DEPARTMENT OF STATISTICS
University of Wisconsin
1300 University Ave.
Madison, WI 53706

TECHNICAL REPORT NO. 1142
July 1, 2008

Estimating Tree-Structured Covariance Matrices via Mixed-Integer Programming with an Application to Phylogenetic Analysis of Gene Expression

Héctor Corrada Bravo[1]
Department of Computer Sciences, University of Wisconsin, Madison WI

Kevin H. Eng[2]
Department of Statistics and Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison WI

Sündüz Keleş[3]
Department of Statistics and Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison WI

Grace Wahba[1]
Department of Statistics, Department of Computer Science and Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison WI

Stephen Wright[4]
Department of Computer Sciences, University of Wisconsin, Madison WI

[1] Research supported in part by NIH Grant EY09946, NSF Grant DMS-0604572 and ONR Grant N0014-06-0095

[2] Research supported in part by NSF Grant DMS-0604572

[3] Research supported in part by PhRMA Foundation Research Starter Grant in Informatics and NIH RO1 grant HG003747

[4] Research supported in part by NSF Grants DMS-0427689, CCF-0430504, CTS-0456694, CNS-0540147, and DOE Grant DE-FG02-04ER25627

Estimating Tree-Structured Covariance Matrices via Mixed-Integer Programming with an Application to Phylogenetic Analysis of Gene Expression

Héctor Corrada Bravo[*], Stephen Wright
Department of Computer Sciences
University of Wisconsin-Madison
Madison, WI 53706-1685, USA

Kevin H. Eng, Sündüz Keleş, Grace Wahba
Department of Statistics and Biostatistics and Medical Informatics
University of Wisconsin-Madison
Madison, WI 53706-1685, USA

Abstract

We present a novel method for estimating tree-structured covariance matrices directly from observed continuous data. A representation of these classes of matrices as linear combinations of rank-one matrices indicating object partitions is used to formulate estimation as instances of well-studied numerical optimization problems.

In particular, we present estimation based on projection where the covariance estimate is the nearest tree-structured covariance matrix to an observed sample covariance matrix. The problem is posed as a linear or quadratic mixed-integer program (MIP) where a setting of the integer variables in the MIP specifies a set of tree topologies of the structured covariance matrix. We solve these problems to optimality using efficient and robust existing MIP solvers. We also show that the least squares distance method of Fitch and Margoliash (1967) can be formulated as a quadratic MIP and thus solved exactly using existing, robust branch-and-bound MIP solvers.

Our motivation for this method is the discovery of phylogenetic structure directly from gene expression data. Recent studies have adapted traditional phylogenetic comparative analysis methods to expression data. Typically, these methods first estimate a phylogenetic tree from genomic sequence data and subsequently analyze expression data. A covariance matrix constructed from the sequence-derived tree is used to correct for the lack of independence in phylogenetically related taxa. However, recent results have shown that the hierarchical structure of sequence-derived tree estimates are highly sensitive to the genomic region chosen to build them. To circumvent this difficulty, we propose a stable method for deriving tree-structured covariance matrices directly from gene expression as an exploratory step that can guide investigators in their modelling choices for these types of comparative analysis.

We present a case study in phylogenetic analysis of expression in yeast gene families. Our method is able to corroborate the presence of phylogenetic structure in the response of expression in a subset of the gene families under particular experimental conditions. Additionally, when used in conjunction with transcription factor occupancy data, our methods show that alternative modelling choices should be considered when creating sequence-derived trees for this comparative analysis.

[*] Corresponding Author, hcorrada@cs.wisc.edu

1 Introduction

Recent studies have adapted existing techniques in population genetics to perform evolutionary analysis of gene expression (Fay and Wittkopp, 2007; Gu, 2004; Oakley et al., 2005; Rifkin et al., 2003; Whitehead and Crawford, 2006). In particular, corrections for evolutionary dependence between taxa, e.g. species or strains, are used in regression (generalized least squares) or other likelihood models. These phylogenetic corrections are well accepted methodologies in phenotypic modeling (Felsenstein et al., 2004), since, without them, statistical analysis is subject to increased false positive rates and decreased power for hypothesis tests. These corrections take the form of a covariance matrix corresponding to a random diffusion process along a phylogenetic tree.

Evolutionary studies of gene expression so far assume that the single phylogenetic tree structure underlying the data is known and typically derived from DNA or amino acid sequence data. While this assumption might be valid for the analysis of *coarse* traits–beak size in birds, for example–as in traditional comparative phylogenetic studies, it might prove too restrictive when carrying out similar analysis at the genomic level. Especially, taking into account recent findings of high variability in tree topology and branch length estimates contingent on the genomic region used to estimate the phylogeny (Frazer et al., 2004; Habib et al., 2007; Yalcin et al., 2004). If we are interested in a particular group of genes, given that they are spread throughout the genome, it makes more sense to develop a covariance estimate appropriate to those genes. We present a principled way of estimating tree-structured covariance matrices directly from sample covariances of observed gene expression data. As an exploratory step, this can help investigators circumvent issues that arise from estimating a global phylogeny from sequence data in an independent previous step.

In this paper, we formulate the problem of estimating a tree-structured covariance matrix as mixed-integer programs (MIP) (Bertsimas and Weismantel, 2005; Wolsey and Nemhauser, 1999). In particular, we look at projection problems that estimate the nearest matrix in the structured class to the observed sample covariance. These problems lead to linear or quadratic mixed integer programs for which algorithms for global solutions are well known and reliable production codes exist. The formulation of these problems hinges on a representation of a tree-structured covariance matrix as a linear expansion of outer products of indicator vectors that specify nested partitions of objects.

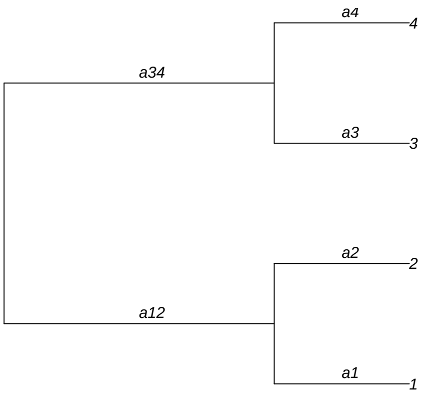
The paper is organized as follows. In Section 2.1 we formulate the representation of tree-structured covariance matrices and give some results regarding the space of such matrices. Section 2.5 shows how to define the constraints that ensure matrices are tree-structured as constraints in mixed-integer programs (MIPs). Projection problems are specifically addressed in Section 2.5.3. We present our results on a case study on phylogenetic analysis of expression in yeast gene families in Section 3. A discussion, including related work, follows in Section 4. Appendix A presents simulation results on estimating the tree topology from observed data that show how our MIP-based method compares favorably to the the well-known Neighbor-Joining method (Saitou, 1987) using distances computed from the observed covariances. Finally, Appendices B and C contain running times and implementation details respectively of the MIP solver used in the experimental results of Section 3.

2 Materials and Methods

2.1 Tree-Structured Covariance Matrices

Our objects of study are covariance matrices of diffusion processes defined over trees (Cavalli-Sforza and Edwards, 1967; Felsenstein et al., 2004). Usually, a Brownian motion assumption is made on the diffusion process where steps are independent and normally distributed with mean zero. However, covariance matrices of diffusion processes with independent steps, mean zero and finite variance will also have the structure we are studying here. We do not make any normality assumptions on the diffusion process and, accordingly, fit covariance matrices by minimizing a projection objective instead of maximizing a likelihood function. Thus, for a tree $\mathcal{T}$ defined over p objects, our assumption is that the observed data are realizations of a random variable $Y \in \mathbb{R}^p$ with $\mathrm{Cov}(Y) = B$, where B is a tree-structured covariance matrix defined by $\mathcal{T}$.

Figure 1 shows a tree with four leaves, corresponding to a diffusion process for four objects. A rooted tree defines a set of nested partitions of objects such that each node in the tree (both interior and leaves) corresponds to a subset of these objects. In our example, the lower branch exiting the root corresponds to subset $\{1, 2\}$. The root of the tree corresponds to the set of all objects while each leaf node corresponds to a singleton set. The subset corresponding to an interior node is the union of the non-overlapping subsets of that node's children. Edges are labeled with nonnegative real numbers indicating tree branch lengths.



