The Geometry of Covariance Selection

by Leon Di Stefano
co-supervised by Prof. Peter Hall and Dr. Alex Ghitza

completed as part of the MSc. program
Department of Mathematics and Statistics,
University of Melbourne

October 2013
## CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>INTRODUCTION</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>GRAPHICAL MODELS</td>
<td>9</td>
</tr>
<tr>
<td>1.1</td>
<td>Basic concepts</td>
<td>9</td>
</tr>
<tr>
<td>1.2</td>
<td>Factorizations and factor graphs</td>
<td>11</td>
</tr>
<tr>
<td>1.3</td>
<td>Markov properties</td>
<td>12</td>
</tr>
<tr>
<td>1.4</td>
<td>Chordal graphs and algorithms</td>
<td>17</td>
</tr>
<tr>
<td>2</td>
<td>COVARIANCE SELECTION AND ITS GEOMETRY</td>
<td>21</td>
</tr>
<tr>
<td>2.1</td>
<td>Gaussian graphical models</td>
<td>21</td>
</tr>
<tr>
<td>2.2</td>
<td>Maximum likelihood estimation</td>
<td>23</td>
</tr>
<tr>
<td>2.3</td>
<td>Computation of the MLE</td>
<td>25</td>
</tr>
<tr>
<td>2.4</td>
<td>Exponential families &amp; the MLE</td>
<td>26</td>
</tr>
<tr>
<td>2.5</td>
<td>The geometry of covariance selection</td>
<td>29</td>
</tr>
<tr>
<td>A</td>
<td>COMPUTER ALGEBRA PRIMER</td>
<td>33</td>
</tr>
<tr>
<td>A.1</td>
<td>Basic concepts</td>
<td>33</td>
</tr>
<tr>
<td>A.2</td>
<td>Bases for ideals</td>
<td>37</td>
</tr>
<tr>
<td>A.3</td>
<td>Monomials and division</td>
<td>40</td>
</tr>
<tr>
<td>A.4</td>
<td>Monomial ideals and Gröbner bases</td>
<td>44</td>
</tr>
<tr>
<td>A.5</td>
<td>Elimination theory</td>
<td>48</td>
</tr>
<tr>
<td>B</td>
<td>A NOTE ON STATISTICS AND LEARNING THEORY</td>
<td>49</td>
</tr>
<tr>
<td>B.1</td>
<td>Inference and estimation</td>
<td>49</td>
</tr>
<tr>
<td>B.2</td>
<td>Statistics and machine learning</td>
<td>49</td>
</tr>
</tbody>
</table>
This project is based on a recent paper of [Uhler, 2012], which was itself based on her PhD thesis [Uhler, 2011]. Both works treat the problem of the existence of the maximum likelihood estimator in Gaussian graphical models, in the case where parameters outnumber observations, using concepts from algebraic geometry and convex optimization, and tools from computational commutative algebra.

The structure of the project is as follows:

- Part 0 presents an overview of the thesis, a non-mathematical outline of graphical models and algebraic statistics, and a discussion of covariance selection, including some examples from the applied sciences.

- Part 1 is an introduction to the theory of graphical models in general, and to that of undirected Gaussian graphical models (also called covariance selection models) in particular.

- Part 2 introduces covariance selection models (also called Gaussian graphical models). It discusses the original results of [Dempster, 1972] in the context of the theory of exponential families, describes the main results of [Uhler, 2012], and suggests some possible extensions.

- Part A is a primer in computational commutative algebra, in particular the theory of Gröbner bases; these are used for computations in Part 2.

**Graphical Models**

Graphical models are statistical models in which marginal and conditional dependencies among variables are summarized by a graph: a mathematical structure consisting of nodes joined by directed or undirected edges. Graphical models developed out of models used in physics, genetics, and economics from the early- to mid-twentieth century, and are still an active area of research [Lauritzen, 2011]. They possess a number of appealing features:

- They are naturally (but not necessarily) represented by pictures, so are often easy to interpret (in this sense, “graphical” has a double meaning).
• They make it easy to reason about causality and confounding; this is particularly true of directed graphical models. Accompanying formalisms include the Rubin causal model, structural equation models, counterfactuals, and Pearl’s “do”-calculus \cite{Wasserman,2004,ch. 16 and 17}.

• They can aid in the interpretation of contingency tables; this is particularly true of representations of graphical models by factor graphs.

• From a modelling perspective, conditional independence assumptions are often natural and \textit{a priori} plausible, for example if the interaction between two physical systems is mediated by a third.

• On the other hand, from the perspective of model selection, introducing conditional independencies can help avoid overfitting. They represent a kind of sparsity, though it is a sparsity in the relationships among a set of random variables, rather than in the number of variables under consideration. In this sense, graphical modelling complements variable selection.

• Assuming conditional independencies among random variables can drastically speed up computations.

• The theory of graphical models is intimately connected with many areas of statistics and applied mathematics, including statistical mechanics, coding theory, bioinformatics, and image processing.

\textbf{Covariance Selection} The particular subject of this project is \textit{Gaussian graphical models}, i.e. undirected graphical models whose variables are jointly normally distributed. These were introduced by \cite{Dempster,1972} as \textit{covariance selection} models, without the associated graphical representation. Gaussian graphical models possess an elegant and appealing theory, particularly with respect to maximum likelihood estimation. Edges in the graph correspond to elements of the inverse covariance matrix (also called the concentration or precision matrix), and represent conditional dependencies among the component random variables, which are represented by vertices.

Covariance selection models must be distinguished from what are sometimes called \textit{covariance graph models}, in which edges correspond to entries of the covariance matrix and represent marginal, rather than conditional, dependencies (i.e. overall correlation rather than partial correlation).

There is a large literature on regularization and model selection in graphical models. Dempster originally suggested adapting the forward- and backward-selection procedures from multivariate
regression [Dempster, 1972, p. 162]. The lasso (a form of $l_1$ regularization) has been extended to the Gaussian graphical model context in [Friedman et al., 2008] and (vertex-wise) in [Meinshausen and Bühlmann, 2006]. Yuan and Lin [2007] explore various penalized likelihood approaches; Drton and Perlman [2008] offer model selection algorithms that can easily incorporate prior information. An objective Bayesian treatment is given by Carvalho and Scott [2009].

On the other hand, the paper of Uhler [2012] which is the basis for this project deals with a problem that is in some sense dual to model selection, namely: given a Gaussian graphical model, what is the minimum number of cases needed to assure the existence of the desired estimator? This question is particularly relevant in the case where a particular sparse model is known or plausible a priori, but in which the dimensionality $p$ of the data is much larger than the number $n$ of available data points (e.g. experimental subjects).

**Algebraic Statistics**  In recent years, graphical models have been studied using techniques from the field of algebraic statistics. This term refers to the application of abstract algebra – especially computational commutative algebra and group theory – to statistical problems.\(^\text{1}\)

In the 1980s and 90s, these problems were primarily in the design and analysis of experiments, and in Markov Chain Monte Carlo sampling from discrete distributions; see [Riccomagno, 2009] for an historical overview along such lines. More recently, similar tools have been applied to graphical models. The geometric objects of study in this case are commonly algebraic varieties of low-rank matrices defined by systems of polynomial equations.

The paper of Uhler [2012] on which this thesis is based is one example of such work; for a recent application of similar ideas to the problem of model selection, see [Chandrasekaran et al., 2012]. For an overview of algebraic statistics, see [Drton et al., 2009].

**Examples from Applied Science**  Rue and Held [2005, pp. 10-13] give a very long list of references to applications of Gaussian graphical models in time series analysis, the analysis of longitudinal and survival data, semiparametric regression and splines, image analysis, spatial statistics, and inverse problems. In this section we focus on their use in biology.

An example from genomics is [Toh and Horimoto, 2002]. The authors first applied a clustering algorithm to 2467 genes of *Saccharomyces cerevisiae*, a variety of yeast, then fitted the obtained clusters to a Gaussian graphical model via a form of backward-

---

\(^1\) The term “algebraic statistics” is a little misleading: linear algebra and statistical theory developed together even in the 19\textsuperscript{th} century; see [Farebrother, 1996].
Figure 1: This diagram, taken from [Toh and Horimoto, 2002], shows a subgraph of an estimated gene network. Associated with each edge is a bootstrap probability. Dotted lines indicate regulatory relationships suggested by experiments, but not estimated by the model. Arrow directions were inferred from experimental results.

selection. They compared this estimated genetic network with the cumulative results of various empirical studies (see figure 1). A more recent applied study along similar lines is [Li and Gui, 2006]. Castelo and Roverato [2006] suggested algorithms for the very high-dimensional application to microarray data.

Gaussian graphical models have also been used in the relatively new field of metabolomics, which aims to measure under given conditions all endogenous metabolites of a tissue or body fluid, and thereby reconstruct the underlying metabolic network of chemical reactions. In Krumsiek et al. [2011], covariance selection was shown to recover known metabolic pathway interactions in samples of human fasting blood serum.

Finally, such models have also been used to explore functional connectivity in the brain based on fMRI data [Ng et al., 2013].
This section develops some of the general theory behind graphical models, in particular Gaussian graphical models. The material here has been adapted in part from [Lauritzen, 1996], a comprehensive monograph on graphical models in general; [Wasserman, 2004], an overview of many topics in statistics; and [Wainwright and Jordan, 2008], a tutorial monograph on variational methods in exponential family graphical models.

1.1 BASIC CONCEPTS

1.1.1. A graph $G$ is a collection $V = V(G)$ of vertices or nodes, together with a set $E = E(G)$ of pairs of vertices called edges. If we consider edges as unordered pairs then $G$ is an undirected graph (or just a graph); otherwise $G$ is a directed graph.

Graphs with both directed and undirected edges are sometimes called chain graphs in the graphical models literature. Pictorially, vertices are represented by dots or the associated indices. An edge $(i, j)$ of a directed graph is usually represented by an arrow $i \rightarrow j$, whereas for undirected graphs, edges are represented by lines $i \sim j$. If $i, j$ are vertices with an edge between them — i.e., $(i, j) \in E$ — then we say that $i$ and $j$ are adjacent. An induced subgraph of a graph $G$ is a subset $U \subseteq V(G)$ of vertices of $G$, together with all of their edges. We let $G^c$ denote the complement of $G$, i.e. the graph defined by

$$V(G^c) = V(G); \quad E(G^c) = \{(i, j) : (i, j) \notin E(G)\}.$$

An undirected graph is called complete if there is an edge between every pair of vertices. A clique is a complete subgraph of an undirected graph; a clique is maximal if it is not contained in any larger clique. We denote by cl$(G)$ the collection of cliques of $G$. A path is a finite sequence of edges in which subsequent edges connect adjacent vertices. In the undirected case, if there exist paths from $i$ to $j$ we say that $i$ and $j$ are connected, otherwise they are separated. A tree is an undirected graph in which every pair of vertices is connected by precisely one path. A cycle is a path beginning and ending at the same vertex. A set $S \subseteq V(G)$ is called an $(i, j)$-separator if every path from $i$ to $j$ passes through $S$.

In a directed graph, if $(i, j) \in E(G)$ then we say that $i$ is a parent of $j$ and $j$ is a child of $i$. We denote by $\pi(i)$ the set of parents of $i$. If there is a path from $i$ to $j$ then we say that $i$ is an ancestor of $j$ and
Figure 2: Undirected and directed graphs on six vertices. In the undirected graph, all vertices are connected; $(2,5,6), (1,3),$ and $(4,5)$ are among the maximal cliques; and a path from 5 to 1 is given by $(5,4,3,1).$ In the directed graph, the children of 3 are 4, 5 and 6, while the parents of 6 are given by $\pi(6) = \{2,3,5\}.$ This digraph is acyclic.

that $j$ is a descendant of $i.$ An important class of directed graphs are those which are acyclic (equivalently, those in which no vertex is its own ancestor).

Associated with any graph $G$ is an adjacency matrix whose $(i,j)$-element counts the number of edges from $i$ to $j.$ For our purposes, this number will always be one or zero. We will also often be entries on the diagonal of the adjacency matrix; for reasons of clarity, though, diagrams of graphical models typically suppress these self-loops.

**GRAPHICAL MODELS** A graphical model is a statistical model – that is, a family of probability distributions – that satisfy (or are adapted to, or are Markov to, or are represented by) some directed or undirected graph. Given a directed or undirected graph $G$ and a probability measure $P$ with distribution (density or mass function) $p,$ there are three concepts that need to be related:

- factorizations of the probability distribution $p;$
- (collections of) conditional independence statements about $P$ (such conditional independence statements are called Markov properties, as they generalize the properties of Markovian stochastic processes); and
- some notion of separation (of subsets of vertices) in $G$

These three notions each differ between directed and undirected graphs. We will consider factorizations first, then give a brief treatment Markov properties and and different forms of separation.

