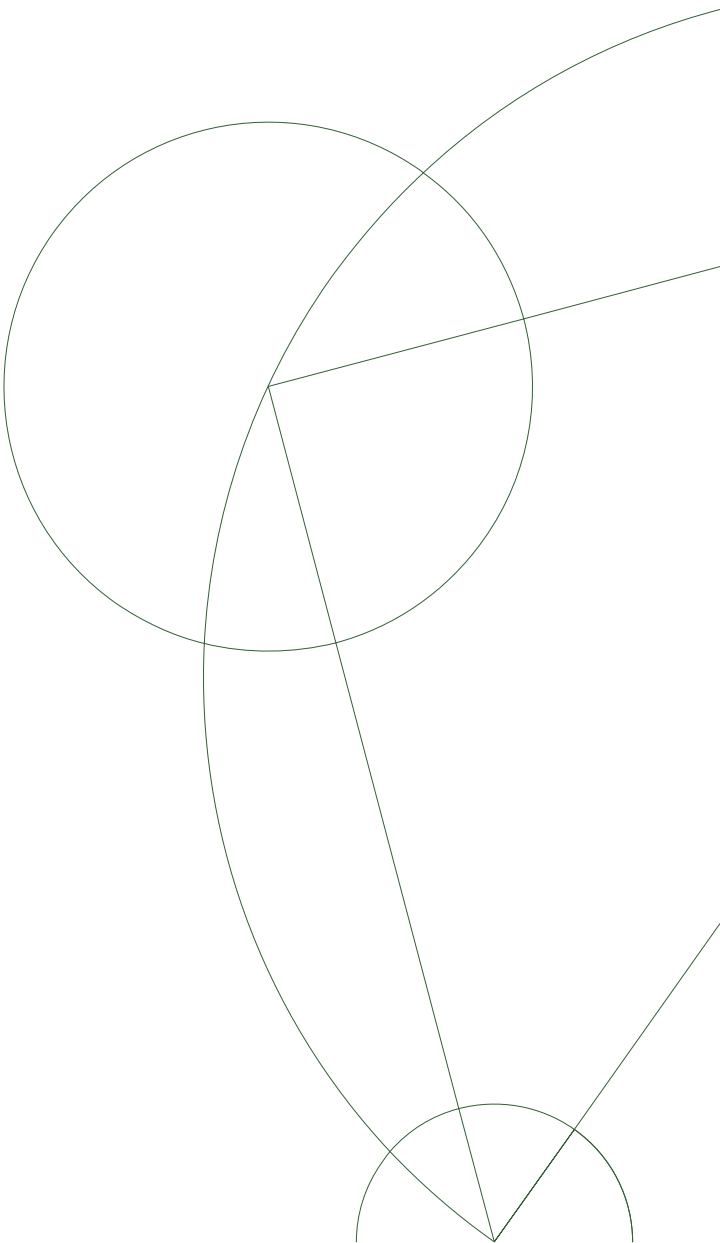
University of Copenhagen
Department of Mathematical Sciences

Tryggvi Sigurpálsson Kamban

Total Positivity in Graphical Models

Master's Thesis in Statistics



Advisor
Steffen L. Lauritzen

2nd of March, 2020

Abstract

We discuss properties of distributions with *multivariate totally positive of order 2* (MTP_2) density functions. In order to do this we first present concepts from convex analysis and optimization theory. We present results regarding the MTP_2 property for arbitrary distributions, and for graphical models. Under MTP_2 we show that if the conditional independence model for some distribution P is a graphoid, then P is automatically faithful to its pairwise independence graph. We go into further detail regarding quadratic exponential families, particularly Gaussian graphical models and Ising models. Notably, we show that an MLE can exist with as few observations as the maximal clique size for binary MTP_2 models on graphs, and an MLE in an MTP_2 Gaussian model exists with probability 1 if there are at least 2 observations. These are both substantial upgrades from the requirements of unrestricted binary and Gaussian models where the number of observations n needs to exceed 2^d and d, respectively, where d is the number of variables. Finally we discuss quadratic exponential families with sample space consisting of binary and continuous variables, and show that they are conditional Gaussian distributions with the associated discrete marginal being an Ising model.

Contents

Introduction

The objective of this master's thesis was to tie together the three articles Fallat et al. (2017), Lauritzen, Uhler, and Zwiernik (2017), and Lauritzen, Uhler, and Zwiernik (2019). The subject of all three articles is *multivariate total positivity of order 2* (MTP_2). A function f on some sample space $\mathcal{X}$ is said to be *multivariate totally positive of order 2* (MTP_2) if

$$f(x)f(y) \leq f(x \wedge y)f(x \vee y) \quad \text{for all } x, y, \in \mathcal{X},$$

where $x \wedge y$ and $x \vee y$ represent the elementwise minimum and maximum, i.e.,

$$x \wedge y = \left(\min\left(x_v, y_v\right), v \in V\right), \qquad x \vee y = \left(\max\left(x_v, y_v\right), v \in V\right).$$

Results in this thesis and their proofs have been taken from these papers unless otherwise indicated.

In chapter 1 we introduce concepts from *convex analysis* and *convex optimization*. The tools described will be immensely helpful to prove some of the central statements regarding maximum likelihood estimation contained in this thesis. This hardly comes as a surprise when one considers that maximum likelihood estimation problems are optimization problems, and if a maximum likelihood estimation problem is convex – which is the case for the problems considered in this thesis – it is certainly a convex optimization problem. The maximum likelihood estimator in a convex optimization problem is, if it exists, uniquely defined, see Barndorff-Nielsen (2014). We introduce the concept of *cones* and important examples of cones, such as the set of all $p \times p$ dimensional positive definite matrices. We discuss primal and dual feasibility and optimality, and finally we introduce *Slater's condition* and the *Karush-Kuhn-Tucker conditions* (KKT) which any solution to a convex optimization problem must satisfy.

In chapter 2 we show general results for distributions that have the MTP_2 property. For instance it is true that any MTP_2 distribution is closed under marginalization and conditioning. The proof that MTP_2 is closed under marginalization is similar to that of the four functions theorem (Ahlswede and Daykin, 1978). It states that given four nonnegative functions f_1, f_2, f_3, f_4 on some sample space $\mathcal{X} \subseteq \mathbb{R}^V$ which satisfy for all $x, y \in \mathcal{X}$ the inequality

$$f_1(x)f_2(y) \leq f_3\left(x \wedge y\right) f_4\left(x \vee y\right)$$

where V is the set of variables and $x \wedge y$ and $x \vee y$ are the coordinatewise minimum and maximum, respectively, then it holds that

$$\int f_1(x)dx \int f_2(x)dx \leq \int f_3(x)dx \int f_4(x)dx.$$

We also present conditional independence models, in particular graphical independence models (graphical models), and make a connection between the MTP_2 property and faithfulness of distributions to graphs.

Chapter 3 is about binary distributions. We discuss the consequences of MTP_2 on such distributions, and give necessary and sufficient criteria for existence of an MLE. The concept of lattices is introduced as they are useful objects in this research. We finish the chapter by the discussion of existence of an MLE in binary models over graphs. The conditions that need to hold for all variables in MTP_2 binary models now need only hold for variables connected by edges in the independence graph.

In chapter 4 we discuss exponential families and in particular *quadratic exponential families*. The quadratic variant has a particular form including a quadratic term, hence the name. The most well-known example of a quadratic exponential familiy is the Gaussian distribution. We will characterize the subfamily of MTP_2 distributions in quadratic exponential families before going into greater detail with Gaussian graphical models. In the spring of 2019, we wrote a project outside course scope about MTP_2 in Gaussian graphical models and a large part of that project is contained in the relevant sections of this chapter. In the climax of the chapter we show that under MTP_2, if we are given at least 2 observations then we can state with full certainty, that the MLE $\hat{\Sigma}$ for the covariance matrix exists for a Gaussian graphical model.

Finally, in chapter 5 we present the Ising model, another example of a quadratic exponential family, as well as *conditional Gaussian distributions* which are Gaussian distributions conditioned on landing in a cell in a discrete system. In this chapter we also have an original result regarding a special case of quadratic exponential families. We show that an MTP_2 quadratic exponential family, whose sample space has values both in $\{-1, 1\}$ and in $\mathbb{R}$, is automatically a conditional Gaussian distribution and the associated discrete marginal is an Ising model.

1. Convex Analysis and Optimization

Many important results contained in this thesis are shown using principles from convex analysis and optimization theory. In this chapter we will introduce these principles and lay the foundation on which further knowledge can be established.

1.1 Introduction to Convex Analysis

We will start off this chapter by giving definitions of basic mathematical objects in convex theory.

We will denote the *standard inner product* as $\langle \cdot, \cdot \rangle$. In $\mathbb{R}^p$ it is given by

$$\langle x, y \rangle = x^T y = \sum_{i=1}^{p} x_i y_i,$$

where $x, y \in \mathbb{R}^p$. We denote the *trace* $\operatorname{tr}(\cdot)$ of a matrix $A =\in \mathbb{R}^{p \times p}$ with entries (a_{ij}) by the sum of its diagonal entries, that is

$$\operatorname{tr}(A) = \sum_{i=1}^{p} a_{ii}.$$

On $\mathbb{R}^{p \times p}$ a standard inner product $\langle \cdot, \cdot \rangle$ is given by the *trace inner product*

$$\langle X, Y \rangle = \operatorname{tr}\left(X^T Y\right) = \sum_{i=1}^{p} \sum_{j=1}^{p} x_{ij} y_{ij}$$

for $X, Y \in \mathbb{R}^{p \times p}$.

Let $\mathcal{V}$ be a vector space. We say that a (non-empty) subset $C \subseteq \mathcal{V}$ is *convex* if for every $\theta \in [0, 1]$ it holds, for two given points $x, y \in C$, that $\theta x + (1 - \theta) y \in C$. In other words the line segment connecting the two points is fully contained in the set C.

It is easy to see that convexity of sets is preserved under intersection: if C_1 and C_2 are convex sets, then $C_1 \cap C_2$ is also convex. Indeed, if we let $x_1, x_2 \in C_1 \cap C_2$,

it holds for any $\theta \in [0,1]$ that $\theta x_1 + (1-\theta)x_2 \in C_1$ and $\theta x_1 + (1-\theta)x_2 \in C_2$, and so $\theta x_1 + (1-\theta)x_2 \in C_1 \cap C_2$.

We say that a function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if its domain, denoted by $\mathbf{dom} f$, is a convex set and if for all $x, y \in \mathbf{dom} f$, and $\theta \in [0,1]$ it holds that

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y).$$

A set $C \in \mathcal{V}$ is called a *cone* if for every $x \in C$ and $\theta \geq 0$ it holds that $\theta x \in C$. A set C is called a *convex cone* if it is both convex and a cone, that is for any $x, y \in C$ and $\theta_1, \theta_2 \geq 0, \theta_1 + \theta_2 > 0$, it holds that

$$\theta_1 x_1 + \theta_2 x_2 \in C$$

Convexity of cones is also preserved under intersection: If C_1 and C_2 are convex cones, then $C_1 \cap C_2$ is also a convex cone. This fact can be seen in much the same way it was with convex sets.

A *hyperplane* is a set of the form

$$\{x \in \mathcal{V} \mid \langle a, x \rangle = b\},$$

where $a \in \mathbb{R}^p \backslash \{\mathbf{0}\}$ and $b \in \mathbb{R}$. Suppose $C \subseteq \mathbb{R}^p$ and let x_0 be a point in its boundary ∂C, i.e.

$$x_0 \in \partial C = \overline{C} \backslash C^\circ$$

where $\overline{C}$ denotes the closure of C, and C° denotes the interior of C. If $a \neq 0$ satisfies $\langle a, x \rangle \leq \langle a, x_0 \rangle$ for all $x \in C$, then the hyperplane $\{x \in C \mid \langle a, x \rangle = \langle a, x_0 \rangle\}$ is called a *supporting hyperplane to C at the point x_0*.

Let C be a cone. The set

$$C^* = \{y \in \mathcal{V} \mid \langle x, y \rangle \geq 0 \text{ for all } x \in C\}$$

is called the *dual cone* of C. It is the set of all supporting hyperplanes to C, it is a cone, and it is always convex even if C is not (Boyd and Vandenberghe, 2004).

A hyperplane divides $\mathbb{R}^p$ into two *halfspaces*. Halfspaces $\mathcal{H}, \tilde{\mathcal{H}}$ are sets of the form

$$\mathcal{H} = \{x \in \mathcal{V} \mid \langle a, x \rangle \leq b\}, \qquad \tilde{\mathcal{H}} = \{x \in \mathcal{V} \mid \langle a, x \rangle > b\}$$

where $a \neq \{\mathbf{0}\}$.

A *polyhedron* is defined as the solution set of a finite number of linear equalities and inequalities:

$$\mathcal{P} = \{x \in \mathcal{V} \mid \langle a_i, x \rangle \leq b_i, i = 1, ..., m, \quad \langle c_j, x \rangle = d_j, j = 1, ..., n\}.$$

A polyhedron is thus the intersection of a finite number of closed halfspaces and hyperplanes. We call a set which is both a cone and a polyhedron a *polyhedral cone*. An example of this is the nonnegative orthant $\mathbb{R}^p_+$:

$$\mathbb{R}^p_+ = \{x \in \mathbb{R}^p \mid x_i \geq 0, i \in 1, ..., p\}.$$

We have now covered most of the basic definitions from convex set theory we will need in this thesis. In the next section we will introduce *M-matrices* – named so in honor of the late mathematician Hermann Minkowski – a type of matrix which is important for the subject of this thesis.

1.2 Positive Definite Matrices and M-matrices

We state the spectral theorem for symmetric matrices without proof, see for example Horn and Johnson (2012).

Theorem 1.2.1 (The spectral theorem for symmetric matrices)**.** *Suppose $A \in \mathbb{R}^{p\times p}$ is symmetric. Then eigenvectors $u_i \in \mathbb{R}^p$ corresponding to distinct eigenvalues λ_p, where $i \in \{1, ..., p\}$ are necessarily orthogonal:*

$$Au_1 = \lambda_1 u_1, \quad Au_2 = \lambda_2 u_2, \quad \lambda_1 \neq \lambda_2 \Rightarrow u_1 \cdot u_2 = 0.$$

In addition A is diagonalizable.

We now give a definition for positive definite and positive semidefinite matrices as they will be critical for the theory to come.

Definition 1.2.2. *A symmetric $p \times p$ matrix A is said to be* **positive (semi)definite** *if for any $x \in \mathbb{R}^p \backslash \{\mathbf{0}\}$ it holds that*

$$\left(x^T Ax \geq 0\right) \qquad x^T Ax > 0.$$

We will write $A \succ 0$ to say that A is positive definite, and $A \succeq 0$ to say that A is positive semidefinite. The following lemma lets us classify symmetric matrices as positive (semi)definite using their eigenvalues.

Lemma 1.2.3. *Let A be a symmetric real-valued $p \times p$ matrix with eigenvalues $\lambda_i, i \in \{1, ..., p\}$. Then $A \succeq 0 \iff \lambda_i \geq 0$ for all $i \in \{1, ..., p\}$.*

Proof.
"$\Leftarrow$": By the spectral theorem for symmetric matrices, eigenvectors x_i corresponding to distinct eigenvalues are necessarily orthogonal, and thus form an orthogonal

basis. Note that A is diagonalizable, so if some eigenvalues are equal we can choose our bases such that they are orthogonal. Thus any $x \in \mathbb{R}^p \backslash \{\mathbf{0}\}$ can be expressed as a unique linear combination of the eigenvectors, that is:

$$x = \sum_{i=1}^{p} a_i x_i \text{ for all } x \in \mathbb{R}^p \backslash \{\mathbf{0}\}$$

where the a_i's are some real coefficients. We get

$$\begin{aligned} x^T A x &= \sum_{i=1}^{p} \sum_{j=1}^{p} a_i a_j x_i^T A x^j \\ &= \sum_{i=1}^{p} \sum_{j=1}^{p} a_i a_j \lambda_i x_i^T x_j \\ &= \sum_{i=1}^{p} a_i^2 \lambda_i \left\| x_i \right\|^2 \geq 0. \end{aligned}$$

The second equality holds since x_i is an eigenvector of A, and the last equality holds since the eigenvectors form an orthogonal basis so $x_i^T x_j = 0$ when $i \neq j$. We conclude that $A \succeq 0$.

"$\Rightarrow$": Assume that x_0 is an eigenvector of A for $\lambda_0 < 0$. Then it holds that

$$x_0^T A x_0 = \lambda_0 \left\| x_0 \right\|^2 < 0$$

and so A is not positive semidefinite. Hence if $A \succeq 0$ its eigenvalues are necessarily nonnegative. □

Note that when determining whether a symmetric matrix is positive definite it is often more convenient to use Sylvester's criterion, namely, that a symmetric matrix A is positive definite if and only if all of the leading principal minors are positive (Gilbert, 1991). We now introduce a type of matrix which will be essential for the topic of total positivity in multivariate Gaussian graphical models.

Definition 1.2.4. *Let $M \in \mathbb{R}^{p \times p}$ be a symmetric matrix. We say that M is an* **M-matrix** *if $M_{ij} \leq 0$ for all $i \neq j$, and all eigenvalues of M are positive.*

A matrix A for which A^{-1} is an M-matrix is said to be an *inverse M-matrix*. If M is an M-matrix, then M is inverse-positive i.e. $(M^{-1})_{ij} \geq 0$ for all $i, j \in V$ Plemmons (1977).

Denote the vector space of all $p \times p$ symmetric matrices $\mathbb{S}^p := \{X \in \mathbb{R}^{p \times p} \mid X = X^T\}$. Using $\mathbb{S}^p$ we can construct the following sets: the set of all symmetric $p \times p$ positive definite matrices

$$\mathbb{S}^p_{\succ 0} := \{X \in \mathbb{S}^p \mid X \succ 0\},$$

the set of all symmetric $p \times p$ positive semidefinite matrices

$$\mathbb{S}^p_{\succeq 0} := \{X \in \mathbb{S}^p \mid X \succeq 0\},$$

and the set of M-matrices

$$\mathcal{M}^p := \{X \in \mathbb{S}^p \mid X \succ 0, X_{ij} \leq 0 \text{ for all } i \neq j\}.$$

These three sets are all convex cones. We show this first for $\mathbb{S}^p_{\succ 0}$ (the proof is analogous for $\mathbb{S}^p_{\succeq 0}$), and then $\mathcal{M}^p$ afterwards. Let $\theta_1, \theta_2 > 0$ and $A, B \in \mathbb{S}^p_{\succ 0}$. Then our claim is that $\theta_1 A + \theta_2 B \in \mathbb{S}^p_{\succ 0}$ as well. This can be seen from the definition of positive definiteness: for any $x \in \mathbb{R}^p \backslash \{\mathbf{0}\}$ it holds that

$$x^T(\theta_1 A + \theta_2 B)x = \theta_1 x^T A x + \theta_2 x^T B x > 0.$$

In order to show that $\mathcal{M}^p$ is a convex cone we define the coordinate half-spaces $\mathcal{H}^p_{ij}$ by

$$\mathcal{H}^p_{ij} = \{X \in \mathbb{S}^p \mid x_{ij} \leq 0 \text{ for all } i \neq j\}.$$

Note that they are polyhedrons by construction. Let $\theta > 0$ and $A \in \mathcal{H}^p_{ij}$. Then it holds that $\theta A_{ij} \leq 0$ for all $i \neq j$, and so $\theta A \in \mathcal{H}^p_{ij}$. Hence, $\mathcal{H}^p_{ij}$ is a cone. It is also convex, since if $A, B \in \mathcal{H}^p_{ij}$, then it holds that the off-diagonal entries in their convex combination will also be negative:

$$(\theta A + (1-\theta)B)_{ij} \leq 0 \quad \text{for} \quad i \neq j.$$

Hence, $\mathcal{H}^p_{ij}$ is a convex polyhedral cone. Since $\mathcal{M}^p = \mathbb{S}^p_{\succ 0} \cap_{ij} \mathcal{H}^p_{ij}$, which is an intersection between countably many convex cones, $\mathcal{M}^p$ is itself a convex cone. By construction it is also a polyhedral cone, a fact we will use later.

Henceforth all matrices mentioned are symmetric, unless specifically stated otherwise. We will now present some important dual cones.

