQDAG10t=0
DatosBreastCancerE$DLDAG10t=0
DatosBreastCancerE$DQDAG10Wv=0
DatosBreastCancerE$DLDAG10Wv=0
DatosBreastCancerE$DQDAG10tv=0
DatosBreastCancerE$DLDAG10tv=0
n1=sum(DatosBreastCancer0$ER.status=="ER+")
n2=sum(DatosBreastCancer0$ER.status=="ER-")
for(j in 1:numren){
index=0
DatosBreastCancer=0
DatosBreastCancerj=0
DatosBreastCancer=DatosBreastCancer0[-c(j),]
DatosBreastCancerj=DatosBreastCancer0[j,]
numcol=ncol(DatosBreastCancer)
index=1:numcol
index=as.data.frame(index)
index$statW[1]=1000000000000000
index$statW[2]=1000000000000000
index$statW[3]=1000000000000000
index$statt[1]=1000000000000000
index$statt[2]=1000000000000000
index$statt[3]=1000000000000000
for(i in 4:numcol){
```

```
index$statW[i]=abs(wilcox.test(DatosBreastCancer[,i] ~ DatosBreastCancer[,1],
                            alternative="two.sided")[["statistic"]])
index$statt[i]=abs(t.test(DatosBreastCancer[,i]~DatosBreastCancer[,1],
                            alternative='two.sided', conf.level=.95,
var.equal=TRUE)[["statistic"]])
}
index <- orderBy(~-statW, data=index )
for(i in 1:numcol){
index$ordenW[i]=i}
index <- orderBy(~-statt, data=index )
for(i in 1:numcol){
index$ordent[i]=i}
index <- orderBy(~+index, data=index )
DatosBreastCancer10W=0
DatosBreastCancerj10W=0
DatosBreastCancer10W=DatosBreastCancer[,index$ordenW<=13]
DatosBreastCancerj10W=DatosBreastCancerj[,(index$ordenW<=13 & index$ordenW>=4)]
DatosBreastCancer10_1W=0
DatosBreastCancer10_2W=0
DatosBreastCancer10_1W=DatosBreastCancer[DatosBreastCancer$ER.status=="ER+",
                        (index$ordenW<=13 & index$ordenW>=4)]
DatosBreastCancer10_2W=DatosBreastCancer[DatosBreastCancer$ER.status=="ER-",
                        (index$ordenW<=13 & index$ordenW>=4)]
media1W=mean(DatosBreastCancer10_1W)
media2W=mean(DatosBreastCancer10_2W)
S1W=cov(DatosBreastCancer10_1W)
S1W=diag(diag(S1W), 10, 10)
S2W=cov(DatosBreastCancer10_2W)
S2W=diag(diag(S2W), 10, 10)
n1W=nrow(DatosBreastCancer10_1W)
n2W=nrow(DatosBreastCancer10_2W)
SW=(S1W*(n1W-1)+S2W*(n2W-1))/(n1W+n2W-2)
DatosBreastCancerE$DQDAG10Wv[j]=(1/2)*log(det(as.matrix(S2W))/det(as.matrix(S1W)))+
                        (1/2)*as.matrix(DatosBreastCancerj10W-media2W)%*%solve(as.matrix(S2W))
                        %*%t(as.matrix(DatosBreastCancerj10W-media2W))-
                        (1/2)*as.matrix(DatosBreastCancerj10W-media1W)%*%solve(as.matrix(S1W))
                        %*%t(as.matrix(DatosBreastCancerj10W-media1W))+log(n1/n2)
if (DatosBreastCancerE$DQDAG10Wv[j]>=0){
DatosBreastCancerE$DQDAG10W[j]="ER+"
}
if (DatosBreastCancerE$DQDAG10Wv[j]<0) {
DatosBreastCancerE$DQDAG10W[j]="ER-"
}
DatosBreastCancerE$DLDAG10Wv[j]=t(as.matrix(media1W-media2W))%*%solve(as.matrix(SW))%*%
                                t(as.matrix(DatosBreastCancerj10W-(1/2)*(media2W+media1W)))
                                +log(n1/n2)
if (DatosBreastCancerE$DLDAG10Wv[j]>=0){