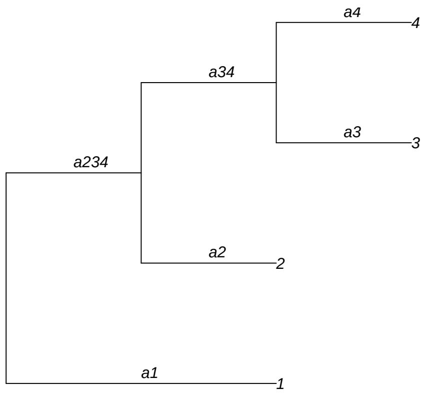
(a)

$$B = \begin{pmatrix} a_{12} + a_1 & a_{12} & 0 & 0 \\ a_{12} & a_{12} + a_2 & 0 & 0 \\ 0 & 0 & a_{34} + a_3 & a_{34} \\ 0 & 0 & a_{34} & a_{34} + a_4 \end{pmatrix}$$

(b)

Figure 1: A schematic example of a phylogenetic tree and corresponding covariance matrix. The root is the leftmost node, while leaves are the rightmost nodes. Branch lengths are arbitrary nonnegative real numbers.

Denoting $B = \mathrm{Cov}(Y)$, entry B_{ij} is the sum of branch lengths for the path starting at the root and ending at the last common ancestor of leaves i and j. In our example, $B_{12} = a_{12}$ is the length of the branch from the root to the node above leaves 1 and 2. For leaf i, B_{ii} is the sum of the branch lengths of the path from root to leaf. The covariance matrix B for our example tree is given in Figure 1(b). If we swap the positions of labels 3 and 4 in our example tree such that label 3 is the topmost label and construct a covariance matrix accordingly we recover the same covariance matrix B. In fact, any tree that specifies this particular set of nested partitions and branch lengths generates the same covariance matrix. All trees that define the same set of nested partitions are said to be of the same topology, and we say that covariance matrices that are generated from trees with the same topology belong to the same class. However, a tree topology that specifies a different set of nested partitions generates a different class of covariance matrices. For example, Figure 2 shows a tree that defines a different set of nested partitions and the matrix it generates.



(a)

$$A = \begin{pmatrix} a_1 & 0 & 0 & 0 \\ 0 & a_{234} + a_2 & a_{234} & a_{234} \\ 0 & a_{234} & a_{234} + a_{34} + a_3 & a_{234} + a_{34} \\ 0 & a_{234} & a_{234} + a_{34} & a_{234} + a_{34} + a_4 \end{pmatrix}$$

(b)

Figure 2: An example phylogenetic tree with different topology and corresponding covariance matrix.

2.2 Representing Tree-Structured Covariance Matrices

Let $d = \begin{bmatrix} a_{12} & a_{34} & a_1 & a_2 & a_3 & a_4 \end{bmatrix}^T$ be a column vector containing the branch lengths of the tree in Figure 1. We can write $B = \sum_{k=1}^{6} d_k M^k$ where M^k is a matrix such that $M^k_{i,j} = 1$ if objects i and j co-occur in the subset corresponding to the node where branch k ends. For the branch with length a_{12}, we have

$$M^1 = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Furthermore, we can use indicator vectors v_k to specify the M^k matrices in the linear expansion of B as outer products of v_k with itself. For example, letting $v_1 = \begin{bmatrix} 1 & 1 & 0 & 0 \end{bmatrix}^T$, we get

$$M^1 = v_1 v_1^T = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 0 \end{bmatrix}.$$

Thus, using vectors v_k we can write $B = \sum_{k=1}^{6} d_k v_k v_k^T$ and defining matrices $V = \begin{bmatrix} v_1 & v_2 & \dots & v_6 \end{bmatrix}$ and $D = \operatorname{diag}(d)$, we can equivalently write

$$B = VDV^T. \tag{1}$$

For Figure 1, the complete expansion is given by

$$V = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad D = \operatorname{diag}(\begin{bmatrix} 0 & a_{12} & a_{34} & a_1 & a_2 & a_3 & a_4 \end{bmatrix}^T). \tag{2}$$

Since the basis matrix V in Equation (1) is determined by the nested partitions defined by the corresponding tree topology, all covariance matrices of the same class are generated by linear expansions of a corresponding matrix V with branch lengths specified in the diagonal matrix D. On the other hand, a distinct basis matrix V corresponds to each distinct tree topology. Matrices spanned by the set of matrices V that correspond to valid partitions are tree-structured covariance matrices. We now characterize this set of valid V matrices by defining a partition property, and give a representation theorem for tree-structured covariance matrices based on this property.

Definition 1 (Partition Property) *A basis matrix V of size p-by-$(2p-1)$ with entries in $\{0, 1\}$ and unique columns has the partition property for trees of size p if it satisfies the following conditions:*

- *V contains the vector of all ones $e = (1, 1, \dots, 1)^T \in \mathbb{R}^p$ as a column, and*
- *for every column w in V with more than one non-zero entry, it contains columns u and v such that $u + v = w$.*

A matrix V with the partition property can be constructed by starting with the column $e \in \mathbb{R}^p$ and splitting it into two nonzero columns u and v with $u + v = e$. These form the next two columns of V. The remaining columns of V are generated by splitting previously unsplit columns recursively into the sum of two nonzero columns, until we finally obtain columns with a single nonzero. It is easy to see that the total number of splits is $p - 1$, with two columns generated at each split. It follows that V does not contain the the zero column, and contains all p vectors that contain $p - 1$ zero terms and a single entry of 1. For example, the V matrix in Equation (2) can be constructed by starting with column 1, splitting into columns 2 and 3, and then splitting each recursively to obtain the remaining four columns.

Theorem 2 (Tree Covariance Representation) *A matrix B is a tree-structured covariance matrix if and only if $B = VDV^T$ where D is a diagonal matrix with nonnegative entries and the basis matrix V has the partition property.*

Proof Assume B is a tree-structured covariance matrix, then construct matrix V using the method above starting from the root, splitting each vector according to the nested partitions at each node. By construction, V will satisfy the partition property and by placing branch lengths in diagonal matrix D we will have $B = VDV^T$. On the other hand, let $B = VDV^T$ with D diagonal and V having the partition property. Then construct a tree by the reverse construction: starting at the root and vector $e \in \mathbb{R}^p$, create a nested partition from the vectors u and v such that $u + v = e$ which must exist since V has the partition property. Define branch lengths from D correspondingly, and continue this construction recursively. B will then be the covariance matrix defined by the resulting tree and therefore be tree-structured. ■

2.3 Characteristics of the Set of Tree-Structured Covariance Matrices

We now state some facts about the set of tree-structured covariance matrices which we make use of in our estimation procedures.

Proposition 3 *The set of tree-structured covariance matrices $B = VDV^T$ generated by a single basis matrix V is convex.*

Proof Let d_1 and d_2 be the branch length vectors of tree-structured covariance matrices $B_1 = V\text{diag}(d_1)V^T$ and $B_2 = V\text{diag}(d_2)V^T$. Let $\theta \in [0,1]$, then $B = \theta B_1 + (1-\theta)B_2 = V\text{diag}(\theta d_1 + (1-\theta)d_2)V^T$. So, B is a tree of the same structure with branch lengths given by $\theta d_1 + (1-\theta)d_2$. ■

We will use this fact to express estimation problems for trees of fixed topology as convex optimization problems. However, estimation of general tree-structured covariance matrices is not so simple, as the set of all tree-structured covariance matrices is *not convex* in general. We can see that this is true in the case $p = 3$ by considering the following example. Defining

$$V_1 = \begin{pmatrix} 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}, \qquad V_2 = \begin{pmatrix} 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 \end{pmatrix},$$

we see that V_1 and V_2 both have the partition property. Therefore by Theorem 2, the matrices $B_1 = V_1\text{diag}(d_1)V_1^T$ and $B_2 = V_2\text{diag}(d_2)V_2^T$ are both tree-structured covariance matrices when d_1 and d_2 contain nonnegative entries. If B is a convex combination of B_1 and B_2, we will have $B_{12} \neq 0$ and $B_{23} \neq 0$ but $B_{13} = 0$. It is not possible to identify a matrix V with the partition property such that $B = VDV^T$, since any such V may contain only a single column apart from the three "unit" columns $(1,0,0)^T$, $(0,1,0)^T$, and $(0,0,1)^T$, and none of the possible candidates for this additional column (namely, $(1,1,0)^T$, $(1,0,1)^T$, and $(0,1,1)^T$) can produce the required nonzero pattern for B. This example can be extended trivially to successively higher dimensions p by expanding V_1 and V_2 appropriately.