Proposition 1.2.5. *The dual cone C^* of the positive definite cone $\mathbb{S}^p_{\succ 0}$ is the positive semidefinite cone $\mathbb{S}^p_{\succeq 0}$.*

Proof.
We need to show that for $X, Y \in \mathbb{S}^p$,

$$\operatorname{tr}(XY) \geq 0 \text{ for all } X \succ 0 \iff Y \succeq 0.$$

Assume that $Y \notin \mathbb{S}^p_{\succeq 0}$. Then there exists $q \in \mathbb{R}^p$ with

$$q^T Y q = \operatorname{tr}(qq^T Y) < 0.$$

Hence the positive definite matrix $X = qq^T \in \mathbb{S}^p_{\succ 0}$ satisfies $\operatorname{tr}(XY) < 0$, and so $Y \notin (\mathbb{S}^p_{\succ 0})^*$.

Assume now that $X \in \mathbb{S}^p_{\succ 0}$ and $Y \in \mathbb{S}^p_{\succeq 0}$. We can express X in terms of its eigenvalue decomposition as $X = \sum_{i=1}^p \lambda_i q_i q_i^T$, where the eigenvalues $\lambda_i > 0$ for $i = 1, ..., p$. Then we have

$$\operatorname{tr}(YX) = \operatorname{tr}\left(Y \sum_{i=1}^p \lambda_i q_i q_i^T\right) = \sum_{i=1}^p \lambda_i q_i^T Y q_i \geq 0.$$

This shows that $Y \in (\mathbb{S}^p_{\succ 0})^*$. □

The following statement holds as a direct consequence of Lemma 1.2.3.

Corollary 1.2.6. *Let A be a symmetric $p \times p$ matrix such that $A_{ij} \leq 0$ for all $i \neq j$. Then $A \in \mathcal{M}^p$ if, and only if, $A \succeq 0$.*

Example 1.2.7 (A 3×3 M-matrix)
Consider the matrix M.

$$M = \begin{bmatrix} 1 & -0.2 & 0 \\ -0.2 & 1 & -0.4 \\ 0 & -0.4 & 1 \end{bmatrix}$$

M is an M-matrix. Indeed, no off-diagonal entry is positive, and it has eigenvalues $\lambda_1 = 1, \lambda_2 = 1/\sqrt{5}, \lambda_3 = 1 - 1/\sqrt{5}$, which are all positive. ∘

We denote the closure of $\mathcal{M}^p$ by $\overline{\mathcal{M}}^p$ as the set of positive semidefinite M-matrices:

$$\overline{\mathcal{M}}^p := \{X \in \mathbb{S}^p \mid X \succeq 0, X_{ij} \leq 0 \text{ for all } i \neq j\}.$$

We now introduce two partial orders on matrices. Let $V = \{1, ..., p\}$, and let A, B be two $p \times p$ matrices. We write $A \geq B$ if $a_{ij} \geq b_{ij}$ for all $(i, j) \in V \times V$, and we write $A \succeq B$ if $A - B \in \mathbb{S}^p_{\succeq 0}$.

We define the cone $\mathcal{N}^p$ as the negative closure of $\mathbb{S}^p_{\succ 0}$, that is

$$\mathcal{N}^p := \{X \in \mathbb{S}^p \mid \exists Y \in \mathbb{S}^p_{\succ 0} \text{ with } Y \geq X \text{ and } \operatorname{diag}(Y) = \operatorname{diag}(X)\}.$$

Lemma 1.2.8. *The closure of $\mathcal{N}^p$ is the dual cone of $\mathcal{M}^p$, that is*

$$\overline{\mathcal{N}}^p = \{S \in \mathbb{S}^p \mid \langle S, K\rangle \geq 0, \text{ for all } K \in \mathcal{M}^p\}.$$

Proof.
Let C_1 and C_2 be two convex cones. Let $\boxplus$ denote the Minkowski sum, that is for sets A, B,

$$A \boxplus B = \{a + b \mid a \in A, b \in B\}.$$

Then by Jeyakumar and Wolkowicz (1992) the following equality holds:

$$(C_1 \cap C_2)^* = C_1^* \boxplus C_2^*. \tag{1.1}$$

Since $(\mathbb{S}^p_{\succ 0})^* = \mathbb{S}^p_{\succeq 0}$ and $\left(\mathcal{H}^p_{ij}\right)^* = \mathcal{H}^p_{ij}$, $\mathcal{M}^p = \mathbb{S}_{\succ 0} \cap_{i<j} \mathcal{H}^p_{ij}$, and (1.1) can be applied inductively to any finite collection of convex cones, the proof is complete. □

1.3 Convex Optimization

A convex optimization problem is one of the form

$$\begin{aligned} \min \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \leq b_i, \quad i = 1, ..., m, \end{aligned} \tag{1.2}$$

where the functions $f_0, ..., f_m : \mathbb{R}^p \to \mathbb{R}$ are convex. Some problems have, in addition to the linear inequality constraints, linear equality constraints. Such problems are irrelevant for the topic of this thesis, and they have therefore not been included. We note that minimizing a convex objective function is equivalent to maximizing a concave objective function.

Convex optimization problems are a core aspect of classical statistics in the form of *maximum likelihood estimation*. Consider a family of probability distributions on $\mathbb{R}^n$ indexed by a vector $\theta \in \mathbb{R}^p$ with densities $f_\theta(\cdot)$. The function $f_\theta(y)$, when considered as a function for a fixed $y \in \mathbb{R}^n$, is called the *likelihood function*. It is often more convenient to consider the *log-likelihood function* denoted l:

$$l(\theta) = \log f_\theta(y).$$

The maximum likelihood estimator (MLE) is the parameter

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmax}}\, l(\theta)$$

i.e. the value of θ that maximizes $l(\theta)$ for the observed value of y. The problem of finding a MLE of the parameter θ can be expressed as

$$\begin{aligned} \max \quad & l(\theta) = \log f_\theta(y) \\ \text{s.t.} \quad & \theta \in C \end{aligned}$$

where $l(\theta)$ is often concave, and $\theta \in C$ conveys some existing information, or other constraints on the parameter vector θ.

1.3.1 Lagrangian Duality

Consider the optimization problem in the standard form:

$$\begin{aligned} \min \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \leq 0, \quad i = 1, ..., m, \end{aligned} \tag{1.3}$$

with $x \in \mathbb{R}^n$. We assume its domain is non-empty, we call f_0 the *object function* and $p^\star$ the optimal value. Define the *Lagrangian* $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ associated with the problem (1.3) as

$$\mathcal{L}(x, \lambda) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x).$$

We refer to $\lambda_i \geq 0$ as the *Lagrange multiplier* associated with the i'th inequality $f_i(x) \leq 0$. The vector $\lambda = (\lambda_1, ..., \lambda_m)$ contains the *dual variables* associated with the problem (1.3).

We define the *Lagrange dual function* $g : \mathbb{R}^m \to \mathbb{R}$ as the minimum value of the Lagrangian over x, that is for $\lambda \in \mathbb{R}^m$,

$$g(\lambda) = \inf_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda) = \inf_{x \in \mathbb{R}^n} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) \right).$$

When the Lagrangian is unbounded below in x, the dual function takes on the value $-\infty$. The dual function yields lower bounds on the optimal value $p^\star$ of the problem (1.3): For any $\lambda \succeq 0$ we have

$$g(\lambda) \leq p^\star. \tag{1.4}$$

This is easily verified. Suppose $\tilde{x}$ is a feasible point for problem (1.3), that is $f_i(\tilde{x}) \leq 0$, and $\lambda \succeq 0$. Then we have

$$\sum_{i=1}^{m} \lambda_i f_i(\tilde{x}) \leq 0$$

since each term is nonpositive. Hence

$$\mathcal{L}(\tilde{x}, \lambda) = f_0(\tilde{x}) + \sum_{i=1}^{m} \lambda_i f_i(\tilde{x}) \leq f_0(\tilde{x}),$$

and therefore

$$g(\lambda) = \inf_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda) \leq \mathcal{L}(\tilde{x}, \lambda) \leq f_0(\tilde{x}).$$

Since $g(\lambda) \leq f_0(\tilde{x})$ is satisfied for any feasible point $\tilde{x}$, the inequality (1.4) holds.

Thus there exists a lower bound that depends on λ, but we wish to know which lower bound that can be obtained from the Lagrange dual function is the best. We get an optimization problem

$$\begin{aligned} \max \quad & g(\lambda) \\ \text{s.t.} \quad & \lambda \succ 0. \end{aligned} \tag{1.5}$$

We call (1.5) the *Lagrange dual problem* associated with *the primal problem* (1.3). If λ is feasible for the dual problem (1.5) we say that λ is *dual feasible*, and if $\lambda^\star$ is optimal for the dual problem (1.5), we say that it is *dual optimal*. The Lagrange dual problem (1.5) is a convex optimization problem, since the objective function to be maximized is concave and the constraint is convex.

The optimal value of the Lagrange dual problem denoted by $d^\star$ is, by definition, the best lower bound on $p^\star$ obtainable from the Lagrange dual function. Hence, we have the inequality

$$d^\star \leq p^\star.$$

This property is called *weak duality*. The *optimal duality gap* of the original problem is defined as the difference $p^\star - d^\star$. Note that it is always nonnegative. If the equality

$$d^\star = p^\star$$

is satisfied, we say that *strong duality* holds. Note that, equivalently, the optimal duality gap is zero. Strong duality means that the best bound obtainable from the Lagrange dual function is tight. It does not hold in general, but there are results that establish conditions on the problem for which strong duality holds, and we call these *constraint qualifications*.

A simple, yet important, constraint qualification is *Slater's condition* which states the following. There exists an x in the interior of the domain of of the primal problem, such that

$$f_i(x) < 0, \quad i = 1, ..., m.$$

We call such a point *strictly feasible*, since the inequality constraints are satisfied with strict inequalities. Slater's theorem states that strong duality holds if the problem is convex and Slater's condition holds.

1.3.2 Karush-Kuhn-Tucker optimality conditions

Assume that $f_0, ..., f_m$ are differentiable. Let $x^\star$ and $\lambda^\star$ be any primal and dual optimal points with zero duality gap. Since $x^\star$ minimizes $\mathcal{L}(x, \lambda^\star)$ over x, it follows that its gradient must vanish at $x^\star$, that is

$$\nabla f_0(x^\star) + \sum_{i=1}^{m} \lambda_i^\star \nabla f_i(x^\star) = 0.$$

Then we have the following conditions

$$\begin{aligned} f_i(x^\star) &\leq 0, \quad i = 1, ..., m \\ \lambda_i^\star &\geq 0, \quad i = 1, ..., m \\ \lambda_i^\star f_i(x^\star) &= 0, \quad i = 1, ..., m \\ \nabla f_0(x^\star) + \sum_{i=1}^{m} \lambda_i^\star \nabla f_i(x^\star) &= 0, \end{aligned} \tag{1.6}$$

which are called the *Karush-Kuhn-Tucker* (KKT) conditions. For any optimization problem with differentiable objective-, and constraint functions for which strong duality holds, any pair of primal and dual optimal points must satisfy the KKT conditions. When the primal problem is convex, the KKT conditions are also sufficient for the points to be primal and dual optimal.

If f_i are convex, and x', λ' are any points that satisfy the KKT conditions, then x' and λ' are primal and dual optimal with zero duality gap. Indeed, note first that the first condition states that x' is primal feasible. Since $\lambda_i' \geq 0$, it holds that $\mathcal{L}(x, \lambda')$ is convex in x, so the last KKT conditions states that its gradient with respect to x is zero when $x = x'$. Hence it follows that x' minimizes $\mathcal{L}(x, \lambda')$ over x. Using the third KKT condition, we can now conclude that

$$\begin{aligned} g(\lambda') &= \mathcal{L}(x', \lambda') \\ &= f_0(x') + \sum_{i=1}^{m} \lambda_i' f_i(x') \\ &= f_0(x'). \end{aligned}$$

This shows that x' and λ' have zero duality gap, and are therefore primal and dual optimal points.

As a final comment on this chapter, Slater's condition and the KKT conditions will prove to be very useful in the problems to come. In particular, we will be interested in showing the existence (with probability 1) of an MLE of the covariance matrix in an MTP_2 Gaussian model so long as we have at least 2 observations. The regular criterion for the existence of a MLE in a Gaussian model without multivariate total positivity is that we need at least as many observations as there are variables, so this upgrade is quite substantial. To this end, both Slater's condition and the KKT conditions will be crucial.

2. Total Positivity

In this chapter we will present general results regarding total positivity. We concern ourselves with statistics and results related to that school of mathematics. As such, we will be exploring probability density functions and point mass functions under MTP_2. For this reason, *totally positive functions*, in particular functions that are *multivariate totally positive of order 2* (MTP_2), will be of importance. Hence, we will cover basic results concerning them. In addition, we will introduce and present results for conditional independence models, both with and without the MTP_2 property.

2.1 Graphs

Before we do anything else, we establish the notation used for graphs in this thesis, along with some basic definitions.

An *undirected graph* $G = (V, E)$ consists of a nonempty set of *vertices* or *nodes* V and a set of undirected edges E. Graphs in this thesis are *simple* meaning that they have no self-loops and no multiple edges. We will write uv for an edge between u and v and say that the vertices u and v are *adjacent* or *neighbours*. A *path* in G is a sequence of nodes $(v_0, v_1, ..., v_k)$ such that $v_i v_{i+1} \in E$ for all $i = 0, ..., k-1$ and such that no node is repeated, that is $v_i \neq v_j$ for all $i, j \in \{0, 1, ..., k\}$ with $i \neq j$. As such, an edge is the shortest type of path. A *cycle* is a path with the modification that $v_0 = v_k$. We say that two distinct nodes $u, v \in V$ are *connected* if there exists a path between u and v. A graph is said to be *connected* if all pairs of distinct nodes are connected. A graph is *complete* if all possible edges are present. We say that two subsets $A, B \subset V$ are *separated* by $S \subset V \backslash (A \cup B)$ if every path between A and B passes through a node in S. A subgraph of G *induced* by a set $A \subset V$ consists of the nodes in A and of the edges in G between nodes in A. A subgraph G^* is *maximal complete* if it is complete and there exists no nodes $u \in G \setminus G^*$ such that $G^* \cup \{u\}$ is a complete graph. We call such subgraphs *cliques*. We say that a graph G is *bipartite* if there exist two subgraphs $A, B \subseteq G$ such that $A \cup B = G$ and all edges E are of the form ij where $i \in A$ and $j \in B$. Note that the maximal size of a clique in a bipartite graph is 2.

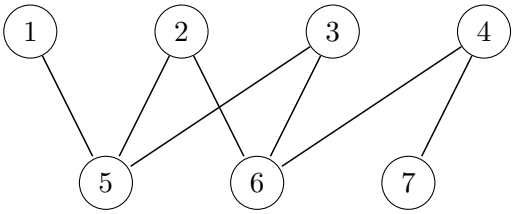


Figure 2.1: Bipartite undirected graph with the two subgraphs $A = \{1, 2, 3, 4\}$ and $B = \{5, 6, 7\}$.

Let $V = \{1, ..., d\}$ be a finite set and let $X = (X_v, v \in V)$ be a random vector, i.e. random variables with labels in V. We consider the product space $\mathcal{X} = \prod_{v \in V} \mathcal{X}_v$, where $\mathcal{X}_v \subseteq \mathbb{R}$ is the state space of X_v, inheriting the order from $\mathbb{R}$. The state spaces $\mathcal{X}_v$ are either discrete finite sets or open intervals on the real line. We can thus partition the set of variables as $V = \Delta \cup \Gamma$, where $\mathcal{X}_v$ is discrete if $v \in \Delta$ and $\mathcal{X}_v$ is an open interval if $v \in \Gamma$.

All distributions are assumed to have density with respect to the product measure $\mu = \bigotimes_{v \in V} \mu_v$, where μ_v is the counting measure for $v \in \Delta$, and μ_v is the Lebesgue measure giving length 1 to the unit interval for $v \in \Gamma$. We refer to μ as the *base measure*.

2.2 Definitions and Basic Results

A function f on $\mathcal{X}$ is said to be *multivariate totally positive of order 2* (MTP_2) if

$$f(x)f(y) \leq f(x \wedge y)f(x \vee y) \quad \text{for all } x, y, \in \mathcal{X}, \tag{2.1}$$

where $x \wedge y$ and $x \vee y$ represent the elementwise minimum and maximum, i.e.,

$$x \wedge y = \left(\min\left(x_v, y_v\right), v \in V\right), \qquad x \vee y = \left(\max\left(x_v, y_v\right), v \in V\right).$$

For $|V| = 2$, a function that is MTP_2 is called *totally positive of order 2* (TP_2). We say that X or the distribution of X is MTP_2 if its density function f is MTP_2.

A function g on $\mathcal{X}$ is *supermodular* if

$$g(x) + g(y) \leq g(x \wedge y) + g(x \vee y) \quad \text{for all } x, y, \in \mathcal{X},$$

submodular if

$$g(x) + g(y) \geq g(x \wedge y) + g(x \vee y) \quad \text{for all } x, y, \in \mathcal{X},$$

and *modular* if it is both supermodular and submodular. Note that g is supermodular on $\mathcal{X}$ if and only if $\exp(g)$ is MTP_2 on $\mathcal{X}$. We will now see some basic properties and results of MTP_2 distributions.

Proposition 2.2.1. *If $f : \mathbb{R}^V \to \mathbb{R}$ is MTP_2, and $y \in \mathbb{R}^V$, then the function $g : \mathbb{R}^V \to \mathbb{R}$ given by*

$$g(y) = \left(\prod_{i=1}^{d} a_v(x_v) \right) f\left(b_v(x_v), v \in V\right) \tag{2.2}$$

where each $a_v : \mathbb{R} \to \mathbb{R}$ is a positive function, and $b_v : \mathbb{R} \to \mathbb{R}$ are all nondecreasing, is MTP_2.

Proof.
Let $x, y \in \mathcal{X}$, and assume $f : \mathbb{R}^V \to \mathbb{R}$ is MTP_2. Then

$$\begin{aligned} g(x)g(y) = &\left(\prod_{v \in V} a_v(x_v) a_v(y_v) \right) \\ &\times f\left(b_v(x_v), v \in V\right) f\left(b_v(y_v), v \in V\right), \end{aligned}$$

and

$$\begin{aligned} g(x \wedge y)g(x \vee y) = &\left(\prod_{v \in V} a_v((x \wedge y)_v) a_v((x \vee y)_v) \right) \\ &\times f\left(b_v((x \wedge y)_v), v \in V\right) f\left(b_v((x \vee y)_v), v \in V\right). \end{aligned}$$

Since for each $v \in V$ it holds that

$$a_v((x \wedge y)_v) a_v((x \vee y)_v) = a_v(x_v) a_v(y_v),$$

the products over all $v \in V$ are also equal. Since b_v is assumed nondecreasing for each $v \in V$, it holds for all $v \in V$ that

$$b_v\left((x \wedge y)_v\right) = b_v(x_v) \wedge b_v(y_v) \quad \text{and} \quad b_v\left((x \vee y)_v\right) = b_v(x_v) \vee b_v(y_v),$$

and since f is MTP_2, it holds that

$$f\left(b_v(x_v), v \in V\right) f\left(b_v(y_v), v \in V\right) \leq f\left(b_v((x \wedge y)_v), v \in V\right) f\left(b_v((x \vee y)_v), v \in V\right).$$

We conclude that the inequality $g(x)g(y) \leq g(x \wedge y)g(x \vee y)$ holds for any $x, y \in \mathcal{X}$, and thus g is MTP_2. □

The following result shows that the MTP_2 property is preserved under increasing coordinate-wise transformations.

Proposition 2.2.2. *Let X be a random vector taking values in $\mathcal{X}$, and let X have density function f. Let $\phi = (\phi_v, v \in V)$ be such that $\phi_v : \mathbb{R} \to \mathbb{R}$ are strictly increasing and differentiable with derivative functions ϕ_v' for all $v \in V$. If X is MTP_2, then $Y = \phi(X)$ is MTP_2.*

Proof.
Consider equation (2.2) and let $b_v(y_v) = \phi_v^{-1}(y_v)$ and let $a_v(y_v) = 1/\phi_v'(\phi_v^{-1}(y_v))$. Then, since the inverse of an increasing function is increasing and by the transformation theorem for multivariate densities, $g(y)$ is the density of $Y = \phi(X)$ and we obtain from proposition 2.2.1 that Y is MTP_2. □

We now state the *four functions theorem* without proof as well as a result which follows from it immediately (see Theorem 2.1 and Corollary 2.2 in Karlin and Rinott, 1980.)

