Estimating Tree-Structured Covariance Matrices via Mixed-Integer Programming

Héctor Corrada Bravo
Department of Biostatistics
Johns Hopkins Bloomberg
School of Public Health
Baltimore, MD 21205

Stephen Wright
Department of Computer Sciences
University of Wisconsin-Madison
Madison, WI 53706

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

Abstract

We present a novel method for estimating tree-structured covariance matrices directly from observed continuous data. Specifically, we estimate a covariance matrix from observations of p continuous random variables encoding a stochastic process over a tree with p leaves. 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, our estimates are 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 present a case study in phylogenetic analysis of gene expression and a simulation study comparing our method to distance-based tree estimating procedures.

1 INTRODUCTION

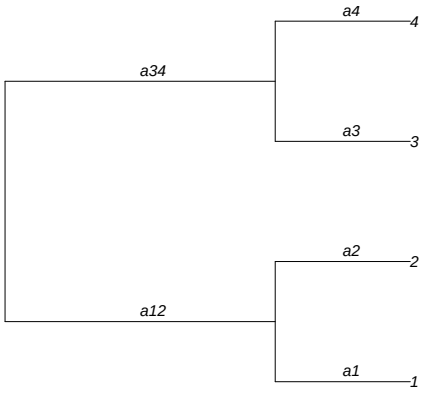
In this paper, we formulate the problem of estimating a tree-structured covariance matrix from observations of multivariate continuous random variables as mixed-integer programs (MIP) (Bertsimas and Weismantel, 2005; Wolsey and Nemhauser, 1999). Specifically, we estimate the covariance matrix of p continuous random variables that encode a stochastic process over a tree where the p variables are observed at the leaves. In particular, we look at estimates that arise from projection problems that compute the nearest tree-structured matrix to the observed sample covariance. These projection 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.

Our setting is similar to the well-known problem of Chow and Liu (1968) except for two key differences: a) in the Chow-Liu setting all variables are observed while in our case we assume observations are made only at the leaves of the tree, b) the stochastic model under which data is assumed to be generated is different in our setting where we assume a continuous stochastic process over a tree (see Section 2) as opposed to the tree specifying a set of second-order conditional independence statements, that is, tree branch lengths are informative in our case. Our motivation for this method is the discovery of phylogenetic structure directly from gene expression data. 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). Typically, these methods first estimate a phylogenetic tree from DNA or amino acid 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 the expression of phylogenetically related objects. However, recent results have shown that the hierarchical structure of sequence-derived tree estimates is highly sensitive to the genomic region chosen to build them. To circumvent this difficulty, we propose a sta-

Appearing in Proceedings of the 12^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5.

ble 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.

The paper is organized as follows. In Section 2 we formulate the representation of tree-structured covariance matrices and give some results regarding the space of such matrices. Section 2.3 shows how to define the constraints that ensure matrices are tree-structured as constraints in mixed-integer programs (MIPs) and formulates projection problems under these constraints. Section 3.1 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. We present our results on a case study on phylogenetic analysis of expression in yeast gene families in Section 3.2. A discussion, including related work, follows in Section 4.



(a)

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

(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.

2 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.

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.

2.1 REPRESENTATION

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.

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 $M^1 = v_1 v_1^T$. 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 $D = \text{diag}(d)$, we can equivalently write $B = VDV^T$. Since the basis matrix V in this expansion 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: 1) V contains the vector of all ones $e = (1,1,\dots,1)^T \in \mathbb{R}^p$ as a column; and 2) for every column w in V with more than one non-zero entry, it contains exactly two 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.

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 defined by tree $\mathcal{T}$, then construct matrix V using the method above starting from the root of $\mathcal{T}$, splitting each vector according to the nested partitions at each node of $\mathcal{T}$. By construction, V will satisfy the partition property and by placing branch lengths, which are non-negative by definition, 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. ■

2.2 CHARACTERISTICS

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 this in the case $p = 3$ by considering the following example. Defining

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

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.

2.3 PROJECTION BY 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.

Every tree-structured covariance matrix satisfies the following properties derived from the linear expansion in Theorem 2: 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 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:

$$B_{ij} \geq B_{ik} = B_{jk} \tag{1a}$$
$$B_{ik} \geq B_{ij} = B_{jk} \tag{1b}$$
$$B_{jk} \geq B_{ij} = B_{ik}.. \tag{1c}$$

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.

