METABOLIC NETWORK INFEERENCE WITH THE GRAPHICAL LASSO

by

JOSEPH KRITTAMETH AICHER
SONG SONG, COMMITTEE CHAIR
LAURA K. REED
YUHUI CHEN
KABE MOEN

A THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Arts
in the Department of Mathematics
in the Graduate School of
The University of Alabama

TUSCALOOSA, ALABAMA

2015
ABSTRACT

Metabolic networks describe the interactions and reactions between different metabolites (e.g. sugars, fatty acids, amino acids) in a biological system, which together give rise to the chemical processes which make life possible. Efforts to further knowledge of the structure of metabolic networks have taken place for well over a century through the efforts of numerous biochemists and have revolutionized our understanding of biology and the capabilities of modern medicine. The introduction in the recent past of metabolomics technologies, which allow for the simultaneous measurement of the concentrations of a significant number of metabolites, has led to the development of mathematical and statistical algorithms that aim to use the information and data that these technologies have made available to make inferences about the structure of metabolic networks.

In this thesis, I investigate the application of the graphical lasso algorithm to metabolomics data for the purposes of metabolic network inference. I use the graphical lasso on a metabolomics dataset collected by gas chromatography-mass spectrometry from *Drosophila melanogaster* to estimate graphical models of varying levels of sparsity that describe the conditional dependence structure of the observed metabolite concentrations. With these estimated models, I describe how they can be chosen from and interpreted in the context of both the data and the underlying biology to inform our knowledge of metabolic network structure.
LIST OF ABBREVIATIONS AND SYMBOLS

$E$ The edges of a graph.

$G$ A graph or network of vertices connected by edges.

GC/MS gas chromatography-mass spectrometry.

GGM Gaussian graphical model.

$\lambda$ The tuning parameter of the graphical lasso.

$\mu$ The mean vector of a multivariate normal distribution.

$n$ The number of observations or measurements.

$p$ The number of variables.

$S$ The sample covariance matrix.

$\Sigma$ The covariance matrix of a multivariate normal distribution.

$\Theta$ The precision matrix (inverse covariance) of a multivariate normal distribution.

$V$ The vertices of a graph.
ACKNOWLEDGMENTS

I would like to thank my advisors at the University of Alabama, Dr. Laura K. Reed and Dr. Song Song, for their willingness to start and help carry out this cross-disciplinary research project. I would also like to thank my brother, Christopher Aicher, and my fiancée, Bernadette K. Bucher, for their support and their helpful conversations about applied mathematics and statistics.

I would also like to acknowledge the scholarships that were generously awarded to me by The University of Alabama and the National Merit Scholarship Corporation. I would also like to acknowledge the National Institute of General Medical Sciences of the National Institutes of Health under award numbers R01GM61600 to Greg Gibson, R01GM098856 to Laura K. Reed, and a Ruth L. Kirschstein NIH-NRSA Postdoctoral Fellowship to Laura K. Reed. Their generous financial support made my education and this research possible.
LIST OF FIGURES

2.1 Lasso constraint is “pointy” and sets coefficients to zero ........................................ 13

4.1 Inferred graphs $\mathcal{G}_k$ from the dataset for different values of $\lambda_k$ show differing
levels of sparsity .................................................................................................................. 19

4.2 Number of edges connected to non-biological compounds over the different
inferred graphs $\mathcal{G}_k$ rules out values of $k$ greater than 27 ........................................ 21

4.3 The proportion of edges that are between metabolites of the same chemical
class for each inferred graph $\mathcal{G}_k$ helps inform its relevance to the underlying
metabolic network .............................................................................................................. 22

4.4 The connected components of $\mathcal{G}_{11}$ with multiple identified metabolites dem-
strate known metabolic relationships .............................................................................. 24

4.5 The four largest components of $\mathcal{G}_{20}$ with thicker edges representing edges
present in earlier graphs are enriched for specific chemical classes .......................... 25
CHAPTER 1
INTRODUCTION

The study of metabolic networks, which dictate the pattern by which metabolism chemically transforms metabolites, including sugars, fats, and proteins, into other metabolites as necessary in biological systems, is interesting from both a biological and mathematical perspective. The structure of metabolic networks can help describe the mechanisms of complex biological processes, including those leading to disease [7]. Thus, understanding metabolic networks can inform fundamental biology and biomedical science. Meanwhile, metabolic networks can be thought of as an implicit mathematical model for the metabolic processes of a system [1]. The development of mathematical methodology to take experimental data from a biological system to infer and propose a model structure for the underlying metabolic network is an active area of applied research.

Metabolism is thought to be well-characterized in several species, especially microbes, and there exist several manually-curated databases attempting to describe the “usual” metabolic networks for these species [8, 16]. However, this is not the case for many other species, including fruit flies (Drosophila melanogaster) [8]. Fruit flies have been successfully used as a model organism to study numerous diseases, including metabolic syndrome [25, 26]. Developing more understanding of the metabolic network structure of Drosophila melanogaster will further inform researchers of the unique aspects of their metabolism, augmenting future studies and allowing the results from these studies to be better translated to humans or other species.