**NOTATION** With each vertex $i \in V(G)$ we denote the associated random variable $X_i;$ for a subset $A \subseteq V(G)$ of the vertices of $G$ we denote by $X_A$ the corresponding random vector; and we denote the
random vector associated with the whole graph by $X$. Similarly, a particular vector of values taken by the random vector $X_A$ will be denoted $x_A$, and those scalar values taken by $X_i$ will be denoted $x_i$. The choice of an ordering on vertices of the graph will usually be left implicit. We also denote the state space of $X_i$ by $\mathcal{X}_i$, that of $X_A$ by $\mathcal{X}_A = \otimes_{i \in A} \mathcal{X}_i$ (the cartesian product of the state spaces of the components), and that of $X$ by $\mathcal{X}$.

Sometimes it will be convenient to conflate a set of vertices with the associated random variables. Moreover, as is often done in Bayesian statistics, we will use the same symbol $p$ for the joint, marginal, and conditional probability densities or mass functions associated with $X$, letting the notation for the arguments distinguish for the reader which function is being referred to. The uppercase $P$ will refer to the associated measure.

Note that many of the statements below could be made more general by writing $p(x) d\mu(x)$ in place of $p(x)$ to denote the probability density of $X$ with respect to some base measure $\mu$.

### 1.2 Factorizations and Factor Graphs

Let’s first consider directed graphs. It is standard to restrict attention in this case to directed acyclic graphs (frequently abbreviated DAGs). Intuitively this makes sense, given that the dependences in a directed graphical model are often interpreted in a causal manner: it is the requirement that an event may not cause itself.

#### 1.2.1. A probability distribution $p$ is said to factorize according to a directed graph $G$ if it can be written in the form

$$p(x_1, \ldots, x_m) = \prod_{i \in V(G)} p_i(x_i | x_{\pi(i)})$$

where $X_{\pi(i)}$ denotes the random vector corresponding to the parents of the random variable $X_i$.

Such directed graphical models are often called Bayes nets in the computer science literature. This terminology can be misleading, though, as one may perform frequentist estimation as well as Bayesian inference over such models.

Undirected graphical models (also called Markov random fields) satisfy a different factorization criterion. To describe this we need a collection of potential functions $\{\psi_C\}_{C \in \operatorname{cl}(G)}$ on the cliques of $G$:

$$\psi_C : \mathcal{X}_C \rightarrow \mathbb{R}_{>0}$$

#### 1.2.2. A Gibbs distribution over an undirected graph $G$ is a probability distribution which satisfies:

$$p(x_1, \ldots, x_m) = \frac{1}{Z_{\operatorname{cl}(G)}} \prod_{C \in \operatorname{cl}(G)} \psi_C(x_C)$$
Figure 3: Example of directed and undirected graphs from Kirshner [2010].

for some (non-unique) choice of the $\psi_C$, where $Z = \int_x d x \prod_C \phi_C$ is a normalizing constant.

Often the set $\text{cl}(G)$ is replaced with the set of maximal cliques of $G$; this leads to no loss of generality, but representations in terms of nonmaximal cliques can be more computationally efficient in some contexts.

To show how these factorizations reduce the complexity of the underlying distribution, consider the example depicted in 1.2. The distribution of the directed graph factorizes as

$$p(x_1, \ldots, x_6) = p(x_1) \times p(x_2 | x_1) \times p(x_3) \times p(x_4 | x_1)$$
$$\times p(x_5 | x_2, x_3, x_4) \times p(x_6 | x_5, x_4)$$

and that of the undirected graph factorizes as

$$p(x_1, \ldots, x_6) \propto \psi(x_1, x_2) \times \psi(x_1, x_4) \times \psi(x_2, x_5)$$
$$\times \psi(x_3, x_5) \times \psi(x_4, x_5, x_6).$$

If the $X_1, \ldots, X_6$ are binary-valued (i.e. Bernoulli) random variables, then their joint distribution takes $2^6 - 1 = 63$ parameters to specify. When adapted to the directed graph, however, the joint distribution can be specified by $2^0 + 2^1 + 2^0 + 2^1 + 2^3 + 2^2 = 18$ parameters. In the case of the undirected graph, things are more complicated (due to, for example, the many possible choices of parametrization) but it turns out there is a canonical way to specify the joint distribution which requires only 13 free variables. The details of this example can be found in the lecture notes of Kirshner [2010, weeks 1 and 10].

Sometimes, many factorizations will be compatible with a single graph $G$. The formalism of factor graphs – in which a second class of vertices represent the terms in the factorization associated to $G$ – remedies this, and also makes various algorithms on graphical models easier to understand in edge-vertex terms. An example is given in figure 1.2.
1.3 Markov Properties

Markov properties are conditional independence statements associated with the structure of a graph.

1.3.1. Given random variables $X, Y, Z$ with joint distribution $P$, the statement $X$ is conditionally independent of $Y$ given $Z$ under $P$ means that for any event $A$ in the sample space of $X$, there exists a version of the conditional probability $P(A \mid Y, Z)$ which is a function of $Z$ alone [Lauritzen, 1996].

We will write this

$$X \not\perp_{P} Y \mid Z,$$

and reserve the notation

$$X \perp Y \mid Z,$$

to refer to (for example) uncorrelatedness. (The underlying distribution $P$ is normally left implicit.)

Many properties of conditional independence are deducible from familiar properties of conditional expectation; for example, as the notation suggests,

$$X \perp Y \mid Z \Leftrightarrow Y \perp X \mid Z. \quad (1)$$

It is also true that if $X \perp Y \mid Z$, and $U = h(X)$ for some measurable function $h$, we can deduce both

$$U \perp Y \mid Z \quad (2)$$

and

$$X \perp Y \mid (Z, U). \quad (3)$$

Finally, one can also derive that

$$X \perp Y \mid Z \text{ and } X \perp W \mid (Y, Z) \Rightarrow X \perp (W, Y) \mid Z. \quad (4)$$
These preceding four results can be reinterpreted as combinatorial axioms, where $X, Y, Z, W$ are replaced by disjoint subsets of a finite set and $U = h(X)$ is replaced by $U \subseteq X$. The resulting structure is called a semi-graphoid. Other important models of semi-graphoids are the relation of separation in an undirected graph (where $X, Y, Z \subseteq V(G)$, and $X \perp_G Y \mid Z$, say, means that $X$ is separated from $Y$ by $Z$ in $G$), and the relation of geometric orthogonality of the subspaces of a finite-dimensional Euclidean space.\(^1\) In Gaussian graphical models, the three models of a semi-graphoid coincide.

To understand how Markov properties and separation in graphs interact, it is worth introducing the following concepts due to \cite{Pearl1988} via \cite{ScutariStrimmer2011}, though our subsequent treatment will be a little less formal.

1.3.2. Given a probability measure $P$ and a graph $G$ with a notion of separation $\perp_G$, we say that $G$ is an independence map or I-map of $P$ if

$$A \perp_G B \mid S \Rightarrow A \perp_P B \mid S,$$

a dependence map or D-map if

$$A \perp_G B \mid S \Leftarrow A \perp_P B \mid S,$$

i.e.

$$\neg(A \perp_G B \mid S) \Rightarrow \neg(A \perp_P B \mid S)$$

and a perfect map or P-map if

$$A \perp_G B \mid S \Leftrightarrow A \perp_P B \mid S.$$

An I-map is said to be minimal if adding any edge would mean it were no longer an I-map. Many of the results which follow could be stated using these terms.\(^2\)

**Undirected Graphs** Let’s first consider possible Markov properties on an undirected graph $G$:

- The global Markov property states that if $A$ and $B$ are subsets of $V(G)$ separated by $S$, then $X_A$ is conditionally independent of $X_B$ given $X_S$.

- The local Markov property states that any vertex is conditionally independent of all others, given its neighbors.

- The pairwise Markov property states that any two non-adjacent vertices are conditionally independent given all other variables.

---

\(^1\) The equivalence is given by $L \perp M \mid N \Leftrightarrow (L \cap N) \perp (M \cap N)$.\(^2\) An alternative notation for independence maps is to let $\mathcal{I}(P)$ denote the set of all conditional independence statements implied by $P$, and $\mathcal{I}(G)$ the set of all separation statements implied by $G$; then $G$ being an I-map of $P$ is equivalent to $\mathcal{I}(G) \subseteq \mathcal{I}(P)$. (This conflates the symbols $\perp_P$ and $\perp_G$ as used above.)
Figure 5: A discrete-time Markov chain, represented as a graph, satisfies the global, local, and pairwise Markov properties.

The global Markov property is strictly stronger than the local property, which is strictly stronger than the pairwise property. On the other hand, the three properties are equivalent under fairly general conditions – for example if $P$ is positive and absolutely continuous with respect to a product measure – due to a theorem of Pearl and Paz [1985]. Moreover, it can be shown that if $P$ factorizes over cliques as described above, the global Markov property must hold [Lauritzen, 1996, p. 35].

The most famous result linking factorizations and conditional independence is the Hammersley-Clifford theorem (which was proved independently by many different authors in the 1970s; see [Speed, 1979]).

1.3.3. (Hammersley-Clifford) If the distribution $p$ of a random vector $X$ factorizes over an undirected graph $G$, then $X$ is Markov with respect to $G$. Conversely, if $X$ is Markov with respect to $G$ and $p > 0$ is strictly positive, then $p$ factorizes over the graph. (See e.g. [Wainwright, 2012, thm. 2], [Lauritzen, 1996].)

Directed Graphs

Specifying Markov properties for directed graphs is more subtle than the undirected case. First, we give a description based on choosing a particular ordering on the vertices. Then we give a more intrinsic definition based on a special notion of separation in directed graphs.

A topological ordering on the vertices of a directed acyclic graph is a map

$$\rho : V(G) \rightarrow \mathbb{N}$$

such that for any $i \in V(G)$ and for any $j \in \pi(i)$, it holds that $\rho(j) < \rho(i)$. In other words, it is an ordering on the vertices of $G$ such that parents come before their children. It is a fundamental result in graph theory that there always exists such at least one such ordering.

Now assume that the canonical ordering on vertices is topological. Then for any $s \neq 1$, its parent set $\pi(s)$ is a subset of $\{1, \ldots, s - 1\}$. Denote by $\nu(s)$ the set $\{1, \ldots, s - 1\} - \pi(s)$.

1.3.4. A random vector $X$ with associated probability measure $P$ is Markov with respect to a directed acyclic graph $G$ if for all $i \in V$,

$$X_i \perp X_{\nu(i)} | X_{\pi(i)}.$$ 

This is equivalent to the distribution $p$ of $X$ factorizing according to $G$ (see [Wainwright, 2012, defn. 3, thm. 1]).
A useful and related concept is the so-called Markov blanket of a vertex \( v \), which is the minimal set of nodes rendering \( v \) conditionally independent from others. In the directed case, it is the set consisting of \( v \)'s parents, its children, and its children's other parents; in the undirected case, it is simply the neighbors of \( v \).

1.3.5. Markov properties of directed graphs may also be formulated in terms of \textit{d-separation} (the “d” is short for “directed”). A \textit{trail} is a sequence of vertices in a directed graph such that successive vertices are linked by edges, though these may have any direction. Let \( t \) be a trail in a directed graph \( G \).

A trail \( t \) is said to be \textit{blocked} by a vertex set \( S \subseteq V(G) \) if any one of the following holds:

1. There is a \textit{chain} in \( t \), i.e. a trail of the form \( i \to s \to j \), with \( s \in S \).

2. There is a \textit{fork} in \( t \), i.e. a trail of the form \( i \leftarrow s \rightarrow j \), with \( s \in S \).

3. There is a \textit{collider} in \( t \), i.e. a trail of the form \( i \rightarrow q \leftarrow j \), such that neither \( q \) nor any descendant of \( q \) is in \( S \).

If every trail between disjoint subsets \( A, B \subseteq V(G) \) is blocked, then we say that \( S \) \textit{d-separates} \( A \) from \( B \).

We can attach a causal semantics to each of the shapes above, imagining probabilistic influence as a kind of flow, and the separators as observed nodes \cite{Koller2007}:

1. \( i \to s \to j \) is a \textit{causal path}, and is active if and only if \( s \) is not observed; similarly, \( i \leftarrow s \rightarrow j \) is called an \textit{evidential path}. The vertices \( i \) and \( j \) are marginally dependent, but conditionally independent given \( s \).

2. \( i \leftarrow s \rightarrow j \) represents a \textit{common cause}, and is active if and only if \( s \) is not observed. The vertices \( i \) and \( j \) are marginally dependent, but conditionally independent given \( s \).

3. \( i \to q \leftarrow j \), represents a \textit{common effect}, and is active if and only if we observe at least one of \( q \) or its descendants. Unlike the previous cases, the vertices \( i \) and \( j \) are marginally independent, but conditionally dependent given \( s \).

The shape in 3. is also sometimes called a \textit{v-structure}, and its associated dependencies correspond to the notion of “explaining away”: in the collider \( x \to q \to y \), for example, \( x \) and \( y \) would be independent without the presence of \( q \), but \( x \) (respectively \( y \)) becomes less likely if we know \( y \) (respectively \( x \)) is the cause of \( q \).