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

$$B_{ij} \geq 0 \ \ \forall i, j \tag{3a}$$
$$B_{ii} \geq B_{ij} \ \ \forall i \neq j \tag{3b}$$
$$B_{ij} \geq B_{ik} - (1 - \rho_{ijk1})M \tag{3c}$$
$$B_{ik} \geq B_{jk} - (1 - \rho_{ijk1})M \tag{3d}$$
$$B_{jk} \geq B_{ik} - (1 - \rho_{ijk1})M \tag{3e}$$
$$B_{ik} \geq B_{ij} - (1 - \rho_{ijk2})M \tag{3f}$$
$$B_{ij} \geq B_{jk} - (1 - \rho_{ijk2})M \tag{3g}$$
$$B_{jk} \geq B_{ij} - (1 - \rho_{ijk2})M \tag{3h}$$
$$B_{jk} \geq B_{ij} - (\rho_{ijk11} + \rho_{ijk2})M \tag{3i}$$
$$B_{ij} \geq B_{ik} - (\rho_{ijk11} + \rho_{ijk2})M \tag{3j}$$
$$B_{ik} \geq B_{ij} - (\rho_{ijk11} + \rho_{ijk2})M \tag{3k}$$
$$\rho_{ijk1} + \rho_{ijk2} \leq 1 \tag{3l}$$
$$\rho_{ijk1}, \rho_{ijk2} \in \{0,1\} \ \ \forall\, i > j > k. \tag{3m}$$

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 (1) in a way that the constraint corresponding to the nonzero-valued binary variable must be satisfied. For example, constraint (1a) is transformed to:

$$B_{ij} \geq B_{ik} - (1 - \rho_{ijk1})M \tag{2a}$$
$$B_{ik} \geq B_{jk} - (1 - \rho_{ijk1})M \tag{2b}$$
$$B_{jk} \geq B_{ik} - (1 - \rho_{ijk1})M, \tag{2c}$$

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

We can now give our MIP formulation to the problem of estimating a tree-structured covariance matrix when tree topology is unknown. 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, $\|B\|_{\text{sav}} = \sum_{ij} |B_{ij}|$.

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:

$$\min_{b\in\mathbb{R}^{p(p+1)/2},\rho\in\mathbb{R}^{\bar{p}}} \quad f(b) = \tfrac{1}{2}b^T b - s_2^T b$$
$$\text{s.t.} \quad \text{constraints (3a)- (3m) hold for } B,$$

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

We can similarly find the nearest tree structured covariance matrix in sum-absolute-value (sav) norm. Letting 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 $\hat{b}$ to the MIP above with objective $f(b) = \|s_1 - b\|_1$.

3 EXPERIMENTAL RESULTS

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. We used CPLEX 9.0 (Ilog, 2003) to solve the mixed-integer programs through an interface to R written by the authors available at `http://cran.r-project.org/web/packages/Rcplex/`.

3.1 SIMULATION STUDY

An alternative method to estimate a tree-structured covariance matrix from an observed sample covariance is to use a distance-based tree-building method such as the Neighbor-Joining (NJ) algorithm (Saitou, 1987) on distances derived from a sample covariance B as $D_{ij} = B_{ii} + B_{jj} - 2B_{ij}$. In this simulation, we compared how well do tree-structured covariance matrix estimates reflect the true underlying tree structure of the data when estimated by this NJ-based method versus our MIP-based projection methods. We measured how close tree structures are using the tree topological distance defined by Penny and Hendy (1985) which essentially counts the number of mismatched nested partitions defined by the trees.

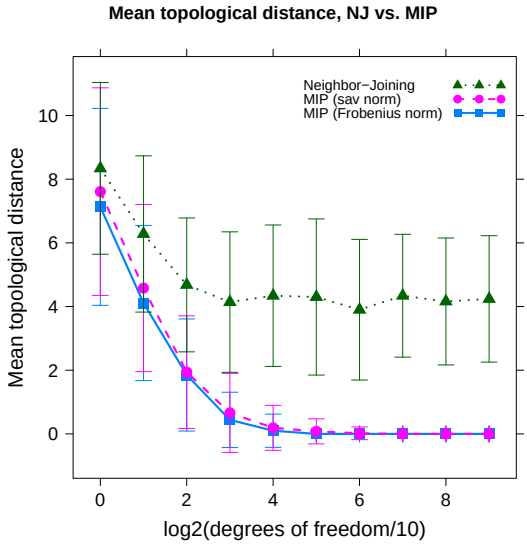
For the simulation we first generated ten trees of size ten at random using the `rtree` function of the `R ape` library (Paradis et al., 2004). This gives ten tree-structured covariance matrices $\{B_1, \ldots, B_{10}\}$ of size 10-by-10. From each tree-structured covariance matrix B_i we drew 10 sample covariances from a Wishart distribution with mean B_i and the desired degrees of freedom df. From each of the 10 sampled matrices we estimated a tree-structured covariance matrix and recorded its topological distance to the true matrix B_i. In Figure 2(a) 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, while the MIP-based method converged to the correct structure for a sample size 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 3.2.