2.4 Fixed Topology Projection Problems

In this section, we address the problem of estimating a tree-structured covariance matrix from a known tree topology by minimizing the distance to an observed sample covariance matrix. That is, given a sample covariance matrix S and a basis matrix V, we find the nearest tree-structured covariance matrix in norm $\|\cdot\|$. We will look at problems using Frobenius norm, $\|B\|_F = \sqrt{\sum_{ij} B_{ij}^2}$, and sum-absolute-value (sav) norm, $\|B\|_{\text{sav}} = \sum_{ij} |B_{ij}|$.

As stated above, the set of covariance matrices corresponding to trees of a particular topology is convex. Since projection problems have convex objective functions, they are convex optimization problems for any norm $\|\cdot\|$. While our emphasis in this paper is optimization over the non-convex set of all tree-structured covariance matrices, it is illustrative to show the convex optimization problem formulations for projection in Frobenius and sum-absolute-value norm with fixed-topologies.

For Frobenius norm, given a covariance matrix S, the nearest tree-structured covariance matrix B in the class determined by basis matrix V is given by the branch length vector that solves the problem

$$\begin{aligned} \min_{d \in \mathbb{R}^{2p-1}} \quad & \|S - V\text{diag}(d)V^T\|_F^2 \\ \text{s.t.} \quad & d \geq 0. \end{aligned}$$

We can simplify this to the following equivalent quadratic problem:

$$\begin{aligned} \min_{d \in \mathbb{R}^{2p-1}} \quad & d^T Q d - 2c^T d \\ \text{s.t.} \quad & d \geq 0, \end{aligned}$$

where $Q = (V^T V) \circ (V^T V)$ and $c = \operatorname{diag}(V^T S V)$ with $\circ$ denoting element-wise (Hadamard) matrix multiplication. For sav norm, the branch lengths d corresponding to the nearest tree-structured matrix in the proper class are given by the solution to the following problem:

$$\begin{aligned} \min_{d \in \mathbb{R}^{2p-1}} \quad & \|S - V \operatorname{diag}(d) V^T\|_{\text{sav}} \\ \text{s.t.} \quad & d \geq 0. \end{aligned}$$

Letting $s \in \mathbb{R}^{p(p+1)/2}$ be the vectorization of symmetric matrix S, we can we can rewrite this as the following linear problem:

$$\begin{aligned} \min_{\substack{d \in \mathbb{R}^{2p-1} \\ p,q \in \mathbb{R}^{p(p+1)/2}}} \quad & e^T (p + q) \\ \text{s.t.} \quad & \begin{bmatrix} H & I & -I \end{bmatrix} \begin{bmatrix} d \\ p \\ q \end{bmatrix} = s \\ & d \geq 0, \;\; p \geq 0, \;\; q \geq 0, \end{aligned}$$

where the row of H corresponding to S_{ij} is $V_{\cdot i} \circ V_{\cdot j}$ and e is the vector of all ones of appropriate length.

2.5 Solving Estimation by Projection for Unknown Tree Topologies using Mixed-Integer Programming

The non-convexity of the set of tree-structured covariance matrices requires estimation procedures that handle the combinatorial nature of optimization over this set. We model these problems as mixed-integer programs (MIPs). In particular, we make use of the fact that algorithms for mixed-integer linear and quadratic programs are well understood and that robust production codes for solving them are available.

2.5.1 Mixed-Integer Programming

Mixed-integer programs (MIPs) place integrality constraints on some of the problem variables. The general statement of a MIP is:

$$\min_{x \in \mathbb{R}^n} f_0(x) \tag{3a}$$

$$\text{s.t. } g_i(x) \leq 0, \;\; i = 1, 2, \ldots, m \tag{3b}$$

$$x_j \in \mathbb{Z}, \;\; j = 1, 2, \ldots, t, \tag{3c}$$

for some $t \leq n$. The functions g_i are (smooth) constraint functions and f_0 is the objective function (also assumed to be smooth), and $\mathbb{Z}$ is the set of integers. When f_0 and g_i, $i = 1, \ldots, m$, are linear, we have a mixed-integer linear program (MILP), and when f_0 is quadratic and g_i, $i = 1, \ldots, m$, are linear, we have a mixed-integer quadratic program (MIQP). We will see that projection problems for tree-structured covariance matrices are MILPs for the sav norm and MIQPs for the Frobenius norm.

Although the problem (3) is intractable in general, many practical instances can be solved, and algorithms for finding solutions have been the subject of intense research for 50 years (see for example Wolsey and Nemhauser (1999)). Current state-of-the-art software combines two methodologies: branch-and-bound and branch-and-cut. Branch-and-bound is based on construction of a tree[1] of relaxations of the problem (3), where each node of the tree contains a subproblem in which some of the integer variables x_j are allowed to take non-integer values (but may be confined to some range). A node is a child of another node in the tree if there is exactly one component x_j that is fixed at an integer value in the current node but that is a continuous variable in the parent node. In the root node of the tree, *all* integer variables are relaxed and allowed to take non-integer values, while at the leaf nodes, all integer variables x_j, $j = 1, 2, \dots, t$ are fixed at certain values. Each node of the tree is therefore a continuous linear program (with real variables), so it can be "evaluated" using the simplex method, usually by modifying the solution of its parent node. The optimal objective at a node gives a lower bound on the optimal objectives of any of its descendants, since the descendants have fewer degrees of freedom (that is, a more restricted feasible set). Hence, if this lower bound is worse than the best integer feasible solution found to date, this node and all its descendants can be "pruned" from the tree; it is not necessary to evaluate them as they cannot contain the solution of (3). The branch-and-bound algorithm traverses this tree judiciously, avoiding evaluation of large parts of the tree that are determined *not* to contain the optimal solution.

Cutting planes are used to enhance the speed of this process. These are additional constraints that exclude from the feasible set those values of x that are determined not to be optimal. Cuts can be valid for the whole tree, or just at a certain node and its descendants.

The branching strategy which determines the order in which the search tree is traversed, and the method of construction of cutting planes, can have significant effects on the efficiency of the MIP solver for particular problems. In Appendix C, we provide details regarding the parameters chosen in our MIP solver for the projection problems we address here.

2.5.2 Mixed-Integer Constraints for Tree Topology

Every tree-structured covariance matrix satisfies the following properties derived from the linear expansion in Equation (1):

1. $B_{ij} \geq 0$ for all i and j, since all entries in V and d are nonnegative.

2. $B_{ii} \geq B_{ij}$ for all i and j, since V has the partition property, every component of d that is added to an off-diagonal entry is added to the corresponding diagonal entries along with the component of d corresponding to the column in V with a single non-zero entry for the corresponding leaves.

3. $B_{ij} \geq \min(B_{ik}, B_{jk})$ for all i, j, and k, with $i \neq j \neq k$. Since V has the partition property, then for every three off-diagonal entries there is one entry that has at least one fewer component of d added in than the other two components.

Since every tree-structured covariance matrix can be expressed as $B = VDV^T$ according to Theorem 2, it is also positive semidefinite, since $VDV^T = \sum_i d_i v_i v_i^T$ is the sum of positive semidefinite matrices. Also, the three properties above follow from the expansion $B = VDV^T$, therefore any

[1] The tree referred to in this paragraph is a tree of related relaxations of the MIP, not a phylogenetic tree.

matrix that satisfies these properties is also positive semidefinite, so we need not add semidefiniteness constraints in the optimization problems below. Therefore, we can solve estimation problems for unknown tree topologies by constraining covariance matrices to satisfy the above properties. However, the third constraint is not convex, so we use integrality constraints to model it.

We begin by rewriting this third constraint for each distinct triplet $i > j > k$ as a disjunction of three constraints:

$$B_{ij} \geq B_{ik} = B_{jk} \tag{4a}$$
$$B_{ik} \geq B_{ij} = B_{jk} \tag{4b}$$
$$B_{jk} \geq B_{ij} = B_{ik}.. \tag{4c}$$

This can be derived by noting that the third property above holds for all orderings of the given i, j, and k thus preventing any one of the values B_{ij}, B_{ik}, B_{jk} from being strictly smaller than the other two values, leading to a tie for the smallest value.