We now touch on the relationship between d-separation and factorizations. In general, d-separation is \textit{sound}, in that if \( p \) factorizes according to \( G \), then

\[ S \text{ d-separates } A \text{ and } B \text{ in } G \Rightarrow A \perp B | S, \]
i.e. d-separation in a directed graphical model $G$ implies the associated conditional independency with respect to any $p$ factorizing over $G$. Equivalently, if $p$ factorizes, then $G$ is an independence map of $P$ with respect to d-separation. Only a weak converse holds, though: if $S$ d-separates $A$ and $B$ in $G$, we can conclude merely that there exists a distribution $p$ which factorizes over $G$ such that $A \indep B \mid S$. For a much more complete survey of Markov properties in directed graphical models, see the review article [Flesch and Lucas, 2007] or the book [Koller and Friedman, 2009].

1.4 CHORDAL GRAPHS AND ALGORITHMS

To motivate what follows, we describe a way to move from a directed to an undirected graphical model.

1.4.1. Let $G$ be a directed graph, and denote by $G^m$ the undirected graph formed by connecting the parents of every vertex, then ignoring the direction of each edge. The undirected graph $G^m$ formed this way is called the moral graph of $G$, and the transformation $G \to G^m$ is called moralization.

Then if $p$ factorizes according to $G$, and we let $C(i)$ denote the clique in $G^m$ formed by $i$ and its parents, we can form the map

$$p_i(x_i \mid x_{\pi(i)}) \to \psi_{C(i)}(x_{C(i)})$$

which, setting $Z = 1$, implies that if $p$ factorizes according to $G$, then it also factorizes according to $G^m$ and hence obeys the global Markov property with respect to $G^m$. It also implies that a subset $S \subset V(G)$ d-separates $A$ from $B$ if and only if $S$ separates $A$ from $B$ in $G^m$ [Lauritzen, 1996, p. 46-53].

Undirected graphs of the form $G^m$ for some directed graph $G$ have a number of equivalent characterizations; they are most commonly called chordal, decomposable or triangular. This family of graphs provide a link between directed and undirected graphical models, both theoretically and via various algorithms used for calculating likelihoods, modes, and marginal distributions. For Gaussian

![Figure 6: The smallest nonchordal graph is the four-cycle, on the left. A minimal chordal cover of this graph is on the right (it is also the smallest chordal graph which is not complete).](image-url)
Figure 7: On the left is a maximal planar graph which is not chordal without the dashed edge (2,8), due to the cycle 2–4–6–8. Including dashed edge gives a minimal chordal cover. On the right is an associated junction tree; maximal cliques are normal nodes, while separators are white-on-black.

graphical models in particular, chordal graphs are connected with the existence of a maximum likelihood estimator of the associated covariance matrix.

There are a number of informative and equivalent characterizations of this class [Bartlett 2003].

1.4.2. A chordal graph is one in which every cycle of length 4 or greater contains a chord, i.e. an edge linking two non-adjacent vertices. Since chordality is defined in terms of forbidden paths, it must be preserved under taking induced subgraphs. Chordal graphs are sometimes called rigid circuit or triangulated; this latter term is both evocative and a little ambiguous: there are maximal planar graphs (which can be drawn in the plane) which are not chordal; see figure 7. A chordal cover $G^+$ of an undirected graph $G$ is a chordal graph satisfying $E(G) \subseteq E(G^+)$. Denote the size of the largest clique of $G$ by $q$, and that of $G^+$ by $q^+$. A minimal chordal cover is a chordal cover whose maximal clique size is as small as possible; we denote this number $q^*$. Then the treewidth of $G$ is defined by

$$\tau(G) = q^* - 1$$

The $-1$ term ensures that trees have treewidth 1.

A decomposable graph is one whose vertices $V(G)$ can be partitioned into disjoint sets $A,B,C$ (with $A$ and $C$ nonempty) satisfying:

- $B$ is a clique,
- $B$ separates $A$ and $C$ in $G$, and
- $A \cup B$ and $B \cup C$ are decomposable.

A vertex is simplicial if its neighbors form a clique; a graph is recursively simplicial if there exists a simplicial vertex whose removal leaves a subgraph which is recursively simplicial.
Finally, a clique tree for a graph $G$ is a tree $T$ with $V(T)$ a set of cliques of $G$ that contains all the maximal cliques of $G$. A junction tree for $G$ is a clique tree satisfying the additional condition that for any $C_1, C_2$ in $V(T)$, every clique on the path connecting $C_1$ and $C_2$ contains $C_1 \cap C_2$ (this is the so-called running intersection property).

1.4.3. [Wainwright, 2012] For an undirected graph $G$, the following are equivalent:

- $G$ is chordal.
- $G$ is decomposable.
- $G$ is recursively simplicial.
- $G$ has a junction tree.
- There is an orientation of the edges of $G$ that give a directed acyclic graph whose moral graph is $G$.
- There is a directed graphical model whose conditional independencies are identical to those implied by $G$.

Junction trees are particularly important in the theory of graphical models because they represent a data structure that allows for efficient computation of modes, marginals, likelihoods, and conditional distributions. For this reason, many such algorithms have complexity that is bounded in the treewidth of the underlying graph.

We close this section with a summary of the differences between directed and undirected graphical models, drawing on [Paskin, 2003, Wainwright, 2012, Kirshner, 2010, Koller et al., 2007].

- Specifying an undirected graphical model is easy, but the factors (clique potentials) don’t have probabilistic interpretations (though they do have a physical interpretation as potentials in statistical
• The notion of separation in undirected models is easy, but the equivalent notion in undirected models (d-separation) is hard.

• Generating samples from directed graphical models is easy, given a vertex ordering. Generating samples from undirected graphical models is not so straightforward.

• Directed and undirected graphical models differ in the conditional independencies they can express. Undirected graphical models can’t represent immoralities; directed models can only represent the dependencies of triangulated undirected graphs. To put it another way, in going from a directed to an undirected graph, one loses those independencies encoded by the V-structures. Moving from an undirected to a directed graph, on the other hand, requires adding triangulating edges to loops. See figure 8.

• Chordal graphs connect directed and undirected graphical models, both theoretically and via various algorithms.
This section begins by explaining the main results of Dempster [1972], who first introduced covariance selection models, and described their theory of maximum likelihood estimation. Next, we situate this theory in the broader context of exponential families. Finally, we give an exposition of the paper of Uhler [2012], who developed a geometric interpretation of the MLE, and an algebraic criterion which under certain circumstances ensures its existence with probability 1.

2.1 GAUSSIAN GRAPHICAL MODELS

The $p$-dimensional random vector $X$ has a multivariate normal distribution with mean $\mu$ and (symmetric and positive definite) covariance matrix $\Sigma$ if it has probability element

$$f(x) \, dx = \left( \frac{1}{2\pi} \right)^{p/2} \left( \frac{1}{\det \Sigma} \right)^{1/2} \exp \left\{ -\frac{1}{2} (x - \mu)^\top K (x - \mu) \right\} \, dx;$$

with $\Sigma = K^{-1}$. The matrix $K = \Sigma^{-1}$ is called the precision or concentration matrix. We denote this by situation $X \sim \mathcal{N}(\mu, \Sigma)$, and refer to the $(i,j)$ element of $\Sigma$ by $\sigma_{ij}$ and the $(i,j)$ element of $K$ by $k_{ij}$. Rescaling the diagonal elements of $\Sigma$ to 1 gives the correlation matrix $R$.

2.1.1. A Gaussian graphical model is a set of $\mathcal{N}(\mu, \Sigma)$ distributions which satisfy the pairwise Markov property with respect to a graph $G$, with the components of the associated random variable $X$ indexed by $V = V(G)$.

The following result shows that this depends only on the zeroes of $K$.

2.1.2. If $X_V \sim \mathcal{N}(\mu, \Sigma)$, then

$$X_i \perp \!
\!
\!
\perp X_j | X_{V \setminus \{i,j\}} \quad \text{if and only if} \quad k_{ij} = 0.$$

So the distribution of $X$ is adapted to $G$ exactly when $k_{ij} = 0$ for all $(i,j) \in E(G)$. 

21
To see why this relationship holds, note that up to an additive constant, the logarithm of the density \( f(x) \) can be written, after the one-to-one change of variables \( x - \mu \rightarrow y \), in the form

\[
-\frac{1}{2} y^\top \Sigma^{-1} y = -\frac{1}{2} \sum_{i=1}^{n} \sum_{m=1}^{n} y_i k_{im} y_m \\
= -\frac{1}{2} \sum_{i \in V} k_{ii} y_i^2 - \sum_{(i,j) \in E} k_{ij} y_i y_j
\]

where \( E = E(G) \). Since we are by convention including all self-loops in our graphical models, this means that the density \( f(x) \) factors as a Gibbs distribution with respect to \( G \), with all edges and vertices of \( G \) forming the cliques. Since \( \Sigma \) is positive definite and hence full rank, so is \( \Sigma^{-1} \). Then \( f \) is strictly positive on \( \mathbb{R}^p \) and we can apply the Hammersley-Clifford theorem. This implies that the measure \( P \) associated with \( X \) is satisfies the global (and hence pairwise) Markov property with respect to \( G \).

We can also show this using well-known properties of the multivariate normal distribution. If \( Y \sim \mathcal{N}(\mu, \Sigma) \) and we divide the component random variables into disjoint subsets \( A \) and \( B \), then the conditional distribution of \( Y_A \mid Y_B = y_B \) is \( \mathcal{N}(\mu_{A|B}, \Sigma_{A|B}) \), where

\[
\mu_{A|B} = \mu_A + \Sigma_{AB} \Sigma_{BB}^{-1} (y_B - \mu_B) \\
\Sigma_{A|B} = \Sigma_{AA} - \Sigma_{AB} \Sigma_{BB}^{-1} \Sigma_{BA}.
\]

The matrix \( \Sigma_{A|B} \) is the Schur complement of \( \Sigma_{AA} \) and therefore is equal to \( K_{AA}^{-1} \). The elements of \( K \) then have the following interpretations [Uhler, 2011]:

- The diagonals \( k_{ii} \) are the reciprocals of the conditional variances \( \text{var}(Y_i \mid Y_{V \setminus \{i\}}) \).

- The conditional covariances \( \text{cov}(Y_i, Y_j \mid Y_{V \setminus \{i,j\}}) \) are given by

\[
-\frac{k_{ij}}{k_{ii} k_{jj} - k_{ij}^2}.
\]

- The partial correlation coefficients are given by

\[
\rho_{ij|V \setminus \{i,j\}} = \frac{\text{cov}(Y_i, Y_j \mid Y_{V \setminus \{i,j\}})}{\sqrt{\text{var}(Y_i \mid Y_{V \setminus \{i\}}) \cdot \text{var}(Y_j \mid Y_{V \setminus \{j\}})}} = \frac{-k_{ij}}{\sqrt{k_{ii} k_{jj}}}
\]

Since uncorrelatedness and independence are equivalent for multivariate Gaussians, the result follows (as in [Speed and Kiiveri, 1986]). We also see that rescaling the diagonal elements of \( K \) to one gives the matrix of negative partial correlation coefficients.
2.2 MAXIMUM LIKELIHOOD ESTIMATION

To estimate $\Sigma$, we will make use of the empirical covariance matrix $S$, defined by

$$S = \frac{1}{n} \sum_{1 \leq i \leq n} (X^i - \bar{X})(X^i - \bar{X})^\top$$

where the $X^i$ are $n$ independent samples from a $\mathcal{N}(0, \Sigma)$-distributed population and $\bar{X} = \frac{1}{n} \sum_i X^i$ is the sample mean.\(^1\)

For a $p \times p$ matrix $M$ with $p = |V(G)|$, define the $G$-partial matrix $M_G$ to be the projection $\pi_G$ of $M$ onto those of its entries corresponding to edges in $G$:

$$\pi_G : M \mapsto M_G := \{M_{ij} : (i, j) \notin E(G)\}.$$  

In general we will associate each element $(i, j)$ of $E(G)$ with the corresponding entries of either $\Sigma$ or $K$. The $G$-partial sample covariance matrix $S_G$ will play a role below.

2.2.1. The following multi-part theorem, due to [Dempster 1972], characterizes the maximum likelihood estimation of $\Sigma$ in Gaussian graphical models. His original proof is quite general and makes use of the theory of exponential families. We first give a direct treatment.

1. When it exists, the maximum likelihood estimator $\hat{\Sigma}$ of $\Sigma$ is a positive definite symmetric matrix such that $\hat{\Sigma}_G = S_G$ and $\hat{K} := \hat{\Sigma}^{-1}$ has zeroes corresponding to the nonedges of $G$, i.e. $\hat{K}_G^c = 0$.