Developing this kind of further knowledge of the metabolic network of a biological system requires taking experimental measurements from the system and inferring from them part or all of the network structure. Efforts in the past century have largely involved sophisti-
cated and specific biochemical techniques to investigate a small section, or pathway, of the metabolic network at a time [17, 21]. These methods have revolutionized our understanding of metabolism and directly provide proof of their inferred pathways. However, they do not consider the whole network at one time and can be expensive and time-consuming to undertake.

Technological advances have recently made available a new approach to metabolic network inference. The advent of metabolomics technologies, such as gas chromatography-mass spectrometry (GC/MS), has introduced the ability to quickly identify a significant number of metabolites in a biological sample and their relative quantities [31]. This has allowed for the development and application of mathematical algorithms that use these data to determine a network or set of networks that best correspond to the data. Numerous approaches for looking at this problem have been suggested and are reviewed in detail in [6]. One promising approach is to apply (probabilistic) graphical models, particularly Gaussian graphical models (GGMs), to data, relating the inferred mathematical structure to the underlying metabolic networks.

Graphical models relate the conditional dependences of a number of random variables following some probability distribution to a graph, or network, structure between variables. GGMs are graphical models, which, by assuming that the variables are drawn from a normal distribution, allow for a strong interpretation of the relationship between variables from the network structure. Furthermore, the network of a GGM is directly related to the parameters of the normal distribution, providing concrete ways of obtaining the network by statistical inference. Classical approaches for constructing GGMs from data are derived from maximum-likelihood inference but can only be used when the number of samples is greater than the number of variables. Newer methods have been created that overcome this limitation. In particular, the graphical lasso is a method which generalizes the classical approaches by maximizing a penalized likelihood function and allows for the inference of GGMs even when the number of variables far exceeds the number of samples [4,9].
GGMs have seen recent interest in the metabolomics community the past few years and have been applied to metabolomics studies, both in silico and in vivo [5, 18, 22]. These past studies inferred GGMs from metabolomics data using classical methods to predict the structure of the underlying metabolic networks. While these past studies demonstrate the utility of GGMs for metabolic network inference, their approach does not extend to the case where the number of metabolites being measured per sample exceeds the number of samples, a situation that is becoming increasingly important with the rapid advancement of metabolomics and other high-throughput technologies.

In this thesis, I investigate the inference of the metabolic network of *Drosophila melanogaster* using a GC/MS dataset. I do so in an approach similar to those taken in [5, 18, 22] by inferring GGMs from the data. I infer these GGMs by using the graphical lasso instead of the classical approaches taken by previous work, extending their methodology to being applicable when the number of metabolites measured per sample exceeds the number of samples in a metabolomic dataset.

The rest of the thesis is structured as follows. First, I review background information, starting with the basics of metabolic networks and metabolomics technologies, continuing with an overview of graphical models, focusing especially on GGMs, and ending with a discussion of the inference of GGMs and how the graphical lasso enables inference even when the number of variables far exceeds the number of samples. In the following chapter, I describe the methods used to collect, process, and analyze my data. Afterwards, I describe and discuss the results from my analysis. Finally, I end the thesis with a brief conclusion considering future possible directions of investigation.
CHAPTER 2
BACKGROUND

Before I discuss how the metabolomic data were obtained from fruit flies and processed in order to infer their metabolic network, I review relevant background information that informed my methods. First, I briefly describe the terminology of metabolic networks and the process by which metabolites are identified and quantified by metabolomics technology. Second, I provide an overview of the theory of graphical models, especially as they relate to Gaussian graphical models (GGMs). Third, I describe the algorithm that I use to infer GGMs, the graphical lasso, and its properties.

2.1 Metabolic Networks and Metabolomics

I review metabolic networks and metabolomics because understanding what is being measured and how it is measured informs the assumptions I make in my analysis.

Metabolic networks are composed of metabolites, which are the chemical intermediates of metabolism. Metabolism is the process by which a biological system constructs or breaks down chemical structures as necessary for life by a sequence of chemical reactions. Metabolites are the intermediates of these reactions and are transformed or combined into new metabolites by these reactions. Examples of metabolites include glucose and fructose, which are sugars, or phenylalanine and other amino acids, which are components of proteins. Metabolic networks describe the relationships between the metabolites and the chemical reactions of metabolism involving them. The model for metabolic networks that I consider in this thesis is the compound graph, which views metabolites as nodes of a network, with connections between any pair of these nodes if there exists a reaction between them [19]. In the language of graph theory, this model views a metabolic network as a graph for which
the vertices are metabolites and the edges are reactions. This observation provides a direct connection to my later discussion of graphical models.

Metabolomics technologies enable the identification and quantification of a significant number of metabolites present in a sample. Not only can these kinds of measurements enable us to consider metabolic network inference [5,18], but they have also been used for functional genomics studies [23] and for biomarker discovery in clinical applications to cancer [28] and cardiovascular disease [27]. Numerous analytical methods have been developed to perform these measurements, typically involving either mass spectrometry or nuclear magnetic resonance [13]. I describe the process of measurement for gas chromatography-mass spectrometry (GC/MS) because my data is obtained from the method.