A standard way of modeling disjunctions is to use $\{0, 1\}$ variables in the optimization problem (Bertsimas and Weismantel, 2005). In our case we can use two integer variables ρ_{ijk1} and ρ_{ijk2}, under the constraint that $\rho_{ijk1} + \rho_{ijk2} \leq 1$, that is, they can both be 0, or, strictly one of the two is allowed to take the value 1. With these binary variables we can write the constraints (4) in a way that the constraint corresponding to the nonzero-valued binary variable must be satisfied. For example, constraint (4a) is transformed to:

$$\begin{aligned} B_{ij} &\geq B_{ik} - (1 - \rho_{ijk1})M \\ B_{ik} &\geq B_{jk} - (1 - \rho_{ijk1})M \\ B_{jk} &\geq B_{ik} - (1 - \rho_{ijk1})M, \end{aligned}$$

where M is a very large positive constant. Constraints (4b) and (4c) are transformed similarly, yielding the full set of mixed-integer constraints in Table 1. When $\rho_{ijk1} = 1$, these constraints imply that constraint 4a is satisfied. However, since $\rho_{ijk1} = 1$ we must have $\rho_{ijk2} = 0$ which implies that constraints 4b and 4c need not be satisfied for a solution to be feasible. When $\rho_{ijk1} = \rho_{ijk2} = 0$, then constraint 4c must be satisfied.

2.5.3 Projection Problems

Let S be a sample covariance matrix, the nearest tree structured covariance matrix in norm $\|\cdot\|$ to S is given by the solution of the mixed-integer problem:

$$\begin{aligned} \min_{B \in \mathcal{S}^p} \quad & \|S - B\| \\ \text{s.t.} \quad & \text{constraints 5a-5m hold for } B. \end{aligned}$$

For Frobenius norm $\|\cdot\|_F$, the problem reduces to a mixed-integer quadratic program. Let s_2 be the vectorization of symmetric matrix S such that $\|S\|_F = \|s_2\|_2$, then the nearest tree-structured covariance matrix in Frobenius norm to matrix S is given by the corresponding matrix representation of solution $\hat{b}$ of the following mixed integer quadratic program:

$$\begin{aligned} \min_{b \in \mathbb{R}^{p(p+1)/2}, \rho \in \mathbb{R}^{\bar{p}}} \quad & \tfrac{1}{2} b^T b - s_2^T b \\ \text{s.t.} \quad & \text{constraints 5a-5m hold for } B, \end{aligned}$$

Table 1: Mixed integer constraints defining tree-structured covariance matrices

$$\begin{aligned}
B_{ij} &\geq 0 \;\; \forall i, j && \text{(5a)}\\
B_{ii} &\geq B_{ij} \;\; \forall i \neq j && \text{(5b)}\\
B_{ij} &\geq B_{ik} - (1 - \rho_{ijk1})M && \text{(5c)}\\
B_{ik} &\geq B_{jk} - (1 - \rho_{ijk1})M && \text{(5d)}\\
B_{jk} &\geq B_{ik} - (1 - \rho_{ijk1})M && \text{(5e)}\\
B_{ik} &\geq B_{ij} - (1 - \rho_{ijk2})M && \text{(5f)}\\
B_{ij} &\geq B_{jk} - (1 - \rho_{ijk2})M && \text{(5g)}\\
B_{jk} &\geq B_{ij} - (1 - \rho_{ijk2})M && \text{(5h)}\\
B_{jk} &\geq B_{ij} - (\rho_{ijk11} + \rho_{ijk2})M && \text{(5i)}\\
B_{ij} &\geq B_{ik} - (\rho_{ijk11} + \rho_{ijk2})M && \text{(5j)}\\
B_{ik} &\geq B_{ij} - (\rho_{ijk11} + \rho_{ijk2})M && \text{(5k)}\\
\rho_{ijk1} + \rho_{ijk2} &\leq 1 && \text{(5l)}\\
\rho_{ijk1}, \rho_{ijk2} &\in \{0, 1\} \;\; \forall\, i > j > k. && \text{(5m)}
\end{aligned}$$

where $\bar{p} = \frac{2p!}{(p-3)!}$.

We can similarly find the nearest tree structured covariance matrix in sum-absolute-value (sav) norm. Let s_1 be the vectorization of symmetric matrix S such that $\|S\|_{sav} = \|s_1\|_1$, then the nearest tree-structured covariance matrix in sum-absolute-value norm is given by the corresponding matrix representation of solution $\hat{b}$ of the following mixed integer linear program:

$$\begin{aligned}
\min_{b \in \mathbb{R}^{p(p+1)/2}, \rho \in \mathbb{R}^{\bar{p}}} \quad & \|s_1 - b\|_1 \\
\text{s.t.} \quad & \text{constraints 5a-5m hold for } B
\end{aligned}$$

3 Results: A Case Study in Gene Family Analysis of Yeast Gene Expression

We applied our methods to the analysis of gene expression in *Saccharomyces cerevisiae* gene families as presented in Oakley et al. (2005) [2]. Following the methodology of Gu et al. (2002), the yeast genome is partitioned into gene families using an amino acid sequence similarity heuristic. The largest 10 of the resulting families are used in this analysis with family sizes ranging from $p = 7$ to $p = 18$ genes. Names and sizes for the gene families used in the analysis are given in Table 3 of Appendix B. We refer to Oakley et al. (2005) for further details.

The gene expression data is from 19 cDNA microarray time course experiments. Each time point in the series is the $\log_2$ ratio of expression at the given time point to expression at the base line under varying experimental conditions. To make our results comparable to the analysis in Oakley et al. (2005), we do not model correlation between measurements at different time points and refer

[2] All data for this analysis was retrieved from `"http://www.lifesci.ucsb.edu/eemb/labs/oakley/pubs/MBE2005data/"`.

to Oakley et al. (2005) and Gu (2004) for a discussion regarding this violation of the independence assumption among measurements.

The analysis in Oakley et al. (2005) proceeded as follows:

1. Phylogenetic trees were derived for each family from DNA sequence using Maximum Likelihood methods. In particular, an alignment of amino acid sequences from the entire gene coding region was used to derive a DNA sequence alignment which was then used to estimate a phylogenetic tree. As stated by the authors (Oakley et al., 2005), this is one of many possible choices, including for example, flanking upstream non-coding regions that could have a significant role in expression regulation.

2. Based on the resulting trees, gene expression data was analyzed using Maximum Likelihood methods under a Brownian diffusion process under two families of models: a phylogenetic class, where the covariance of the diffusion process has a tree structure, and a non-phylogenetic class where the covariance of the diffusion process is diagonal. The AIC score of the resulting ML estimate is used to classify each gene family-experiment pair as evolving under a phylogenetic or non-phylogenetic model.

For each gene family and experiment we have a matrix Y_{gi} of size n_i-by-p where n_i is the number of time points in the ith experiment and p is the gene family size. We partition the experiments of each gene family into two disjoints sets $P = \{1, \ldots, l\}$ and $NP = \{l{+}1, \ldots, 19\}$ where l is the number of experiments classified as phylogenetic in Oakley et al. (2005). This partition yields two matrices of measurements for each gene family $Y_{gP} = \begin{bmatrix} Y_{g1}^T & \cdots & Y_{gl}^T \end{bmatrix}^T$ and similarly for Y_{gNP}, obtained by concatenating the measurement matrices of experiments in the corresponding set. The idea of concatenating gene expression measurement matrices directly to estimate covariance was sparked by the success of Stuart et al. (2003) where gene expression measurements were concatenated directly to measure correlation between genes. Since we will treat the rows of these two matrices as samples from distributions with $\mathbb{E}Y = 0$, we center each row independently to have mean 0.