Ignoring the constant term, the log-likelihood function for a multivariate Gaussian on the independent observations $X^1, \ldots, X^n$ is given by

$$l(\mu, \Sigma) = -\frac{n}{2} \log \det \Sigma - \frac{1}{2} \text{tr} \left\{ K \sum_{i=1}^n (X^i - \mu)(X^i - \mu)^\top \right\}$$

which can be rewritten in the form

$$-\frac{n}{2} \log \det \Sigma - \frac{n}{2} \text{tr} (SK) - \frac{n}{2} (\bar{X} - \mu) K (\bar{X} - \mu)$$

by putting $X^i - \mu = (X^i - \bar{X}) - (\mu - \bar{X})$ and simplifying. The third term contains all of the $\mu$-dependence. Since $K$ is positive definite, this term is equal to zero exactly when $\mu = \bar{X}$ and positive otherwise, so we can take $\hat{\mu} = \bar{X}$. We need to find a $\hat{\Sigma}$, then, which will maximize

$$-\frac{n}{2} \log \det \Sigma - \frac{n}{2} \text{tr} (SK)$$

[Dempster 1972] used the sample covariance $S' = (n - 1)^{-1} \sum_i (X^i - \bar{X})(X^i - \bar{X})^\top$, which I don’t believe works as a maximum likelihood estimate.
which is the same, termwise, as
\[
\frac{n}{2} \log \det K - \frac{n}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} s_{ij} k_{ij}
\]
(6)
since \(\log \det \Sigma = -\log \det K\) and \(\text{tr}(SK) = \sum_i \sum_j s_{ij} k_{ji}\). Using the symmetry of the matrices, this expression for \(\text{tr}(SK)\) implies that the \(G\)-partial sample covariance matrix \(S_G\) is a sufficient statistic for \(K\) (and it is in fact minimal sufficient). Setting the partial derivatives of the above expression to zero and cancelling the \(n/2\) terms gives a set of critical equations of the form
\[
\frac{1}{\det K} \cdot \frac{\partial}{\partial k_{ij}} \det K = (2 - \delta_{ij}) s_{ij}
\]
(7)
for \((i,j) \in E\). The derivative \(\frac{\partial}{\partial k_{ij}} \det K\) gives the \((i,j)\) cofactor of \(K\), and so by Cramer’s rule the right hand side is simply \((K^{-1})_{ij}\). So we must have \(\hat{\Sigma}_{ij} = s_{ij}\). The zeroes follow from the fact that \(\hat{K}\) must be part of the parameter space for the model. (The proof here is adapted from [Drton et al., 2009, ch. 2].)

The uniqueness of the MLE follows from the fact that the log-likelihood \(l(K)\) expressed as a function of the precision matrix (as in equation 6) is a strictly concave function on \(\mathbb{S}_+^p\), the set of positive definite \(p \times p\) matrices. It also follows from more general results in the theory of exponential families, as we discuss below.

2. If there is any positive definite symmetric matrix agreeing with \(S\) in the positions corresponding to \(E(G)\) then there is exactly one such matrix \(\hat{\Sigma}\) with the additional property that \(\hat{K}\) as defined above has zeroes corresponding to the nonedges of \(G\).

This result is proven by [Dempster, 1972], exploiting the theory of canonical/mean parameter duality in exponential families, which we develop in the next section. It is also a general property of positive definite matrix completion problems of this type, and can be derived from the discussion that follows combined with e.g. [Grone et al., 1984, theorem 2].

3. Among all multivariate normal distributions such that \(\Sigma_G = S_G\), i.e. in which \(\Sigma\) agrees with \(S\) on the edges of \(G\), the estimator \(\hat{\Sigma}\) yields the distribution with maximum entropy.

In general, for a distribution defined by a density \(f\) with respect to some base measure \(\mu\), the entropy is defined by
\[
H(f) = -\int_{f>0} f(x) \log f(x) \mu(dx).
\]
This is a generalization of the quantity \(-\sum_i p_i \log p_i\), which for a discrete distribution represents the expected self-information (or
“surprise”) associated with each of the outcomes \(i\) (by convention the summand is zero whenever \(p_i\) is zero). Entropy is a measure of smoothness: on a finite set or fixed real interval, the maximum entropy distribution is uniform; on the half-line, the maximum entropy distribution with a given mean is exponential; and on the real line, the maximum entropy distribution with a given mean and variance is Gaussian. The entropy of a mean-zero multivariate Gaussian is given by

\[
- \left\{ \frac{1}{2} p \log 2\pi + \frac{1}{2} \log \det \Sigma + \frac{1}{2} p \right\};
\]

and so the claim is essentially that \(\hat{\Sigma}\) maximizes \(-\log \det \Sigma\). This can be seen, though, from equation \(\mathbf{B}\), where the second term is fixed by the requirement that \(\Sigma_G = S_G\).

As [Dempster, 1972] points out,

\[
\det \Sigma = \sigma_{11}\sigma_{22}\cdots\sigma_{pp} \times \det R,
\]

where \(R\) is the correlation matrix, and so the estimator with maximum entropy corresponds in this case to the estimator with minimum overall correlation, in some sense.

It may seem unnatural at first to take \(\hat{\Sigma}\) to be a positive definite completion of \(S_G\), rather than \(S_G\) itself, particularly given that \(S_G = S\) in the \(n = p\) case. However, the above theoretical properties of \(\hat{\Sigma}\), along with the fact that \(S_G\) may not be positive definite, are compelling reasons to favor \(\hat{\Sigma}\), particularly if the conditional independencies corresponding to \(G\) are known a priori.

### 2.3 Computation of the MLE

There have been a number of algorithms proposed for computing the MLE in practice:

- **Dempster [1972]** originally suggested two algorithms. The first is of Newton-Raphson type, based on a second-order approximation to the log-likelihood. The second cycles through entry-wise adjustments to the non-fixed elements of the estimated covariance or concentration matrix.

- **Speed and Kiiveri [1986]** put forward a form of iterative proportional scaling, and suggest a geometric interpretation in terms of I-divergence (more commonly known nowadays as Kullback-Leibler divergence) ([Csiszar, 1975]; see also [Lauritzen, 1998, p. 134]). Their algorithm cycles through cliques of the underlying graph.
• More recently, Dahl et al. [2008] proposed adaptations of Newton’s method and the conjugate gradient method that take advantage of the structure of chordal graphs by first finding a chordal cover of $G$.

All of the above algorithms can be understood in the framework of conic optimization as solving the feasibility problem for a semidefinite program [Boyd and Vandenberghe, 2004, p. 204]. More specifically, our scenario corresponds to a positive definite matrix completion problem under a rank constraint corresponding to the sample size $n$ [Uhler, 2011].

For many of these algorithms, due to the convex nature of the problem, convergence to the MLE $\hat{\Sigma}$ is guaranteed, and so in practice, there should be little need to determine whether $\hat{\Sigma}$ exists independently of trying to compute it. In many modern applications, moreover, it has become common to perform estimation and (graphical) model selection in the one step (for example in the regularization framework).

From a theoretical point of view, though, it is illuminating to view the problem from the more general perspective of the theory of exponential families. In that context, one can obtain necessary and sufficient conditions for the existence and uniqueness of $\hat{\Sigma}$, as well as generalizations of all three of the above results. We give a brief description of this theory here.

2.4 EXponential FAmilies & the mLE

An exponential family is a collection of parametrized probability densities whose logarithms are linear in the parameters. They are central to statistical theory for a number of reasons. According to the Pitman-Koopman-Darmois theorem, for example, among families of probability distributions whose support does not vary with the parameter to be estimated, only in exponential families is there a sufficient statistic with bounded dimension as the sample size grows. Exponential families also naturally appear as conjugate priors in Bayesian inference.

In our case, the theory of exponential families furnishes necessary and sufficient conditions for the existence of the MLE. Through the conjugate dual relationship between natural parameters and mean parameters (and thereby maximum likelihood and maximum entropy) the theory also sheds light on both Dempster’s results, and the geometric interpretation of them given in the section to follow. We begin with some definitions.

2.4.1. Given a random vector $X$ defined on $\mathbb{R}^n$ (more generally on some space $\mathcal{X}^n$), a vector of canonical or exponential parameters $\theta \in \mathbb{R}^p$, a collection of potential functions or sufficient statistics $\phi_1, \ldots, \phi_p$ mapping $\mathbb{R}^n \to \mathbb{R}^p$, and a base measure $\nu$, the exponential family
associated with the vector of sufficient statistics \( \phi(x) \) is the collection of distributions

\[
\exp \{ \{ \theta, \phi(x) \} - A(\theta) \} \ d\nu,
\]

where \( A(\theta) \) is the normalization factor

\[
\log \int_{\mathbb{R}^n} \ d\nu \exp \{ \theta^\top \phi(x) \},
\]

ensuring that the distributions are probabilities.

Other names for \( A(\theta) \) are the *cumulant function* or *log-partition function*. It is convex in \( \theta \), and has derivatives of all orders on \( \Omega \). This implies that the *natural parameter space*

\[
\Omega := \{ \theta \in \mathbb{R}^p : A(\theta) < \infty \}
\]

is also convex. If \( \Omega \) is an open set, the exponential family is called *regular*; we will assume this to hold. If there exists no nonzero \( a \in \mathbb{R}^p \) such that \( \langle a, \phi(x) \rangle \) is equal to a constant \( \mu \)-almost everywhere, then the family is called *minimal*; in this instance, there is a unique \( \theta \) associated with each density, ensuring identifiability. A family that is not minimal is called *overcomplete*.

Under many circumstances, an exponential family may also be parametrized by its *mean parameters*. If \( f \) is a probability measure absolutely continuous with respect to a base measure \( \nu \) (where \( f \) at this point does not necessarily belonging to an exponential family), then the mean parameter \( \mu_a \) associated with the sufficient statistic \( \phi_a \) is defined to be

\[
\mu_a = E_f \phi_a(X) = \int \phi_a(x)f(x) \ d\nu(x),
\]

leading to a \( p \)-dimensional vector \( \mu \) of *mean parameters*. As in [Wainwright and Jordan, 2008, sec. 3.4.1] we define the set

\[
\mathcal{M} := \{ \mu \in \mathbb{R}^p : \exists f \text{ s.t. } E_f \phi_a(X) = \mu_a \quad \forall a \in \mathcal{I} \}
\]

where \( |\mathcal{I}| = p \). This corresponds, for a given set of sufficient statistics \( \{ \phi_a \}_{a \in \mathcal{I}} \), to all values of the mean parameters realizable by any density \( f \). Like \( \Omega \), the set \( \mathcal{M} \) is convex. It is not too hard to show that

\[
\frac{\partial A}{\partial \theta_a}(\theta) = \int \phi_a(x)f_\theta(x)\nu(dx) = E_\theta \phi_a(X),
\]

which yields a map

\[
\nabla A : \Omega \to \mathcal{M}
\]

\[
\theta \to E_\theta \phi(X)
\]

from canonical parameters to mean parameters.
2.4.2. The mapping \( r_A \) is one-to-one if and only if the exponential representation is minimal; and minimality also implies that \( r_A \) maps onto the interior of \( \mathcal{M} \), denoted \( \mathcal{M}^\circ \). ([Wainwright and Jordan, 2008, prop. 3.2 and thm. 3.3.])

From these results it can be seen that given a set of sufficient statistics (and disregarding the boundary of \( \mathcal{M} \)), all mean parameters that are realizable by some density \( f \) are realizable by a member of the exponential family – namely, the density \( f_{\theta(\mu)} \) corresponding to the unique element \( \theta(\mu) \) satisfying the moment matching condition

\[
E_{\theta(\mu)} [\phi(X)] = \nabla A(\theta(\mu)) = \mu.
\] (8)

Pairs \((\theta, \mu)\) satisfying this condition are said to be dually coupled.

We are now ready to discuss the conjugate duality of maximum likelihood and maximum entropy. The conjugate dual of \( A \) is the function given by

\[
A^*(\mu) := \sup_{\theta \in \Omega} \{ \langle \mu, \theta \rangle - A(\theta) \}.
\]

If we pass as an argument to \( A^* \) a vector of estimated mean parameters \( \hat{\mu} \in \mathcal{M} \), then \( A^* \) returns the value of \( \theta \in \Omega \) which maximizes the log-likelihood. Making this more precise,

2.4.3. ([Barndorff-Nielsen, 1978, p. 153]) In a regular exponential family, the maximum likelihood estimator exists if and only if the sufficient statistic \( \phi(x) \) lies in the interior of its convex support.

On the other hand, we also have the following theorem:

2.4.4. ([Wainwright and Jordan, 2008, thm 3.4 and discussion]) Assume our exponential family has a minimal representation. For any \( \mu \in \mathcal{M}^\circ \), denote by \( \theta(\mu) \) unique canonical parameter satisfying the matching condition [8]. Then the following hold:

1. The conjugate dual function \( A^* \) takes the form

\[
A^*(\mu) = \begin{cases} 
-H(f_{\theta(\mu)}) & \text{if } \mu \in \mathcal{M}^\circ \\
+\infty & \text{if } \mu \notin \overline{\mathcal{M}}
\end{cases},
\]

where \( \overline{\mathcal{M}} \) denotes the (topological) closure of \( \mathcal{M} \) and \( H \) as a function of \( \mu \) returns the entropy of the density \( f_{\theta(\mu)} \). For any boundary point \( \mu \in \overline{\mathcal{M}} \setminus \mathcal{M}^\circ \), \( A^*(\mu) = \lim_{n \to \infty} A(\mu_n) \) for any sequence \( \mu_n \) in \( \mathcal{M}^\circ \) converging to \( \mu \).