GC/MS identifies and quantifies the metabolites present in a sample by coupling together the techniques of gas chromatography and mass spectrometry [15]. Gas chromatography takes a mixture of different compounds and separates them according to physical/chemical differences, measuring for each separated compound the time it spends in the system, called the retention time. Meanwhile, mass spectrometry takes a chemical input and breaks it into fragments whose mass-charge ratios can be measured simultaneously to create a spectrum detailing the intensities of the different fragments that is often able to identify the compound. Combining the two methods allows the metabolites in a biological sample to be separated by gas chromatography followed by their individual analysis by mass spectrometry, yielding an intensity spectrum as a function of retention time and mass-charge ratio. Using the heights/areas of the peaks in the obtained spectrum, I can identify the compounds present in the sample and their relative concentrations [14].
2.2 Graphical Models

Graphical models associate a graph with the conditional dependence structure of random variables. Mathematically, if we say that the random variables

$$X = (X_1, X_2, \cdots, X_p)$$

have some joint distribution $P$, we associate with this distribution some graph $G = (V, E)$, where $V = \{1, 2, \cdots, p\}$ is the set of vertices of the graph, which correspond to the $p$ random variables, and $E \subseteq V \times V$ is the set of edges of the graph. A possible edge between the $i$ and $j$-th vertices ($(i, j) \in V \times V$) either exists ($(i, j) \in E$) or does not exist ($(i, j) \notin E$) if the random variables $X_i$ and $X_j$ satisfy some property with respect to their distribution, $P$. We call the pair $(G, P)$ a graphical model.

Graphical models can be subdivided into directed graphical models or undirected graphical models [30]. The distinction corresponds exactly with the difference between directed and undirected graphs: an edge $(i, j) \in E$ for a directed graph is thought of as an edge from $i$ to $j$ and does not imply the existence of an edge $(j, i)$, while edges in undirected graphs are thought of as between vertices and the edge $(i, j)$ is in $E$ if and only if the edge $(j, i)$ is in $E$. For our problem of metabolic network inference, we will focus on undirected graphical models. Directed graphical models typically model the joint distribution of $X$ as being able to be factored into a product of conditional probabilities as specified by the edges and have successfully been applied to problems in bioinformatics in the past [10]. However, this factorization property requires the networks associated with directed graphical models to be acyclic, which is an undesirable property for considering metabolic networks, which are known to contain loops [2, 20]. Undirected graphical models are capable of modeling these loops.

A conditional independence graph is a type of undirected graphical model that satisfies one of the Markov properties, which describe a relationship between the set of edges of a
graphical model and the conditional independence structure of the variables. The Markov
properties are defined as follows:

**Definition 2.1** (Markov Properties [4]). Consider a graphical model \((G, P)\).

\(P\) satisfies the **pairwise Markov property** with respect to \(G\) if for all edges \((j, k)\) that
are not in \(E\) with \(j \neq k\), \(X_j\) and \(X_k\) are conditionally independent given the values of the
other variables \(\{X_i : i \in V \setminus \{j, k\}\}\).

\(P\) satisfies the **global Markov property** with respect to \(G\) if, given any disjoint sets
of vertices \(A, B, C \subseteq V\) such that \(C\) separates \(A\) from \(B\), the sets of random variables
\(X^{(A)} = \{X_j : j \in A\}\) and \(X^{(B)} = \{X_k : k \in B\}\) are conditionally independent given the set
of variables \(X^{(C)} = \{X_i : i \in C\}\).

It can be easily shown that the global Markov property implies the pairwise Markov
property in all cases. Under certain conditions, including when \(P\) follows a normal distri-
bution, the reverse is true [4]. Knowing the Markov properties makes defining a conditional
independence graph easy:

**Definition 2.2** (Conditional Independence Graph). A **conditional independence graph**
is a graphical model \((G, P)\) where the pairwise Markov property holds.

A Gaussian graphical model (GGM) is a conditional independence graph \((G, P)\) for which
the distribution \(P\) follows a multivariate normal distribution. GGMs have particular utility
in the fact that the edges of the conditional independence graph are directly related to the
parameters of the normal distribution. Furthermore, the edges of the graph follow a stricter
condition than the Markov properties. Recall that a \(p\)-dimensional normal distribution is
entirely determined by its \(p\)-dimensional mean vector, \(\mu\), and its positive definite \(p \times p\)
covariance matrix, \(\Sigma\). A random variable \(X\) with such a distribution satisfies:

\[
\mu_j = \mathbb{E}[X_j],
\]

\[
\Sigma_{jk} = \text{Cov}(X_j, X_k) = \mathbb{E}[(X_j - \mu_j)(X_k - \mu_k)],
\]
with joint probability density:

\[
f(x|\mu, \Sigma) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp \left( -\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu) \right).
\]

The following result holds for GGMs:

**Theorem 2.1** (Edges of a GGM [4]). For any \((j, k) \in \{1, 2, \cdots, p\} \times \{1, 2, \cdots, p\}\), the following statements are equivalent:

1. \((j, k)\) and \((k, j)\) are not in \(E\).
2. \(X_j\) and \(X_k\) are conditionally independent given the other \(p - 2\) variables.
3. \(\Sigma^{-1}_{jk} \neq 0\).