One of the constraints in Section 2.5.2 that characterize tree-structured covariance matrices is the nonnegativity of their entries. Therefore, to initialize our projection solvers, we first estimate Maximum-Likelihood covariance matrices B_{gP}^+ and B_{gNP}^+ constrained to have nonnegative entries from sample matrices Y_{gP} and Y_{gNP}. Treating the rows of n-by-p matrix Y as independent samples from a multivariate normal distribution $N(0, B^+)$ the goal is to find matrix B^+ that maximizes likelihood, where B^+ is constrained to have nonnegative entries. Following the constrained maximum-likelihood formulation in Vandenberghe et al. (1998), we define the following convex determinant maximization problem

$$\begin{aligned} \max_{R \in \mathcal{S}^p} \quad & n \log\det R - \operatorname{tr}(RS) && \text{(6a)} \\ \text{s.t.} \quad & R_{ij} \le 0, \ \forall i \neq j && \text{(6b)} \\ & R \succ 0, && \text{(6c)} \end{aligned}$$

where $\mathcal{S}^p$ is the space of p-by-p symmetric matrices, n is the number of samples in matrix Y, and $S = YY^T$ its sample covariance matrix. The expression $R \succ 0$ denotes that R is positive definite and we take variable R to be the inverse of the estimate $B^+ = R^{-1}$. By the nonpositivity element-wise constraints (6b), along with the positive definiteness constraint (6c), feasible solutions to Problem (6) will be members of the class of M-matrices (Horn and Johnson, 1991) which have the property that their inverse are matrices with nonnegative entries (Theorem 2.5.3 in Horn and

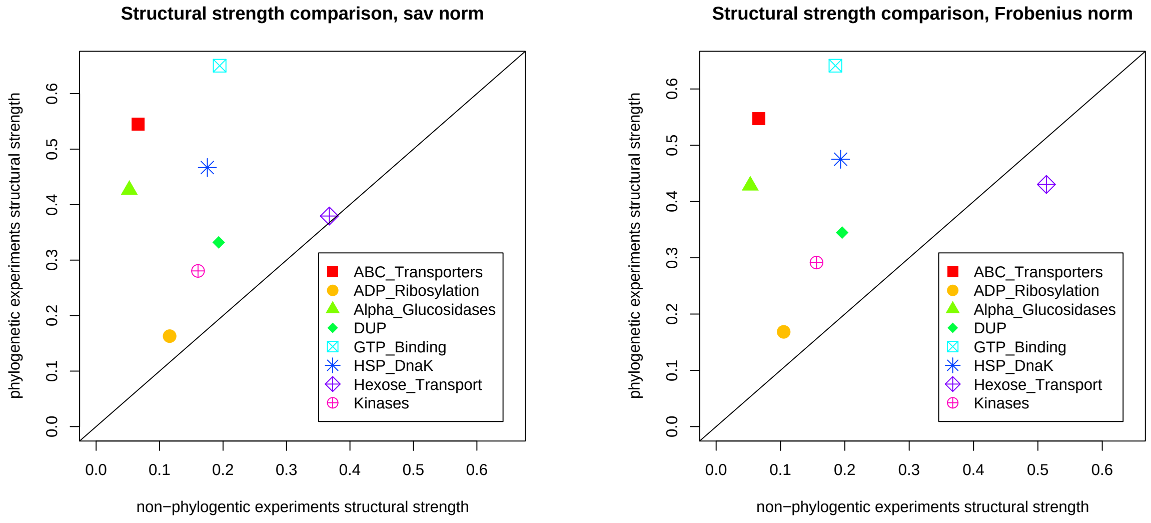
Johnson (1991)). Therefore, the constraints in Problem (6) imply that estimate $\hat{B}^+$ will be the maximum likelihood estimate with nonnegative entries.

From estimates $\hat{B}^+_{gP}$ and $\hat{B}^+_{gNP}$ we estimate tree-structured covariance matrices $\hat{B}_{gP}$ and $\hat{B}_{gNP}$ using our MIP projection methods. To describe the strength of the hierarchical structure of these estimated covariances we define the *structural strength* metric as follows:

$$SS(B) = \frac{1}{p} \sum_{i=1}^{p} \frac{\max_{i \neq j} B_{ij}}{B_{ii}}. \tag{7}$$

The term $\max_{i \neq j} B_{ij}$ is the largest covariance between gene i and a different gene j. This is the length of the path from the root to the immediate ancestor of leaf i in the corresponding tree. Therefore, the ratio in $SS(B)$ compares the length of the path from the root to leaf i to the length of the subpath from the root to i's immediate ancestor. A value of $SS(B)$ near zero means that on average objects have zero covariance, values near one means that the tree is strongly hierarchical where objects spend very little time taking independent steps in the diffusion process.

Under the classification of experiments as undergoing phylogenetic versus non-phylogenetic evolution we expect that the structural strength metric should be quite different for estimated tree-structured covariance matrices $\hat{B}_{gP}$ and $\hat{B}_{gNP}$. That is, we expect that $SS(\hat{B}_{gP}) \geq SS(\hat{B}_{gNP})$ for most gene families g. We show our results in Figure 3 which validate this hypothesis. We plot $SS(\hat{B}_{gP})$ versus $SS(\hat{B}_{gNP})$ for each gene family g. The diagonal is the area where $SS(\hat{B}_{gP}) = SS(\hat{B}_{gNP})$. We see that in fact $SS(\hat{B}_{gP}) > SS(\hat{B}_{gNP})$ for all gene families g except the Hexose Transport Family.

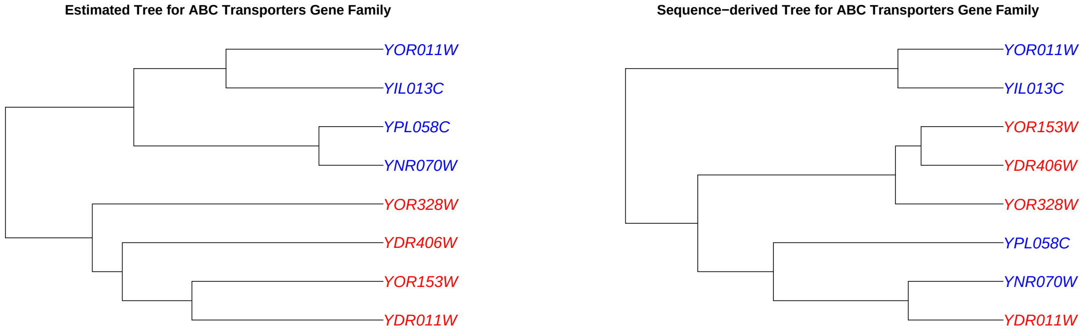


(a)

(b)

Figure 3: Comparison of structural strengths for tree-structured covariance estimates $\hat{B}_{gP}$ and $\hat{B}_{gNP}$ for projection under sav (a) and Frobenius (b) norms. Each point represents a gene family. The x-axis is $SS(\hat{B}_{gNP})$. We can see that for all, except the Hexose Transport gene family, $SS(\hat{B}_{gP}) > SS(\hat{B}_{gNP})$. Only eight families are shown since the Putative Helicases and Permeases families did not have any experiments classified as phylogenetic.

We next look at the resulting tree for the ABC (ATP-binding cassette) Transporters gene family (see Jungwirth and Kuchler (2006) for a short literature review) in more detail. In particular, the eight genes included in this group are members of the subfamily conferring pleiotropic drug resistance (PDR) and are all located in the plasma membrane. A number of transcription factors have been found for the PDR subfamily, including the PDR3 factor considered one of the master regulators of the PDR network (Delaveau et al., 1994). Figure 4 shows the tree estimated by the MIP projection method for this family along with the sequence-derived tree reported by Oakley et al. (2005). We can notice topological differences between the two trees, in particular, the subtree in Figure 4(a) containing genes YOR328W, YDR406W, YOR153W and YDR011W.



(a)

(b)

Figure 4: (a) Tree estimated by the MIP projection method using Frobenius norm for the ABC Transporters gene family. (b) Sequence-derived tree reported by Oakley et al. (2005) for the ABC Transporters gene family. The red tips correspond to genes YOR328W, YDR406W, YOR153W and YDR011W which form a subtree in (a) but not in (b).

In order to elucidate this topological difference, we turn to the characteristics of the promoter (regulatory) regions of the genes and ask whether transcription factor (TF) binding site contents of the upstream regions could account for this difference. We compiled a list of known yeast transcription factor binding site consensus sequences using Gasch et al. (2004) and the Promoter Database of *Saccharomyces cerevisiae* (SCPD) (`http://rulai.cshl.edu/SCPD/`). Then, we generated a transcription factor binding site occurrence vector for each gene by simply counting the number of occurrences of each consensus sequence in the 1000 base pairs upstream of the coding region. Putting these profiles together we obtained a 8-by-128 matrix where rows represent the 8 genes in the ABC Transporters gene family and columns represent 128 transcription factors. Inspection of this matrix once the rows are permuted to follow the hierarchy in the tree estimated by the MIP projection method (Figure 4(a)) immediately revealed that the presence or absence of the PDR3 transcription factor binding site in the flanking upstream region may account for the topological difference apparent in the two estimated trees. Table 2 shows the number of times the motif for the PDR3 factor was detected in the upstream region of each gene.