2. The log partition function \( A \) has the representation

\[
A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \}.
\] (9)
3. For all $\theta \in \Omega$, the supremum in equation 9 above is attained uniquely at the vector $\mu \in \mathcal{M}^\circ$ specified by the moment matching conditions

$$\mu = E_{\theta} \phi(X).$$

4. Where they are both defined, $\nabla A^\ast$ is the inverse of $\nabla A$.

These propositions imply that $A^\ast(\mu)$ corresponds to the optimum of the following maximum entropy problem:

$$\text{find } \arg \max_{p \in \mathcal{P}} H(p) \text{ subject to } E_p \phi(X) = \mu$$

where $\mathcal{P}$ is the set of all ($\nu$-absolutely continuous) probability distributions over the random variable $X$ and $\phi$ is a vector of sufficient statistics.

With this framework in mind, the result given in the appendix of [Dempster 1972] can be described as follows:

2.4.5. In a minimal, regular exponential family, if we hold certain moment parameters $\theta_A$ fixed while letting the corresponding exponential parameters $\mu_A$ vary, and at the same time hold the exponential parameters $\theta_B$ fixed while letting the corresponding $\mu_B$ vary (here the index set $\mathcal{I}$ is the disjoint union of $A$ and $B$) then the maximum likelihood and maximum entropy estimators correspond.

2.5 THE GEOMETRY OF COVARIANCE SELECTION

Let’s return to the multivariate Gaussian setting. In the general $\mathcal{N}(\mu, \Sigma)$ normal model, $\mu$ and $\Sigma$ cannot be estimated independently, so at this point we will follow both Dempster and Uhler in restricting our attention to the case where the mean $\mu$ is known (and will without loss of generality take $\mu = 0$). A recent paper of Gehrmann and Lauritzen [2012] explores equality constraints on the edges of $G$ (equivalently, the entries of $K$) which ensure estimability of $\mu$ independently of the true value of $\Sigma$.

2.5.1. A cone $C$ is a subset of a real vector space $V$ which is closed under $\mathbb{R}_+^\ast$ multiplication. A subset $S$ of $V$ is convex if for any $p, p' \in S$, the line segment between $p$ and $p'$ lies entirely in $S$, i.e.

$$\lambda p + (1 - \lambda)p' \in S$$

$^2$A recent paper of Gehrmann and Lauritzen [2012] explores equality constraints on the edges of $G$ (equivalently, the entries of $K$) which ensure estimability of $\mu$ independently of the true value of $\Sigma$. 
for $\lambda \in [0, 1]$. A subset $C'$ of a real vector space is a *convex cone* if it is convex and a cone, or equivalently, if for any $q, q' \in C'$ and $\alpha, \beta \geq 0$, 

$$aq + \beta q' \in C'.$$

The most important convex cones for our purposes are the cones $\mathbb{S}^p_>$ (respectively $\mathbb{S}^p_\succeq$) of positive definite (respectively positive semi-definite) matrices. These are full-dimensional in the $\frac{1}{2}p(p + 1)$-dimensional $\mathbb{R}$-vector space of $p \times p$ symmetric matrices, which we denote $\mathbb{S}^p$.

One can introduce an inner product on this space distinct from the usual Euclidean inner product, namely the Frobenius inner product defined by 

$$(A, B) = \text{tr}(AB).$$

The dual of a cone $C$ is defined by 

$$C^* = \{\langle X, Y \rangle \geq 0 : Y \in C\};$$

geometrically, $y \in C^*$ if and only if $-y$ is normal to a hyperplane supporting $C$ at the origin. Under the Frobenius inner product, $\mathbb{S}^p_\succeq$ is self-dual.

Notice, then, that the multivariate normal density can be written 

$$f(x) = \exp\left\{\langle \mu, x \rangle - \frac{1}{2}(K, xx^\top)\right\}.$$ 

Since we are assuming that $\mu = 0$, this simplifies to 

$$f(x) = \exp\left\{-\frac{1}{2}(K, xx^\top)\right\}.$$ 

and so in the notation of the previous section we have 

$$\Omega = \mathbb{S}^p_\succ \quad \text{and} \quad \mathcal{M}^\circ = \mathbb{S}^p_\succ$$

If we impose that our model be adapted to some undirected graph $G$, the situation changes somewhat. Our new $\Omega$ in this case is the convex cone of concentration matrices in the model, defined by 

$$\mathcal{K}_G := \{K \in \mathbb{S}^p_\succ : k_{ij} = 0 \text{ whenever } (i, j) \notin E\}.$$ 

The set $\mathcal{K}_G^{-1}$ of all covariance matrices in the model consists of all inverses of the elements in $\mathcal{K}_G$. Since each such element is positive definite, so is every element of $\mathcal{K}_G^{-1}$. Moreover, by Cramer’s rule and the fact that $\det K > 0$ for all $K \in \mathcal{K}_G$, the set $\mathcal{K}_G^{-1}$ must be an algebraic variety intersected with $\mathbb{S}^p_\succ$.

Define the projection 

$$\mathcal{E}_G := \pi_G(\mathbb{S}^p_\succ)$$
2.5 The Geometry of Covariance Selection

Figure 9: Diagram taken from [Uhler, 2012]; she uses \( m \) where we use \( p \) for the number of components of \( X \). Note that, despite how it is drawn, the set \( \mathcal{C}_G \) is also a cone, though its elements may not be positive (semi)definite.

of the cone of positive definite matrices; then by 2.5.2, \( \mathcal{C}_G \) represents the set of all sufficient statistics in the model for which the MLE exists. Note that \( \mathcal{C}_G \not\subset \mathbb{S}_+^p \) in general, since the \( S_G \) may not be positive definite. The set \( \mathcal{C}_G \) is a convex cone, and is dual to \( \mathcal{K}_G \), as was proven in [Sturmfels and Uhler, 2010].

This leads to the following geometric interpretation of the MLE (see figure 10):

2.5.2. The MLEs \( \hat{\Sigma} \) and \( \hat{K} \) exist for a given sample covariance matrix \( S \) if and only if

\[
\text{fiber}_\phi(S) := \{ \Sigma \in \mathbb{S}_+^m : \Sigma G = S_G \}
\]

is nonempty, in which case \( \text{fiber}_\phi(S) \) intersects \( \mathcal{K}_G^{-1} \) in exactly one point, namely \( \hat{\Sigma} \).

In particular, this implies that the MLE \( \hat{\Sigma} \) has an algebraic description in terms of the sufficient statistics \( S_G \), i.e. can be represented as a solution of polynomial equations in its entries.

Dempster’s results (2.2.1) show that the MLE exists in any Gaussian graphical model when \( n \geq p \) (and this is a well-known fact besides). More specifically, since \( S \) in that case has rank \( p \) with probability one, it is strictly positive definite, and hence \( S_G \) has (among others) the positive definite completion \( \hat{S} \); the assertion then follows from the second part of 2.2.1.
Based on the structure of the graph $G$, it turns out that we can improve this lower bound:

**2.5.3** *(Grone et al. [1984]).* For a graph $G$, the following are equivalent:

- a $G$-partial matrix has a positive definite completion if and only if all submatrices corresponding to maximal cliques in $M_G$ are positive definite)

- the graph $G$ is chordal.

This implies the following upper and lower bounds (recall that $q$ denotes the maximum clique size of $G$, and $q^*$ denotes its treewidth):

**2.5.4.** *(Buhl, 1993, corr. 3.3):* If $n \geq q^*$, the MLE exists with probability 1. If $n < q$, the MLE does not exist.

Given this result, only the case $q \geq n \leq q^*$ is of theoretical interest. The following theorem in theory sharpens these bounds still further, but only with great computational effort.

**2.5.5.** *(Uhler, 2012, Thm. 3.3) (Algebraic elimination criterion.)* Let $I_{G,n}$ be the elimination ideal obtained from the ideal of $(n+1) \times (n+1)$ minors of a symmetric $m \times m$ matrix $S$ of unknowns by eliminating all unknowns corresponding to nonedges of the graph $G$. If $I_{G,n}$ is the zero ideal, then the MLE exists with probability one for $n$ observations.

**Proof.** The variety associated with the ideal of $(n+1) \times (n+1)$ minors of a $p \times p$ symmetric matrix consists of all $p \times p$ matrices of rank at most $n$ in $\mathbb{S}^p$; this follows from the characterization of the rank of a matrix by the size of the largest nonvanishing minor. Eliminating the unknowns corresponding to nonedges of the graph $G$ results in the elimination ideal $I_{G,n}$, and is the geometric equivalent of projecting the variety of matrices of rank onto the cone of sufficient statistics $\mathcal{C}_G$. Let $V$ be the ideal corresponding to the elimination ideal $I_{G,n}$. Denote by $k$ its dimension and by $\mu$ a $k$-dimensional Lebesgue measure. The MMLE exists with probability one for $n$ observations if

$$\mu(V \cap \partial \mathcal{C}_G) = 0,$$

where $\partial \mathcal{C}_G$ denotes the boundary of the cone of sufficient statistics $\mathcal{C}_G$. If $I_{G,n}$ is the zero ideal, then the variety $V$ is full-dimensional, and its dimension is $k = \dim \mathcal{C}_G$. If we then assume that $\mu(V \cap \partial \mathcal{C}_G) > 0$, then $\mu(\partial \mathcal{C}_G) > 0$, contradicting that $\dim \mathcal{C}_G < k$. \(\blacktriangleright\)
Algebraic geometry grew out of the classical study of curves and surfaces which occur as the solution sets (also called loci) of systems of polynomial equations, such as points and lines, circles and spheres, conic sections, and elliptic curves. From the 1930s through the 1960s and 70s, the abstract-algebraic aspects of the subject (collected under the heading commutative algebra) grew increasingly important. The level of abstraction consequently increased, but the language and concerns of geometry persisted. See [Dieudonné, 1972] for an overview.

The concept of a Gröbner basis was independently developed by Hironaka (in less generality) and Buchburger in the mid 1960s, though the germs of the idea can in hindsight be traced as far back as a 1900 paper of Gordan. Buchburger, in his PhD thesis, also devised an algorithm to compute them. Combined with the availability of sufficient computing power, these revitalized the more combinatorial and algorithmic aspects of the subject [Sturmfels, 2005].

This part of the project aims to give a brisk introduction to the field, in order to make the Gröbner basis algorithms used in part 2 more transparent. The presentation is mostly adapted from [Cox et al., 1997] and [Schenck, 2003]. The former text is an extremely thorough and computational introduction to the subject; the latter is more brisk and abstract.

### A.1 Basic Concepts

**A.1.1. A ring** is a set whose elements can be added and multiplied. In other words, a ring is a commutative group under $+$, which also possesses an associative operation $\cdot$ which distributes over addition. It is customary to restrict attention to rings in which multiplication is commutative with identity element 1. A **field** is such a ring in which every nonzero element has a multiplicative inverse.

Examples of rings include the familiar number systems $\mathbb{Z}, \mathbb{Q}, \mathbb{R},$ and $\mathbb{C}$ (the latter two are fields), integers mod $n$, and most importantly for our purposes, the ring of polynomials in finitely many indeterminates $x_1, \ldots, x_n$ with coefficients in a field $k$, denoted $k[x_1, \ldots, x_n]$.

The main idea behind algebraic geometry is to identify the elements of this ring with polynomial functions on the n-dimensional affine space $k^n$ (or on n-dimensional projective space $P^n(k)$) and to
identify collections of these polynomials with the zeroes of the cor-
responding functions.

For example, letting \( k = \mathbb{R} \) and considering the ring \( \mathbb{R}[x, y] \), polyno-
imals of the form \( ax + by + c \) correspond, via equations of the form
\( ax + by + c = 0 \), to straight lines in \( \mathbb{R}^2 \). Two such equations consid-
ered together may correspond to a line, a single point, or no points,
depending on whether associated lines are coincident, cross at a
point, or are parallel and distinct. The element \( y - x^2 \) corresponds
to a parabola; the polynomials \( \{ y - x, y - x^2 \} \) considered jointly cor-
respond to the points \( (0,0), (0,1) \). The element 1 corresponds to \( \emptyset \);
the element 0 corresponds to the whole of \( \mathbb{R}^2 \).

**A.1.2.** An **affine variety** is a set of the form
\[
\{ x \in k^n : f(x) = 0 \text{ for all } f \in S \}
\]
for some set \( x \) of polynomials \( S \subseteq k[x_1,\ldots,x_n] \).

We use the term affine to distinguish varieties in \( k^n \) from those
in projective space, and we say that the set of polynomials **cut out**
the associated variety. We denote by \( V(\cdot) \) the map that takes a set
of polynomials to the associated affine variety.

To give some sense of the breadth of algebraic geometry, we
could also imagine working over the finite field \( \mathbb{F}_2 \), in which \( 1 + 1 = 0 \)
(which is a field since \( 1^{-1} = 1 \)). In this case, polynomials in indeter-
minates \( x_1,\ldots,x_n \) can be naturally identified with indicator func-
tions (in the sense of probability theory) of subsets of \( \{x_1,\ldots,x_n\} \).
For example, \( x_1 \cdot x_2 \) takes the value 1 when \( x_1 = x_2 = 1 \) and zero oth-
erwise. There are many subtleties involved in working over finite
fields; over \( \mathbb{F}_2 \), for example, \( x^2 \) and \( x \) are identical as functions. The
flipside is that working over fields larger than \( \mathbb{R} \) can make things
easier: in \( \mathbb{C}[x] \), for example, the fundamental theorem of algebra
guarantees that each polynomial has at least one root. Many of the
concepts of algebraic geometry apply when working with more gen-
eral rings of functions on a space (such as the ring of continuous
\( \mathbb{R} \)-valued functionals on a topological space).