Theorem 2.1 is a stronger condition on the edges of a GGM than specified by the Markov properties. First, note that statement 1 \(\Rightarrow\) statement 2 is just the local Markov property, which implies the global Markov property since \(P\) follows a normal distribution. It is the fact that statement 2 \(\Rightarrow\) statement 1, as well, that provides an even stronger interpretation of the edges of a GGM – not only does the absence of an edge imply a kind of conditional independence as described by the Markov properties, but the presence of an edge implies a kind of conditional dependence. This interpretation makes GGMs particularly attractive for modeling metabolic networks. We want edges between metabolites if there exists a chemical reaction between them, while the steady state concentrations of a metabolite in a biological system should be conditionally dependent only on the concentrations of metabolites involved in reactions with that metabolite. Thus, we can argue that the network structure for a GGM for the concentrations of metabolites, which can be measured using metabolomics technologies, should closely relate to the underlying metabolic network, as supported by the evidence presented in this thesis and the work of other authors [5, 18].

Theorem 2.1, through the equivalence of statements 3 and 1, also provides a concrete way of determining the network structure of a GGM by using the inverse of the covariance matrix,
which we call the precision matrix and denote by $\Theta \equiv \Sigma^{-1}$. An edge between two variables in a GGM exists if and only if the corresponding element of the precision matrix is nonzero. Thus, determining the structure of a GGM reduces to identifying the nonzero elements of the corresponding precision matrix. We describe in the next section some methods for inferring from observed data these nonzero elements, thus allowing us to also infer the full network structure.

### 2.3 Inference of GGMs and the Graphical Lasso

The inference of GGMs from data is equivalent by Theorem 2.1 to the inference of the nonzero elements of $\Theta$. In this section, we first describe the classical method for inferring $\Theta$. Then, we explain some of the limitations of this method, especially for metabolic network inference, and how these limitations can lead to the method being unusable altogether. We then conclude the section with alternatives for inferring the nonzero elements of the precision matrix that, given some assumptions that are reasonable for metabolic networks, overcome the classical method’s limitations by applying regularization, focusing on a method called the graphical lasso.

We set the stage for describing methods for inference by setting up notation and our data model. Suppose we have $n$ observations of $p$ real-valued variables. We denote these data using an $n \times p$ matrix which we call $X$. Our $i$-th observation (for $i \in \{1, \cdots, n\}$) is represented by the $i$-th row of $X$, which we denote by $x_i$. We assume that each $x_i$ is independent and identically distributed, according to a normal distribution. Without loss of generality, we can assume that they have zero mean, so that $x_i \sim \mathcal{N}(0, \Sigma = \Theta^{-1})$ for each $i$.

The classical approach for inferring graphical models is done by determining the maximum likelihood estimator of $\Theta$, $\hat{\Theta}_{\text{MLE}}$, and choosing some threshold for which the elements of $\hat{\Theta}_{\text{MLE}}$ are considered zero [12]. The maximum likelihood estimator of $\Theta$ is derived as follows. The likelihood function of the precision matrix, $\Theta$, given the observations $\{x_i : i \in 1, \cdots, n\}$,
is the product of the underlying densities evaluated at points of the observations, that is:

\[
\mathcal{L}(\Theta|x_1, \cdots, x_n) \equiv \prod_{i=1}^{n} f(x_i|0, \Theta^{-1})
\]

\[
= \prod_{i=1}^{n} |\Theta|^{1/2} \left(\frac{1}{2\pi}p\right)^{p/2} \exp \left( \frac{-1}{2} x_i^{T} \Theta x_i \right)
\]

\[
= |\Theta|^{1/2} \left(\frac{1}{2\pi}p\right)^{p/2} \exp \left( \frac{-1}{2} \sum_{i=1}^{n} x_i^{T} \Theta x_i \right).
\]

Maximizing the likelihood is equivalent to maximizing the logarithm of the likelihood (called the log-likelihood), which equals:

\[
\ell(\Theta|x_1, \cdots, x_n) \equiv \log \mathcal{L}(\Theta|x_1, \cdots, x_n)
\]

\[
= -\frac{np}{2} \log (2\pi) + \frac{n}{2} \log \det \Theta - \frac{1}{2} \sum_{i=1}^{n} x_i^{T} \Theta x_i
\]

\[
= -\frac{np}{2} \log (2\pi) + \frac{n}{2} \log \det \Theta - \frac{1}{2} \text{tr} (X^{T}X \Theta)
\]

\[
= -\frac{np}{2} \log (2\pi) + \frac{n}{2} (\log \det \Theta - \text{tr} (S \Theta)),
\]

where the quantity \(S\) is the estimator for \(\Sigma\) called the sample covariance matrix and defined as \(S \equiv \frac{1}{n} X^{T}X\).