It is known (Delaveau et al., 1994) that the four genes in Table 2 with multiple PDR3 binding sites are, as opposed to the other four genes, targets of this transcription factor which controls the pleiotropic drug resistance phenomenon. The structure of the subtree in Figure 4(a) corresponding to the PDR3 target genes essentially follows the frequency of PDR3 occurrences. On the other hand, the structure of the subtree for the non-PDR3 target genes follows that of the sequence-

Table 2: Number of occurrences of the PDR3 transcription factor motif in the 1000 bp upstream region for each gene in the ABC Transporters family. Colors match those of Figure 4.

	gene	Occurrences of PDR3
1	YOR011W	0
2	YIL013C	0
3	YPL058C	0
4	YNR070W	0
5	YDR406W	3
6	YOR328W	4
7	YDR011W	6
8	YOR153W	9

derived tree of Figure 4(b). Namely, pairs (YOR011W,YIL013C) and (YPL058C,YNR070W) are near each other in both the sequence-derived and the MIP-derived trees. Therefore, after taking into account the initial split characterized by the presence of the PDR3 transcription factor, the MIP estimated tree (Figure 4(a)) is similar to the sequence-derived tree (Figure 4(b)).

We reiterate the observation of Oakley et al. (2005) that the choice of sequence region to create the reference phylogenetic trees used in their analysis plays a crucial role and results could vary accordingly. From our methods, we have found evidence that using upstream sequence flanking the coding region might yield a tree that is better suited to explore the influence of evolution in gene expression for this particular gene family. We believe that finding a good estimate for tree-structured covariance matrices directly from expression measurements can help investigators guide their choices for downstream comparative analysis like that of Oakley et al. (2005).

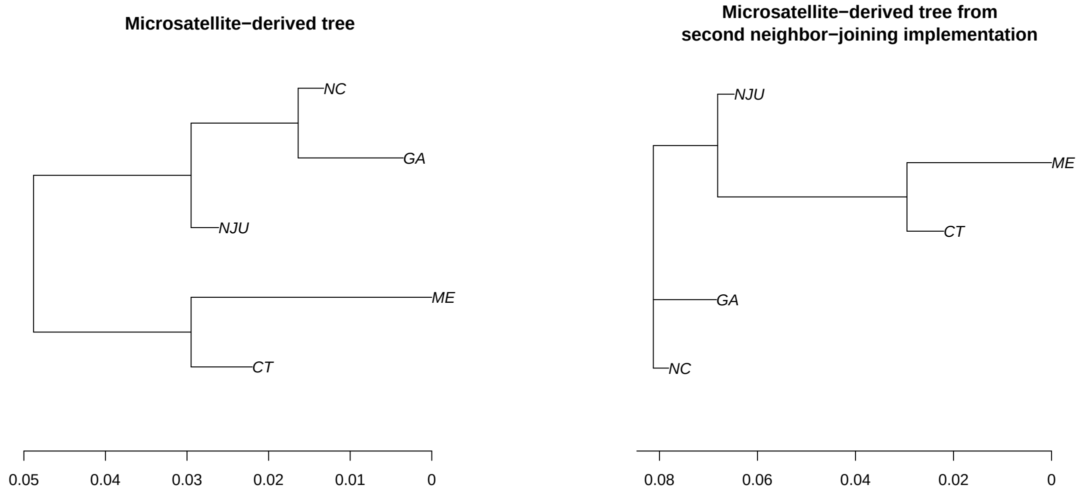
Appendices C and B detail implementation choices and running times of our mixed-integer estimation procedure.

4 Discussion

The issues we hope to address by estimating tree-structured covariance matrices directly from observed sample covariances from gene expression data can be illustrated using the work of Whitehead and Crawford (2006) who characterize evolution patterns of the expression of 329 genes in five strains of the *Fundulus heteroclitus* fish. One of their analyses uses generalized least squares regression of gene expression on habitat temperature using a tree-structured covariance matrix for correction. This structured covariance matrix is derived from a phylogeny constructed from five microsatellite markers (short repeating strings) which are random characters expected not to be influenced by selection and to evolve at the same base rate as the whole genome. The tree is constructed with the greedy neighbor-joining algorithm (Saitou, 1987) from Cavalli-Sforza and Edward's (CSE) chord distances between the five microsatellite markers. We reproduce this microsatellite-derived tree in Figure 5(a). The neighbor-joining algorithm is a greedy algorithm susceptible to generating different solutions depending on how the algorithm is implemented. For example, the implementation of this algorithm in the ape R package [3] yields a different tree (Figure 5(b)) given the CSE distances. For the purpose of generalized least squares, and therefore the

[3] Version 1.10-2. We thank Dr. Andrew Whitehead for providing the distance data through personal communication.

evolutionary statements asserted as a result, this difference in topology can be significant. Considering this instability of the resulting neighbor-joining tree and the importance it plays in the authors' analyses, we posit that deriving tree-structured covariance matrices directly from the expression data can guide investigators in comparing sequence-derived phylogenetic trees for use in subsequent comparative analysis.


Microsatellite−derived tree
NC
GA
NJU
ME
CT
0.05
0.04
0.03
0.02
0.01
0
Microsatellite−derived tree from second neighbor−joining implementation
NJU
ME
CT
GA
NC
0.08
0.06
0.04
0.02
0

(a)

(b)

Figure 5: Microsatellite-derived trees built by two implementations of the neighbor-joining algorithm from Cavalli-Sforza and Edward's chord distances. Figure 5(a) is the tree reported in Whitehead and Crawford (2006), and Figure 5(b) was obtained by the `ape` `R` package.

To address these shortcomings and motivated by what we think is a problem of genomic resolution as described in the Introduction, we have described a method for estimating tree-structured covariance matrices directly from observed sample covariance matrices by projection methods. We showed that projection problems for known topologies are linear or quadratic programs depending on the approximation norm used. For unknown topology problems, we proposed and evaluated a mixed-integer formulation which can be solved to optimality by existing branch-and-bound solvers.

The work of McCullagh (McCullagh, 2006) on tree-structured covariance matrices is the closest to our work. He proposes the *minimax projection* to estimate the structure of a given sample covariance matrix. Given this structure, likelihood is maximized as in Anderson (1973). The *minimax projection* is independent of the estimation problem being solved as opposed to our MIP method which minimizes the estimation objective while finding tree structure simultaneously. Furthermore, the MIP solver guarantees optimality upon completion, at the cost of longer execution in difficult cases where the optimal trees in many tree topologies have similar objective values.

Rifkin et al. (2003) use expression directly to estimate phylogenetic structure, but use a distance-based method utilizing the number of pairwise differentially expressed genes as the source of distances. They observe that for the resulting distance matrix the neighbor joining tree-building algorithm (Saitou, 1987) produces a tree estimate that matches the sequence derived tree for a subgroup of Drosophila species.

Using the MIP formulation to model tree-structured matrix constraints, we can also address the need to solve existing tree estimation problems exactly. In particular, the least squares method of Fitch and Margoliash (1967) estimates a tree that minimizes the least-squares deviation of the distance between objects in the tree and a given distance matrix D. However, given a covariance matrix B, we can compute squared distances between objects using the linear expression $D_{ij}^2 = B_{ii} + B_{jj} - 2B_{ij}$. This implies that the least squares distance-deviance objective is a quadratic function of the entries of covariance matrix B. Therefore, using the MIP formulation of Section 2.5 and the quadratic least squares distance-deviance objective, we can express the least-squares method of Fitch and Margoliash (1967) as a MIQP. Thus, generic branch-and-bound solvers of quadratic MIPs fill the gap observed in Felsenstein et al. (2004) which states that no branch-and-bound method to solve the least-squares problem exactly has been proposed.

Along the same line, MIPs have been used to solve phylogeny estimation problems for haplotype data (Brown and Harrower, 2006; Huang et al., 2005; Sridhar et al., 2008; Wang and Xu, 2003). The observed data from the tree leaves in this case is haplotype variation represented as sequences of ones and zeros. Although our MIP formulation is related, the data in our case is assumed to be observations from a diffusion process along a tree, suitable for continuous traits like gene expression.

We can place the problem of estimating tree-structured covariance matrices in the broader context of structured covariance matrix estimation (Anderson, 1973; Li et al., 1999; Schulz, 1997). The work of Anderson (1973) is especially relevant since an iterative procedure is used to fit matrices, or matrix inverses, which can be expressed as linear combinations of known symmetric matrices. For known topologies, this method solves likelihood maximization problems where a normality assumption is made on the diffusion process underlying the data. However, for unknown topologies, maximum likelihood problems require that we extend our computational methods to, for example, determinant maximization problems. Solving these and similar types of nonlinear MIPs is an active area of research in the optimization community (Lee, 2007). In recent years, the problem of structured covariance matrix estimation has been mainly addressed in its application to sparse Gaussian Graphical Models (Banerjee and Natsoulis, 2006; Chaudhuri et al., 2007; Drton and Richardson, 2003, 2004; Yuan and Lin, 2007). In this instance, sparsity in the inverse covariance matrix induces a set of conditional independence properties that can be encoded as a sparse graph (not necessarily a tree).