DatosBreastCancerE$DLDAG10W[j]="ER+"
}
if (DatosBreastCancerE$DLDAG10Wv[j]<0) {
DatosBreastCancerE$DLDAG10W[j]="ER-"
}
```

```
DatosBreastCancer10t=0
DatosBreastCancerj10t=0
DatosBreastCancer10t=DatosBreastCancer[,index$ordent<=13]
DatosBreastCancerj10t=DatosBreastCancerj[,(index$ordent<=13 & index$ordent>=4)]
DatosBreastCancer10_1t=0
DatosBreastCancer10_2t=0
DatosBreastCancer10_1t=DatosBreastCancer[DatosBreastCancer$ER.status=="ER+",
                        (index$ordent<=13 & index$ordent>=4)]
DatosBreastCancer10_2t=DatosBreastCancer[DatosBreastCancer$ER.status=="ER-",
                        (index$ordent<=13 & index$ordent>=4)]
media1t=mean(DatosBreastCancer10_1t)
media2t=mean(DatosBreastCancer10_2t)
S1t=cov(DatosBreastCancer10_1t)
S1t=diag(diag(S1t), 10, 10)
S2t=cov(DatosBreastCancer10_2t)
S2t=diag(diag(S2t), 10, 10)
n1t=nrow(DatosBreastCancer10_1t)
n2t=nrow(DatosBreastCancer10_2t)
St=(S1t*(n1t-1)+S2t*(n2t-1))/(n1t+n2t-2)
DatosBreastCancerE$DQDAG10tv[j]=(1/2)*log(det(as.matrix(S2t))/det(as.matrix(S1t)))+
                        (1/2)*as.matrix(DatosBreastCancerj10t-media2t)%*%solve(as.matrix(S2t))
                                %*%t(as.matrix(DatosBreastCancerj10t-media2t))-
                        (1/2)*as.matrix(DatosBreastCancerj10t-media1t)%*%solve(as.matrix(S1t))
                                %*%t(as.matrix(DatosBreastCancerj10t-media1t))+log(n1/n2)
if (DatosBreastCancerE$DQDAG10tv[j]>=0){
DatosBreastCancerE$DQDAG10t[j]="ER+"
}
if (DatosBreastCancerE$DQDAG10tv[j]<0) {
DatosBreastCancerE$DQDAG10t[j]="ER-"
}
DatosBreastCancerE$DLDAG10tv[j]=t(as.matrix(media1t-media2t))%*%solve(as.matrix(St))%*%
                                t(as.matrix(DatosBreastCancerj10t-(1/2)*(media2t+media1t)))
                                +log(n1/n2)
if (DatosBreastCancerE$DLDAG10tv[j]>=0){
DatosBreastCancerE$DLDAG10t[j]="ER+"
}
if (DatosBreastCancerE$DLDAG10tv[j]<0) {
DatosBreastCancerE$DLDAG10t[j]="ER-"
}
}
table(DatosBreastCancerE$DLDAG10t,DatosBreastCancerE$ER.status)
table(DatosBreastCancerE$DLDAG10W,DatosBreastCancerE$ER.status)
table(DatosBreastCancerE$DQDAG10t,DatosBreastCancerE$ER.status)
table(DatosBreastCancerE$DQDAG10W,DatosBreastCancerE$ER.status)
```

Leave-one-out internal cross-validation

```
rm(list = ls())
#library(gRc)
library(doBy)
DatosBreastCancer0 <-
```

```
  read.table("C:/Users/TOSHIBA/Documents/GONZALO/Doctorado/Datos_breastCancer_West2001_Speed2003
           /Base4.csv",header=TRUE, sep=",", na.strings="NA", dec=".", strip.white=TRUE)
DatosBreastCancerE=DatosBreastCancer0[,1:3]
numren=nrow(DatosBreastCancer0)
DatosBreastCancerE$DQDAG10W=0
DatosBreastCancerE$DLDAG10W=0
DatosBreastCancerE$DQDAG10t=0
DatosBreastCancerE$DLDAG10t=0
DatosBreastCancerE$DQDAG10Wv=0
DatosBreastCancerE$DLDAG10Wv=0
DatosBreastCancerE$DQDAG10tv=0
DatosBreastCancerE$DLDAG10tv=0
n1=sum(DatosBreastCancer0$ER.status=="ER+")