Let’s return to the ring \( \mathbb{R}[x,y] \), and consider the example
\( V(y - x, y - x^2) = \{(0,0),(0,1)\} \). Notice that enlarging our set of polynomials
by an \( \mathbb{R}[x,y] \)-multiple of a polynomial which is already in the set
will leave the associated variety unchanged: for example, \( V(y - x, y - x^2) = V(y - x, y - x^2, x(y - x)) \). This suggests the following definitions.

**A.1.3.** An **ideal** in a ring \( R \) is a subset of \( R \) which is closed un-
der addition and \( R \)-multiplication. The ideal **generated** by a set
of elements \( A \subseteq R \) is the set of finite \( R \)-linear combinations of these
elements, denoted \( \langle A \rangle \). An **\( R \)-module** is any set that is closed under
\( R \)-linear combinations of its elements (i.e., under the ring action of
\( R \)), and so an ideal is just an \( R \)-module contained in \( R \).
Figure 10: Three varieties plotted in $\mathbb{R}^3$. Let $f_1 = x^2 + y^2 - 1$ and $f_2 = x^2 + z^2 - 1$. Then $V(f_1)$ and $V(f_2)$ are orthogonal cylinders in $\mathbb{R}^3$, and their union, given by $V(f_1 : f_2)$, is plotted on the left. Their intersection is $V(f_1) \cap V(f_2)$ since $f_1^2 + f_2^2 = 0 \Rightarrow f_1 = f_2 = 0$; the middle image shows a perturbation of this, given by $V(f_1^2 + f_2^2 - \alpha)$ with $\alpha = 0.005$. The image on the right shows the perturbation $V(f_1 : f_2 - \alpha)$ with $\alpha = 1$.

As suggested by the examples above, an important consequence of these definitions is that for any subset $A \subseteq k[x_1, \ldots, x_n]$ we have that $V(A) = V(\langle A \rangle)$. For some more intuition behind this fact, consider that $\langle A \rangle$ may be thought of as the closure of $A$ under “polynomial consequences” — for $f \in I, g \in R, f = 0$ always implies $fg = 0$.

So far we have considered a map

$$V : k[x_1, \ldots, x_n] \ni A \mapsto \{ \text{zero set of } A \} \subseteq \mathbb{k}^n,$$

defined images under this map to be varieties, and discovered it makes sense to restrict this map to ideals of $k[x_1, \ldots, x_n]$. Let’s now go in the other direction, and consider the map

$$I : \mathbb{k}^n \ni B \mapsto \{ \text{polynomials vanishing on } B \} \subseteq k[x_1, \ldots, x_n].$$

As the notation suggests, images under this map are ideals in $k[x_1, \ldots, x_n]$; to see this, notice that if $f$ and $f'$ vanish on $B$ then so does their sum, as well as the polynomial $gf$ for any $g \in k[x_1, \ldots, x_n]$.

Given this correspondence, it is natural to consider, given a set $S \subseteq \mathbb{k}^n$ or an ideal $I'$, the variety $V(I(S))$ and the ideal $I(V(I'))$.

**A.1.4. The process of passing between algebraic and geometric objects is inclusion-reversing — i.e.,

$$I_1 \subseteq I_2 \Rightarrow V(I_2) \subseteq V(I_1)$$

and

$$S_1 \subseteq S_2 \Rightarrow I(S_2) \subseteq I(S_1)$$

— though the reverse implications do not hold.**

Working in $\mathbb{R}[x]$, for example, the ideal $I(\mathbb{R}) = I(\langle j \rangle_{j \in \mathbb{Z}}) = \langle 0 \rangle$ (since any nonzero polynomial has finitely many roots) but $\mathbb{R} \not\subseteq$
On the other hand, $V\langle x \rangle = V\langle x^2 \rangle = \langle 0 \rangle$, but $\langle x \rangle \not\subseteq \langle x^2 \rangle$. It is easy to argue that when $S$ is any subset of $k^n$, $V(I(S))$ is the smallest affine variety containing $S$.

The second counterexample given above easily generalizes: over any field, for $m \geq 1$, $V\langle x^m \rangle = V\langle x \rangle = 0$, since no field contains zero divisors (if $ab = 0$, multiply by either $a^{-1}$ or $b^{-1}$). Moreover, for any $f \in k[x_1, \ldots, x_n]$, $f(x)^m = 0$ if and only if $f(x) = 0$, so that if $f^m \in I(V)$ then $f \in I(V)$, and $V\langle \ldots, f, \ldots \rangle = V\langle \ldots, f^m, \ldots \rangle$.

These facts motivate the following definition: an ideal $I$ in a ring $R$ is radical if whenever $f^m \in I$ for some $m \geq 1$, it is also true that $f \in I$. The radical of an ideal $J \subseteq R$ is the ideal

$$\sqrt{J} := \{ f : f^m \in J \text{ for some } m \geq 1 \}.$$

From the definition and the facts above, it follows that for any variety $V$ the ideal $I(V)$ is always radical. For some additional intuition (and justification of the $\sqrt{\cdot}$ notation), notice that in the integers, radical ideals are precisely those generated by squarefree integers, i.e. integers whose prime factorizations contain no powers greater than 1.

This brings us to Hilbert’s Nullstellensatz, one of the most important results in basic algebraic geometry.

**A.1.5.** If $k$ is algebraically closed, and $J$ is an ideal in $k[x_1, \ldots, x_n]$, then

$$I(V(J)) = \sqrt{J}.$$

The Nullstellensatz has a fairly involved proof, which is given in [Cox et al. 1997]. An easy consequence of this result is that the maps $I(\cdot)$ and $V(\cdot)$ restrict (on radical ideals) to an inclusion-reversing bijection:

$$\{ \text{radical ideals} \} \leftrightarrow \{ \text{affine varieties} \}.$$

This correspondence, intimately connected with the fundamental theorem of algebra, helps to explain and motivate the centrality of algebraically closed fields in algebraic geometry.

To close this section, we collect some basic facts about combining and decomposing ideals and varieties.

**A.1.6.** Working in with polynomial ring $k[x_1, \ldots, x_n]$,

1. The sum of two ideals

$$I + J := \{ f + g : f \in I, g \in J \}$$

is an ideal, as is the product

$$I \cdot J := \langle fg : f \in I, g \in J \rangle.$$

2. The intersection of two ideals is an ideal, and $I \cap J \subseteq I \cdot J$. 
3. The union of two (hence finitely many) affine varieties is again an affine variety.

4. The intersection of any number of affine varieties is again an affine variety.

5. \( V(I + J) = V(I) \cap V(J) \)

6. \( V(I \cdot J) = V(I \cap J) = V(I) \cup V(J) \)

The angle brackets in the definition of the product of ideals are essential: consider \( \langle x \rangle \cdot \langle y \rangle \), and \( x + y \) in \( k[x, y] \). Also, the inclusion in 2. is proper in general: consider \( \langle x \rangle \subset k[x] \), whose intersection with itself is \( \langle x \rangle \) but whose product with itself is \( \langle x^2 \rangle \).

From points 3. and 4., and the fact that \( k^n = \langle 0 \rangle \) and \( \langle 1 \rangle = \emptyset \), we can let affine varieties form the closed sets of a topology on \( k^n \). This is called the \textit{Zariski topology}; in general it is not Hausdorff.

The following concepts and result hint at (but does not establish) the possibility of decomposing an arbitrary affine variety into a union of indecomposable components.

\textbf{A.1.7.} An affine variety \( V \subseteq k^n \) is \textit{irreducible} if it cannot be expressed as a union of proper subvarieties. An ideal \( I \subseteq k[x_1, \ldots, x_n] \) is \textit{prime} if whenever \( f, g \in k[x_1, \ldots, x_n] \) and \( fg \in I \), then either \( f \in I \) or \( g \in I \).

\textbf{A.1.8.} An affine variety \( V \subseteq k^n \) is irreducible exactly when \( I(V) \) is a prime ideal.

\section*{A.2 Bases for Ideals}

So far we have seen that in order to study the solution sets of systems of polynomial equations – affine algebraic varieties – the algebraic objects most of interest are ideals in the associated ring of polynomial functions, which are exactly the submodules of that ring. We have also seen that with any set of elements in a ring we can associate the ideal generated by that set.

A generating set for an ideal is called a \textit{basis}. These are not entirely analogous with bases in linear algebra.

For example, for an ideal \( I \subseteq R \) (considered as a module with an \( R \)-action), the \( R \)-linear independence of basis elements does not hold in general. Consider the ideal \( \langle x, y \rangle \subset k[x, y] \); a nontrivial \( k[x, y] \)-linear combination of basis elements is given by \( x \cdot y - y \cdot x \). Such a relation among generators of a module (equivalently, the kernel of a homomorphism of a free module) is called a \textit{syzygy}.

Another potential pathology is that \( \langle x, y \rangle \) is infinite-dimensional as a \( k \)-vector space. A basis is called \textit{minimal} if it has no proper subset that is also a basis; a given ideal can have minimal bases...
consisting of different numbers of elements (e.g. \( \langle x \rangle = \langle x^2 + x, x^3 \rangle \subset k[x] \)).

In linear algebra, to solve a system of linear equations normally means to find a parametric or implicit description of the solution set. Things are not so simple in the polynomial case; for example, only so-called unirational varieties admit rational parametrizations.

We would like to find conditions ensuring the existence of finite generating sets for the ideals in a ring \( R \) (and more generally, for submodules of an \( R \)-module \( M \)).

**A.2.1.** A ring is said to be Noetherian, or to satisfy the ascending chain condition, if every increasing sequence of ideals eventually terminates, i.e. there exist no sequences of ideals of the form

\[
I_1 \subset I_2 \subset I_3 \subset \ldots
\]

where the inclusions are proper.

**A.2.2.** A ring \( R \) is Noetherian if and only if every ideal is finitely generated.

*Proof.* First, suppose that every ideal is finitely generated, but that there exists an ascending chain \( \{I_k\}_{k \geq 1} \). Form the union \( \bigcup_k I_k \); this is itself an ideal. It has some finite generating set \( \langle f_1, \ldots, f_n \rangle \), each of whose members \( f_j \) appears in some ideal \( I_{l(j)} \) in the chain. Take the largest such \( l(j) \); then

\[
\langle f_1, \ldots, f_n \rangle = I_{l(j)} = I_{l(j)+1} = \ldots,
\]

contradicting that the chain is ascending.

On the other hand, assume that there exists some ideal that is not finitely generated. Then we can find a sequence of generators with \( f_i \notin \langle f_1, \ldots, f_{i-1} \rangle \) for each \( i \). Then

\[
\langle f_1 \rangle \subset \langle f_1, f_2 \rangle \subset \ldots
\]

is an infinite ascending chain. \( \square \)

One can show using very similar reasoning to show that an \( R \)-module \( M \) is Noetherian – i.e., contains no infinite ascending chain of submodules – if and only if every submodule of \( M \) is finitely generated.

Every field \( k \) is Noetherian, since the only ideals are \( \langle 0 \rangle = \{0\} \) and \( \langle 1 \rangle = k \) (every nonzero ideal contains 1, by invertability). It turns out that every polynomial ring \( k[x_1, \ldots, x_n] \) is also Noetherian, as an inductive consequence of the following theorem.

**A.2.3. (Hilbert’s Basis Theorem)** If \( R \) is a Noetherian ring, so is \( R[x] \).
Proof. (via [Schenck, 2003]) Let \( I \) be an ideal in \( R[x] \); it suffices to show that \( I \) is finitely generated. The leading coefficients of the elements of \( I \) generate an ideal \( I' \) of \( R \), which has a finite basis \( a_1, \ldots, a_k \in R \), say (since \( R \) is Noetherian). We can associate with each of the generators \( a_i \) an element \( f_i \) in \( I \) with leading term of the form \( a_i x^{m(i)} \). Let \( I'' \) be the ideal of \( R[x] \) generated by the \( f_i \), and let \( m \) be the largest of the \( m(i) \).

Now given any \( f \in I \), we can decompose \( f \) into a sum of elements from \( I'' \), plus terms of degree less than \( m \). Consider the \( R \)-module \( M \) generated by \( \{x_1, \ldots, x_m\} \); it is finitely generated, and hence so is the \( R \)-module \( M \cap I \). Then the generators of \( M \cap I \), together with those of \( I'' \), generate \( I \).  

It is now easy to show the following:

A.2.4. Every affine variety \( V \subseteq k^n \) has a decomposition into irreducible varieties:  
\[ V = V_1 \cup \cdots \cup V_m \]
where \( V_i \not\subseteq V_j \) for \( i \neq j \). (Such a decomposition is called minimal.) Moreover, this decomposition is unique (up to permutations of the \( 1, \ldots, m \)).