Although we presented a descriptive metric of structural strength in our estimates in Section 3, future work will concentrate on leveraging these methods in principled hypothesis testing frameworks that better assess the presence of hierarchical structure in observed data. We expect that the resulting methods are likely to impact how evolutionary analysis of gene expression traits is conducted.

References

T.W. Anderson. Asymptotically Efficient Estimation of Covariance Matrices with Linear Structure. *The Annals of Statistics*, 1(1):135–141, 1973.

O. Banerjee and G. Natsoulis. Convex optimization techniques for fitting sparse Gaussian graphical models. *Proceedings of the 23rd international conference on Machine learning*, pages 89–96, 2006.

D. Bertsimas and R. Weismantel. *Optimization over integers*. Dynamic Ideas, 2005.

D.G. Brown and I.M. Harrower. Integer programming approaches to haplotype inference by pure parsimony. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3(2):141–154, 2006.

L.L. Cavalli-Sforza and AWF Edwards. Phylogenetic Analysis: Models and Estimation Procedures. *Evolution*, 21(3):550–570, 1967.

S. Chaudhuri, M. Drton, and T.S. Richardson. Estimation of a covariance matrix with zeros. *Biometrika*, 94(1):199, 2007.

T. Delaveau, A. Delahodde, E. Carvajal, J. Subik, and C. Jacq. PDR3, a new yeast regulatory gene, is homologous toPDR1 and controls the multidrug resistance phenomenon. *Molecular Genetics and Genomics*, 244(5):501–511, 1994.

M. Drton and T.S. Richardson. A New Algorithm for Maximum Likelihood Estimation in Gaussian Graphical Models for Marginal Independence. *UAI (Uffe Kjærulff and Christopher Meek, eds.), San Francisco: Morgan Kaufmann*, pages 184–191, 2003.

M. Drton and T.S. Richardson. Iterative conditional fitting for Gaussian ancestral graph models. *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 130–137, 2004.

J.C. Fay and P.J. Wittkopp. Evaluating the role of natural selection in the evolution of gene regulation. *Heredity*, 1:9, 2007.

J. Felsenstein et al. *Inferring phylogenies.* Sinauer Associates Sunderland, Mass., USA, 2004.

W.M. Fitch and E. Margoliash. Construction of Phylogenetic Trees. *Science*, 155(3760):279–284, 1967.

K.A. Frazer, C.M. Wade, D.A. Hinds, N. Patil, D.R. Cox, and M.J. Daly. Segmental Phylogenetic Relationships of Inbred Mouse Strains Revealed by Fine-Scale Analysis of Sequence Variation Across 4.6 Mb of Mouse Genome. *Genome Research*, 14:1493–1500, 2004.

A.P. Gasch, A.M. Moses, D.Y. Chiang, H.B. Fraser, M. Berardini, and M.B. Eisen. Conservation and evolution of cis-regulatory systems in ascomycete fungi. *PLoS Biol*, 2(12):e398, 2004.

X. Gu. Statistical Framework for Phylogenomic Analysis of Gene Family Expression Profiles. *Genetics*, 167(1):531–542, 2004.

Z. Gu, A. Cavalcanti, F.C. Chen, P. Bouman, and W.H. Li. Extent of Gene Duplication in the Genomes of Drosophila, Nematode, and Yeast. *Molecular Biology and Evolution*, 19(3):256–262, 2002.

F. Habib, A.D. Johnson, R. Bundschuh, and D. Janies. Large scale genotype-phenotype correlation analysis based on phylogenetic trees. *Bioinformatics*, 23(7):785, 2007.

R.A. Horn and C.R. Johnson. *Topics in Matrix Analysis.* Cambridge University Press, 1991.

Y.T. Huang, K.M. Chao, and T. Chen. An approximation algorithm for haplotype inference by maximum parsimony. *Journal of Computational Biology*, 12(10):1261–1274, 2005.

SA Ilog. Ilog Cplex 9.0 Users Manual, 2003.

H. Jungwirth and K. Kuchler. Yeast ABC transporters–A tale of sex, stress, drugs and aging. *FEBS Letters*, 580(4):1131–1138, 2006.

J. Lee. Mixed-integer nonlinear programming: Some modeling and solution issues. *IBM JOURNAL OF RESEARCH AND DEVELOPMENT*, 51(3/4):489, 2007.

H. Li, P. Stoica, and J. Li. Computationally efficient maximum likelihood estimation of structured covariance matrices. *Signal Processing, IEEE Transactions on [see also Acoustics, Speech, and Signal Processing, IEEE Transactions on]*, 47(5):1314–1323, 1999.

P. McCullagh. Structured covariance matrices in multivariate regression models. Technical report, Department of Statistics, University of Chicago, 2006.

T.H. Oakley, Z. Gu, E. Abouheif, N.H. Patel, and W.H. Li. Comparative Methods for the Analysis of Gene-Expression Evolution: An Example Using Yeast Functional Genomic Data. *Molecular Biology and Evolution*, 22(1):40–50, 2005.

E. Paradis, J. Claude, and K. Strimmer. Ape: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20:289–290, 2004.

D. Penny and M.D. Hendy. The use of tree comparison metrics. *Syst. Zool*, 34(1):75–82, 1985.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2007. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

S.A. Rifkin, J. Kim, and K.P. White. Evolution of gene expression in the Drosophila melanogaster subgroup. *Nature Genetics*, 33(2):138–144, 2003.

N. Saitou. The neighbor-joining method: a new method for reconstructing phylogenetic trees, 1987.

T.J. Schulz. Penalized maximum-likelihood estimation of covariance matrices with linear structure. *Signal Processing, IEEE Transactions on [see also Acoustics, Speech, and Signal Processing, IEEE Transactions on]*, 45(12):3027–3038, 1997.

S. Sridhar, F. Lam, G. Blelloch, R. Ravi, and R. Schwartz. Mixed Integer Linear Programming for Maximum Parsimony Phylogeny Inference. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2008.

J.M. Stuart, E. Segal, D. Koller, and S.K. Kim. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. *Science*, 302(5643):249–255, 2003.

R.H. Tütüncü, K.C. Toh, and M.J. Todd. Solving semidefinite-quadratic-linear programs using SDPT3. *Mathematical Programming*, 95(2):189–217, 2003.

L. Vandenberghe, S. Boyd, and S.P. Wu. Determinant Maximization with Linear Matrix Inequality Constraints. *SIAM Journal on Matrix Analysis and Applications*, 19(2):499–533, 1998.

L. Wang and Y. Xu. Haplotype inference by maximum parsimony, 2003.

A. Whitehead and D.L. Crawford. Neutral and adaptive variation in gene expression. *Proceedings of the National Academy of Sciences*, 103(14):5425–5430, 2006.

L.A. Wolsey and G.L. Nemhauser. *Integer and Combinatorial Optimization*. Wiley-Interscience, 1999.

B. Yalcin, J. Fullerton, S. Miller, DA Keays, S. Brady, A. Bhomra, A. Jefferson, E. Volpi, RR Copley, J. Flint, et al. Unexpected complexity in the haplotypes of commonly used inbred strains of laboratory mice. *Proc Natl Acad Sci US A*, 101(26):9734–9739, 2004.

M. Yuan and Y. Lin. Model Selection and Estimation in the Gaussian Graphical Model. *Biometrika*, 94(1):19–35, 2007.

A Simulation Study: Comparing MIP Projection Methods and Neighbor-Joining

An alternative method to estimate a tree-structured covariance matrix from an observed sample covariance is to use a distance-matrix method such as the Neighbor-Joining (NJ) algorithm (Saitou, 1987) as follows: given sample covariance B, create a distance matrix D such that $D_{ij} = B_{ii} + B_{jj} - 2B_{ij}$, and use the NJ algorithm to estimate a tree and its corresponding tree-structured covariance matrix. In this simulation, we compare the closeness of the correct tree structure to the estimated tree-structured covariance matrix when using this NJ-based method against using our MIP-based projection methods. Specifically, we measure how close the structure of estimated tree-structured matrices are to the true structure of true matrices by using the tree topological distance defined by Penny and Hendy (1985) which essentially counts the number of mismatched nested partitions defined by the trees.