n2=sum(DatosBreastCancer0$ER.status=="ER-")
index=0
numcol=ncol(DatosBreastCancer0)
index=1:numcol
index=as.data.frame(index)
index$statW[1]=1000000000000000
index$statW[2]=1000000000000000
index$statW[3]=1000000000000000
index$statt[1]=1000000000000000
index$statt[2]=1000000000000000
index$statt[3]=1000000000000000
for(i in 4:numcol){
index$statW[i]=abs(wilcox.test(DatosBreastCancer0[,i] ~ DatosBreastCancer0[,1],
                        alternative="two.sided")[["statistic"]])
index$statt[i]=abs(t.test(DatosBreastCancer0[,i]~DatosBreastCancer0[,1],
                        alternative='two.sided', conf.level=.95,
var.equal=TRUE)[["statistic"]])
}
index <- orderBy(~-statW, data=index )
for(i in 1:numcol){
index$ordenW[i]=i}
index <- orderBy(~-statt, data=index )
for(i in 1:numcol){
index$ordent[i]=i}
index <- orderBy(~+index, data=index )
DatosBreastCancer0W=DatosBreastCancer0[,index$ordenW<=13]
DatosBreastCancer0t=DatosBreastCancer0[,index$ordent<=13]
library(MASS)
DatosBreastCancer0W.scal <- cmdscale(dist(DatosBreastCancer0W[,4:13]), k = 2, eig = T)
X11()
plot(DatosBreastCancer0W.scal$points, type = "n")
text(DatosBreastCancer0W.scal$points, labels = as.character(DatosBreastCancer0W[,1]),
     col = 1 + unclass(DatosBreastCancer0W[,1]), cex = 0.8)
DatosBreastCancer0t.scal <- cmdscale(dist(DatosBreastCancer0t[,4:13]), k = 2, eig = T)
X11()
plot(DatosBreastCancer0t.scal$points, type = "n")
text(DatosBreastCancer0t.scal$points, labels = as.character(DatosBreastCancer0t[,1]),
     col = 1 + unclass(DatosBreastCancer0t[,1]), cex = 0.8)
write.csv(DatosBreastCancer0W, file =
```

```
        "C:/Users/TOSHIBA/Documents/GONZALO/Doctorado/Datos_breastCancer_West2001_Speed2003/
        DatosBreastCancer0W.csv",row.names=FALSE)
write.csv(DatosBreastCancer0t, file =
        "C:/Users/TOSHIBA/Documents/GONZALO/Doctorado/Datos_breastCancer_West2001_Speed2003/
        DatosBreastCancer0t.csv",row.names=FALSE)
for(j in 1:numren){
DatosBreastCancerW=0
DatosBreastCancerjW=0
DatosBreastCancerW=DatosBreastCancer0W[-c(j),]
DatosBreastCancerjW=DatosBreastCancer0W[j,]
DatosBreastCancert=0
DatosBreastCancerjt=0
DatosBreastCancert=DatosBreastCancer0t[-c(j),]
DatosBreastCancerjt=DatosBreastCancer0t[j,]
DatosBreastCancer10W=0
DatosBreastCancerj10W=0
DatosBreastCancer10W=DatosBreastCancerW
DatosBreastCancerj10W=DatosBreastCancerjW[,-c(1,2,3)]
DatosBreastCancer10_1W=0
DatosBreastCancer10_2W=0
DatosBreastCancer10_1W=DatosBreastCancerW[DatosBreastCancerW$ER.status=="ER+",-c(1,2,3)]
DatosBreastCancer10_2W=DatosBreastCancerW[DatosBreastCancerW$ER.status=="ER-",-c(1,2,3)]
media1W=mean(DatosBreastCancer10_1W)
media2W=mean(DatosBreastCancer10_2W)
S1W=cov(DatosBreastCancer10_1W)
S1W=diag(diag(S1W), 10, 10)
S2W=cov(DatosBreastCancer10_2W)
S2W=diag(diag(S2W), 10, 10)
n1W=nrow(DatosBreastCancer10_1W)
n2W=nrow(DatosBreastCancer10_2W)
SW=(S1W*(n1W-1)+S2W*(n2W-1))/(n1W+n2W-2)