Theorem 2.2.3 (The four functions theorem)**.** *Let f_1, f_2, f_3, f_4 be nonnegative functions on $\mathcal{X}$ satisfying for all $x, y \in \mathcal{X}$ the inequality*

$$f_1(x)f_2(y) \;\leq\; f_3(x \wedge y)f_4(x \vee y). \tag{2.3}$$

Then

$$\int f_1(x)dx \int f_2(x)dx \;\leq\; \int f_3(x)dx \int f_4(x)dx.$$

Corollary 2.2.4. *Let $A, B \subseteq \mathcal{X}$ and define*

$$A \vee B = \{a \vee b : a \in A, b \in B\}, \qquad A \wedge B = \{a \wedge b : a \in A, b \in B\}.$$

Let f_1, f_2, f_3, f_4 be nonnegative functions on $\mathcal{X}$ satisfying (2.3) for all $x, y \in \mathcal{X}$. Then

$$\int_A f_1(x)dx \int_B f_2(x)dx \;\leq\; \int_{A\wedge B} f_3(x)dx \int_{A\vee B} f_4(x)dx.$$

We say that a function $\phi_v(x) : \mathcal{X}_v \to \mathbb{R}$ is *piecewise constant* if the image $\phi_v(\mathcal{X}_v)$ has finite size. We can then show that the MTP_2 property is preserved under piecewise constant and nondecreasing transformations.

Proposition 2.2.5. *Let X be a random vector taking values in $\mathcal{X}$. For $A \subseteq V$, let $\phi = (\phi_v, v \in V)$ be such that $\phi_v : \mathcal{X}_v \to \mathbb{R}$ is piecewise constant and nondecreasing for all $v \in A$ and $\phi_v(x_v) = x_v$ for $v \notin A$. If X is MTP_2, then $Y = \phi(X)$ is MTP_2.*

Proof.
If f denotes the density function for X and g the density function for Y, both with respect to a standard base measure μ, we have that

$$g(y) = \int_{\phi_A^{-1}(y_A)} f(x) d\mu_A(x_A).$$

Since ϕ is nondecreasing, we have

$$\phi_A^{-1}\left(y_A^1\right) \wedge \phi_A^{-1}\left(y_A^2\right) = \phi_A^{-1}\left(y_A^1 \wedge y_A^2\right), \quad \text{and} \quad \phi_A^{-1}\left(y_A^1\right) \vee \phi_A^{-1}\left(y_A^2\right) = \phi_A^{-1}\left(y_A^1 \vee y_A^2\right),$$

Applying corollary 2.2.4, and letting $g = f_1 = f_2 = f_3 = f_4$, shows that Y is MTP_2. □

A *monotone coarsening* is an operation on a finite discrete state space $\mathcal{X}_i$ that identifies a collection of neighbouring states. For example, if $\mathcal{X}_i = \{i_1, ..., i_n\}$ then $\mathcal{X}_i' = \{\{i_1, ..., i_j\}, i_{j+1}, ..., i_k, \{i_{k+1}, ..., i_n\}\}$ is a monotone coarsening.

Proposition 2.2.6. *The* MTP_2 *property is preserved under taking products, conditioning, marginalization, and monotone coarsening, that is:*

1. *If the functions* f *and* g *on* $\mathcal{X}$ *are* MTP_2*, then their product* fg *is* MTP_2*.*
2. *If* X *has an* MTP_2 *distribution, then for every* $C \subseteq V$ *the conditional distribution of* $X_C \mid X_{V \setminus C}$ *is* MTP_2 *for almost all* $x_{V \setminus C}$*.*
3. *If* X *has an* MTP_2 *distribution, then for every* $A \subset V$*, the marginal distribution of* $X_A \subseteq X$ *is* MTP_2*.*
4. *If* X *is* MTP_2 *and discrete, and* Y *is obtained from* X *by monotone coarsening, then* Y *is* MTP_2*.*

Proof.

1. This property follows directly from the definition of MTP_2.
2. This property follows directly from the definition of MTP_2.
3. We prove this by induction. Let f_1, f_2, f_3, f_4 be nonnegative functions on $\mathcal{X}$ satisfying for all $x, y \in \mathcal{X}$

$$f_1(x) f_2(y) \leq f_3(x \vee y) f_4(x \wedge y). \tag{2.4}$$

Consider the marginals

$$\varphi_{1,j}(x_2,...,x_n) = \int_{\mathcal{X}_1} f_j(x_1,...,x_n)\, d\mu_1(x_1),\ j = 1,...,4$$

for all $x, y \in \prod_{i=2}^n \mathcal{X}_i$. Consider fixed $x, y \in \prod_{i=2}^n \mathcal{X}_i$. Then

$$\begin{aligned}
\varphi_{1,1}(x)\varphi_{1,2}(y) &= \int_{\mathcal{X}_1} f_1(x,x_1)d\mu_1(x_1) \int_{\mathcal{X}_1} f_2(y,y_1)d\mu_1(y_1) \\
&= \int_{x_1<y_1}\int_{\mathcal{X}_1} f_1(x,x_1)f_2(y,y_1)d\mu_1(x_1)d\mu_1(y_1) \\
&\quad + \int_{y_1<x_1}\int_{\mathcal{X}_1} f_1(x,x_1)f_2(y,y_1)d\mu_1(x_1)d\mu_1(y_1) \\
&= \int_{x_1<y_1}\int_{\mathcal{X}_1} f_1(x,x_1)f_2(y,y_1)d\mu_1(x_1)d\mu_1(y_1) \\
&\quad + \int_{x_1<y_1}\int_{\mathcal{X}_1} f_1(x,y_1)f_2(y,x_1)d\mu_1(x_1)d\mu_1(y_1) \\
&= \int_{x_1<y_1}\int_{\mathcal{X}_1} f_1(x,x_1)f_2(y,y_1) + f_1(x,y_1)f_2(y,x_1)d\mu_1(x_1)d\mu_1(y_1).
\end{aligned}$$

Likewise, we see that

$$\begin{aligned}
\varphi_{1,3}(x\vee y)\varphi_{1,4}(x\wedge y) &= \int_{x_1<y_1}\int_{\mathcal{X}_1} f_3(x\vee y, x_1\vee y_1)f_4(x\wedge y, x_1\wedge y_1)d\mu_1(x_1)d\mu_1(y_1) \\
&\quad + \int_{y_1<x_1}\int_{\mathcal{X}_1} f_3(x\vee y, x_1\vee y_1)f_4(x\wedge y, x_1\wedge y_1)d\mu_1(x_1)d\mu_1(y_1) \\
&= \int_{x_1<y_1}\int_{\mathcal{X}_1} f_3(x\vee y, y_1)f_4(x\wedge y, x_1) \\
&\quad + f_3(x\vee y, x_1)f_4(x\wedge y, y_1)d\mu_1(x_1)d\mu_1(y_1).
\end{aligned}$$

Define $a := f_1(x,x_1)f_2(y,y_1)$, $b := f_1(x,y_1)f_2(y,x_1)$, $c := f_3(x\vee y, x_1)f_4(x\wedge y, y_1)$, $d := f_3(x\vee y, y_1)f_4(x\wedge y, x_1)$. We wish to show that

$$\varphi_{1,1}(x)\varphi_{1,2}(y) \le \varphi_{1,3}(x\vee y)\varphi_{1,4}(x\wedge y). \tag{2.5}$$

To this end, it suffices to show that $a + b \le c + d$.

Suppose $x_1 \le y_1$. Then we know by assumption that

$$d = f_3(x\vee y, y_1)f_4(x\wedge y, x_1) \ge f_1(x,x_1)f_2(y,y_1) = a,$$

and $d \geq f_1(x, y_1)f_2(y, x_1) = b$. In addition, (2.4) implies that

$$f_1(x, x_1)f_2(x, x_1) \leq f_3(x, x_1)f_4(x, x_1).$$

Then we see that

$$c + d - (a + b) = \frac{1}{d}\left((d - a)(d - b) + (cd - ab)\right) \geq 0.$$

Each term in the second expression is nonnegative, and sums and products of nonnegative terms are nonnegative. Hence (2.5) is shown, which means that $\varphi_{1,j}$, $j = 1, ..., 4$ satisfy (2.4) on $\prod_{i=2}^{n} \mathcal{X}_i$.

Repeating the process for the marginals

$$\varphi_{m+1,j}\left(x_{m+2}, ..., x_n\right) = \int_{\mathcal{X}_{m+1}} \varphi_{m,j}\left(x_{m+1}, ..., x_n\right) d\mu_m\left(x_m\right),$$

where $j = 1, ..., 4$ and $m = 1, ..., k$, and letting $f = f_1 = f_2 = f_3 = f_4$ be the density of X yields the result.

4. This property is an instance of a nondecreasing and piecewise constant transformation, and so it follows from Proposition 2.2.5.

□

We say that f has *interval support* if for any $x, y \in \mathcal{X}$

$$f(x)f(y) \neq 0 \quad \Rightarrow \quad f(z) \neq 0 \quad \text{for any} \quad x \wedge y \leq z \leq x \vee y.$$

We say that the pair $x, y \in \mathcal{X}$ is *elementary* if they are not comparable, that is $x \not\leq y$ and $y \not\leq x$, and if they differ in exactly two coordinates, and we denote the set of all elementary pairs by $\mathcal{E} \subset \mathcal{X} \times \mathcal{X}$. The following proposition shows, that under the interval support condition, MTP_2 is a pairwise property, in the sense that it can be checked on the level of two variables when the remaining variables are held fixed.

Proposition 2.2.7. *If f has interval support and $f : \mathcal{X} \to \mathbb{R}$ is TP_2 in every pair of arguments when the remaining arguments are held constant, then f is MTP_2.*

Proof.
Suppose without loss of generality that

$$x = (x_1^*, ..., x_k^*, x_{k+1}, ..., x_n)\,, \; y = \left(y_1, ..., y_k, y_{k+1}^*, ..., y_n^*\right),$$

where $x_i^* \geq y_i, i = 1, ..., k$ and $x_j \leq y_j^*, j = k+1, ..., n$. Then it holds that

$$\begin{aligned}\frac{f(x \vee y) f(x \wedge y)}{f(x) f(x)} &= \frac{f\left(x_1^*, ..., x_k^*, y_{k+1}^*, ..., y_n^*\right) f\left(y_1, ..., y_k, x_{k+1}, ..., x_n\right)}{f\left(x_1^*, ..., x_k^*, x_{k+1}, ..., x_n\right) f\left(y_1, ..., y_k, y_{k+1}^*, ..., y_n^*\right)} \\ &= \frac{f\left(x_1^*, ..., x_k^*, y_{k+1}^*, ..., y_n^*\right) f\left(x_1^*, y_2, ..., y_k, x_{k+1}, ..., x_n\right)}{f\left(x_1^*, ..., x_k^*, x_{k+1}, ..., x_n\right) f\left(x_1^*, y_2, ..., y_k, y_{k+1}^*, ..., y_n^*\right)} \\ &\quad \times \frac{f\left(x_1^*, y_2, ..., y_k, y_{k+1}^*, ..., y_n^*\right) f\left(y_1, ..., y_k, x_{k+1}, ..., x_n\right)}{f\left(x_1^*, y_2, ..., y_k, x_{k+1}, ..., x_n\right) f\left(y_1, ..., y_k, y_{k+1}^*, ..., y_n^*\right)}.\end{aligned}$$

In the third expression we have multiplied by 1 in a clever way to achieve a product of two terms, which either exceed or is equal to one by the TP_2 assumption: the first term by fixing x_1^* and applying the TP_2 assumption to the remaining $n-1$ variables, and the second by fixing $y_2, ..., y_k$. The denominators are nonzero, since we have assumed that f has interval support, and so they contain terms of the form $f(z)$ with $x \wedge y \leq z \leq x \vee y$. □

This is a strong result, as it is much easier for us to check whether a distribution is MTP_2 if we only need to consider two variables at a time, rather than having to consider all variables at once. Not only is it powerful, it is also practical. If a density function f has full support, it holds trivially that f will have interval support. A large portion of our work in the chapters to come, revolve around the Gaussian distribution, which has full support everywhere. In addition, if a distribution p is strictly positive, it holds for any $x \in \mathcal{X}$ that $p(x) > 0$. Hence, it has interval support.

Theorem 2.2.8. *A distribution of the form*

$$p(x) = \frac{1}{Z} \prod_{uv \in E} \psi_{uv}\left(x_u, x_v\right),$$

where ψ_{uv} are strictly positive functions and Z is a normalizing constant, is MTP_2 *if and only if each ψ_{uv} is an* MTP_2 *function.*

Proof.
Since the distribution p is strictly positive, by Proposition 2.2.7, p satisfies the MTP_2 property if and only if it does so for $x, y \in \mathcal{X}$, that differ in two coordinates, say with indices u, v. We write E_u for the set of edges that contain u but not v, and E_v for the set of edges that contain v but not u. Consider the case where $uv \in E$. Then we have that $p(x \wedge y)p(x \vee y) - p(x)p(y) \geq 0$ if and only if

$$\psi_{uv}\left((x \wedge y)_{uv}\right) \psi_{uv}\left((x \vee y)_{uv}\right) \prod_{st \in E_u \cup E_v} \psi_{st}\left((x \wedge y)_{st}\right) \psi_{st}\left((x \vee y)_{st}\right)$$

$$\geq \quad \psi_{uv}(x_{uv})\,\psi_{uv}(y_{uv}) \prod_{st \in E_u \cup E_v} \psi_{st}(x_{st})\,\psi_{st}(y_{st})$$

All other terms cancel out because of the assumption that $x_w = y_w$ for $w \in V \backslash \{u, v\}$.

For $st \in E_u \cup E_v$ we have $\{x_{st}, y_{st}\} = \{(x \wedge y)_{st}(x \vee y)_{st}\}$ and so the above inequality holds if and only if

$$\psi_{uv}((x \wedge y)_{uv})\,\psi_{uv}((x \vee y)_{uv}) \geq \psi_{uv}(x_{uv})\,\psi_{uv}(y_{uv}),$$

in other words, if and only if ψ_{uv} is MTP_2.

Consider now the case where $uv \notin E$. Using the same argument, one concludes that in this case the above inequalities are equalities instead, and so the proof is complete. □

2.3 Conditional Independence Models

Graphs are extremely useful, as it turns out that instead of writing down a long list of independence statements, e.g. $X_1 \perp\!\!\!\perp X_5 \mid \{X_2, X_3\}$, we can simply draw a graph to represent them. With a bit of practice, one is able to quickly determine the conditional independences of the variables in the data by looking at the corresponding graph. Graph separation is an important example of *conditional independence model*, which we now define.

Definition 2.3.1. *An* **independence model** $\mathcal{J}$ *over a finite set* V *is a set of triples* $\langle A, B \mid C\rangle$ *of disjoint subsets of* V *called* **independence statements**. *The independence statement* $\langle A, B \mid C\rangle$ *is read as "*A **is independent of** B **given** C**."** *We call the independence model a* **semi-graphoid** *if it holds for all mutually disjoint subsets* A, B, C, D *that:*

(S1) $\langle A, B \mid C\rangle \in \mathcal{J} \iff \langle B, A \mid C\rangle \in \mathcal{J}$;
(S2) $\langle A, B \cup D \mid C\rangle \in \mathcal{J}, \Rightarrow \langle A, B \mid C\rangle \in \mathcal{J} \Rightarrow \langle A, D \mid C\rangle \in \mathcal{J}$;
(S3) $\langle A, B \cup D \mid C\rangle \in \mathcal{J}, \Rightarrow \langle A, B \mid C \cup D\rangle \in \mathcal{J}$ *and* $\langle A, D \mid C \cup B\rangle \in \mathcal{J}$;
(S4) $\langle A, B \mid C \cup D\rangle \in \mathcal{J}$ *and* $\langle A, D \mid C\rangle \in \mathcal{J} \iff \langle A, B \cup D \mid C\rangle$.

The conditions (S1)-(S4) are also known as **symmetry, decomposition, weak union,** *and* **contraction**, *respectively. A semi-graphoid that also satisfies the reverse implication of (S3), that is*

(S5) $\langle A, B \mid C \cup D\rangle \in \mathcal{J}$ *and* $\langle A, D \mid C \cup B\rangle \in \mathcal{J}, \Rightarrow \langle A, B \cup D \mid C\rangle \in \mathcal{J}$,

also known as **intersection** *is called a* **graphoid**. *A graphoid or semi-graphoid satisfying the reverse implication of (S2), that is*

$$\textit{(S6)} \quad \langle A, B \mid C\rangle \in \mathcal{J} \text{ and } \langle A, D \mid C\rangle \in \mathcal{J}, \text{ then } \langle A, B \cup D \mid C\rangle \in \mathcal{J},$$

is said to be **compositional**. *Some independence models have additional properties. Below we write singleton sets $\{u\}$ as u, etc.*

$$\textit{(S7)} \quad \langle u, v \mid C\rangle \in \mathcal{J} \text{ and } \langle u, v \mid C \cup w\rangle \in \mathcal{J}, \Rightarrow \langle u, w \mid C\rangle \in \mathcal{J} \text{ or } \langle v, w \mid C\rangle \in \mathcal{J};$$
$$\textit{(S8)} \quad \langle A, B \mid C\rangle \in \mathcal{J} \text{ and } D \subseteq V\backslash(A \cup B), \Rightarrow \langle A, B \mid C \cup D\rangle \in \mathcal{J}.$$

(S7) and (S8) are known as **singleton-transitivity** *and* **upward-stability**, *respectively.*

Note that the properties above are not independent, in the sense that they sometimes imply one another. For example, upward stability often implies composition, as we will now show.

Lemma 2.3.2. *Let $\mathcal{J}$ be a semi-graphoid. If $\mathcal{J}$ is upwards stable, then $\mathcal{J}$ is also compositional.*

Proof.
We have assumed that (S1)-(S4) and (S8) hold, and wish to show that then (S6) holds. If $\langle A, B \mid C\rangle \in \mathcal{J}$, then by (S8) it holds that $\langle A, B \mid C \cup D\rangle \in \mathcal{J}$ for some $D \subseteq V\backslash(A \cup B)$, and hence by (S4) we get that if $\langle A, D \mid C\rangle \in \mathcal{J}$ then $\langle A, B \cup D \mid C\rangle \in \mathcal{J}$, which is exactly what (S6) says. □

Consider a set V and associated random variables $X = (X_v)_{v\in V}$. Let A, B and C be disjoint subsets of V. We say that X_A is *conditionally independent of X_B given X_C* if for any measurable subset $\Omega \subseteq \mathcal{X}_A$ and P-almost all x_B and x_C,

$$P\left(X_A \in \Omega \mid X_B = x_B, X_C = x_C\right) = P\left(X_A \in \Omega \mid X_C = x_C\right),$$

and we write $A \perp\!\!\!\perp B \mid C$. We can now induce an independence model $\mathcal{J}(P)$ by letting

$$\langle A, B \mid C\rangle \in \mathcal{J}(P) \quad \Longleftrightarrow \quad A \perp\!\!\!\perp B \mid C \quad \text{w.r.t. } P.$$

We say that an independence model $\mathcal{J}$ is *probabilistic* if there is a distribution P such that $\mathcal{J} = \mathcal{J}(P)$.

Proposition 2.3.3. *Let $\mathcal{J}(P)$ be a probabilistic independence model. Then $\mathcal{J}(P)$ is always a semi-graphoid, and if P has strictly positive density f, $\mathcal{J}(P)$ is always a graphoid.*

Proof.
We first need to show that $\mathcal{J}(P)$ has the properties (S1)-(S4).

(S1) follows immediately from the definition.

$\langle A, B \cup D \mid C\rangle \in \mathcal{J}(P) \iff A \perp\!\!\!\perp B \cup D \mid C \Rightarrow A \perp\!\!\!\perp B \mid C$ and $A \perp\!\!\!\perp D \mid C \iff \langle A, B \mid C\rangle \in \mathcal{J}(P)$ and $\langle A, D \mid C\rangle \in \mathcal{J}(P)$. Hence (S2) is satisfied.

$\langle A, B \cup D \mid C\rangle \in \mathcal{J}(P) \iff A \perp\!\!\!\perp B \cup D \mid C \iff P(A \mid B, C, D) = P(A \mid C)$
$\Rightarrow P(A \mid B, C) = P(A \mid C)$ and $P(A \mid C, D) = P(A \mid C) \iff \langle A, B \mid C \cup D\rangle \in \mathcal{J}(P)$ and $\langle A, D \mid C \cup B\rangle \in \mathcal{J}(P)$. Hence (S3) is satisfied.

Let $\langle A, B \mid C \cup D\rangle \in \mathcal{J}(P)$ and $\langle A, D \mid C\rangle \in \mathcal{J}(P)$. Then $A \perp\!\!\!\perp B \mid C \cup D$ and $A \perp\!\!\!\perp D \mid C$, and so $P(A \mid B, C, D) = P(A \mid C, D)$ and $P(A \mid C, D) = P(A \mid C)$. Thus $P(A \mid B, C, D) = P(A \mid C)$, hence $A \perp\!\!\!\perp B \cup D \mid C$, so $\langle A, B \cup C \mid D\rangle \in \mathcal{J}(P)$. The converse implication holds by (S2) and (S3) and thus (S4) is satisfied.