Proof. Existence is a consequence of the inclusion-reversing relationship between varieties and ideals, and \( k[x_1, \ldots, x_n] \) satisfying the ascending chain condition. To show uniqueness, let
\[ V = V_1 \cup \cdots \cup V_m = V'_1 \cup \cdots \cup V'_l \]
be two distinct minimal decompositions. Then for any \( V_i \),
\[ V_i = V_i \cap V = (V_i \cap V'_1) \cup \cdots \cup (V_i \cap V'_l) \]
and since \( V_i \) is irreducible we must have \( V_i = V_i \cap V'_j \Rightarrow V_i \subseteq V'_j \) for some \( j \) between 1 and \( l \). Applying a similar argument to \( V'_j \) implies that \( V'_j \subseteq V_k \) for some \( k \) between 1 and \( m \). But then
\[ V_i \subseteq V'_j \subseteq V_k, \]
contradicting minimality.

We close this section with a brief discussion of the dimension of an affine variety, which is needed for the discussion in part 2. This notion is more subtle for varieties than it is for, say, smooth manifolds or vector spaces: for example, \( Vxz, zy) \) corresponds to the union of the \( x-y \) plane and the \( z \) axis in \( \mathbb{R}^3 \).

A.2.5. The \textit{Krull dimension} of a ring \( R \) is the maximum of lengths of chains of prime ideals of \( R \). The \textit{coordinate ring} of an affine variety \( V \subseteq k^n \) is the quotient ring
\[ k[x_1, \ldots, x_n]/I(V). \]
The dimension of an affine variety is defined as the Krull dimension of its coordinate ring. This coincides with the maximum of lengths of chains of proper irreducible subvarieties of $V$. [Perrin, 2008, ch. IV]

The last assertion follows directly from the bijection which exists between irreducible varieties and prime ideals, and the fact that any ring homomorphism maps ideals containing the kernel to ideals contained in the image. Finally, the following alternative characterization of dimension is also useful:

A.2.6. (Cox et al. [1997] corr. 9.5.4) Let $V \subseteq k^n$ be an affine variety. Then the dimension of $V$ is equal to the largest integer $r$ such that $I(V) \cap k[x_{i(1)}, \ldots, x_{i(r)}] = \{0\}$.

A.3 MONOMIALS AND DIVISION

We have proven that every ideal in a polynomial ring has a finite basis. Gröbner bases are finite bases satisfying properties which among other things make certain computations easy. For example, given a polynomial $f$ and an ideal $I = \langle f_1, \ldots, f_n \rangle$, Gröbner basis methods allow one to determine whether $f \in I$, to describe the solution set $V(I)$ of the equations $f_1 = \ldots = f_n = 0$, and to implicitize (i.e., give a system of implicit equations for) a variety defined parametrically.

The method of Gröbner bases is a generalization of both Gaussian elimination with respect to a basis $x_1, \ldots, x_n$ in linear algebra, and the division algorithm for univariate polynomials in $k[x]$. In each of these algorithms an ordering of terms is important: in the linear case, we need an ordering of the basis elements with respect to which our matrix is expressed; and in $k[x]$, we have a natural partial ordering of polynomials by degree.

Importantly, the partial ordering by degree in $k[x]$ is compatible with multiplication, which guarantees that the division algorithm in $k[x]$ eventually terminates. This compatibility is a consequence of the obvious but important equivalence

$$\deg f \leq \deg g \iff \text{lt } f \text{ divides } \text{lt } g$$

where we let $\text{lt } f$ denote the leading term of $f$, and $\deg f$ its degree. Recall that this algorithm, when given input polynomials $f, g$, returns the unique $q$ and $r$ such that

$$f = qg + r,$$

where $r$ is either 0 or has degree strictly less than that of $g$. Here is a description in pseudocode, where $a \mid b$ denotes “$a$ divides $b$”:
• input $f, g$
• set $q := 0, r := f$
• while $r \neq 0$ and $\text{lt } g \mid \text{lt } r$,
  – set $q \leftarrow q + \frac{\text{lt } r}{\text{lt } g}$, and
  – set $r \leftarrow r - \frac{\text{lt } r}{\text{lt } g} g$.

The algorithm outputs $q$ and $r$. At each iteration of the while-loop, the equality $f = qg + r$ holds, while $r$ must either be set to zero or have its degree decrease.

In the multivariate setting of $k[x_1, \ldots, x_n]$, the two kinds of order mentioned above need to be combined and generalized. We require compatibility with division — so that the analogous multivariate division algorithm terminates — but we also need to ensure uniqueness of the results.

First, let’s generalize the notion of the degree of a univariate polynomial.

A3.1. Given ordered indeterminates $x_1, \ldots, x_n$, a monomial is an element of the form $x_1^{a_1} \cdots x_n^{a_n}$, and its multidegree is the element $\alpha = (a_1, \ldots, a_n) \in \mathbb{N}^n$. (Here $\mathbb{N} = \mathbb{Z}_{\geq 0}$.) Its total degree is $\sum \alpha_i$.

A monomial ordering on $k[x_1, \ldots, x_n]$ is a total well-ordering $\prec$ on $\mathbb{N}^n$ (equivalently, on the set of monic monomials $x^\alpha$, $\alpha \in \mathbb{N}^n$) such that $\alpha \prec \beta$ implies $\alpha + \gamma \prec \beta + \gamma$ for any $\alpha, \beta, \gamma \in \mathbb{N}^n$. This means that $\prec$ is compatible with the multiplication of monomials, and that every nonempty subset of $\mathbb{N}^n$ has a least element. The well-ordering property is equivalent to the statement that $\alpha \geq 0$ for all $\alpha \in \mathbb{N}^n$.

Given a monomial ordering $\succ$ and a polynomial $f$, we generalize the notion of leading term to refer to its largest term with respect to $\succ$ (we write $\text{lt } \succ f$ or, say, $\text{lt}_{\text{lex }} f$ when the ordering needs to be specified). We then define the multidegree of $f$ to be the multidegree of its leading term, the leading coefficient $\text{lc } f$ to be the coefficient of its leading term, and the leading monomial $\text{lm } f$ to be the (monic) monomial corresponding to its leading term.

Here are some examples of monomial orderings. Lexicographic or lex order means that $\alpha \succ \beta$ exactly when the left-most nonzero entry of $\alpha - \beta \in \mathbb{Z}$ is positive; it corresponds to the usual ordering of words in a dictionary. Graded lexicographic or grlex order means that $\alpha \succ \beta$ whenever $\sum_i \alpha_i > \sum_i \beta_i = |\beta|$ or $|\alpha| = |\beta|$ and $\alpha \succ_{\text{lex }} \beta$;

it corresponds to an ordering by total degree with ties broken lexicwise. A less intuitive but occasionally computationally useful order-
ing is graded reverse lexicographic order or `grevlex`, under which \( \alpha > \beta \) exactly when

\[
|\alpha| = \sum_i \alpha_i > \sum_i \beta_i = |\beta| \quad \text{or} \quad |\alpha| = |\beta|
\]

and in \( \alpha - \beta \), the rightmost nonzero entry is negative. In this case polynomials are again partially ordered by total degree, but ties are broken by a kind of double-reversal of the lex ordering.

Given a monomial ordering, we can unambiguously order the monomials in a multivariate polynomial. For example, the polynomial \( 4xyz^2 + 4z^2 - 5x^3 + 7x^2z^2 \) would be written

\[
-5x^3 + 7x^2z^2 + 4xy^2z + 4z^2
\]

with respect to lex order,

\[
7x^2z^2 + 4xy^2z - 5x^3 + 4z^2
\]

with respect to grlex, and

\[
4xy^2z + 7x^2z^2 - 5x^3 + 4z^2
\]

with respect to grevlex.

The multivariate division algorithm generalizes the univariate case in two ways: firstly (and most obviously) in that we are dividing multivariate polynomials; and secondly, in that we may divide by more than one polynomial. We begin with an example.

Set \( f = x^2y + xy^2 + y^2 \); we wish to divide this by \( f_1 = xy - 1 \), \( f_2 = y^2 - 1 \) using lex order. Our output will be of the form

\[
f = a_1f_1 + \ldots + a_sf_s + r.
\]

To begin with we set the remainder, \( r \), to 0.

- If \( f_1 \mid \text{lt } f \), so we compute \( \text{lt } f/f_1 = x \) and put \( f \leftarrow f - x \cdot f_1 = xy^2 + x + y^2 \).
- Again, \( f_1 \mid \text{lt } f \); this time \( \text{lt } f/f_1 = y \) so we set \( f \leftarrow f - y \cdot f_1 = x + y^2 + y \).
- Now \( f_1 \nmid \text{lt } f \) \( = x \), so we need to check \( f_2 \). We discover, though, that \( f_2 \nmid \text{lt } f \) \( = y \) — a situation which never arises in the univariate case. To deal with it, we simply add \( x \) to the remainder and continue as before — i.e., \( r \leftarrow r + x \), \( f \leftarrow f - x = y^2 + y \).
- Now \( f_1 \nmid \text{lt } f \), but \( f_2 \mid \text{lt } f \) and \( \text{lt } f/f_2 = 1 \). Then \( f \leftarrow f - 1 \cdot f_2 = y^2 + y - y^2 + 1 = y + 1 \).
- If \( f_1 \nmid \text{lt } f \) \( = y \) and \( f_2 \nmid \text{lt } f \), so we set \( r \leftarrow r + y \) and \( f \leftarrow f - y = 1 \).
Again, \( \text{lt}\ f_1 \nmid \text{lt}\ f \) and \( \text{lt}\ f_2 \nmid \text{lt}\ f \), so \( r \leftarrow r + 1 \) and \( f \leftarrow f - 1 = 0 \) and the algorithm terminates. Its conclusion is that

\[
f = x \cdot (x y - 1) + y \cdot (x y - 1) + 1 \cdot (y^2 - 1) + (x + y + 1).
\]

\[
= (x + y) \cdot f_1 + 1 \cdot f_2 + r.
\]

Notice that \( r \) is a \( k \)-linear combination of monomials, none of which is divisible by the leading terms of the divisors \( f_1, f_2 \).

Here is the multivariate division algorithm, in pseudocode and in greater generality:

- **inputs** \( f_1, \ldots, f_s, f \); 
- set \( a_1, \ldots, a_s, r \leftarrow 0 \), and set \( p \leftarrow f \); 
- while \( p \neq 0 \),
  - set \( i \leftarrow 1 \); while \( i \leq s \) and division hasn’t occurred,
    - if \( \text{lt}\ f_i \mid \text{lt}\ p \),
      - set \( a_i \leftarrow a_i + \frac{\text{lt}\ p}{\text{lt}\ f_i} \) and 
      - set \( p \leftarrow p - \frac{\text{lt}\ p}{\text{lt}\ f_i} \cdot f_i \); 
      - division has occurred; 
    - otherwise increment \( i \leftarrow i + 1 \). 
  - if division did not occur, then 
    - set \( r \leftarrow r + \text{lt}\ p \) and \( p \leftarrow p - \text{lt}\ p \). 

There are a few things to note about this algorithm. Each time through the main while-loop, either a division occurs (in which case the multidegree of either \( p \) decreases or becomes zero) or \( \text{lt}\ p \) is added to the remainder. Also, the equality

\[
f = a_1 f_1 + \ldots + a_s f_s + p + r
\]

holds before and after each block of reassignments. Each coefficient \( a_i \) is of the form \( \text{lt}\ p/\text{lt}\ f_i \) for some value of the variable \( p \), from which one can deduce that the multidegree of each \( a_i \) is less than or equal to that of \( f \), when \( a_i f_i \neq 0 \).

On the other hand, completing the example given with \( f_1 \) and \( f_2 \) interchanged yields

\[
f = (x + 1) \cdot (y^2 - 1) + x \cdot (x y - 1) + 2x + 1,
\]

showing that the remainder \( r \) is not uniquely determined, and that the output of the algorithm is sensitive to the order of the divisors \( f_1, \ldots, f_s \). In particular, the multivariate division algorithm on its own yields only a partial solution to the ideal membership problem: if the remainder of \( f \) on division by \( f_1, \ldots, f_s \) is zero, we may conclude that \( f \in \langle f_1, \ldots, f_s \rangle \), but a nonzero remainder is inconclusive.

Gröbner bases solve this problem (and are in fact characterized by solving this problem, as we shall see).
A.4 MONOMIAL IDEALS AND GRÖBNER BASES

To develop the theory of Gröbner bases, we first focus on monomials and the ideals they generate.

A.4.1. A monomial ideal $I \subseteq k[x_1, \ldots, x_n]$ is an ideal generated by a (possibly infinite) set of monomials $\{x^\alpha\}_{\alpha \in A}$, where $A \subseteq \mathbb{N}^n$.

A monomial $x^\beta$ lies in a monomial ideal $I$ if and only if $x^\alpha | x^\beta$ for some $\alpha \in A$: to see why, notice that both sides of the equality $x^\beta = \sum_i g_i x^{\alpha(i)}$ are divisible by any of the $x^{\alpha(i)}$; this proves the forward implication, and the other direction is trivial.