The simulation setting was the following: 1) we first generated 10 $\{\mathcal{T}_1, \dots, \mathcal{T}_{10}\}$ trees with 10 leaves each at random using the `rtree` function of the `R` `ape` library (Paradis et al., 2004), which gives 10 tree-structured covariance matrices $\{B_1, \dots, B_{10}\}$ of size 10-by-10; 2) from each tree-structured covariance matrix B_i, we draw 10 sample covariances randomly $\{B_i^1, \dots, B_i^{10}\}$ using a Wishart distribution with mean B_i and the desired degrees of freedom df. This corresponds to the sample covariance matrix of a sample with df observations from a multivariate normal random variable distributed as $N(0, B_i)$. Note that the resulting sample covariances are not necessarily tree-structured. Then, we estimate a tree-structured covariance matrix $\hat{B}_i^j$ from each sample covariance matrix B_i^j and record its topological distance to the true matrix B_i. In Figure 6 we report the mean topological distance of the resulting 100 estimates as a function of the degrees of freedom df, or number of observations. The values of the x-axis are defined to satisfy $df = 10 \times 2^x$, so for $x = 0$ there are 10 observations in each sample and so on.

We can see that the method based on NJ is unable to recover the correct structure even for large numbers of observations. On the other hand the MIP-based method is able to converge to the correct structure for both loss functions when the sample size is 16 times the number of taxa. Although the topological distances even for smaller sample sizes are not too large, this simulation also illustrates that, as expected, having a large number of replicates is better for this method. This observation is partly the reason for concatenating different experiments in the yeast gene-family analysis of Section 3.

B Running Times in Gene Family Analysis

family	p	norm	class	n	time	gap
ABC_Transporters	8	sav	phy	13	0.49	
ABC_Transporters	8	sav	nonphy	148	0.66	
ABC_Transporters	8	sav	all	161	0.26	
ABC_Transporters	8	fro	phy	13	2.01	
ABC_Transporters	8	fro	nonphy	148	0.70	

ABC_Transporters	8	fro	all	161	0.72	
ADP_Ribosylation	7	sav	phy	44	0.17	
ADP_Ribosylation	7	sav	nonphy	100	0.02	
ADP_Ribosylation	7	sav	all	144	0.07	
ADP_Ribosylation	7	fro	phy	44	0.05	
ADP_Ribosylation	7	fro	nonphy	100	0.09	
ADP_Ribosylation	7	fro	all	144	0.33	
Alpha_Glucosidases	6	sav	phy	20	0.02	
Alpha_Glucosidases	6	sav	nonphy	148	0.02	
Alpha_Glucosidases	6	sav	all	168	0.00	
Alpha_Glucosidases	6	fro	phy	20	0.11	
Alpha_Glucosidases	6	fro	nonphy	148	0.01	
Alpha_Glucosidases	6	fro	all	168	0.01	
DUP	10	sav	phy	15	112.21	
DUP	10	sav	nonphy	106	27.81	
DUP	10	sav	all	121	19.91	
DUP	10	fro	phy	15	34.86	
DUP	10	fro	nonphy	106	294.61	
DUP	10	fro	all	121	600.02	0.29%
GTP_Binding	11	sav	phy	9	22.92	
GTP_Binding	11	sav	nonphy	152	55.05	
GTP_Binding	11	sav	all	161	63.36	
GTP_Binding	11	fro	phy	9	20.93	
GTP_Binding	11	fro	nonphy	152	600.02	0.55%
GTP_Binding	11	fro	all	161	106.19	
HSP_DnaK	10	sav	phy	61	31.71	
HSP_DnaK	10	sav	nonphy	75	81.72	
HSP_DnaK	10	sav	all	136	26.49	
HSP_DnaK	10	fro	phy	61	21.60	
HSP_DnaK	10	fro	nonphy	75	412.33	
HSP_DnaK	10	fro	all	136	34.45	
Hexose_Transport	18	sav	phy	96	600.05	75.89%
Hexose_Transport	18	sav	nonphy	12	600.02	68.78%
Hexose_Transport	18	sav	all	108	600.02	76.78%
Hexose_Transport	18	fro	phy	96	600.04	2.64%
Hexose_Transport	18	fro	nonphy	12	600.08	7.39%
Hexose_Transport	18	fro	all	108	600.11	4.93%
Kinases	7	sav	phy	31	0.65	
Kinases	7	sav	nonphy	100	0.08	
Kinases	7	sav	all	131	0.09	
Kinases	7	fro	phy	31	1.04	
Kinases	7	fro	nonphy	100	0.81	
Kinases	7	fro	all	131	0.81	
Permeases	17	sav	nonphy	97	600.04	76.92%
Permeases	17	sav	all	97	600.06	76.92%
Permeases	17	fro	nonphy	97	600.01	4.49%
Permeases	17	fro	all	97	600.03	4.49%
Putative_Helicases	11	sav	nonphy	96	481.55	

Putative_Helicases	11	sav	all	96	481.50	
Putative_Helicases	11	fro	nonphy	96	600.01	0.42%
Putative_Helicases	11	fro	all	96	600.02	0.42%

Table 3: Run times for gene family analysis tree fitting. Each row corresponds to the MIP approximation problem for the given family and approximation norm. p is the size of the gene family, n is the number of replicates in the data matrix, and *class* indicates which class of experiments are included in the data matrix. Time reported is CPU user time in seconds. For those MIPs reaching the 10 minute time limit, we report the relative optimality gap of the returned solution.

C Implementation Details

In this paper we use CPLEX 9.0 (Ilog, 2003) to solve the mixed-integer programs described above. This solver allows the user to specify a number of options to control the behavior of the branch-and-cut algorithm. Some of the options that we found to be very useful to solve these projection problems are the following:

1. `MIP_EMPHASIS`: The default behavior in CPLEX is to balance the traversal of the search tree to both tighten the lower bound of the optimum and find integer-feasible solutions. Since the set of tree-structured covariance matrices is non-empty, we know there exists an integer-feasible solution. Therefore, we specify that the emphasis should be solely in tightening the lower bound.

2. `VARSEL` and `NODESEL`: These parameters determine the order in which the search tree is traversed. `VARSEL` determines which variables are branched on while `NODESEL` determines the order in which nodes in the search tree are explored. We set `VARSEL` to *strong branching* so that a small number of branches are explored quickly before deciding which one to take. We set `NODESEL` to *best estimate* where an estimate of the optimum value for integer-feasible solutions under this node is used to determine order.

3. `DISJCUTS` and `FLOWCOVERS`: These parameters controls how often *disjunctive* and *flowcover* cutting planes are generated. We set both to *generate aggressively.*

4. `PROBE` *Probing* is a preprocessing step where the logical implications of setting binary variables to 1 or 0 are explored. We set this parameter to the maximum level of probing.

The determinant maximization Problem (6) is solved using the SDPT3 Tütüncü et al. (2003) semidefinite programming solver. Except for this problem, all experiments and analyses were carried out in R (R Development Core Team, 2007), and many utilities of the `ape` package (Paradis et al., 2004) were used. CPLEX was used through an interface to R written by the authors available at `http://cran.r-project.org/web/packages/Rcplex/`. An R package including the MIP projection solvers will be made available by the authors. Since CPLEX is proprietary software, our published code will also allow the use of the Rsymphony interface (`http://cran.r-project.org/web/packages/Rsymphony/index.html`) to the SYMPHONY MILP solver (`http://www.coin-or.org/SYMPHONY/`).

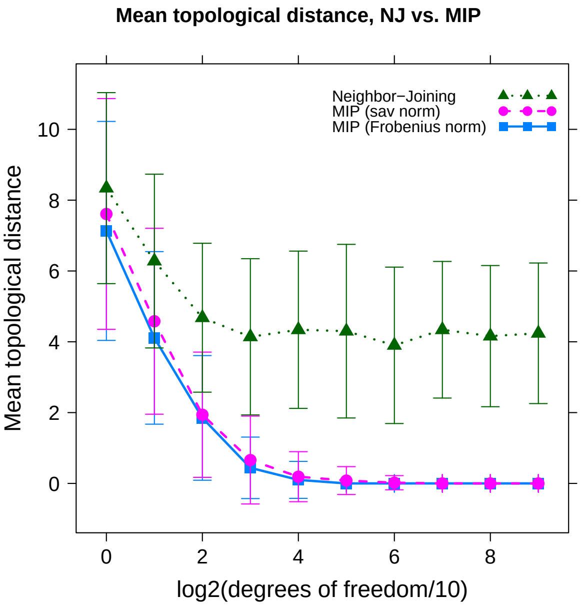


Figure 6: Mean topological distance between estimated and true tree-structured covariance matrices.