Assume that the variables have density $p(x_A, x_B, x_C, x_D) > 0$ and that $A \perp\!\!\!\perp B \mid C \cup D$ and $A \perp\!\!\!\perp D \mid C \cup B$. Then, by Proposition 2.21 in Lauritzen (2018),

$$p(x_A, x_B, x_C, x_D) = k(x_A, x_D, x_C)\, l(x_B, x_D, x_C) = g(x_A, x_B, x_D)\, h(x_B, x_D, x_C)$$

where k, l, g, h are suitable strictly positive functions. Hence we have

$$g(x_A, x_B, x_D) = \frac{k(x_A, x_D, x_C)\, l(x_B, x_D, x_C)}{h(x_B, x_D, x_C)}.$$

Fixing $x_D = \tilde{x}_D$ we have

$$g(x_A, x_B, x_D) = \pi(x_A, x_C)\, \rho(x_B, x_C)$$

where $\pi(x_A, x_C) = k(x_A, \tilde{x}_D, x_C)$ and $\rho(x_A, x_C) = l(x_B, \tilde{x}_D, x_C)/h(x_B, \tilde{x}_D, x_C)$. Thus

$$p(x_A, x_B, x_C, x_D) = \pi(x_A, x_C)\, \rho(x_B, x_C)\, h(x_B, x_C, x_D),$$

and hence $A \perp\!\!\!\perp B \cup D \mid C$. □

An important example of an independence model is induced by separation in an undirected graph $G = (V, E)$, denoted by $\mathcal{J}(G)$:

$$\langle A, B \mid S\rangle \in \mathcal{J}(G) \iff S \text{ separates } A \text{ from } B.$$

This independence model behaves very nicely, as shown in the following proposition.

Proposition 2.3.4. *An independence model $\mathcal{J}(G)$ induced by separation in an undirected graph $G = (V, E)$ is a compositional graphoid which is upward-stable, and singleton-transitive.*

Proof.
We need to show that $\mathcal{J}(G)$ has all the properties (S1)-(S8).

- (S1): $\langle A, B \mid C\rangle \in \mathcal{J}(G) \iff C$ separates A from $B \iff C$ separates B from $A \iff \langle B, A \mid C\rangle \in \mathcal{J}(G)$.
- (S2): $\langle A, B \cup D \mid C\rangle \in \mathcal{J}(G) \iff C$ separates A from $B \cup D \Rightarrow C$ separates A from B and C separates A from $D \iff \langle A, B \mid C\rangle \in \mathcal{J}(G)$ and $\langle A, D \mid C\rangle \in \mathcal{J}(G)$.
- (S3): $\langle A, B \cup D \mid C\rangle \in \mathcal{J}(G) \iff C$ separates A from $B \cup D \Rightarrow C$ separates A from $B \Rightarrow C \cup D$ separates A from $B \iff \langle A, B \mid C \cup D\rangle \in \mathcal{J}(G)$.
- (S4) " $\Rightarrow$ ": $\langle A, B \mid C \cup D\rangle \in \mathcal{J}(G)$ and $\langle A, D \mid C\rangle \in \mathcal{J}(G) \iff C \cup D$ separates A from B and C separates A from D. The only path from A to D goes through C, hence any path from A to B either goes through C to B or through C to D to B. Hence, it is implied that C separates A from B and C separates A from $D \iff \langle A, B \mid C\rangle \in \mathcal{J}(G)$ and $\langle A, D \mid C\rangle \in \mathcal{J}(G) \iff \langle A, B \cup D \mid C\rangle \in \mathcal{J}(G)$. (S4) "$\Leftarrow$": Assume $\langle A, B \cup D \mid C\rangle \in \mathcal{J}(G)$. (S2) yields $\langle A, D \mid C\rangle \in \mathcal{J}(G)$, and (S3) yields $\langle A, B \mid C \cup D\rangle \in \mathcal{J}(G)$.
- (S5): $\langle A, B \mid C \cup D\rangle \in \mathcal{J}(G)$ and $\langle A, D \mid C \cup B\rangle \in \mathcal{J}(G) \iff C \cup D$ separates A from B and $C \cup B$ separates A from D. For both of these statements to hold simultaneously, it is necessarily true that the only path from B to D goes through C. Hence, it is implied that C separates A from B and C separates A from $D \iff C$ separates A from $B \cup D \iff \langle A, B \cup D \mid C\rangle \in \mathcal{J}(G)$.
- (S7): $\langle u, v \mid C\rangle \in \mathcal{J}(G)$ and $\langle u, v \mid C \cup w\rangle \in \mathcal{J}(G) \Rightarrow C$ separates u from v and $C \cup w$ separates u from v. Assume that $\langle u, w \mid C\rangle \notin \mathcal{J}(G)$, so u and w are *not* separated by C. Then, since we have assumed there is no path from u to v bypassing C via w, it must hold that v and w are separated by C, and so $\langle v, w \mid C\rangle \in \mathcal{J}(G)$. Similar arguments are made to show that $\langle v, w \mid C\rangle \notin \mathcal{J}(G) \Rightarrow \langle u, w \mid C\rangle \in \mathcal{J}(G)$.

- (S8): $\langle A, B \mid C\rangle \in \mathcal{J}(G)$ and $D \subseteq V\backslash(A \cup B) \Rightarrow C$ separates A from $B \Rightarrow C \cup D$ separates A from $B \iff \langle A, B \mid C \cup D\rangle \in \mathcal{J}(G)$.
- (S6): Since (S1)-(S4) all hold, $\mathcal{J}(G)$ is a semi-graphoid, and since $\mathcal{J}(G)$ satisfies (S8), by lemma 2.3.2, (S6) is also satisfied.

□

2.4 Markov Properties, Faithfulness, and Total Positivity

In this section we consider relationships between probabilistic independence models and *graphical independence models*. Relationships of this sort take the form of *Markov properties*, which are statements saying that certain graph separations imply conditional independence statements in the probabilistic independence model.

We shall henceforth write $A \perp_G B \mid C$ for the graph separation $\langle A, B \mid C\rangle \in \mathcal{J}(G)$, and $A \perp\!\!\!\perp_P B \mid C$ for the relation $\langle A, B \mid C\rangle \in \mathcal{J}(P)$.

Definition 2.4.1. *Let $G = (V, E)$ be an undirected graph, and let $(X_u)_{u\in V}$ be a collection of random variables taking values in Borel spaces $(\mathcal{X}_u)_{u\in V}$. A distribution P on $\mathcal{X} = \prod_{u\in V} \mathcal{X}_u$ is said to satisfy*

(P) *the* **pairwise Markov property** *w.r.t. an undirected graph $G = (V, E)$ if for any pair $u, v \in V$*

$$uv \notin E \;\Rightarrow\; u \perp\!\!\!\perp_p v \mid V\backslash\{u, v\};$$

(L) *the* **local Markov property** *w.r.t. an undirected graph $G = (V, E)$ if for any vertex $u \in V$*

$$u \perp\!\!\!\perp_P V\backslash\mathrm{cl}(u) \mid \mathrm{bd}(u);$$

(G) *the* **global Markov property** *w.r.t. an undirected graph $G = (V, E)$ if separation in G implies conditional independence*

$$A \perp_G B \mid C \;\Rightarrow\; A \perp\!\!\!\perp_P B \mid C.$$

Proposition 2.37 in Lauritzen, 2018 states that for any undirected graph G, and any probability distribution on $\mathcal{X}$, it holds that (**G**) ⇒ (**L**) ⇒ (**P**). Proposition 2.38 in the same book states, that the three Markov properties are equivalent if the conditional independence relation $\perp\!\!\!\perp_P$ induced by P is a graphoid.

For a graph $G = (V, E)$, an independence model $\mathcal{J}$ defined over V is said to satisfy the global Markov property w.r.t. a graph G, if for disjoint subsets A, B, and C of V the following holds:

$$A \perp_G B \mid C \quad \Rightarrow \quad \langle A, B \mid C \rangle \in \mathcal{J}.$$

If $\mathcal{J}(P)$ satisfies the global Markov property w.r.t. a graph G, we say that P is *Markov* w.r.t. G. If conditional independence in $\mathcal{J}(P)$ implies separation in G we say that the graph G is Markov w.r.t. the distribution P. Recall that an independence model $\mathcal{J}$ is probabilistic if there is a distribution P such that $\mathcal{J} = \mathcal{J}(P)$. We then also say that P is *faithful* to $\mathcal{J}$. If P is faithful to $\mathcal{J}(G)$ for a graph G, we say that P is *faithful to* G. A more formal definition is:

Definition 2.4.2. *P is said to be* **faithful** *to the graph G if for all disjoint subsets A, B, C of V*

$$A \perp_G B \mid C \iff A \perp\!\!\!\perp_P B \mid C$$

i.e. G is Markov w.r.t. P and vice versa.

Let P denote a distribution on $\mathcal{X}$. The *pairwise independence graph* of P is the undirected graph $G(P) = (V, E(P))$ with

$$uv \notin E(P) \iff u \perp\!\!\!\perp v \mid V \backslash \{u, v\}.$$

Any distribution P satisfies the pairwise Markov property w.r.t. its pairwise independence graph $G(P)$. Indeed, since $G(P)$ is the smallest graph that makes P pairwise Markov.

If P has continuous density f, Peters (2015) showed that the induced independence model is a graphoid if and only if the support is *coordinatewise connected*, i.e. all connected components of the support of the density can be connected by axis-parallel lines. This applies in particular to the discrete case, since any function over a discrete space is continuous.

Theorem 2.4.3. *If P is* MTP_2 *and its independence model is a graphoid, then P is faithful to its pairwise independence graph $G(P)$.*

Proof.
By Theorem 2.38 in Lauritzen (2018) it follows that P is globally Markov w.r.t. $G(P)$.

We need to show that $G(P)$ is also Markov w.r.t. P. To this end, consider disjoint subsets $A, B, C \subseteq V$ such that C does not separate A from B in $G(P)$. We need to show that $A \not\perp\!\!\!\perp B \mid C$.

Let $uv \in E$. Then $u \not\perp v \mid V \backslash \{u, v\}$ and by upward-stability $u \not\perp v \mid C$ for any $C \subset V \backslash \{u, v\}$. Since C does not separate A from B, there exists $u \in A$ and $v \in B$ such that $uv \notin E$ and a path $u = v_1, v_2, ..., v_r = v$ such that $v_k \notin C$ for all $k = 1, ..., r$, and $v_k v_{k+1} \in E$ for all $k = 1, ..., r-1$. Thus we obtain $v_k \not\perp v_{k+1} \mid C$ for all $k = 1, ..., r-1$. By singleton-transitivity $v_1 \not\perp v_2 \mid C$ and $v_2 \not\perp v_3 \mid C$ imply that $v_1 \not\perp v_3 \mid C$. Repeating this argument yields $u \not\perp v \mid C$ and hence $A \not\perp B \mid C$. □

The results we have established thus far apply to MTP_2 distributions in general. In the chapters to come we will make assumptions about the state space of the distribution and present results restricted to them.

3. Binary Distributions

This chapter will be dedicated to MTP_2 distributions with binary sample spaces. We briefly introduce exponential families and present some results concerning them. With these results established we define order theoretical lattices, and make a connection between the elements of the sample space $\mathcal{X}$ and the elements of the set of all subsets of $\{1, ..., d\}$. We will discuss the conditions for the existence of the MLE in these models, and we end the chapter by discussing binary models over graphs. We shall see that in a binary model under MTP_2 it becomes possible for an MLE to exist with fewer observations than in one not under MTP_2, and even fewer still in a binary MTP_2 graphical model.

3.1 Exponential Families

We start off by giving a formal definition of an exponential family.

Definition 3.1.1 (Exponential family)**.** *Consider a class $\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$ of probability measures on the measure space $(\mathcal{X}, \mathcal{A}, \mu)$, where μ is a σ-finite measure. Suppose P_θ is absolutely continuous w.r.t. μ for all $\theta \in \Theta$, and that there exist functions $\phi = (\phi_1, ..., \phi_k) : \Theta \to \mathbb{R}$ and $a : \Theta \to (0, \infty)$ and measurable functions $T = (T_1, ..., T_k) : \mathcal{X} \to \mathbb{R}$ and $b : \mathcal{X} \to \mathbb{R}_+$ such that for all $\theta \in \Theta$*

$$\frac{dP_\theta}{d\mu}(x) = a(\theta)b(x)\exp(\phi(\theta)t(x)). \tag{3.1}$$

If (3.1) is satisfied, we call $\mathcal{P}$ an **exponential family** *with* **canonical statistic** *$t = T(X)$ and* **canonical parameter** *$\phi(\theta)$. The smallest k for which a representation of the form (3.1) is possible, is called the* **order** *of the exponential family. If the representation is* **minimal***, i.e. if k is the order of the family, then T is called a* **minimal canonical statistic** *and ϕ is called a* **minimal canonical parameter**.

We often write $p(x; \theta) := \frac{dP_\theta}{d\mu}(x)$. We wish to give the definition of a *sufficient statistic*. To do this we first need to give the definition of a *Markov kernel*.

Definition 3.1.2 (Markov kernel)**.** *Let $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$ be measure spaces. A* **Markov kernel** *$\pi : \mathcal{X} \to \mathcal{Y}$ with source $(\mathcal{X}, \mathcal{A})$ and target $(\mathcal{Y}, \mathcal{B})$ is a map with the following properties:*

- $\pi(\,\cdot \mid T\,)$ *is a probability measure on* $(\mathcal{X}, \mathcal{Y})$ *for all* $T \in \mathcal{Y}$,
- $\pi(\,A \mid \cdot\,)$ *is* $\mathcal{B}$*-measurable for all* $A \in \mathcal{A}$.

Definition 3.1.3 (Sufficient statistic)**.** *A statistic* $t = T(X)$ *is said to be* **sufficient** *for* $\mathcal{P}$ *if there exists a Markov kernel* $\pi(A \mid T)$*, where* $A \in \mathcal{A}$ *and* $t \in \mathcal{Y}$ *such that* π *is a regular conditional probability given* T *under* P *for any* $P \in \mathcal{P}$*, i.e. if*

$$\int_B \pi(A \mid T) dP_T(t) = P\big(A \cap T^{-1}(B)\big)\,, \quad \forall A \in \mathcal{A},\ \forall B \in \mathcal{B},\ \forall P \in \mathcal{P}.$$

A less formal definition is that a statistic $t = T(X)$ for $\mathcal{P}$ is said to be sufficient if and only if the conditional probability distribution of data X given T does not depend on $P \in \mathcal{P}$.

Let $x \in \mathcal{X}$, $\theta \in \mathbb{R}^k$, $g : \mathbb{R}^n \to \mathbb{R}$, $A : \mathbb{R}^k \to \mathbb{R}$. Consider an exponential family with density $p(x;\theta)$ satisfying

$$\log p(x;\theta) = \langle \theta, T(x) \rangle - A(\theta) + g(x) \tag{3.2}$$

with sample space $\mathcal{X}$, sufficient statistics $T : \mathcal{X} \to \mathbb{R}^k$, base measure μ, and standard inner product on $\mathbb{R}^k$. Assume that the family is minimally represented, and that the family is regular so that the space of canonical parameters

$$\mathcal{K} = \big\{\theta \in \mathbb{R}^k : A(\theta) < \infty\big\}$$

is an open convex set. Assume that there exists θ_0 such that $p\,(x;\theta_0)$ is a product distribution, or equivalently,

$$p(x \vee y;\theta_0)p(x \wedge y;\theta_0) = p(x;\theta_0)p(y;\theta_0) \quad \text{for all } x, y \in \mathcal{X}.$$

Every distribution in an exponential family can be used as the base distribution, and as such we can pick $p(x;\theta_0)$ as the base measure. It then holds that

$$g(x \wedge y) + g(x \vee y) - g(x) - g(y) = 0.$$

We say that such an exponential family *has a product base*. For an exponential family of the form (3.2) and any two $x, y \in \mathcal{X}$ we define

$$\Delta(x, y;\theta) := \log\left(\frac{p(x \vee y;\theta)p(x \wedge y;\theta)}{p(x;\theta)p(y;\theta)}\right)$$

The density $p(x;\theta)$ is MTP_2 if and only if $\Delta(x, y;\theta) \geq 0$ for all elementary pairs in $\mathcal{E}$. Indeed, since $p(x;\theta)$ is the density of an exponential family, hence it has

interval support, and so applying Proposition 2.2.7 yields the result. It is straight forward to see that for exponential families with a product base it holds that

$$\Delta(x, y; \theta) = \langle \theta, T(x \wedge y) + T(x \vee y) - T(x) - T(y) \rangle \,,$$

which is a linear function in θ for all $x, y \in \mathcal{X}$. Let $\mathcal{K}_2 \subset \mathcal{K}$ denote the subset of canonical parameters for which the density $p(x; \theta)$ is MTP_2.

Theorem 3.1.4. *$\mathcal{K}_2$ is a convex set that is closed relative to $\mathcal{K}$.*

Proof.
We have that

$$\mathcal{K}_2 = \{\theta \in \mathcal{K} : \Delta(x, y; \theta) \geq 0 \; \forall (x, y) \in \mathcal{E}\}$$

so by definition, $\mathcal{K}_2$ is closed relative to $\mathcal{K}$, and since $\Delta(x, y; \theta)$ is a linear function in θ, $\mathcal{K}_2$ is convex. □

Theorem 3.1.5. *Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a closed convex cone such that its dual cone $\mathcal{C}^*$ is the closure of the cone generated by the set*

$$\{T(x \wedge y) + T(x \vee y) - T(x) - T(y) : x, y \in \mathcal{E}\} \,.$$

Then $\mathcal{K}_2 = \mathcal{K} \cap \mathcal{C}^*$.

Proof.
The set of inequalities $\Delta(x, y; \theta) \geq 0$, one for each elementary pair $x, y \in \mathcal{E}$, defines a convex cone in $\theta \in \mathbb{R}$. We have $\Delta(x, y; \theta) \geq 0$ for all $x, y \in \mathcal{E}$ if and only if $\langle \theta, v \rangle \geq 0$ for all v in the cone $\mathcal{C}^\vee$ generated by the set

$$\{T(x \wedge y) + T(x \vee y) - T(x) - T(y) : x, y \in \mathcal{E}\} \,.$$

By definition $(\mathcal{C}^\vee)^* = \mathcal{C}$, and so $\mathcal{C}^* = (\mathcal{C}^\vee)^{**}$. Since dual cones always are closed and convex, the latter is equal to the closure of $\mathcal{C}^\vee$. □

When $\mathcal{X}$ is finite – i.e. for log-linear models – Proposition 3.1.5 implies that $\mathcal{C}$ is polyhedral. As we will see, in the Gaussian setting where $\mathcal{X} \subseteq \mathbb{R}^p$, and $\mathcal{C} = \mathcal{H}^p_{ij}$, is a closed and convex polyhedral cone, yet $\mathcal{X}$ might not be finite. This shows that finiteness of $\mathcal{X}$ is not a necessity for $\mathcal{C}$ to be polyhedral. As a matter of fact, in Chapter 4 we show in Proposition 4.1.1 that $\mathcal{C}$ is a polyhedral cone for any quadratic exponential family.

We close this section with the discussion of the MLE and its existence. An important consequence of Theorem 3.1.4 is that any MTP_2 exponential family is a

convex exponential family. Hence, if the maximum likelihood estimator exists, it is uniquely defined. Let $x_1, ..., x_n$ denote a random sample of size n and let $\overline{T} := \frac{1}{n}\sum_{i=1}^{n} T(x_i)$ be the average of the corresponding sufficient statistics. Let $\mathcal{S}$ denote the interior of the convex support of the sufficient statistics, i.e.

$$\mathcal{S} = \operatorname{conv}\left(\operatorname{supp}\left(\mu \circ T^{-1}\right)\right).$$

By the general theory of exponential families, the MLE $\hat{\theta}$ exists if and only if $\overline{T}$ lies in $\mathcal{S}$ (Barndorff-Nielsen, 2014). If this is the case, it is uniquely defined by

$$\nabla A(\hat{\theta}) = \mathbb{E}_\theta[T(X)] = \overline{T}.$$

By Proposition 3.1.5 there exists a closed convex cone $\mathcal{C}$ such that $\mathcal{K}_2 = \mathcal{K} \cap \mathcal{C}$. Define

$$\mathcal{S}_2 := \mathcal{S} \boxplus (-\mathcal{C})^*$$

as the Minkowski sum of $\mathcal{S}$ and the dual of $-\mathcal{C}$.