This criterion can be given a neat geometric interpretation: $x^\alpha | x^\beta$ if and only if $\bar{x}$ lies in the translated positive orthant $\alpha + \mathbb{N}^n := \{\alpha + \gamma : \gamma \in \mathbb{N}^n\} \subseteq \mathbb{N}^n$ with apex at $\alpha$.

A similarly-derived property of monomial ideals is that for a polynomial $f$, the condition that $f \in I$ is equivalent to each term of $f$ lying in $I$, which is equivalent to $f$ being a $k$-linear combination of monomials in $I$. This implies that monomial ideals are exactly determined by the monomials they contain.

We can say a little more though.

A.4.2. Any monomial ideal $I$ in a polynomial ring $k[x_1, \ldots, x_n]$ is generated by some finite set of monomials from the generating set $\{x^\alpha : \alpha \in A\}$.

This is equivalent, it turns out, to Dickson’s lemma in combinatorics: any set $S \subseteq \mathbb{N}^n$ has finitely many minimal elements under the product partial order. The proof of this statement closely parallels that of the Hilbert basis theorem given above.

\textbf{Proof.} Let’s proceed by induction on $n$, where $A \subseteq \mathbb{N}^n$ is the set of indices of the generators. The $n = 1$ case is clear, since in $k[x]$ we need only take $x$ to the power of the smallest index in $A$; this term generates $I$. Now assume that $n > 1$ and that the result holds for $n - 1$, and let $I$ be a monomial ideal in $k[x_1, \ldots, x_{n-1}, y]$.

Let $J$ be the ideal in $k[x_1, \ldots, x_{n-1}]$ generated by the monomials $x^\alpha$ for which $x^\alpha y^m \in I$ for some $m \geq 0$. It can be thought of as the projection of $I$ onto $k[x_1, \ldots, x_{n-1}]$ (and in fact, the associated operation on varieties is a kind of projection). Then by the inductive hypothesis, $J$ is generated by finitely many of the $x^\alpha$ terms, say $x^{\alpha(1)}, \ldots, x^{\alpha(l)}$.

For each $i = 1, \ldots, l$, there exists a term $x^{\alpha(i)} y^{m(i)}$ in $I$ for some $m(i) \geq 0$ by the construction of $J$; let $m$ be the largest of the $m(i)$. For each $k = 0, \ldots, m-1$, form the ideal $J_k \subseteq k[x_1, \ldots, x_{n-1}]$ generated by the monomials $x^\beta$ such that $x^\beta y^k \in I$. Then $J_k$ is a “slice” of $I$ generated by monomials containing $y$ to the $k$th power. Again by
the inductive hypothesis, \( J_k \) is generated by finitely many terms in 
\( k[x_1, \ldots, x_{n-1}] \), which we denote \( x^{a_{1(k)}}, \ldots, x^{a_{b(k)}} \).

Summarizing the above, we have the following monomials from 
\( I \): associated with 
\( J \),
\[
x^{a_{1(k)}}, \ldots, x^{a_{b(k)}};
\]
and associated with each of the 
\( J_k \) in 
\( J_1, \ldots, J_{m-1} \),
\[
x^{a_{1(k)}}, \ldots, x^{a_{b(k)}}.
\]
A monomial \( x^a y^p \) from 
\( I \), by the constructions above, is divisible by 
one of the terms from 
\( J \) if \( p \geq m \), and by one of the terms from 
\( J_k \) if \( p = k < m \).

We now bridge the gap between monomial ideals and arbitrary ideals.

Define \( \langle \text{lt } I \rangle \) to be the set of leading terms of elements of 
\( I \), and 
\( \langle \text{lt } I \rangle \) to be the ideal generated by those terms. (We're implicitly assuming a choice of monomial ordering.) Then if 
\( I = \langle f_1, \ldots, f_m \rangle \), it is clear that in general, 
\( \langle \text{lt } f_1, \ldots, \text{lt } f_m \rangle \subseteq \langle \text{lt } I \rangle \). This inclusion can be strict: for example, let 
\( I \) be
\[
\langle x^3 - 2xy, x^2y - 2y^2 + x \rangle;
\]
then \( x^2 = -y \cdot f_1 + x \cdot f_2 \in I \) but, using graded lexicographic order, 
\( \text{lt } f_1 = x^3 \) and \( \text{lt } f_2 = x^2y \).

On the other hand, as a direct consequence of Dickson’s lemma we have that 
\( \langle \text{lt } I \rangle = \langle \text{lt } g_1, \ldots, \text{lt } g_m \rangle \) for some choice of \( g_1, \ldots, g_m \in I \).

A.4.3. A set of elements \( G = \{g_1, \ldots, g_m\} \) satisfying the above property is called a \textit{Gröbner basis} for \( I \). More formally, \( G \subseteq k[x_1, \ldots, x_n] \) is a Gröbner basis for an ideal \( I \) if

\[
\langle \text{lt } G \rangle = \langle \text{lt } I \rangle
\]
i.e. if for any \( f \in I \) there exists a \( g \in G \) such that \( \text{lt } g \) divides \( \text{lt } f \).

As the terminology suggests, it turns out (rather surprisingly) that any Gröbner basis \( G \) of \( I \) satisfies \( \langle G \rangle = I \). Notice that this fact plus Dickson’s lemma slightly strengthen the Hilbert basis theorem: every nonzero ideal has a finite Gröbner basis.

To see why, first note that the inclusion \( \langle g_1, \ldots, g_m \rangle \subseteq I \) is easy, since the \( g_j \) are elements of \( I \). Conversely, for any \( f \in I \) we can apply the division algorithm to yield
\[
f = a_1g_1 + \ldots + a_m g_m + r
\]
where none of the lt $g_j$ divide $r$. If $r = 0$ then $f \in I$. To see that this is true, assume otherwise and write
\[
r = f - a_1 g_1 - \ldots - a_m g_m.
\]
Then lt $r \in \langle \text{lt } I \rangle = \langle \text{lt } g_1, \ldots, \text{lt } g_m \rangle$, and one of the lt $g_j$ must divide $r$, a contradiction. ▷

We now state without proof some equivalent characterizations of a Gröbner basis $G = \{g_1, \ldots, g_m\}$ of an ideal $I \subseteq k[x_1, \ldots, x_n]$:

A.4.4. the leading term of any polynomial in $I$ is divisible by one of the $g_j$;

• multivariate division of any polynomial in $k[x_1, \ldots, x_n]$ by the $g_j$ gives a unique remainder;

• multivariate division of any polynomial in $I$ by the $g_j$ gives remainder 0.

To give an easy example, \{1\} is a Gröbner basis for the whole of $k[x_1, \ldots, x_n]$, while \{(x_1, x_1 - 1)\} is not.

As defined above, and as the example shows, there is no minimality requirement on being a Gröbner basis, and in fact any finite subset of an ideal which contains a Gröbner basis is a Gröbner basis itself. A reduced Gröbner basis $G$ additionally satisfies that for all $p \in G$, $\text{lc } p = 1$ and no monomial of $p$ lies in $\langle \text{lt } G - \{p\} \rangle$. The second condition is equivalent to the set lt $G$ minimally generating $\langle \text{lt } I \rangle$ and no trailing term of any $g \in G$ lying in $\langle \text{lt } I \rangle$. Reduced Gröbner bases, it turns out, are unique, up to a choice of monomial ordering.

The following theory will establish an algorithmic way of determining a Gröbner basis for the ideal generated by an arbitrary finite set of polynomials $\langle f_1, \ldots, f_m \rangle$. Notice that what prevents this set from being a Gröbner basis is that some $k[x_1, \ldots, x_n]$-linear combination of the $f_j$ has a leading term which is not in the ideal $\langle \text{lt } f_1, \ldots, \text{lt } f_m \rangle$.

This behavior is captured by $S$-polynomials (short for syzygy or subtraction polynomials). First, we define the least common multiple of monomials $x^\alpha$ and $x^\beta$ to be the monomial $\text{lcm } (x^\alpha, x^\beta) := x^\gamma$, where $\gamma = (\max(\alpha_1, \beta_1), \ldots, \max(\alpha_n, \beta_n))$. Then the $S$-polynomial of $f, g \in k[x_1, \ldots, x_n]$ is
\[
S(f, g) := \frac{x^\gamma}{\text{lt } f} \cdot f - \frac{x^\gamma}{\text{lt } g} \cdot g
\]
where $x^\gamma = \text{lcm } (\text{lm } f, \text{lm } g)$ is the least common multiple of the leading terms of $f$ and $g$. Notice that, by design, the $S$-polynomial cancels the leading terms of $f$ and $g$.

If we have an ideal $I = \langle f_1, \ldots, f_m \rangle$, then for any $f_i, f_j$, the polynomial $S(f_i, f_j)$ must lie in $I$. Then if the $f_1, \ldots, f_m$ form a Gröbner
basis, the remainder of $S(f_i, f_j)$ on division by these polynomials will be zero. It turns out, though the proof is quite lengthy, that the converse also holds, namely that if for all $i \neq j$ the $S$-polynomial $S(f_i, f_j)$ has remainder zero on division by the $f_1, \ldots, f_m$, then these polynomials form a Gröbner basis.

The original algorithm given for computing Gröbner bases was Buchberger’s algorithm, given in pseudocode below:

- input {$f_1, \ldots, f_s$};
- set $G \leftarrow F$;
- while $G' \neq G$:
  - set $G' \leftarrow G$
  - for each $(p, q), p \neq q$ in $G'$:
    * set $S \leftarrow$ the remainder of $S(p, q)$ on division by $G'$;
    * if $S \neq 0$ then set $G \leftarrow G \cup \{S\}$

Notice that the algorithm enlarges the generating set it is given. To show that it eventually terminates, we first observe that on each pass through the main loop, $\langle \text{lt } G' \rangle \subseteq \langle \text{lt } G \rangle$ and the inclusion is strict if $G' \neq G$; this follows from the fact that $S$ is a remainder on division by the terms of $G'$, and hence $\text{lt } S$ is not divisible by any of the elements of $S$. Buchberger’s algorithm has been improved in various ways. For example, it is known to be more efficient in most cases to compute a Gröbner basis with respect to grevlex order, then convert this to one with respect to lex order, than to compute a lex-respecting Gröbner basis directly.

Here is an example of such an algorithm in action. The ideal associated with the system of equations

\[
\begin{align*}
  x^2 + y + z &= 1 \\
  x + y^2 + z &= 1 \\
  x + y + z^2 &= 1
\end{align*}
\]

has the following Gröbner basis with respect to lex order:

\[
\{x + y + z^2 - 1, y^2 - y - z^2 + z, 2yz^2 + z^4 - z^2, z^6 - 4z^4 + 4z^3 - z^2\}
\]

Generating a Gröbner basis depends on a choice of monomial ordering. One consequence of this that is particularly relevant to what follows is that the Gröbner basis associated with a generating set of polynomials may lose much of the symmetry of the original set of polynomials.
A.5 ELIMINATION THEORY

Gröbner bases have many applications in commutative algebra and algebraic geometry, but what most concerns us here is their application to elimination theory. Elimination theory, as one might expect, concerns eliminating variables from sets of polynomial equations in a systematic way. For example, consider the equations given above, and let \( k = \mathbb{C} \). The last polynomial in the Gröbner basis is entirely in terms of \( z \). One can find its roots and substitute the associated \( z \) values into the second and third equations to find values of \( y \), and then substitute these into the first equation to find values of \( x \). In this case, we get back a complete description of the associated variety, which consists of the five real points.

Generalizing this example and passing from equations to the associated ideals leads to the following definition.

**A.5.1.** Given \( I \subseteq k[x_1, \ldots, x_n] \), the \( k \)-th elimination ideal, denoted \( I_k \), is the ideal

\[
I \cap k[x_{k+1}, \ldots, x_n].
\]

It consists of all polynomial consequences of \( f_1 = \ldots = f_m = 0 \) eliminating the variables \( x_1, \ldots, x_k \).

Elimination exactly corresponds to Gaussian elimination in the degree 1 case, with lex order.

The significance of elimination more generally, though, comes from the following two theorems:

**A.5.2.** ([Cox et al., 1997] [3.1.2]) Let \( I \subseteq k[x_1, \ldots, x_n] \) be an ideal, and \( G \) a Gröbner basis with respect to lex order where \( x_1 > \ldots > x_n \). Then the set

\[
G_l = G \cap k[x_{l+1}, \ldots, x_n]
\]

is a Gröbner basis for the \( l \)-th elimination ideal \( I_l \).

**A.5.3.** If \( k \) is algebraically closed, then \( V(I_l) = \pi_l(V(I)) \), where \( \pi_l \) denotes projection onto the last \( n - l \) coordinates of \( k^n \).
BIBLIOGRAPHY


Judea Pearl and Azaria Paz. *Graphoids: A Graph-Based Logic for Reasoning: About Relevance Relations*. Computer Science Department, University of California, 1985.


Bernd Sturmfels. What is a gröbner basis. 52(10):11991200, 2005.