DatosBreastCancerE$DQDAG10Wv[j]=(1/2)*log(det(as.matrix(S2W))/det(as.matrix(S1W)))+(1/2)*
                            as.matrix(DatosBreastCancerj10W-media2W)%*%solve(as.matrix(S2W))
                            %*%t(as.matrix(DatosBreastCancerj10W-media2W))-(1/2)*
                            as.matrix(DatosBreastCancerj10W-media1W)%*%solve(as.matrix(S1W))
                            %*%t(as.matrix(DatosBreastCancerj10W-media1W))+log(n1/n2)
if (DatosBreastCancerE$DQDAG10Wv[j]>=0){
DatosBreastCancerE$DQDAG10W[j]="ER+"
}
if (DatosBreastCancerE$DQDAG10Wv[j]<0) {
DatosBreastCancerE$DQDAG10W[j]="ER-"
}
DatosBreastCancerE$DLDAG10Wv[j]=t(as.matrix(media1W-media2W))%*%solve(as.matrix(SW))%*%
                                t(as.matrix(DatosBreastCancerj10W-(1/2)*(media2W+media1W)))
                                +log(n1/n2)
if (DatosBreastCancerE$DLDAG10Wv[j]>=0){
DatosBreastCancerE$DLDAG10W[j]="ER+"
}
if (DatosBreastCancerE$DLDAG10Wv[j]<0) {
DatosBreastCancerE$DLDAG10W[j]="ER-"
}
DatosBreastCancer10t=0
```

```
DatosBreastCancerj10t=0
DatosBreastCancer10t=DatosBreastCancert
DatosBreastCancerj10t=DatosBreastCancerjt[,-c(1,2,3)]
DatosBreastCancer10_1t=0
DatosBreastCancer10_2t=0
DatosBreastCancer10_1t=DatosBreastCancert[DatosBreastCancert$ER.status=="ER+",-c(1,2,3)]
DatosBreastCancer10_2t=DatosBreastCancert[DatosBreastCancert$ER.status=="ER-",-c(1,2,3)]
media1t=mean(DatosBreastCancer10_1t)
media2t=mean(DatosBreastCancer10_2t)
S1t=cov(DatosBreastCancer10_1t)
S1t=diag(diag(S1t), 10, 10)
S2t=cov(DatosBreastCancer10_2t)
S2t=diag(diag(S2t), 10, 10)
n1t=nrow(DatosBreastCancer10_1t)
n2t=nrow(DatosBreastCancer10_2t)
St=(S1t*(n1t-1)+S2t*(n2t-1))/(n1t+n2t-2)
DatosBreastCancerE$DQDAG10tv[j]=(1/2)*log(det(as.matrix(S2t))/det(as.matrix(S1t)))+(1/2)*
                    as.matrix(DatosBreastCancerj10t-media2t)%*%solve(as.matrix(S2t))%*%
                    t(as.matrix(DatosBreastCancerj10t-media2t))-(1/2)*
                    as.matrix(DatosBreastCancerj10t-media1t)%*%solve(as.matrix(S1t))%*%
                    t(as.matrix(DatosBreastCancerj10t-media1t))+log(n1/n2)
if (DatosBreastCancerE$DQDAG10tv[j]>=0){
DatosBreastCancerE$DQDAG10t[j]="ER+"
}
if (DatosBreastCancerE$DQDAG10tv[j]<0) {
DatosBreastCancerE$DQDAG10t[j]="ER-"
}
DatosBreastCancerE$DLDAG10tv[j]=t(as.matrix(media1t-media2t))%*%solve(as.matrix(St))%*%
                                t(as.matrix(DatosBreastCancerj10t-(1/2)*(media2t+media1t)))
                                +log(n1/n2)
if (DatosBreastCancerE$DLDAG10tv[j]>=0){
DatosBreastCancerE$DLDAG10t[j]="ER+"
}
if (DatosBreastCancerE$DLDAG10tv[j]<0) {
DatosBreastCancerE$DLDAG10t[j]="ER-"
}
}
table(DatosBreastCancerE$DLDAG10t,DatosBreastCancerE$ER.status)
table(DatosBreastCancerE$DLDAG10W,DatosBreastCancerE$ER.status)
table(DatosBreastCancerE$DQDAG10t,DatosBreastCancerE$ER.status)
table(DatosBreastCancerE$DQDAG10W,DatosBreastCancerE$ER.status)
```