The maximum-likelihood estimator of the precision matrix, then, is defined by the maximization of the log-likelihood with respect to positive definite matrices \(\Theta\):

\[
\hat{\Theta}_{\text{MLE}} \equiv \arg \max_{\Theta > 0} \{\ell(\Theta|x_1, \cdots, x_n)\}
\]

\[
= \arg \max_{\Theta > 0} \left\{-\frac{np}{2} \log (2\pi) + \frac{n}{2} (\log \det \Theta - \text{tr} (S \Theta))\right\}
\]

\[
= \arg \max_{\Theta > 0} \{\log \det \Theta - \text{tr} (S \Theta)\}.
\]

To maximize this, we take the derivative with respect to \(\Theta\) and set it to zero, which implies
that

\[ \Theta^{-1} = S. \]

Thus, the maximum-likelihood estimator of the precision matrix is:

\[ \hat{\Theta}_{\text{MLE}} = S^{-1}. \]

Note that this solution for the maximum-likelihood estimator of the precision matrix only makes sense when the sample covariance matrix is invertible. This requires the sample covariance matrix to be full rank \( \text{rank}(S) = p \). However, by the definition of the sample covariance as proportional to \( X^T X \), it follows that

\[ \text{rank}(S) \leq \min(n, p). \]

If we are able to collect a lot of observations so that \( p \leq n \), this is not necessarily a problem. However, for applications to biology, the number of variables we observe, \( p \), is very large while each observation is expensive to make, so that the we have the high-dimensional case \( p \gg n \) [3]. Metabolomics studies are no exception. If \( p > n \), then the maximum-likelihood estimator of the precision matrix is ill-defined. Thus, for more widely applicable use in biological studies, we need a different method to infer GGMs that is applicable even when \( p > n \).

One approach to make estimating GGMs tractable for the case where \( p > n \) to assume that the network structure is sparse. That is, most variables are conditionally independent, and the elements of the precision matrix are mostly zero. Sparsity is not an unreasonable assumption; we do not expect each metabolite to be involved in reactions with all the others. We must apply this assumption in some way to how we estimate the nonzero entries of the precision matrix.

The approach we take is to modify the maximum-likelihood approach by subtracting what
is known as a lasso penalty from the log-likelihood function, which enables the maximization procedure to have a unique and sparse solution even in the high-dimensional case where \( p \gg n \). This penalty takes the form

\[
\lambda \|\Theta\|_1 = \lambda \sum_{j=1}^{p} \sum_{k=1}^{p} |\Theta_{jk}|,
\]

where \( \lambda > 0 \) is a tuning parameter that relates to how sparse we think the true precision matrix is, and \( \|\Theta\|_1 \) is the vector \( L^1 \) norm of \( \Theta \). This yields the penalized maximum likelihood estimate for a given value of the tuning parameter \( \lambda \):

\[
\hat{\Theta}_{\text{PMLE}} \equiv \arg \max_{\Theta \succ 0} \{ \log \det \Theta - \text{tr} (S\Theta) - \lambda \|\Theta\|_1 \}.
\]

This maximization problem has a finite and unique solution because it is the maximization of a concave function and is the Lagrangian dual form of a constrained maximization problem [33]. This equivalent constrained problem is the same as maximizing the unpenalized log-likelihood with the constraint that

\[
\|\Theta\|_1 \leq t
\]

for some \( t > 0 \). This constraint helps illustrate why the lasso penalty helps shrink coefficients to zero, encouraging a sparse solution. If the value of \( \Theta \) that maximizes the likelihood lies outside this constraint, the maximum value of the likelihood that satisfies the constraint is attained at a point on the maximum-valued level set of the likelihood that intersects with the surface \( \|\Theta\|_1 = t \). But, the surface \( \|\Theta\|_1 = t \) is “pointy” where coefficients are set to zero, which leads to these maxima corresponding to solutions with coefficients set exactly to zero, as illustrated in Figure 2.1 for \( \Theta \) with two coefficients being considered [11].

One truly remarkable aspect of this penalized likelihood approach is that the underlying optimization problem is concave. Thus, we can apply efficient general-purpose convex optimization algorithms to solve for \( \hat{\Theta}_{\text{PMLE}} \). However, specific algorithms have been developed
Figure 2.1: Lasso constraint is “pointy” and sets coefficients to zero to take advantage of particular structure in the problem. Most notable of these algorithms is the graphical lasso, which, when published, claimed to be one to three orders of magnitude faster than the existing methods at the time [9]. Further work has identified additional improvements to the original graphical lasso algorithm [32], and its popularity has led to the entire above-described penalized maximum-likelihood approach for constructing a sparse estimator for the precision matrix to be called the graphical lasso. The algorithm solves the penalized problem by iterative block coordinate ascent one column at a time until convergence. The algorithm also solves for the penalized estimators of a set of different values for $\lambda$ quite efficiently. Starting from the largest values of $\lambda$ corresponding to the sparsest estimates of the precision matrix, the results of the coordinate ascent from one $\lambda$ are used to start the coordinate ascent for decreasing values of the regularization parameter (an increasingly denser precision matrices), allowing for work done in previous steps not having to be repeated. Thus, the graphical lasso can be used to infer multiple possible network structures for a GGM, each with different possible levels of sparsity, simultaneously.
As we can now see, the graphical lasso is a powerful algorithm for inferring sparse GGMs from data even in the high-dimensional case where $p \gg n$. The edges of these GGMs relate to a strong form of conditional independence which provide an attractive interpretation of the edges with respect to our ultimate goal of inferring the metabolic networks of *Drosophila melanogaster*.
I constructed predictions for the metabolic network of *Drosophila melanogaster* by inferring GGMs from GC/MS data using the graphical lasso. To do so, I considered GC/MS data collected for a previous experiment by Reed et. al. [24], cleaned and normalized the data, and applied the graphical lasso to infer GGMs from the data.