Theorem 3.1.6. *Let $p(x;\theta)$ be a minimally represented regular exponential family. Then the MLE $\hat{\theta}$ based on $\overline{T}$ exists in the* MTP_2 *submodel if and only if $\overline{T} \in \mathcal{S}_2$, in which case $\hat{\theta}$ is uniquely defined by*

(a) *primal feasibility:* $\hat{\theta} \in \mathcal{K}_2$,

(b) *dual feasibility:* $\hat{\sigma} := \nabla A(\hat{\theta}) \in \mathcal{S}$ *with* $\hat{\sigma} - \overline{T} \in \mathcal{C}^*$,

(c) *complementary slackness:* $\langle \hat{\theta}, \hat{\sigma} - \overline{T} \rangle = 0$.

Proof.
The maximum likelihood estimation problem can be formulated as the following optimization problem:

$$\begin{aligned} \max_{\theta \in \mathcal{K}} \quad & \left\langle \theta, \overline{T} \right\rangle - A(\theta) \\ \text{s.t.} \quad & \theta \in \mathcal{C}. \end{aligned} \tag{3.3}$$

Since $A(\theta)$ is convex on $\mathcal{K}$, (3.3) is a convex optimization problem. The Lagrangian is

$$\mathcal{L}(\theta, \lambda) = \left\langle \theta, \overline{T} \right\rangle - A(\theta) + \langle \theta, \lambda \rangle,$$

where $\lambda \in \mathcal{C}^*$. Let A^* denote the conjugate dual of A with domain $\mathcal{S}$. Then

$$\max_{\theta \in \mathcal{K}} \mathcal{L}(\theta, \lambda) = A^*\left(\overline{T} + \lambda\right),$$

and hence the dual optimization problem is given by

$$\begin{aligned} \min_{\sigma \in \mathcal{S}} \quad & A^*(\sigma) \\ \text{s.t.} \quad & \sigma - \overline{T} \in \mathcal{C}^*. \end{aligned} \tag{3.4}$$

The MLE exists if and only if the primal and dual problems are feasible. The primal problem 3.3 is feasible by the assumption $\mathcal{K}_2 \neq \emptyset$. The dual problem 3.4 is feasible if and only if $\overline{T} \in \mathcal{S}_2$. The characterization of the MLE then follows from the KKT conditions. □

3.2 Binary Distributions

For unrestricted binary distributions the MLE exists only if all 2^d states are observed at least once. We will see in this section that in MTP_2 binary distributions the MLE *can* exist with a considerably smaller sample size. Before delving deeper into the subject we need first establish some concepts.

A *partial order* is any binary relation $\sim$ that is reflexive, antisymmetric and transitive.

Example 3.2.1
$\leq$ is a partial order of a set A. Indeed, let $a, b, c \in A$. Then the following statements hold.

1. $a \leq a$ is true, so $\leq$ is reflexive;
2. $a \leq b$ and $b \leq a \Rightarrow a = b$, so $\leq$ is antisymmetric;
3. $a \leq b$ and $b \leq c \Rightarrow a \leq c$, so $\leq$ is transitive. ∘

A *partially ordered set* (or *poset* for short) is defined as an ordered pair $P = (X, \leq)$ where X is called the *ground set* of P and $\leq$ is the *partial order* of P.

A poset L, in which every pair of elements has both a unique supremum and a unique infimum, is called a *lattice*. For a lattice L we say that a subset L' of L forms a *sublattice* of L if for any two elements $x, y \in L$ it holds that $x \wedge y \in L'$ and $x \vee y \in L'$

In this section we turn our attention to binary distributions, i.e. distributions over the sample space $\mathcal{X} = \{-1, 1\}^d$. To simplify notation, we will often use the following bijection between $\mathcal{X}$ and the set $\mathbf{B}_d$ of all subsets of $\{1, ..., d\}$. An

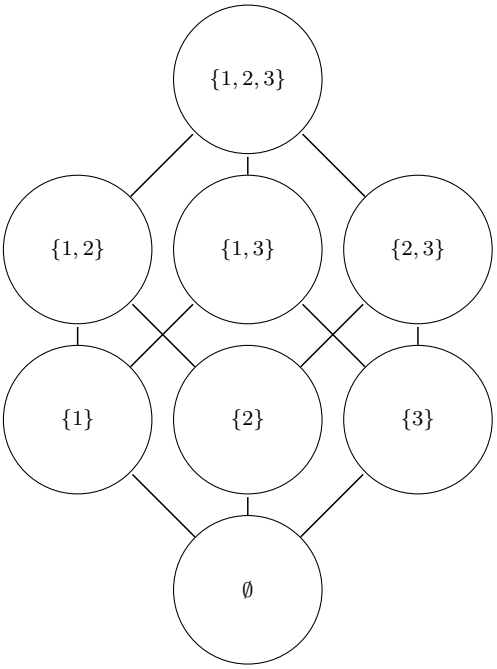


Figure 3.1: Lattice of all subsets of the set $\{1, 2, 3\}$ ordered by "is subset of"

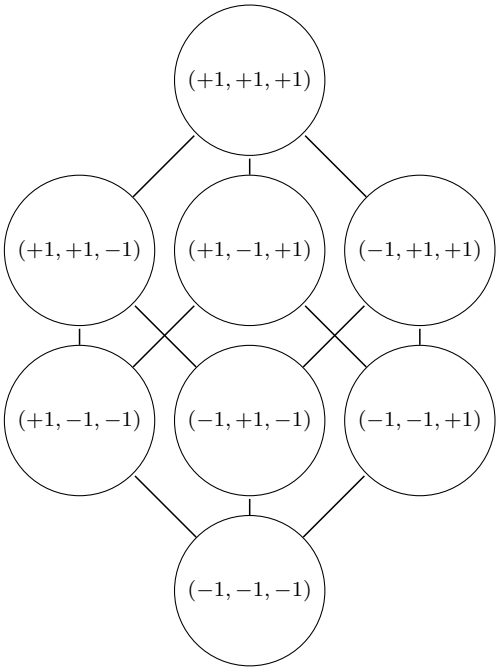


Figure 3.2: Lattice of points in $\{-1, 1\}^3$ ordered by the min-max operators

element $x \in \mathcal{X}$ maps to the subset of all $i \in \{1, ..., d\}$ for which $x_i = 1$. This is illustrated in Figure 3.1 and Figure 3.2, where for example the subset $\{1, 3\} \in \mathbf{B}_3$ maps to the point $(1, -1, 1) \in \mathcal{X}$. Note that $\mathcal{X}$ and $\mathbf{B}_d$ are also isomorphic as lattices because the minimum and maximum operators $\wedge, \vee$ on $\mathcal{X}$ correspond to the set of operations $\cap, \cup$ in $\mathbf{B}_d$. Note that any subset $A \subseteq \mathcal{L}$ of a lattice $\mathcal{L}$ can be written as

$$\bigvee_{\alpha \in A} \{\alpha\},$$

which implies that the full lattice $\mathcal{L}$ is generated by the set of all singleton sets $\{\alpha\} \in \mathcal{L}$.

Let $\mathbb{P}(\mathcal{X})$ denote the set of all probability distributions over $\mathcal{X}$ and let $\mathcal{P}_2$ be the set of all totally positive binary distributions, that is

$$\mathcal{P}_2 = \{p \in \mathbb{P}(\mathcal{X}) \mid \forall x, y \in \mathcal{X} : p(x \wedge y)p(x \vee y) \geq p(x)p(y)\}.$$

Note that $\mathcal{P}_2$ is a closed set, and it contains its limit points, so it is a compact set. In addition, it is *geometrically convex*, that is

$$p_1, p_2 \in \mathcal{P}_2 \Rightarrow c^{-1}\sqrt{p_1 p_2} \in \mathcal{P}_2$$

where

$$c := \sum_{x \in \mathcal{X}} \sqrt{p_1(x) p_2(x)} \leq 1$$

and by the Cauchy-Schwarz inequality $c < 1$ unless $p_1 = p_2$. Note also for any $p \in \mathcal{P}_2$ its support $\text{supp}(p) = \{x : p(x) > 0\}$ is always a sublattice of $\mathcal{X}$. Indeed, since

$$p(x) > 0 \text{ and } p(y) > 0 \Rightarrow p(x \wedge y) p(x \vee y) \geq p(x) p(y) > 0,$$

and hence $p(x \wedge y) > 0$ and $p(x \vee y) > 0$, showing that $x \wedge y, x \vee y \in \text{supp}(p)$.

Consider a sample $U = \{x_1, ..., x_n\}$ with likelihood function

$$L(p) = \prod_{i=1}^{n} p(x_i)$$

and let $\mathcal{L}(U)$ be the smallest sublattice of $\mathcal{X}$ containing U. We now show that the support of the MLE of $\mathcal{P}_2$ is given by $\mathcal{L}(U)$.

Theorem 3.2.2. *The likelihood function $L(p)$ attains its maximum over $\mathcal{P}_2$ in a unique point $\hat{p}$. In addition, it holds that* $\text{supp}(\hat{p}) = \mathcal{L}(U)$.

Proof.
Since $L(p)$ is continuous and since $\mathcal{P}_2$ is compact, the maximum is attained with probability 1. To show uniqueness, assume for contradiction that $\hat{p}_1 \neq \hat{p}_2$ both maximize L. Then

$$L\left(c^{-1} \sqrt{\hat{p}_1 \hat{p}_2}\right) = c^{-n} \sqrt{L(\hat{p}_1)\, L(\hat{p}_2)} > L\left(\hat{p}_i\right)$$

for $i = 1, 2$ in contradiction with our assumption that $\hat{p}_1, \hat{p}_2$ were maximizers.

For the final part of the theorem, note that $U \subseteq \text{supp}(\hat{p})$ and so $\mathcal{L}(U) \subseteq \text{supp}(\hat{p})$. We show that $\mathcal{L}(U) \supseteq \text{supp}(\hat{p})$ by contradiction. Assume $\mathcal{L}(U) \subsetneq \text{supp}(\hat{p})$, i.e. there exists an element in $\text{supp}(\hat{p})$ that is not contained in $\mathcal{L}(U)$. Then we can construct $\tilde{p} \in \mathcal{P}_2$ such that $L(\tilde{p}) > L(\hat{p})$, which contradicts the fact that $\hat{p}$ is the MLE. Indeed, let $\tilde{p}$ be $\hat{p}$ projected onto $\mathcal{L}(U)$ and rescaled to be a probability mass function, that is $\tilde{p}(x) \propto p(x) \mathbf{1}_{\mathcal{L}(U)}$. By construction

$$\tilde{p}(x) = \frac{p(x)}{p(x \in \mathcal{L}(U))},$$

hence $\tilde{p} \in \mathcal{P}_2$ and $L(\tilde{p}) \geq L(\hat{p})$ where equality holds if and only if $\mathcal{L}(U) = \mathcal{X}$. □

A *pair-marginal* U_{ij} of a sample U is the $2 \times n$ dimensional matrix where the first row is the i'th entries in each sample in U, and the second row is the j'th entries in each sample in U, that is

$$U_{ij} = \begin{bmatrix} (x_1)_i & (x_2)_i & \cdots & (x_n)_i \\ (x_1)_j & (x_2)_j & \cdots & (x_n)_j \end{bmatrix}.$$

When we say that the pair-marginal *has* (a, b) *represented* we mean that U_{ij} has a column equal to $(a, b)^T$.

Example 3.2.3
Let $U = (x, y, z)$ where

$$x = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \quad y = \begin{pmatrix} -1 \\ -1 \\ -1 \end{pmatrix}, \quad z = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}.$$

Then

$$U_{12} = \begin{pmatrix} 1 & -1 & 1 \\ 1 & -1 & -1 \end{pmatrix}.$$

We see that $(1, 1)$, $(-1, -1)$, and $(1, -1)$ are represented in U_{12} but $(-1, 1)$ is not. ∘

The following result provides us with conditions for the existence of the MLE for binary MTP_2 distributions.

Proposition 3.2.4. *Let $\mathcal{X} = \{-1, 1\}^d$. The MLE exists within $\mathcal{K}_2$ if and only if $\mathcal{L}(U) = \mathcal{X}$. Furthermore, $\mathcal{L}(U) = \mathcal{X}$ if and only if every pair-marginal U_{ij} for $i, j \in V$ has both of $(1, -1)$ and $(-1, 1)$ represented.*

Proof.
We start by showing the first statement. The MLE exists in the binary exponential family if and only if the estimator $\hat{p}$ in the extended family $\mathbb{P}(\mathcal{X})$ has full support. Then, by Theorem 3.2.2, the MTP_2 MLE exists if and only if $\mathcal{L}(U) = \mathcal{X}$.

We show the second statement, first taking the backwards implication using the identification between $\mathcal{X}$ and subsets of V. Suppose every pair-marginal U_{ij} for $i, j \in V$ has both of $(1, -1)$ and $(-1, 1)$ represented. Then for every i there is a set $x_{ij} \in U$ with $i \in x_{ij}$ and $j \notin x_{ij}$ (equivalently, the i'th entry of x equals 1, and the j'th entry if x equals -1.) But then

$$\{i\} = \bigcap_{j \in V \setminus i} x_{ij} \in \mathcal{L}(U) \quad \text{for all } i.$$

Since the set of all singletons $\{i\}$ for $i \in V$ generates the full lattice $\mathcal{X}$, we obtain $\mathcal{L}(U) = \mathcal{X}$ as desired.

The forward implication is shown by proving its contrapositive. Suppose there is a pair ij such that all sets $x \in U$ have the property that

$$i \in x \Rightarrow j \in x.$$

The set of subsets y satisfying this property form a proper sublattice $\mathcal{L}' \subset \mathcal{X}$. Since $\mathcal{L}(U) \subseteq \mathcal{L}'$ we obtain that $\mathcal{L}(U) \neq \mathcal{X}$, which completes the proof. □

For general d, a minimal generating set of $\{-1, 1\}^d$ is of order $\mathcal{O}(d)$ and there always exists a minimal generating set of size exactly d. Indeed, let $d \in \mathbb{N}$ be given. Then the vectors $(1, -1, -1, ..., -1), (-1, 1, -1, -1, ..., -1), ..., (-1, -1, ..., -1, 1)$ generate all of $\{-1, 1\}^d$. This set cannot be reduced, i.e. none of its subsets generates $\{-1, 1\}^d$. As such every sample supported on these points will admit a unique MLE under the MTP_2 constraint. Therefore for binary MTP_2 distributions d samples can be sufficient for an MLE to exist, considerably fewer than the otherwise required 2^d different observed states for unrestricted binary distributions.

We now give conditions for when $p \in \mathcal{P}_2$ has full support, as it will be crucial in order to analyze when the MLE of a binary distribution over a graph has full support.

Proposition 3.2.5. *Let $p \in \mathcal{P}_2$ and let $x \in \mathcal{X}$. If $p_{ij}(x_i, x_j) > 0$ for all pairs (i, j) then $p(x) > 0$.*

Proof.
For every i, j let $y^{(ij)} \in \text{supp}(p)$ such that $y^{(ij)}_{ij} = (x_i, x_j)$. Let $A \backslash B$ be the partition of V such that $x_i = -1$ for $i \in A$ and $x_i = 1$ for $i \in B$. For each $i \in A$ define $z^{(i)} = \max_{j \in B} y^{(ij)}$. By construction $z^{(i)}_i = -1$ and $z^{(i)}_B = (1, ..., 1)$. In addition, $z^{(i)} \in \text{supp}(p)$ because $\text{supp}(p)$ is a lattice. Since $x = \min_{i \in A} z^{(i)}$, it holds that $x \in \text{supp}(p)$ because $\text{supp}(p)$ is a lattice. □

We get as an immediate result the following.

Corollary 3.2.6. *If $p \in \mathcal{P}_2$ then p has full support on $\mathcal{X}$ if and only if each pair-margin p_{ij} has full support.*

3.3 Binary Models over Graphs

Let $G = (V, E)$ be a graph, and let $\mathcal{P}_2(G)$ denote the set of distributions in $\mathcal{P}_2$ that lie in the completion of the exponential family for the graphical model over G, i.e.

$$\mathcal{P}_2(G) = \mathcal{P}_2 \cap \overline{\mathcal{P}},$$

where $\overline{\mathcal{P}}$ denotes the set of extended Markov distributions. Note that $\mathcal{P}_2(G)$ is compact and geometrically convex. As such the MLE over $\mathcal{P}_2(G)$ exists and is unique. We wish to extend Proposition 3.2.4 to binary graphical models. In order to do so we need a few results.

Lemma 3.3.1. *If the sample edge-margin U_{ij} has both of $(1,-1)$ and $(-1,1)$ represented for all $ij \in E$, then $\operatorname{supp}(\hat{p}_{ij}) = \{-1,1\}^2$ for all $ij \in E$.*

Proof.
Since $\hat{p}$ is MTP_2, so are its marginals $\hat{p}_{ij}$. Therefore $\operatorname{supp}(\hat{p}_{ij})$ is a lattice containing U_{ij}. It then holds that if U_{ij} has both of $(1,-1)$ and $(-1,1)$ represented, then $\operatorname{supp}(\hat{p}_{ij}) = \{-1,1\}^2$. □

Let $\text{ne}(i)$ denote the neighbours of node $i \in V$ in G.

Lemma 3.3.2. *Assume that every edge-margin U_{ij} has both of $(1,-1)$ and $(-1,1)$ represented. If $\hat{p}_{\text{ne}(i)}\big(x_{\text{ne}(i)}\big) > 0$ for some $x_{\text{ne}(i)}$, then $\hat{p}_{i\cup\text{ne}(i)}\big(x_{i\cup\text{ne}(i)}\big) > 0$ for every x_i.*

Proof.
Since $\hat{p}_{\text{ne}(i)}\big(x_{\text{ne}(i)}\big) > 0$ it holds that $\hat{p}_{i\cup\text{ne}(i)}\big(x_{i\cup\text{ne}(i)}\big) > 0$ for some x_i, for example $x_i = 1$. We need to show that $\hat{p}_{i\cup\text{ne}(i)}\big(y_{i\cup\text{ne}(i)}\big) > 0$ also if $y_i = -1$ and $y_{\text{ne}(i)} = x_{\text{ne}(i)}$. Let $z_{i\cup\text{ne}}$ be such that $z_i = -1$ and $z_{\text{ne}(i)} = (1, ..., 1)$. Since $\hat{p} \in \mathcal{P}_2(G)$, its support is a lattice and the same applies to each margin of $\hat{p}$. Since

$$y_{i\cup\text{ne}(i)} = x_{i\cup\text{ne}(i)} \wedge z_{i\cup\text{ne}(i)},$$

in order to show that $y_{i\cup\text{ne}(i)} \in \operatorname{supp}\big(\hat{p}_{i\cup\text{ne}(i)}\big)$ it is enough to show that this holds for $z_{i\cup\text{ne}(i)}$. By assumption for each $j \in \text{ne}(i)$ the edge-margin U_{ij} has $(-1,1)$ represented. Then, a fortiori, there exists a point $u^{(j)} \in \mathcal{X}$ such that $u_i^{(j)} = -1$ and $u_j^{(j)} = 1$. The support of $\hat{p}$ contains all elements in U and therefore $\hat{p}\big(u^{(j)}\big) > 0$ for all $j \in \text{ne}(i)$. Let u be the element-wise maximum of all $u^{(j)}$. This point lies in $\operatorname{supp}(\hat{p})$ because it forms a lattice. By construction, $u_{i\cup\text{ne}(i)} = z_{i\cup\text{ne}(i)}$, which proves that $z_{i\cup\text{ne}(i)}$ and also $y_{i\cup\text{ne}(i)}$ lie in the support of $\hat{p}_{i\cup\text{ne}(i)}$. The proof is analogous for the case where $x_i = -1$. □

We are now ready to state and prove the extension to Proposition 3.2.4 to binary graphical models.

Theorem 3.3.3. *If every sample edge-margin $U_{ij}, ij \in E$ has both $(1,-1)$ and $(-1,1)$ represented, then the MLE $\hat{p} \in \mathcal{P}_2(G)$ is unique and has full support.*

Proof.
From Corollary 3.2.6 we see that $\hat{p}$ has full support if and only if the marginal support $\operatorname{supp}(\hat{p}_{ij})$ is full for all $i, j \in V$. When $ij \in E$ this follows from Lemma 3.3.1. Consider a pair $ij \notin E$. Since $\hat{p} \in \overline{\mathcal{P}}$, it satisfies the local Markov property with respect to G. Hence for any $x_i, x_j \in \{-1, 1\}$, it holds that

$$\begin{aligned}\hat{p}_{ij}(x_i, x_j) &= \sum_{x_{\mathrm{ne}(i)\cup\mathrm{ne}(j)}} \hat{p}\big(x_i, x_j \mid x_{\mathrm{ne}(i)\cup\mathrm{ne}(j)}\big)\, \hat{p}\big(x_{\mathrm{ne}(i)\cup\mathrm{ne}(j)}\big) \\ &= \sum_{x_{\mathrm{ne}(i)\cup\mathrm{ne}(j)}} \hat{p}\big(x_i \mid x_{\mathrm{ne}(i)}\big)\, \hat{p}\big(x_j \mid x_{\mathrm{ne}(j)}\big)\, \hat{p}\big(x_{\mathrm{ne}(i)\cup\mathrm{ne}(j)}\big)\,.\end{aligned}$$

Since there is at least one $x_{\mathrm{ne}(i)\cup\mathrm{ne}(j)}$ contained in $\operatorname{supp}\big(\hat{p}_{\mathrm{ne}(i)\cup\mathrm{ne}(j)}\big)$, then by Lemma 3.3.2 both $\hat{p}\big(x_i, x_{\mathrm{ne}(i)}\big)$ and $\hat{p}\big(x_j, x_{\mathrm{ne}(j)}\big)$ are strictly positive and therefore also the corresponding summand is strictly positive. It then follows that $\hat{p}_{ij}(x_i, x_j) > 0$. □

Theorem 3.3.3 implies that

Corollary 3.3.4. *If G is bipartite, then the minimal sample size required for existence of the MLE is $n = 2$. More generally, for arbitrary graphs the minimal sample size for existence of the MLE is of the order of the maximal clique size.*

In summary, we have seen that the minimal sample size for existence of the MLE goes from 2^d for unrestricted binary distributions to d for MTP_2 binary distributions to $\mathcal{O}(M)$ where M is the maximal clique size for MTP_2 binary distributions on graphs.

4. Quadratic Exponential Families

In this chapter we give an introduction to *quadratic exponential families*, as they have some properties relevant to the subject of total positivity. We characterize MTP_2 quadratic exponential families, and discuss in great detail a famous and important example of such families, namely *Gaussian graphical models* in the continuous setting. In chapter 5 we present quadratic exponential families with binary sample spaces.

Author's disclaimer: In the spring of 2019 I wrote a project on MTP_2 in Gaussian distributions titled *Multivariate Total Positivity of Order 2 in Gaussian Models*, supervised by Steffen L. Lauritzen (who also supervises this thesis.) The sections 4.2 - 4.5 in this chapter, regarding totally positive Gaussian graphical models, are more or less a repetition of that project.

4.1 Quadratic Exponential Families

The density function of a *quadratic exponential family* is of the form

$$p(x; h, J) = \exp\left(h^T x + x^T J x/2 - A(h, J)\right)$$

with $h \in \mathbb{R}^d$ and $J \in \mathbb{S}^d$. In the binary setting (when $\mathcal{X}_v = \{-1, 1\}$ for all $v \in V$) we require $J_{ii} = 0$ for all i in order to obtain a minimally represented exponential family. Quadratic exponential families only model the pairwise interactions betweens nodes, in other words interactions are only on the edges of the underlying graph G. In the binary setting we refer to a quadratic exponential family as an *Ising model*. This type of model, along with multivariate Gaussian graphical models which are also quadratic exponential families, will be examined in much greater detail in this thesis.

Proposition 4.1.1. *The subfamily of* MTP_2 *distributions in a quadratic exponential family is obtained by intersecting* $\mathcal{K}$ *with a polyhedral cone* $\mathcal{C}$*, namely the cone*

$$\mathbb{S}^d_+ = \{J \in \mathbb{S}^d \mid J_{ij} \geq 0 \quad \textit{for all} \quad i \neq j\}.$$

Proof.
The density of function of a symmetric QEF can be written as

$$p(x;h,J) = \frac{\exp\left(x^T Jx/2\right)}{\exp\left(A(h,J)\right)} = \frac{\exp\left(\sum_{uv\in E} x_u x_v J_{uv}\right)}{\exp\left(A(h,J)\right)} = \frac{\prod_{uv\in E}\exp\left(x_u x_v J_{uv}\right)}{\exp\left(A(h,J)\right)} \tag{4.1}$$

We can now apply Theorem 2.2.8 to see that a quadratic exponential family is MTP_2 if and only if $\psi(x) := \exp\left(J_{uv}x_u x_v\right)$ is MTP_2 for all $u \neq v$. This is the case if and only if for every x, y that differ in two coordinates u, v with $x_u < y_u$ and $x_v > y_v$, it holds that

$$\begin{aligned}\psi(x\wedge y)\psi(x\vee y) \geq \psi(x)\psi(y) &\iff y_u x_v J_{uv} + y_v x_u J_{uv} - x_u x_v J_{uv} - y_u y_v J_{uv} \geq 0\\ &\iff J_{uv}\left(y_u - x_u\right)\left(x_v - y_v\right) \geq 0,\end{aligned}$$

which is equivalent to $J_{uv} \geq 0$. □

4.2 Gaussian Graphical Models

A random vector $X \in \mathbb{R}^p$ is distributed according to the multivariate Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ with mean vector $\mu \in \mathbb{R}^p$ and covariance matrix $\Sigma \in \mathbb{S}^p_{\succ 0}$ if it has density function

$$f_{\mu,\Sigma}(x) = (2\pi)^{-p/2}(\det\Sigma)^{-1/2}\exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}, \quad x \in \mathbb{R}^p.$$

We denote the inverse covariance matrix, also known as the *concentration matrix* by K. In terms of $K = \Sigma^{-1}$ and using the trace inner product on $\mathbb{S}^p$, the density $f_{\mu,\Sigma}(x)$ can equivalently be formulated as:

$$f_{\mu,K}(x) = \exp\left\{\mu^T Kx - \frac{1}{2}x^T Kx - \frac{p}{2}\log(2\pi) + \frac{1}{2}\log\det(K) - \frac{1}{2}\mu^T K\mu\right\}.$$

The Gaussian distribution is a quadratic exponential family with canonical parameters $\left(-\mu^T K, K\right)$ and sufficient statistics $\left(x, \frac{1}{2}xx^T\right)$. Indeed, it has the form (4.1) where $J = -K$, $h^T = \mu^T K$, and $A(h,J) = \frac{p}{2}\log(2\pi) + \frac{1}{2}\mu^T K\mu - \log\det(K)$.

Let $G = (V, E)$ be an undirected graph with vertices $V = \{1, ..., p\}$ and edges E. A random vector $X \in \mathbb{R}^p$ is said to satisfy the Gaussian graphical model with graph G if X has a multivariate Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ with

$$\left(\Sigma^{-1}\right)_{i,j} = 0 \quad \text{for all } (i,j) \notin E.$$

The graph G therefore describes the sparsity pattern of the concentration matrix, and G is known as the *concentration graph*. Missing edges in G also correspond to conditional independence relations in the corresponding Gaussian graphical model, which we will see in Corollary 4.2.2.

Below is a result that will help us illustrate the consequences of MTP_2 (a full proof can be found in Anderson (1962)).

Proposition 4.2.1. *Let $X \in \mathbb{R}^p$ be distributed as $\mathcal{N}(\mu, \Sigma)$ and partition X into two components $X_A \in \mathbb{R}^a$ and $X_B \in \mathbb{R}^b$ such that $a + b = p$. Let μ and Σ be partitioned correspondingly as*

$$\mu = \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_{A,A} & \Sigma_{A,B} \\ \Sigma_{B,A} & \Sigma_{B,B} \end{pmatrix},$$

where $\Sigma_{A,A} \in \mathbb{S}^a_{\succ 0}$ and $\Sigma_{B,B} \in \mathbb{S}^b_{\succ 0}$. Then the following two statements hold:

1. *The marginal distribution of X_A is $\mathcal{N}(\mu_A, \Sigma_{A,A})$.*
2. *The conditional distribution of $X_A \mid X_B = x_B$ is $\mathcal{N}\left(\mu_{A,A|B}, \Sigma_{A,A|B}\right)$, where*

$$\mu_{A,A|B} = \mu_A + \Sigma_{A,B}\left(\Sigma_{B,B}\right)^{-1}(x_B - \mu_B), \quad \text{and}$$

$$\Sigma_{A,A|B} = \Sigma_{A,A} - \Sigma_{A,B}\left(\Sigma_{B,B}\right)^{-1}\Sigma_{B,A}.$$

It is often more convenient for us to view this proposition in terms of K, that is

$$K = \begin{pmatrix} K_{A,A} & K_{A,B} \\ K_{B,A} & K_{B,B} \end{pmatrix}.$$

As seen in Lauritzen (2018), by using Schur's complement it is possible to obtain

$$\left(K_{A,A}\right)^{-1} = \Sigma_{A,A} - \Sigma_{A,B}\left(\Sigma_{B,B}\right)^{-1}\Sigma_{B,A} = \Sigma_{A|B} \tag{4.2}$$

and likewise

$$\left(\Sigma_{A,A}\right)^{-1} = K_{A,A} - K_{A,B}\left(K_{B,B}\right)^{-1}K_{B,A}.$$

Thus the marginal distribution of X_A can also be written as

$$\mathcal{N}\left(\mu_A, \left(\Sigma_{A,A}\right)^{-1}\right) = \mathcal{N}\left(\mu_A, K_{A,A} - K_{A,B}\left(K_{B,B}\right)^{-1}K_{B,A}\right)$$

and the conditional distribution of $X_A|X_B = x_B$ can be written as $\mathcal{N}\left(\mu_{A|B}, K_{A|B}\right)$ where

$$\mu_{A|B} = \mu_A - \left(K_{A,A}\right)^{-1}K_{A,B}\left(x_B - \mu_B\right) \quad \text{and} \quad K_{A|B} = K_{A,A}.$$

These properties have interesting implications to the interpretation of zeros in the covariance matrix and the concentration matrix. Zeros correspond to conditional independence relations. For disjoint subsets $A, B, C \subset V = \{1, ..., p\}$, we denote the statement that X_A is conditionally independent of X_B given X_C by $X_A \perp\!\!\!\perp X_B \mid X_C$, and if $C = \emptyset$ we write $X_A \perp\!\!\!\perp X_B$.

Corollary 4.2.2. *Let $X \in \mathbb{R}^p$ be distributed as $\mathcal{N}(\mu, \Sigma)$ and let $i, j \in \{1, ..., p\}$ with $i \neq j$. Then the following statements hold:*

(a) $X_i \perp\!\!\!\perp X_j$ if and only if $\Sigma_{i,j} = 0$;

(b) $X_i \perp\!\!\!\perp X_j \mid X_{V\setminus\{i,j\}}$ if and only if $K_{i,j} = 0$;

(c) $K_{i,j} = 0$ if and only if $\det\left(\Sigma_{V\setminus\{i\},V\setminus\{j\}}\right) = 0$.

Proof.
Statement (a) follows directly from the expression for the conditional mean in Proposition 4.2.1 (b), as $\mu_{i|j} = \mu_i$ will hold.

From the expression for the conditional covariance in Proposition 4.2.1 (b) it follows that $\Sigma_{\{i,j\}|V\setminus\{i,j\}} = \left(K_{\{i,j\},\{i,j\}}\right)^{-1}$. To prove statement (b), note first that it follows from (a) that $X_i \perp\!\!\!\perp X_j \mid X_{V\setminus\{i,j\}}$ if and only if the 2×2 conditional covariance matrix $\Sigma_{\{i,j\}|(V\setminus\{i,j\})}$ is diagonal. This is the case if and only if $K_{\{i,j\},\{i,j\}}$ is diagonal or in other words, if $K_{i,j} = 0$.

Statement (c) holds as a consequence of the cofactor formula for matrix inversion, since

$$K_{i,j} = \left(\Sigma^{-1}\right)_{i,j} = (-1)^{i+j} \frac{\det\left(\Sigma_{V\setminus\{i\},V\setminus\{j\}}\right)}{\det(\Sigma)}$$

which is zero if and only if the numerator of the right hand side fraction is zero, completing the proof. □

This shows that for undirected Gaussian graphical models a missing edge (i, j) in the concentration graph corresponds to the conditional independence relation $X_i \perp\!\!\!\perp X_j \mid X_{V\setminus\{i,j\}}$.

4.3 Gaussian Likelihood and Convex Optimization

Suppose we are given n i.i.d. observations $X_1, ..., X_n$ from $\mathcal{N}(\mu, \Sigma)$, and let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean. We define the sample covariance matrix

as

$$S = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(X_i - \overline{X}\right)^T.$$

We can write the log-likelihood function in terms of the sufficient statistics $\overline{X}$ and S. Ignoring the normalizing constant, the Gaussian log-likelihood expressed as a function of (μ, Σ) is

$$\begin{aligned} l(\mu,\Sigma) &\propto -\frac{n}{2}\log\det(\Sigma) - \frac{1}{2}\sum_{i=1}^{n}\left(X_i-\mu\right)^T\Sigma^{-1}\left(X_i-\mu\right) \\ &= -\frac{n}{2}\log\det(\Sigma) - \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\left(\sum_{i=1}^{n}\left(X_i-\mu\right)\left(X_i-\mu\right)^T\right)\right) \\ &= -\frac{n}{2}\left(\log\det(\Sigma) - \mathrm{tr}\left(S\Sigma^{-1}\right) - \left(\overline{X}-\mu\right)^T\Sigma^{-1}\left(\overline{X}-\mu\right)\right) \end{aligned}$$

where we expanded $X_i - \mu = \left(X_i - \overline{X}\right) + \left(\overline{X} - \mu\right)$ for the last equality, and we used the fact that $\sum_{i=1}^{n}\left(X_i - \overline{X}\right) = 0$. Hence, it can be seen that in the saturated model where $(\mu, \Sigma) \in \mathbb{R}^p \times \mathbb{S}^p_{\succ 0}$, the MLE is given by

$$\hat{\mu} = \overline{X} \quad \text{and} \quad \hat{\Sigma} = S,$$

assuming that $S \in \mathbb{S}^p_{\succ 0}$.

Let $(\mu, \Sigma) \in \mathbb{R}^p \times \Theta$ where $\Theta \subseteq \mathbb{S}^p_{\succ 0}$. The maximum likelihood estimation problem for Σ is

$$\begin{aligned} \max_{\Sigma} \quad & -\log\det(\Sigma) - \mathrm{tr}\left(S\Sigma^{-1}\right) \\ \text{s.t.} \quad & \Sigma \in \Theta. \end{aligned} \tag{4.3}$$

It is often convenient to write this optimization problem in terms of the concentration matrix K:

$$\begin{aligned} \max_{K} \quad & \log\det(K) - \mathrm{tr}\left(SK\right) \\ \text{s.t.} \quad & K \in \mathcal{K}, \end{aligned} \tag{4.4}$$

where $\mathcal{K} = \Theta^{-1}$. For a Gaussian graphical model with graph $G = (V, E)$ the constraints are given by $K \in \mathcal{K}_G$, where

$$\mathcal{K}_G = \left\{K \in \mathbb{S}^p_{\succ 0} \mid K_{i,j} = 0 \text{ for all } i \neq j \text{ with } (i,j) \notin E\right\}.$$

Note that $\mathcal{K}_G$ is a convex cone obtained by intersecting the convex cone $\mathbb{S}^p_{\succ 0}$ with a linear subspace. We call $\mathcal{K}_G$ the *cone of concentration matrices*. We now show that the objective function (4.4) as a function of K is concave over its full domain $\mathbb{S}^p_{\succ 0}$. Since $\mathcal{K}_G$ is a convex cone, this implies that maximum likelihood estimation for Gaussian graphical models is a convex optimization problem. The proof of the following proposition can be found in Boyd and Vandenberghe (2004).

Proposition 4.3.1. *The function $f(Y) = \log\det(Y) - \operatorname{tr}(SY)$ is concave on its domain $\mathbb{S}^p_{\succ 0}$.*

As a consequence of Proposition 4.3.1 we can study the dual of (4.4) with $\mathcal{K} = \mathcal{K}_G$. The Lagrangian of this convex optimization problem is given by:

$$\begin{aligned}\mathcal{L}(K,\nu) &= \log\det(K) - \operatorname{tr}(SK) - 2\sum_{(i,j)\notin E,\ i\neq j} \nu_{i,j}K_{i,j} \\ &= \log\det(K) - \sum_{i=1}^{p} S_{i,i}K_{i,i} - 2\sum_{(i,j)\in E} S_{i,j}K_{i,j} - 2\sum_{(i,j)\notin E,\ i\neq j} \nu_{i,j}K_{i,j},\end{aligned}$$

where $\nu = (\nu_{i,j})_{(i,j)\notin E}$ are the Lagrangian multipliers. We omit the constraint $K \in \mathbb{S}^p_{\succ 0}$ for simplicity, and this can be done since we have assumed that K is in the domain of $\mathcal{L}$. Maximizing $\mathcal{L}(K,\nu)$ with respect to K gives

$$\left(\hat{K}^{-1}\right)_{i,j} = \begin{cases} S_{i,j} & \text{if } i = j \text{ or } (i,j) \in E \\ \nu_{i,j} & \text{otherwise.} \end{cases}$$

The Lagrange dual function is obtained by plugging in $\hat{K}$ for K in $\mathcal{L}(K,\nu)$, which results in

$$g(\nu) = \log\det(\hat{K}) - \operatorname{tr}(\hat{K}^{-1}\hat{K}) = \log\det(\hat{K}) - p.$$

The dual problem becomes

$$\begin{aligned} \min_{\Sigma\in\mathbb{S}^p_{\succ 0}} \quad & -\log\det(\Sigma) - p \\ \text{s.t.} \quad & \Sigma_{i,j} = S_{i,j} \ \text{ for all } \ i = j \text{ or } (i,j) \in E. \end{aligned} \tag{4.5}$$

4.4 Totally Positive Gaussian Graphical Models

An MTP_2 Gaussian distribution has some attractive qualities. For instance, it is straightforward to see, from the definition of MTP_2, that a multivariate Gaussian distribution with a positive definite covariance matrix Σ is MTP_2 if and only if the

concentration matrix K is a symmetric M-matrix. Hence, all partial correlations will be non-negative:

$$\rho_{X_i X_j | V \setminus \{i,j\}} = -\frac{K_{ij}}{\sqrt{K_{ii} K_{jj}}} \geq 0 \quad \text{for all} \quad i, j \in \{1, ..., p\}.$$

In addition, since K is an M-matrix, $K^{-1} = \Sigma$ will have positive values in all entries. This is how we see that the MTP_2 property is closed under conditioning and marginalization in the Gaussian case. Indeed, since any marginal covariance matrix $\Sigma_{M,M}$ of Σ also has positive values in all entries. Moreover, by proposition 4.2.1 for a partitioning of Σ it holds that $\Sigma_{A|B} = K_{A,A}^{-1}$ and since $K_{A,A}$ is an M-matrix, its inverse will be a matrix with positive values in all entries, so the correlation coefficient between any two variables will be positive.

Example 4.4.1
Consider a multivariate Gaussian distribution of $X = (X_1, ..., X_5)$ with mean μ and covariance matrix

$$\Sigma = \begin{bmatrix} 1 & 0.2 & 0.1 & 0.1 & 0.1 \\ 0.2 & 1.04 & 0.02 & 0.02 & 0.52 \\ 0.1 & 0.02 & 1.01 & 1.01 & 0.01 \\ 0.1 & 0.02 & 1.01 & 2.01 & 0.01 \\ 0.1 & 0.52 & 0.01 & 0.01 & 1.26 \end{bmatrix}$$

and concentration matrix

$$K = \Sigma^{-1} = \begin{bmatrix} 1.05 & -0.2 & -0.1 & 0 & 0 \\ -0.2 & 1.25 & 0 & 0 & -0.5 \\ -0.1 & 0 & 2 & -1 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & -0.5 & 0 & 0 & 1 \end{bmatrix}$$

which is an M-matrix, so this is an MTP_2 distribution. The corresponding concentration graph is given below

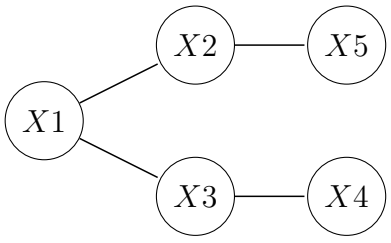


Figure 4.1: The concentration graph corresponding to the concentration matrix K.