My data were obtained by GC/MS as part of a previous experiment by Reed et. al. As described in [24], data for 425 GC/MS samples were collected from 20 different genetic lines of *Drosophila melanogaster* being fed 4 different diets so that each of the 80 possible combinations of line and diet had measurements for at least 5 distinct biological samples. The raw GC/MS data for these samples were aligned to match metabolites present in the majority of samples and processed to obtain the relative log$_2$-transformed concentrations of 189 common chemicals for each sample. Of these chemicals, 187 of them were identified as putative metabolites, while the other 2 were identified as non-biological compounds. The exact identities of 60 metabolites and the chemical class (e.g., disaccharide or unsaturated fatty acid) of another 124 metabolites were determined. The exact identity of a metabolite was determined for 60 of the metabolites, and another 124 of the metabolites were identified as a member of a chemical class (e.g., disaccharides or unsaturated fatty acids). Thus, my original dataset was composed of $n = 425$ different samples of the relative log-transformed concentrations of $p = 189$ different chemicals, 187 of which were metabolites.

However, 7.7 percent of these measurements were recorded as missing data, while the graphical lasso can only be used on datasets with no missing values whatsoever. To make it possible to use the graphical lasso, I followed the following procedure to either remove or impute the missing data:
1. First, I assessed the frequency of missing data on a per-metabolite and per-sample basis and removed individual metabolites and samples for which too many measurements were missing. Specifically, if the measurement of a metabolite was missing for more than 40 percent of the samples, I removed the metabolite from further consideration in my analysis. Likewise, if a particular sample had more than 40 percent of its concentrations recorded as missing, I removed the sample from further consideration.

2. Second, I considered the structure of the dataset with respect to the samples coming from different combinations of genetic lines and diets and removed individual metabolites for which too many measurements were missing in one of the line/diet combinations. If, for a particular metabolite, there existed a line/diet combination such that the metabolite was missing for more than 60 percent of its samples, I removed the metabolite from further consideration.

3. Third, I imputed the remaining missing values by again considering the structure of the dataset. To do so, I first calculated the mean values for each metabolite over different line/diet subpopulations. Then, I replaced each missing value with the mean value for the corresponding metabolite and line/diet subpopulation.

Following these steps led to the final reduced dataset on which I performed my analysis. Step 1 removed 3 metabolites and 4 samples from the original dataset. Step 2 removed an additional 22 metabolites. These two steps for removing missing data thus led to a reduced dataset with \( n = 421 \) different samples of the concentrations of \( p = 164 \) different chemicals, including 162 metabolites. This reduced dataset was missing only 4.3 percent of its values. Step 3 imputed these remaining missing values and resulted in the final reduced dataset on which I performed my analysis.

Finally, I constructed estimators of the precision matrix \( \Theta \) corresponding to the final reduced dataset using the graphical lasso as implemented in the “huge” package [34] in the R statistical environment. To ensure that each of the metabolites were treated equally by the
lasso penalty, I normalized the values for each metabolite in my dataset by autoscaling, which amounts to centering and scaling the data to zero sample mean and unit sample standard deviation [29], while constructing these estimators. I solved for estimates of the precision matrix, $\hat{\Theta}_{\text{PMLE}}$, for 50 different values of $\lambda$ which were chosen by the “huge” package to correspond to varying levels of sparsity. These estimates of the precision matrix, along with their corresponding GGMs through Theorem 2.1 and their interpretation with respect to the underlying metabolic network of *Drosophila melanogaster*, form the basis of discussion in the next chapter.
I estimated precision matrices from my final *Drosophila melanogaster* metabolomics dataset with $n = 421$ different observations of the relative log$_2$-transformed concentrations of $p = 164$ chemicals (of which 162 were metabolites) with varying levels of sparsity using the graphical lasso with 50 different values of $\lambda$ set automatically by the “huge” package. To discuss these estimates, I say that these 50 tuning parameters took on the values of the ordered set:

$$\{ \lambda_k : k = 1, 2, \cdots, 50 \},$$

such that

$$\lambda_1 > \lambda_2 > \cdots > \lambda_{50}.$$  

The graphical lasso, using these different values of $\lambda_k$, estimated the precision matrices:

$$\left\{ \hat{\Theta}_k : k = 1, 2, \cdots, 50 \right\},$$

which correspond by Theorem 2.1 to estimates of the GGM structure underlying the observed data:

$$\left\{ \hat{G}_k = (V, \hat{E}_k) : k = 1, 2, \cdots, 50 \right\}.$$

Smaller values of $k$ correspond to sparser precision matrices $\Theta_k$ and sparser graphs $\hat{G}_k$, as visualized in Figure 4.1, which illustrates the inferred network structure for four different values of $k$.