◦

In chapter 1 we defined the cone $\mathcal{N}^p$ as the negative closure of $\mathbb{S}^p_{\succ 0}$,

$$\mathcal{N}^p = \{X \in \mathbb{S}^p \mid \exists Y \in \mathbb{S}^p_{\succ 0} \text{ with } X \leq Y \text{ and } \operatorname{diag}(X) = \operatorname{diag}(Y)\}.$$

Since M-matrices form a convex subset of $\mathbb{S}^p_{\succ 0}$, and since $l(K;S) = \log\det(K) - \operatorname{tr}(SK)$ is a strictly concave function of $K \in \mathbb{S}^p_{\succeq 0}$, the optimization problem for finding the MLE for MTP_2 Gaussian models is a convex optimization problem formulated as

$$\begin{aligned} \max_{K} \quad & \log\det(K) - \operatorname{tr}(SK) \\ \text{s.t.} \quad & K \in \mathcal{M}^p. \end{aligned} \tag{4.6}$$

Note that the objective function is proportional to the log-likelihood function. The dual optimization problem to (4.6) was shown in Slawski and Hein (2015) to be given by

$$\begin{aligned} \min_{\Sigma \succeq 0} \quad & -\log\det(\Sigma) - p \\ \text{s.t.} \quad & (\Sigma_{ii} - S_{ii}) = 0, \quad \text{for all } i \in V, \\ & (\Sigma_{ij} - S_{ij}) \geq 0, \quad \text{for all } i \neq j. \end{aligned} \tag{4.7}$$

The constraints in (4.7) hold if and only if $S \in \mathcal{N}^p$, by the definition of $\mathcal{N}^p$.

Using $\mathcal{M}^p$ and $\mathcal{N}^p$ we can now determine conditions for the existence of the MLE in Gaussian MTP_2 models and give a characterization of the MLE. We say that the MLE does *not* exist if the likelihood function does not attain the global maximum. Note that the identity matrix is a strictly feasible point for (4.6). Hence, the MLE does not exist if and only if the likelihood is unbounded. In the following proposition, we give the Karush-Kuhn-Tucker (KKT) conditions, which any pair of primal and dual optimal solutions must satisfy.

Proposition 4.4.2. *Consider a Gaussian MTP_2 model. Then the MLE $\hat{\Sigma}$ (and $\hat{K}$) exists for a given sample covariance matrix S on V if and only if $S \in \mathcal{N}^p$. It is then equal to the unique element $\hat{\Sigma} \succ 0$ that satisfies the following system of equations and inequalities*

$$\left(\hat{\Sigma}^{-1}\right)_{ij} \leq 0 \quad \text{for all } i \neq j, \tag{4.8}$$

$$\left(\hat{\Sigma}_{ii} - S_{ii}\right) = 0 \quad \text{for all } i \in V, \tag{4.9}$$

$$\left(\hat{\Sigma}_{ij} - S_{ij}\right) \geq 0 \quad \text{for all } i \neq j, \tag{4.10}$$

$$\left(\hat{\Sigma}_{ij} - S_{ij}\right)\left(\hat{\Sigma}^{-1}\right)_{ij} = 0 \quad \text{for all } i \neq j \tag{4.11}$$

Proof.
The MLE $\hat{\Sigma}$ needs to be both primal feasible and dual feasible and must therefore satisfy the constraints of both the primal problem (4.6) and the dual problem (4.7). In addition it must attain the global maximum. Inequality (4.8) ensures primal feasibility, (4.9) and (4.10) ensure dual feasibility, and (4.11) ensures that the derivative is zero in $\hat{\Sigma}$, so by the convexity of the problem the unique global maximum will have been reached in $\hat{\Sigma}$.

For Slater's constraint qualification to hold, we require a strictly primal feasible point, i.e. a matrix X contained in the interior of the domain $\mathcal{M}^p$ of the primal problem, to exist. The $p \times p$ dimensional identity matrix is one such point, as it is a strictly positive definite M-matrix. It now holds that the MLE does not exist if and only if the likelihood is unbounded in $\mathcal{M}^p$. As a strictly feasible point exists, Slater's theorem states that strong duality holds for the convex optimization problems (4.6) and (4.7) and therefore it holds that the MLE does not exist if and only if $S \notin \mathcal{N}^p$. □

4.5 Slawski and Hein's Theorem

For any MTP_2 Gaussian model, in order to secure the existence of an MLE, we are in need of a primal feasible point and a dual feasible point. We already have a primal feasible point in the identity matrix (it is an M-matrix), but we have yet to be convinced that S is a dual feasible point. A certain tool has been proven to be invaluable to our studies, namely *ultrametric matrices*.

Definition 4.5.1. *A nonnegative symmetric matrix U is said to be* **ultrametric** *if*

1. $U_{ii} \geq U_{ij}$ *for all* $i, j \in V$, *and*
2. $U_{ij} \geq \min\{U_{ik}, U_{jk}\}$ *for all* $i, j, k \in V$.

There is a connection between M-matrices and ultrametric matrices as shown by Dellacherie, Martinez, and San Martin (2014). We state the theorem here

Theorem 4.5.2. *Let U be an ultrametric matrix with strictly positive entries on the diagonal. Then U is invertible if and only if no two rows in U are equal. In addition, if U is invertible, then U is an inverse M-matrix.*

Ultrametric matrices are highly relevant. Indeed, let R be a symmetric positive semidefinite $p \times p$ matrix such that $R_{ii} = 1$ for all $i \in V$. Consider the weighted

graph $G^+ = G^+(R)$ over V with an edge between i and j whenever $R_{ij} > 0$, and assign to each edge the corresponding positive weight R_{ij}. Define a $p \times p$ matrix Z by setting $Z_{ii} = 1$ for all $i \in V$ and

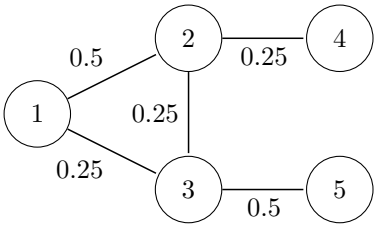
$$Z_{ij} := \max_{p} \min_{uv \in P} R_{uv},$$

for all $i \neq j$, where the maximum is taken over all paths in G^+ between i and j and is set to zero if no such path exists. Z is called the *single-linkage matrix based on* R.

Example 4.5.3
Suppose that

$$R = \begin{bmatrix} 1 & 0.5 & 0.25 & -0.2 & -0.25 \\ 0.5 & 1 & 0.25 & 0.25 & -0.1 \\ 0.25 & 0.25 & 1 & -0.25 & 0.5 \\ -0.2 & 0.25 & -0.25 & 1 & -0.5 \\ -0.25 & -0.1 & 0.5 & -0.5 & 1 \end{bmatrix}$$

Then G^+ and Z are given by



$$Z = \begin{bmatrix} 1 & 0.5 & 0.25 & 0.25 & 0.25 \\ 0.5 & 1 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 1 & 0.25 & 0.5 \\ 0.25 & 0.25 & 0.25 & 1 & 0.25 \\ 0.25 & 0.25 & 0.5 & 0.25 & 1 \end{bmatrix}$$

For instance, in order to find Z_{12} we follow the two paths $1-2$ and $1-3-2$, take the smallest weight along them, compare them, and set Z_{12} to be equal to the larger of the two. We see that the minimum of R_{uv} along the first path is 0.5 and 0.25 along the second path, which gives us that $Z_{12} = 0.5$. ∘

We note that in the example above $R \in \mathbb{S}^p_{\succ 0}$, Z is an ultrametric matrix, $Z \geq R$, Z is invertible, and Z is an inverse M-matrix. This is an example of a general result which we will now show.

Proposition 4.5.4. *Let $R \in \mathbb{S}^p_{\succeq 0}$ such that $R_{ii} = 1$ for all $i \in V$. Then the single-linkage matrix Z based on R is an ultrametric matrix with $Z_{ij} \geq R_{ij}$ for all $i \neq j$. Furthermore, if it holds that $R_{ij} < 1$ for all $i \neq j$, then Z is invertible and therefore an inverse M-matrix.*

Proof.
We first show that Z is an ultrametric matrix, in other words, that Z satisfies the two conditions 1, and 2 from Definition 4.5.1. Z is symmetric by definition, and since R is positive semidefinite, it holds that $R_{ij} \leq 1$ for all i, j. By construction, it holds that $Z_{ij} \leq 1$, such that $Z_{ii} \geq Z_{ij}$ for all i, j, satisfying condition 1.

For condition 2, let $i, j, k \in V$. There are two cases we need to examine. Suppose first that i, j, k all lie in the same connected component of G^+. Let P_1, P_2 be the paths in G^+ such that $Z_{ik} = \min_{uv \in P_1} R_{uv}$ and $Z_{jk} = \min_{uv \in P_2} R_{uv}$. Let P_{12} be the path between i and j obtained by connecting P_1 and P_2 end to end (they must necessarily meet in k.) Then, since P_{12} is one of the possible paths between i and j, the largest of all possible paths' minimum weights between i and j must be at least as large as the minimum weight in P_{12}. More formally,

$$Z_{ij} = \max_{P} \min_{uv \in P} R_{uv} \geq \min_{uv \in P_{12}} R_{uv} = \min \{Z_{ik}, Z_{jk}\}.$$

Now suppose that i, j, k are not all in the same connected component of G^+. In that case at least two of the values Z_{ij}, Z_{ik}, Z_{jk} equals 0 and therefore $Z_{ij} \geq \min \{Z_{ik}, Z_{jk}\} = 0$. Hence, Z is an ultrametric matrix. Note that the edge ij forms a path between i and j, so $Z_{ij} \geq R_{ij}$ for all i and j.

Assume now that $R_{ij} < 1$ for all $i \neq j$. Clearly it also holds that $Z_{ij} < 1$ for all $i \neq j$. Then it follows that no two rows of Z can be equal, since if the i'th row is equal to the j'th row for some $i \neq j$, then it holds that $Z_{ij} = Z_{ii} = Z_{jj}$ which, since $Z_{ii} = Z_{jj} = 1$, contradicts the assumption of $Z_{ij} < 1$. Then by Theorem 4.5.2 it follows that Z is invertible, and thus an inverse M-matrix. □

We obtain the following result as a consequence.

Proposition 4.5.5. *Let $S \in \mathbb{S}^p_{\succeq 0}$ such that $S_{ii} > 0$ for all $i \in V$, and $S_{ij} < \sqrt{S_{ii} S_{jj}}$ for all $i \neq j$. Then there exists an inverse M-matrix Z such that $Z \geq S$ and $Z_{ii} = S_{ii}$ for all $i \in V$.*

Proof.
We apply Proposition 4.5.4 to the normalized version R of S, with entries $R_{ij} := S_{ij} / \sqrt{S_{ii} S_{jj}}$. Since $R_{ij} < 1$ for all $i \neq j$, the corresponding single-linkage matrix Z' is ultrametric with $Z' \geq R$ and Z' is an inverse M-matrix. Define Z by $Z_{ij} = \sqrt{S_{ii} S_{jj}} Z'_{ij}$. Then $Z \geq S$ and $Z_{ii} = S_{ii}$ for all $i \in V$. In addition, Z is an inverse M-matrix because it is simply a rescaling of the inverse M-matrix Z'. □

This matrix property may ring a bell. Indeed, if there exists a matrix Z satisfying the properties from Proposition 4.5.5, the matrix S will be an element of $\mathcal{N}^p$, and so by Proposition 4.4.2 an MLE $\hat{\Sigma}$ will exist. We are now ready to state and prove the main result, namely Slawski and Hein's theorem.

Theorem 4.5.6 (Slawski and Hein)**.** *Consider a Gaussian MTP$_2$ model and let S be the sample covariance matrix. If $S_{ij} < \sqrt{S_{ii}S_{jj}}$ for all $i \neq j$ then the MLE $\hat{\Sigma}$ (and $\hat{K}$) exists and it is unique. In particular, if the number n of observations satisfies $n \geq 2$, then the MLE exists with probability 1.*

Proof.
Since S is a covariance matrix, it holds that $S \in \mathbb{S}^p_{\succeq 0}$ such that $S_{ii} > 0$ for all $i \in V$. We can apply Proposition 4.5.5 to obtain an inverse M-matrix Z such that $Z \geq S$ and $Z_{ii} = S_{ii}$ for all $i \in V$. It then follows that $(Z^{-1})_{ij} \leq 0$ for all $i \neq j$, so Z satisfies the primal feasibility constraint (4.8). Furthermore, $Z_{ii} - S_{ii} = 0$ and $Z_{ij} - S_{ij} \geq 0$, so Z also satisfies the dual feasibility constraints (4.9) and (4.10). By Proposition 4.4.2 the MLE exists, and it is unique by the convexity of the problem.

For the second part of the theorem, denote an arbitrary 2×2 submatrix around the diagonal by

$$S_2 := \begin{bmatrix} S_{ii} & S_{ij} \\ S_{ji} & S_{jj} \end{bmatrix}.$$

Consider n independent observations and let $x_i \in \mathbb{R}^2$ be their subvectors restricted to S_2. Assume without loss of generality that $\mu = 0$. We will show that S_2 is positive definite, since then $S_{ij} < \sqrt{S_{ii}S_{jj}}$ holds.

$$S_2 = \frac{1}{n}\sum_{1=1}^{n} x_i x_i^T = \frac{1}{n}\left(x_1 x_1^T + \cdots x_n x_n^T\right).$$

Fix an arbitrary vector $\lambda \in \mathbb{R}^2$. Then

$$\begin{aligned} \lambda^T S_2 \lambda &= \frac{1}{n}\left(\lambda^T x_1 x_1^T \lambda + \cdots \lambda^T x_n x_n^T \lambda\right) \\ &= \frac{1}{n}\left(\|\lambda^T x_1\|^2 + ... + \|\lambda^T x_n\|^2\right) \geq 0. \end{aligned}$$

Equality holds if and only if $\|\lambda^T x_i\|^2 = 0$ for all $i = 1, ..., n$, which holds if and only if λ and x_i are orthogonal for all $i = 1, ..., n$, which holds if and only if x_i and x_j are linearly dependent for all $i \neq j$. Since we have assumed that the observations are independent, we conclude that $S_{ij} < \sqrt{S_{ii}S_{jj}}$ if and only if $n \neq 1$. It then follows that $S \in \mathcal{N}^p$. □

We have seen that it is easy for us to verify whether data indicates that the underlying Gaussian distribution is MTP_2, and if it does then it also becomes easy for us to verify whether an MLE exists. Algorithms converging to the MLE $\hat{K} = \hat{\Sigma}^{-1} \in \mathcal{M}^p$ have been developed, see for example Lauritzen, Uhler, and Zwiernik (2017).

We end this chapter by showing that in the Gaussian setting any concentration graph is realizable by an MTP_2 distribution. This is an important result for the completeness of Theorem 2.4.3.

Proposition 4.5.7. *Any undirected graph G is realizable as the concentration graph $G(P) = (V, E(P))$ of some MTP_2 Gaussian distribution.*

Proof.
Let A be an adjacency matrix of G, that is, $A_{ij} = 1$ if and only if $ij \in E(P)$. Since G is undirected, A is necessarily symmetric, and since $\mathbb{S}^p_{\succ 0}$ is an open set and the identity matrix $I \in \mathbb{S}^p_{\succ 0}$, it follows that $K = I - \varepsilon A \in \mathbb{S}^p_{\succ 0}$ if $\varepsilon > 0$ is sufficiently small. By construction, K is an M-matrix and its nonzero elements correspond to the edges of the graph G. □

5. Ising Models and Conditional Gaussian Distributions

In this chapter we will introduce *Ising models* and discuss them under MTP_2. We will subsequently introduce the notion of *conditional Gaussian distributions (CG-distributions)* and we will characterize them under MTP_2. In addition, we will concentrate some effort on the special case where the sample space $\mathcal{X} = \{-1, 1\}^\Delta \times \mathbb{R}^\Gamma$, where we establish a link between MTP_2 quadratic exponential families and conditional Gaussian distributions.

5.1 Ising Models

Ising models are named after the german physicist Ernst Ising, though invented by his mentor Wilhelm Lenz (Lenz, 1920). They are mathematical models of ferromagnetism in statistical physics that consist of $|\Delta|$ binary variables $\omega_\delta \in \{-1, 1\}, \delta \in \Delta$, representing magnetic dipole moments of atomic spins. The atoms are usually arranged in a grid, and all interactions between spins are pairwise, i.e., each atom interacts only with its immediate neighbouring atoms. The energy of neighbouring spins that are in agreement is lower than the energy of those in disagreement, and the system tends towards the lowest energy. Higher temperatures disrupt this tendency, causing the spins to depend less on their neighbouring spins, which is why the magnets on your refrigerator lose their magnetic properties when you heat them up.

We shall not delve deeper into the realm of physics in this thesis. However, as mentioned in chapter 4, a property of quadratic exponential families is that the only interactions between variables are pairwise. For this reason, we have decided to steal the name "Ising models" to refer to quadratic exponential families with binary sample space $\mathcal{X} = \{-1, 1\}^{|\Delta|}$. Since Ising models form a quadratic exponential family, their probability mass function is of the form

$$p(x; h, J) = \exp\left(h^T x + x^T J x/2 - A(h, J)\right),$$

with $h \in \mathbb{R}^p$ representing the external field, and since the sample space is binary (hence discrete), $J \in \mathbb{S}_0^d$ represents interaction potential between variables,

where $\mathbb{S}_0^d$ is the set of symmetric matrices in $\mathbb{R}^{d\times d}$ with $J_{ii} = 0$ for all i to ensure minimality of the representation.

Let $\theta = (h, J)$ denote the canonical parameters. For any $i, j \in V$ let $A = V\backslash\{i, j\}$. Let $x, y \in \mathcal{X}$ be any two points satisfying $x_A = y_A$, $x_i = y_j = 1$ and $x_j = y_i = -1$. Equivalently, $x_i > y_i$ and $x_j < y_j$ in the binary setting. Then (see chapter 4) it holds that

$$\begin{aligned}\log\left(\frac{p(x\wedge y)p(x\vee y)}{p(x)p(y)}\right) &= J_{ij}\left(x_iy_j + x_jy_i - x_ix_j - y_iy_j\right)\\ &= 4J_{ij}.\end{aligned}$$

Hence, the conditional log-odds ratios are all equal. We already showed that quadratic exponential families are MTP_2 if and only if $J \in \mathbb{S}_+^d$. This is merely another way of showing the same thing for an Ising model defined by (h, j).

A special example of a symmetric binary distribution is the Ising model with no external field (sometimes called *the palindromic Ising model*, i.e. an Ising model with $h = \mathbf{0}$. The sample space $\mathcal{X} = \{-1, 1\}^{|\Delta|}$ is binary as before, and the family of distributions is of the form

$$p(x) = \frac{1}{c(J)}\exp\left(x^TJx/2\right). \tag{5.1}$$

The space of canonical parameters is the set $\mathbb{S}_0^d$. The mean parameter is, since $\mathbb{E}X_i = 0$, $\Sigma = \Xi = \mathbb{E}XX^T$ which is the correlation matrix. This is a quadratic exponential family, so by Proposition 4.1.1 it is MTP_2 if and only if $J_{ij} \geq 0$ for all $i \neq j$. Note that by letting $J = -K = -\Sigma^{-1}$ (5.1) becomes almost identical to the Gaussian density.

As we saw in chapter 4, a Gaussian distribution is MTP_2 if and only if its covariance matrix is an inverse M-matrix. Since palindromic Ising models are almost identical to Gaussian distributions, it may be interesting to see if there is a similar result for them. As it turns out, there is no such result for MTP_2 Ising models in general. However, we will show that Σ is an inverse M-matrix for MTP_2 palindromic Ising models when the underlying graph $G(J)$ is a cycle.

Before we are able to do that, we need first establish the following. Let (X, Y) be two multivariate random variables with finite second moments and a regular joint covariance matrix Σ. We denote the partial covariance matrix by $\Sigma_{X,X\cdot Y} = \Sigma_{X,X} - \Sigma_{X,Y}\Sigma_{Y,Y}^{-1}\Sigma_{Y,X}$ and the covariance of X in the conditional distribution of X given Y by $\Sigma_{X,X|Y}$. The following result was shown by Baba, Shibata, and Sibuya (2004).

Theorem 5.1.1. *For any multivariate random variables (X, Y) where $X = (X_1, ..., X_m)$ and $Y = (Y_1, ..., Y_k)$, the following two statements are equivalent.*

(i) $\mathbb{E}[X|Y] = a + BY$, *where* $a \in \mathbb{R}^m$ *and* $B \in \mathbb{R}^{k\times k}$, *i.e.,* $\mathbb{E}[X|Y]$ *is an affine function of* Y,

(ii) $\Sigma_{X,X\cdot Y} = \mathbb{E}\left[\Sigma_{X,X|Y}\right]$.

Proposition 5.1.2. *Suppose $p(x; J)$ is an* MTP_2 *palindromic Ising model. Let (X, Y) be two multivariate random variables with finite second moments and a regular joint covariance matrix*

$$\Sigma = \begin{bmatrix} \Sigma_{X,X} & \Sigma_{X,Y} \\ \Sigma_{Y,X} & \Sigma_{Y,Y} \end{bmatrix}.$$

If $J \geq 0$ and $G(J)$ is a cycle, then Σ is an inverse M-matrix.