The ultimate goal of my analysis is to infer the structure of the underlying metabolic network of *Drosophila melanogaster*. The graphs $\{ \hat{G}_k \}$ obtained from the graphical lasso do
Figure 4.1: Inferred graphs $\hat{G}_k$ from the dataset for different values of $\lambda_k$ show differing levels of sparsity

(a) $k = 5$, $\lambda_5 = 0.8257$

(b) $k = 20$, $\lambda_{20} = 0.4080$

(c) $k = 35$, $\lambda_{35} = 0.2016$

(d) $k = 50$, $\lambda_{50} = 0.0996$
not directly represent this metabolic network, but instead estimate the conditional dependence structure of simultaneous measurements of the 164 different chemicals by GC/MS by the procedure outlined in [24]. As I discussed in chapter 2 and as supported by [18], these inferred GGMs should relate to the underlying network. However, it is possible that the graph structures could also include conditional dependences due to non-biological relationships or noise resulting from the process of measurement, especially as the tuning parameter $\lambda$ is decreased.

For the purpose of relating these graphs to a metabolic network, I want graphs for which the non-biological contributions to the inferred dependence structure is minimal. While there is no obvious way of quantifying the degree to which a particular $\hat{G}_k$ describes the biological versus non-biological components of the conditional dependence structure, the 2 non-biological compounds in the data provide one test which can determine when the inferred graphs reflect conditional dependence structure unrelated to the metabolic network. These two compounds, which in the dataset are called Target 0121 and Target 0755, measure the presence of silicone contamination from the GC/MS apparatus and a standard spike used to help quantify the other metabolites. Since they do not come from *Drosophila melanogaster*, they should not be connected to any metabolic network, and any connections present in a graphical model between one of them and an actual metabolite reflect structure arising from a non-biological process.

Figure 4.2 shows the number of edges connected to the two non-biological compounds, Target 0121 and Target 0755, for the different inferred graphs $\hat{G}_k$. As discussed before, graphs for which either number of edges is nonzero reflect non-biological structure. Thus, since at least one edge is connected to Target 0755 from a metabolite for the graphs $\hat{G}_k$ with $k > 27$, I can safely omit these graphs from consideration in my goal of metabolic network inference.

Biochemical information can also be used to inform which inferred graphs should be considered in relating them to the metabolic network. The reactants and products of metabolic
Figure 4.2: Number of edges connected to non-biological compounds over the different inferred graphs $\hat{G}_{k}$ rules out values of $k$ greater than 27.
Figure 4.3: The proportion of edges that are between metabolites of the same chemical class for each inferred graph $\hat{G}_k$ helps inform its relevance to the underlying metabolic network. Reactions typically have similar chemical structures and are thus often from the same chemical class. So, in the underlying metabolic network, I expect there to be more edges between metabolites of the same or similar chemical classes than between metabolites of different classes. Therefore, I argue that inferred graphs with more edges between metabolites of the same or similar chemical classes can be thought of as reflecting a biologically more likely network. I quantify this by calculating, for each graph $\hat{G}_k$, the proportion of the edges present in the graph that are between metabolites of the same identified chemical class.

Figure 4.3 illustrates how this proportion of edges between metabolites of the same class changes over the different inferred graphs $\hat{G}_k$. This metric supports the previous argument that graphs corresponding to $k > 27$ do not adequately represent biological structure through an entirely different line of reasoning – the proportion of edges for these graphs is less than all the others except for $\hat{G}_2$, which has exactly two edges. The maximum value of these proportions corresponds to $\hat{G}_{11}$, and the graphs corresponding to $11 \leq k \leq 20$ all have
similarly valued proportions, suggesting that these graphs may be particularly relevant to looking at the underlying metabolic network.

I note that this metric, as calculated here, is particularly crude. It penalizes edges between different chemical classes equally without taking into account the fact that different pairs of chemical classes are more similar. For example, monosaccharides and disaccharides (particular forms of sugars) are far more chemically similar than monosaccharides and amino acids. For the purposes of this analysis, the metric simply guides my focus to the graphs:

$$\left\{ \mathcal{G}_k : k = 11, 12, \ldots, 20 \right\}.$$

It is possible to interpret these inferred graphs together towards our ultimate goal of inferring the underlying metabolic network of *Drosophila melanogaster*. Instead of picking one of them over all the others, I observe that:

$$\hat{E}_{11} \subset \hat{E}_{12} \subset \cdots \subset \hat{E}_{20},$$

that is, edges in the sparser graphs stay in the denser graphs. Thus, I view edges present in smaller values of $k$ as corresponding to carrying a higher significance or signal from the data. We interpret these graphs with respect to the underlying metabolic network, starting with $\hat{G}_{11}$.

The inferred graph $\hat{G}_{11}$ has two connected components with multiple identified metabolites which reflect known metabolic pathways, as shown in Figure 4.4. The component on the left in Figure 4.4 shows tetradecanoic acid connected to hexadecanoic acid connected to octodecanoic acid. These metabolites are saturated fatty acids with 14, 16, and 18 carbons. The KEGG Pathway database lists them as components of the fatty acid biosynthesis reference pathway connected by six intermediate reactions each [16]. Meanwhile, the component on the right shows the branched-chain amino acids valine, leucine, and isoleucine along with another unidentified compound. These branched-chain amino acids are chemically very sim-
Figure 4.4: The connected components of $\hat{G}_{11}$ with multiple identified metabolites demonstrate known metabolic relationships and are found very close together in the valine, leucine, and isoleucine biosynthesis and degradation reference pathways in KEGG.

As larger values of $k$ are considered, more such components form and grow in size. Figure 4.5 displays the four largest components of $\hat{G}_{20}$, scaling the thickness of the edges so that more significant edges between metabolites are emphasized. The left-most component includes the fatty acid metabolites as described in $\hat{G}_{11}$ in addition to newer metabolites. Meanwhile, the bottom-right component is composed primarily composed of amino acids, including the branched-chain amino acids from the earlier graphs. The other two components show similar enrichment for one particular chemical class, in particular, the middle component is primarily composed of disaccharides, while the top-right component is composed primarily of saturated or unsaturated fatty acids.

The inferred $\{\hat{G}_k\}$, particularly for $11 \leq k \leq 20$, relate directly to known biological relationships. Non-biological measurements in the dataset and known chemical information about the metabolites inform the analysis, suggesting which of the structures are most relevant to the underlying metabolic network. Despite the challenge of further interpreting the results given the lack of information about the identities of more than half of the metabolites, this analysis relating Drosophila melanogaster GC/MS data to the species’ underlying metabolic network demonstrates the potential of applying the graphical lasso to infer GGMs
Figure 4.5: The four largest components of $\hat{G}_{20}$ with thicker edges representing edges present in earlier graphs are enriched for specific chemical classes from metabolomics data for the purposes of metabolic network inference.
CHAPTER 5
CONCLUSION

In this thesis, I applied the graphical lasso to *Drosophila melanogaster* metabolomics data to infer Gaussian graphical models (GGMs) describing the conditional dependence structure of the measurements of the relative concentrations of different metabolites. I then described how doing so could be used to address the biologically-oriented problem of inferring information about the underlying metabolic network.

Previous studies have investigated the application of GGMs to metabolomics data and their relevance to metabolic network inference [5,18,22]. These studies estimate GGMs from data by computing the sample covariance matrix $S$, inverting it to obtain the maximum-likelihood estimator of the precision matrix, and setting a threshold for which entries of the estimate whose absolute value exceed the threshold correspond to edges. This procedure requires that the number of samples exceeds the number of metabolites being measured per sample, which is restrictive given the high-throughput nature of metabolomics technology. By instead applying the graphical lasso, which does not suffer this limitation, to estimate GGMs from metabolomics data, my thesis illustrates how the results of these previous papers could be extended to the high-dimensional case.

The dataset I considered in this thesis does not actually make the maximum-likelihood approach to estimating network structure impossible. After cleaning the data, the number of samples, $n = 421$, exceeds the number of different chemicals per sample, $p = 164$. My analysis demonstrates approaches for using and interpreting the graphical lasso for metabolic network inference with this dataset and provides support for the hypothesis that the network structures obtained by such analyses are biologically meaningful and relevant to metabolic network inference. However, the driving reason for extending the existing methodology for
inferring GGMs is so that these analyses can still be made when $p > n$.

One possibility for extending the work in this thesis would be to repeat the analysis with a dataset where $p > n$. One particularly interesting possible source of such future datasets is the dataset that I used for this analysis, looking specifically at the experimentally imposed subpopulations in the data. The subpopulations of the samples that come from one of the 20 different genetic lines or that have been fed one of the 4 different diets are all smaller than the number of metabolites. Thus, such a future analysis would necessarily require a newer approach, such as the graphical lasso, while also revealing how these experimental interventions or genetic differences impact the underlying metabolic network.

Another possibility for future research is to reevaluate the assumption that the different observations are normally distributed. While this assumption is theoretically useful for interpreting the model and the inferred graphs’ relationship to the underlying metabolic networks, it is not quite realistic. Future work could investigate nonparametric or semiparametric methods for estimating graphical models and compare with the graphical lasso and the underlying metabolic networks.

Ultimately, the graphical lasso extends previous methodology in the metabolomics literature by enabling the inference of GGMs even when $p > n$. My thesis shows how this algorithm can be applied to a real metabolomics dataset and interpreted to develop understanding about the underlying metabolic network being measured. In doing so, my work can help inform future studies of metabolic networks, especially as metabolomics technologies continue to improve and allow us to simultaneously measure an even greater number of metabolites.
REFERENCES