Proof.
Showing that Σ is an inverse M-matrix is equivalent to showing that for every $i, j \in V$, denoting $Z = (X_i, X_j)$ and $Y = X_{V\setminus\{i,j\}}$, the partial covariance matrix $\Sigma_{Z,Z\cdot Y}$ has only non-negative entries. Since the MTP_2 property is closed under conditioning, it is certainly the case that the entries in $\mathbb{E}\left[\Sigma_{Z|Y}\right]$ will all be non-negative. We will show that $\Sigma_{Z,Z\cdot Y} = \mathbb{E}\left[\Sigma_{Z|Y}\right]$, which will complete the proof. By Theorem 5.1.1, this is equivalent to showing that $\mathbb{E}[Z|Y]$ is an affine function of Y. We show this for $\mathbb{E}[X_i|Y]$.

Denote by $u, v \in V\setminus\{i, j\}$ the two vertices that separate i from the remaining vertices in $V\setminus\{i, j\}$. By the global Markov property, $\mathbb{E}[X_i|Y] = \mathbb{E}[X_i|X_u, X_v]$. Since $X_u, X_v \in \{-1, 1\}$, it holds that

$$\mathbb{E}[X_i|X_u, X_v] = a + bX_u + cX_v + dX_uX_v \tag{5.2}$$

for some $a, b, c, d \in \mathbb{R}$. By the tower property, taking expectations on both sides yields $0 = a + d\rho_{uv}$. If we instead multiply by X_uX_v and then take expectations, we get

$$\begin{aligned} &\mathbb{E}[X_uX_v\mathbb{E}(X_i \mid X_u, X_v)] &&= \mathbb{E}[X_uX_v(a + bX_u + cX_v + dX_uX_v)] \\ \Rightarrow\ &\mathbb{E}[X_iX_uX_v] &&= a\rho_{uv} + b\mathbb{E}\left[X_u^2X_v\right] + c\mathbb{E}\left[X_uX_v^2\right] + d\mathbb{E}\left[X_u^2X_v^2\right] \\ \Rightarrow\ &0 &&= a\rho_{uv} + d, \end{aligned}$$

where $\mathbb{E}[X_i X_u X_v] = -\mathbb{E}[X_i X_u X_v] = 0$ by symmetry of the distribution. This yields a system of two equalities and two unknowns a, d:

$$\begin{array}{l} a + d\rho_{uv} = 0 \\ a\rho_{uv} + d = 0 \end{array} \quad \Rightarrow \quad \begin{array}{l} a = a\rho_{uv}^2 \\ d = d\rho_{uv}^2 \end{array}$$

It is clear, that if $\rho_{uv}^2 \neq 1$ the only solution to this system is $a = d = 0$. If $\rho_{uv}^2 = 1$ then it must necessarily hold that $X_u X_v = 1$ or $X_u X_v = -1$, in particular $X_u X_v$ must be constant. In either case it holds that the quadratic term in equation (5.2) disappears, which implies that $\mathbb{E}[X_i | X_u, X_v] = \mathbb{E}[X_i | Y]$ is an affine function, completing the proof. □

5.2 Conditional Gaussian Distributions

For the remainder of the thesis let $\mathcal{X} = \mathcal{X}_\Delta \times \mathcal{X}_\Gamma = \{-1, 1\}^\Delta \times \mathbb{R}^\Gamma$. The density of a CG-distribution is given by specifying a strictly positive distribution $p(i)$ over the discrete variables $i \in \mathcal{X}_\Delta$ to be the probability of the discrete variables falling in cell i. Then the conditional distribution of the continuous variables $y \in \mathcal{X}_\Gamma$ given that the discrete variables landed in cell i has density $f(y \mid i)$ of a multivariate Gaussian distribution $\mathcal{N}_\Gamma(\mu(i), \Sigma(i))$, where $\mu(i) \in \mathbb{R}^\Gamma$ is the mean vector, and $\Sigma(i)$ is the covariance matrix. The density takes the form

$$f(i, y) = p(i)(2\pi)^{-q/2} \det(\Sigma)^{-1/2} \exp\left\{ -\frac{1}{2}(y - \mu(i))^T \Sigma^{-1} (y - \mu(i)) \right\}. \tag{5.3}$$

CG-distributions can also be represented by the set of canonical characteristics (g, h, K) as

$$\log f(x) = \log f(i, y) = g(i) + h(i)^T y - \frac{1}{2} y^T K(i) y.$$

$g(i) = \log p(i) - \frac{1}{2}\log\det(\Sigma) - \frac{1}{2}\mu(i)^T\Sigma^{-1}\mu(i) - \frac{q}{2}\log(2\pi)$ is a scalar called *the discrete canonical parameter*, $K(i) = \Sigma^{-1}$ is the conditional concentration matrix, and $h(i) = \Sigma^{-1}\mu(i)$ is a q-vector called *the linear canonical parameter*. We can write up an expression for the marginal $p(i)$ Lauritzen (1996) as

$$\log p(i) = \zeta + g(i) + \frac{1}{2} h(i)^T \Sigma h(i), \tag{5.4}$$

where ζ is a constant.

We say that a function $u(i)$ is *additive* if it has the form

$$u(i) = \sum_{\delta \in \Delta} \alpha_\delta(i_\delta),$$

In order to show the next lemma we will state and prove the *Möbius inversion lemma*, see for instance (Lauritzen, 2018).

Lemma 5.2.1 (Möbius Inversion)**.** *Let Ψ and Φ be functions defined on the set of all subsets of a finite set V, taking values in $\mathbb{R}$. Then the following two statements are equivalent:*

1. *for all $a \subseteq V : \Psi(a) = \sum_{b:b\subseteq a} \Phi(b)$,*
2. *for all $a \subseteq V : \Phi(a) = \sum_{b:b\subseteq a} (-1)^{|a\setminus b|}\Psi(b)$.*

Proof.
We assume (2) and show (1). We see that

$$\begin{aligned}\sum_{b:b\subseteq a} \Psi(b) &= \sum_{b:b\subseteq a}\sum_{c:c\subseteq b} (-1)^{|b\setminus c|}\Psi(c) \\ &= \sum_{c:c\subseteq a} \Psi(c) \left(\sum_{b:c\subseteq b\subseteq a} (-1)^{|b\setminus c|} \right) \\ &= \sum_{c:c\subseteq a} \Psi(c) \left(\sum_{h:h\subseteq a\setminus c} (-1)^{|h|} \right).\end{aligned}$$

Unless $a\setminus c = \emptyset$, i.e. if $c = a$ the latter sum equals zero, because any finite and non-empty set has the same number of subsets of even as of odd cardinality. The opposite implication is performed analogously. □

Lemma 5.2.2. *A function $u : \mathcal{X}_\Delta \to \mathbb{R}$ is additive if and only if it is modular.*

Proof.
If u is additive it is easy to see that it is also modular. We show that if u is modular, then it is additive. Let $\mathcal{D}$ denote the set of subsets of Δ. Any function $u : \mathcal{X}_\Delta \to \mathbb{R}^n$ of $\mathcal{X}^\Delta$ can be expanded as

$$u(x) = \sum_{D\in\mathcal{D}} \eta_D(x)$$

where η_D are functions on $\mathcal{X}$ that only depend on x through x_D, i.e. η_D satisfy $\eta_D(x) = \eta_D(x_D)$. Without loss of generality we assume that $\min \mathcal{X}_\delta = 0, \delta \in \Delta$, and to assure that the representation is unique we require that $\eta_D(x) = 0$ whenever $x_d = 0$ for some $d \in D$. With this convention the sum may be rewritten so it only extends over such $D \in \mathcal{D}$ which are contained in the support $S(x)$ of x where

$d \in S(x) \iff x_d \neq 0$. For a fixed pair $\alpha, \beta \in V$, we define a function $\gamma_{\alpha\beta}$ on $\mathcal{X}$ by

$$\gamma_{\alpha\beta}(x) = \sum_{D:\{\alpha,\beta\}\subseteq D\subseteq S(x)} \eta_D(x).$$

We see that $\gamma_{\alpha\beta}(x) = 0$ unless $\alpha, \beta \in S(x)$, hence in particular whenever $|S(x)|\leq 1$. We will show that if u is modular, then $\eta_D(x) = 0$ whenever $|D|\geq 2$.

If for $C \subseteq \Delta$ we let

$$\omega_C(x) = u\left(x_C, \mathbf{0}_{\Delta\setminus C}\right),$$

it follows from the Möbius inversion lemma that

$$\eta_D(x) = \sum_{A:A\subseteq D} (-1)^{|D\setminus A|}\omega_A(x).$$

If $|D|\geq 2$, we can for distinct $i, j \in D$ rewrite this as

$$\eta_D(x) = \sum_{A:A\subseteq D\setminus\{i,j\}} (-1)^{|D\setminus A|}\left(\omega_{A\cup\{i,j\}}(x) - \omega_{A\cup\{i\}}(x) - \omega_{A\cup\{j\}}(x) + \omega_A(x)\right).$$

Note that $x_{A\cup\{i\}} \wedge x_{A\cup\{j\}} = x_{A\cup\{i,j\}}$ and $x_{A\cup\{i\}} \wedge x_{A\cup\{j\}} = x_A$, hence if u is modular, all terms inside the brackets are zero. Then $\eta_D(x)$ can only be different from zero if $|D|= 1$, i.e. if D is a singleton set which, by definition, shows that u is additive. □

Below we give an original example, showing that a quadratic exponential family on the sample space $\mathcal{X} = \{-1, 1\}^\Delta \times \mathbb{R}^\Gamma$ is automatically a conditional Gaussian distribution.

Example 5.2.3
Let $x = (i, y)$ where $i \in \{-1, 1\}^\Delta$, $y \in \mathbb{R}^\Gamma$, hence $x \in \{-1, 1\}^\Delta \times \mathbb{R}^\Gamma$. Let a quadratic exponential family be given as

$$\log f(x; a, B) = a^T x + \frac{1}{2}x^T Bx - A(a, B). \tag{5.5}$$

The matrix B can be separated into four block matrices corresponding to the discrete, the continuous, and the mixed variables: $B = \begin{bmatrix} B_{\Delta,\Delta} & B_{\Delta,\Gamma} \\ B_{\Gamma,\Delta} & B_{\Gamma,\Gamma} \end{bmatrix}$. Note that in the marginal distribution of Δ we are in the binary setting, why we require $\operatorname{diag}(B_{\Delta,\Delta}) = \mathbf{0}$. $B_{\Gamma,\Gamma}$ is positive definite since it is the marginal covariance matrix restricted to the continuous variables. We can then rewrite the expression (5.5) as

$$\log f(x; a, B) = a_\Delta^T i + a_\Gamma^T y + \frac{1}{2}(i, y)\begin{bmatrix} B_{\Delta,\Delta} & B_{\Delta,\Gamma} \\ B_{\Gamma,\Delta} & B_{\Gamma,\Gamma} \end{bmatrix}\begin{pmatrix} i \\ y \end{pmatrix} - A(a, B)$$

$$= a_\Delta^T i + a_\Gamma^T y + \frac{1}{2} i^T B_{\Delta,\Delta} i + i^T B_{\Delta,\Gamma} y + \frac{1}{2} y^T B_{\Gamma,\Gamma} y - A(a, B).$$

We see that this is a conditional Gaussian distribution. Indeed, since we can let $g(i) = a_\Delta^T i + \frac{1}{2} i^T B_{\Delta,\Delta} i - A(a, B)$; $h(i) = a_\Gamma + B_{\Gamma,\Delta} i$; and $K(i) = K = -B_{\Gamma,\Gamma}$ and see that

$$\log f(i, y; a, B) = g(i) + h(i)^T y - \frac{1}{2} y^T K(i) y$$

which is the desired form. If we further assume that the quadratic exponential family is MTP_2, proposition 4.1.1 states that $B_{ij} \geq 0$ for all $i \neq j$, where $i, j = 1, ..., (\Delta + \Gamma)$.

In this case, $-B_{\Gamma,\Gamma}$ will be an M-matrix, as it is positive definite and has negative values in every off-diagonal entry.

Furthermore, let $\gamma \in \Gamma$ be given. Then

$$h(i)_\gamma = a_\gamma + \sum_{\delta \in \Delta} b_{\gamma\delta} i_\delta$$

so $h(i)$ is an additive function. Now since $f(i, y; a, B)$ is MTP_2, it holds that $b_{\gamma\delta} \geq 0$ for all $\gamma \in \Gamma, \delta \in \Delta$, and so $h(i)$ is increasing. Indeed, since taking two vectors $i^{(1)}$ and $i^{(2)}$ and letting them differ in only one entry, say, the first such that $i_1^{(1)} = 1$ and $i_1^{(2)} = -1$ yields $h\big(i^{(1)}\big) \geq h\big(i^{(2)}\big)$.

Finally, $g(i)$ is supermodular. This can be seen in much the same way one would show that Gaussian graphical models are MTP_2 if and only if the concentration matrix K is an M-matrix. ∘

In the example above we also saw properties of the canonical characteristics of the conditional Gaussian distribution under MTP_2. This a general phenomenon, as seen by the following proposition shown by Lauritzen, Uhler, and Zwiernik (2019).

Proposition 5.2.4. *A CG-distribution P with canonical characteristics (g, h, K) is MTP_2 if and only if:*

(i) *$g(i)$ is supermodular;*

(ii) *$h(i)$ is additive and nondecreasing;*

(iii) *$K(i) = K$ for all i where K is an M-matrix.*

We now present an original result.

Theorem 5.2.5. *Let P be a quadratic exponential family under* MTP_2 *where $x \in \{-1,1\}^\Delta \times \mathbb{R}^\Gamma$. Then P is a conditional Gaussian distribution where the marginal $p(i)$ is an Ising model.*

Proof.
In Example 5.2.3 we saw that P is a conditional Gaussian distribution. It remains to be shown that $p(i)$ is an Ising model. We already know that $i \in \{-1,1\}^\Delta$, so we need to show that $p(i)$ is a quadratic exponential family.

Consider the form (5.4). It can be written as

$$\begin{aligned}\log p(i) &= \zeta_1 + a_\Delta^T + \frac{1}{2} i^T B_{\Delta,\Delta} i + \frac{1}{2} \left(a_\Gamma + B_{\Gamma,\Delta} i\right)^T \Sigma \left(a_\Gamma + B_{\Gamma,\Delta} i\right) \\ &= \zeta_2 + a_\Delta^T i + \left(a_\Gamma^T \Sigma B_{\Gamma,\Delta}\right) i + \frac{1}{2} i^T \left(B_{\Delta,\Gamma} \Sigma B_{\Gamma,\Delta} + B_{\Delta,\Delta}\right) i,\end{aligned}$$

where ζ_1, ζ_2 are constants. We have the linear term

$$m^T i := \left(a_\Delta^T + a_\Gamma^T B_{\Gamma,\Delta}\right)^T i$$

and the quadratic term

$$\begin{aligned}&i^T \left(B_{\Delta,\Gamma} \Sigma B_{\Gamma,\Delta} + B_\Delta\right) i \\ =&\zeta_3 + i^T \left(B_{\Delta,\Gamma} \Sigma B_{\Gamma,\Delta} + B_\Delta\right)_{\mathrm{diag}=0} i,\end{aligned}$$

Where ζ_3 is a constant. Letting $J := \left(B_{\Delta,\Gamma} \Sigma B_{\Gamma,\Delta} + B_\Delta\right)_{\mathrm{diag}=0} \in \mathbb{S}^\Delta$ and ζ be a constant gives us the form

$$\log p(i) = \zeta + m^T i + \frac{1}{2} i^T J i,$$

which is a quadratic exponential family. Hence $p(i)$ is an Ising model. □

6. Conclusion and Future Work

We have covered a lot of material in this thesis. In chapter 2 we saw many results for MTP_2 distributions that hold in general. Perhaps the most notable of these was Proposition 2.2.7 which implies that for strictly positive distributions, MTP_2 can be verified by checking that the MTP_2 inequality 2.1 holds for any $(x, y) \in \mathcal{E}$. We saw also that if a distribution P is MTP_2 and its independence model is a graphoid, then P is automatically faithful to its pairwise independence graph. We added to this in Chapter 4 that any undirected graph is realizable as the concentration graph of a Gaussian distribution.

In chapter 3 we saw that the MLE of an MTP_2 binary distribution exists if and only if every sample edge-marginal has both $(1, -1)$ and $(-1, 1)$ represented. In binary graphical models we went a little further. Here we require that every sample edge-marginal over edges in the underlying graph have $(1, -1)$ and $(-1, 1)$ represented.

We learned in chapter 4 that a Gaussian distribution is MTP_2 if and only if its concentration matrix is a symmetric M-matrix. In addition, given at least 2 observations the MLE for the covariance matrix in a Gaussian MTP_2 model (and therefore the concentration matrix) exists with absolute certainty.

Finally, in chapter 5 we introduced Ising models and combined our established knowledge about them as well as Gaussian models to enable the discussion of conditional Gaussian distributions under MTP_2. We ended the chapter with an original result stating that under MTP_2, a quadratic exponential family on sample space $\mathcal{X} = \{-1, 1\}^{\Delta} \times \mathbb{R}^{\Gamma}$ is automatically a conditional Gaussian distribution whose associated discrete marginal distribution is an Ising model. We saw a characterization of CG-distributions under MTP_2. An obvious next step in our research is to determine conditions for the existence of an MLE in MTP_2 CG-distributions. We could start in the special case considered in this thesis, and once a result has been established, we could search for conditions in the general case.

MTP_2 distributions have an abundance of desirable properties. Certainly, if there's one thing we have learned, it is that *our dreams may come true with MTP$_2$*.

Bibliography

Ahlswede, Rudolf and David E Daykin (1978). "An inequality for the weights of two families of sets, their unions and intersections". In: *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 43.3.

Anderson, Theodore Wilbur (1962). *An introduction to multivariate statistical analysis*.

Baba, Kunihiro, Ritei Shibata, and Masaaki Sibuya (2004). "Partial correlation and conditional correlation as measures of conditional independence". In: *Australian & New Zealand Journal of Statistics* 46.4, pp. 657–664.

Barndorff-Nielsen, Ole (2014). *Information and exponential families: in statistical theory*. John Wiley & Sons.

Boyd, Stephen and Lieven Vandenberghe (2004). *Convex optimization*. Cambridge university press.

Dellacherie, Claude, Servet Martinez, and Jaime San Martin (2014). *Inverse M-matrices and ultrametric matrices*. Vol. 2118. Springer.

Fallat, Shaun et al. (2017). "Total positivity in Markov structures". In: *The Annals of Statistics* 45.3, pp. 1152–1184.

Gilbert, George T (1991). "Positive definite matrices and Sylvester's criterion". In: *The American Mathematical Monthly* 98.1, pp. 44–46.

Horn, Roger A and Charles R Johnson (2012). *Matrix analysis*. Cambridge university press.

Jeyakumar, V. and Henry Wolkowicz (1992). "Generalizations of Slater's constraint qualification for infinite convex programs". In: *Mathematical Programming* 57.1, pp. 85–101. DOI: 10.1007/BF01581074. URL: https://doi.org/10.1007/BF01581074.

Karlin, Samuel and Yosef Rinott (1980). "Classes of orderings of measures and related correlation inequalities. I. Multivariate totally positive distributions". In: *Journal of Multivariate Analysis* 10.4, pp. 467–498.

Lauritzen, Steffen, Caroline Uhler, and Piotr Zwiernik (2017). "Maximum likelihood estimation in Gaussian models under total positivity". In: *arXiv preprint arXiv:1702.04031*.

— (2019). "Total positivity in structured binary distributions". In: *arXiv preprint arXiv:1905.00516*.

Lauritzen, Steffen L (1996). *Graphical models*. Vol. 17. Clarendon Press.

Lauritzen, Steffen L (2018). *Lectures on Graphical Models*. Department of Mathematical Sciences, Faculty of Science, University of Copenhagen.

Lenz, Wilhelm (1920). "Beitrag zum Verständnis der magnetischen Erscheinungen in festen Körpern". In: *Z. Phys.* 21, pp. 613–615.

Peters, Jonas (2015). "On the intersection property of conditional independence and its application to causal discovery". In: *Journal of Causal Inference* 3.1, pp. 97–108.

Plemmons, Robert J (1977). "M-matrix characterizations. I—nonsingular M-matrices". In: *Linear Algebra and its Applications* 18.2, pp. 175–188.

Slawski, Martin and Matthias Hein (2015). "Estimation of positive definite M-matrices and structure learning for attractive Gaussian Markov random fields". In: *Linear Algebra and its Applications* 473, pp. 145–179.