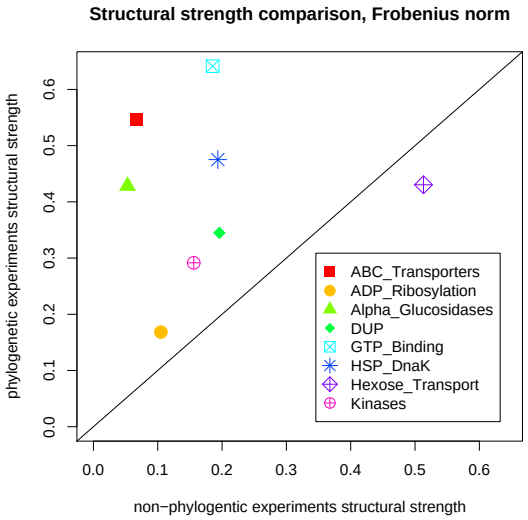
3.2 GENE FAMILY ANALYSIS OF YEAST GENE EXPRESSION

We applied our methods to the analysis of gene expression in *Saccharomyces cerevisiae* gene families data presented in Oakley et al. (2005) and retrieved from `"http://www.lifesci.ucsb.edu/eemb/labs/oakley/pubs/MBE2005data/"`. 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. 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. We refer to Oakley et al. (2005) for further details.

The analysis in Oakley et al. (2005) uses phylogenetic trees estimated from the coding regions of genes in a likelihood model to determine if a gene family shows a phylogenetic effect in each of the 19 experiments. Therefore, 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



(a) Mean topological distance between estimated and true tree-structured covariance matrices.



(b) Comparison of structural strengths for tree-structured covariance estimates $\hat{B}_{gP}$ and $\hat{B}_{gNP}$ for projection under Frobenius norm.

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.

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

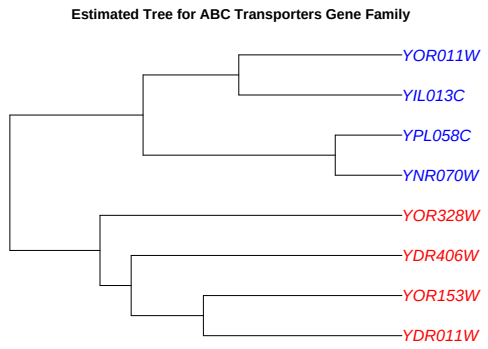
$$SS(B) = \frac{1}{p} \sum_{i=1}^{p} \frac{\max_{i \neq j} B_{ij}}{B_{ii}}. \tag{4}$$

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.

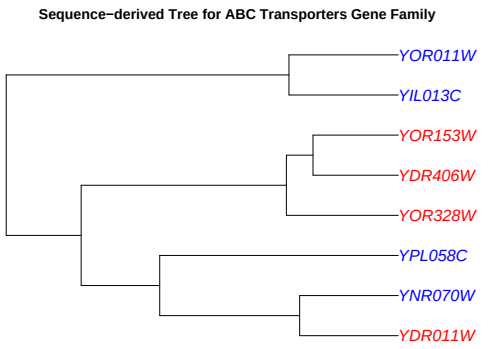
Under the classification of experiments as undergoing phylogenetic versus non-phylogenetic evolution we expect that the structural strength metric should be quite different for estimated tree-structured covariance matrices $\hat{B}_{gP}$ and $\hat{B}_{gNP}$. That is, we expect that $SS(\hat{B}_{gP}) \geq SS(\hat{B}_{gNP})$ for most gene families g. We show our results in Figure 2(b) 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.

For illustrative purposes, we examine the resulting tree for the ABC (ATP-binding cassette) Transporters gene family (see Jungwirth and Kuchler (2006) for a short literature review). The eight genes included in this family 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 2 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 2(c) containing genes YOR328W, YDR406W, YOR153W and YDR011W indicated in red.

To elucidate this topological difference, we turned to the characteristics of the promoter (regulatory) regions of the genes and asked whether transcription factor (TF) binding site contents of the upstream regions



(c)



(d)

Figure 2: (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.

could account for this difference. For each gene, we generated a transcription factor binding site occurrence vector by simply counting the number of occurrences in the 1000 bp upstream of the coding region of each of 128 known yeast transcription factor binding site consensus sequences compiled using Gasch et al. (2004) and the Promoter Database of *Saccharomyces cerevisiae* (SCPD) (`http://rulai.cshl.edu/SCPD/`). From this data we saw 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.

It is known (Delaveau et al., 1994) that the four genes in red in Figure 2 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 2(c) 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-derived tree of Figure 2(d). 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 2(c)) is similar to the sequence-derived tree (Figure 2(d)).

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).

4 DISCUSSION

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.

MIPs have been used to solve phylogeny estimation problems for haplotype data (Brown and Harrower, 2006; Sridhar et al., 2008). 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). In our setting, 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 re-

search 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, 2004). 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).

While we presented the structural strength metric in Section 3.2, 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.

Acknowledgements

Research supported in part by NIH Grant EY09946, NSF Grant DMS-0604572, ONR Grant N0014-06-0095 (HCB, KE, GW); PhRMA Foundation Research Starter Grant in Informatics and NIH RO1 grant HG003747 (SK); NSF Grants DMS-0427689, CCF-0430504, CTS-0456694, CNS-0540147 and DOE Grant DE-FG02-04ER25627 (SW).

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.

C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 14(3):462–467, 1968.

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. 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.

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.

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. *IEEE Transactions on Signal Processing*, 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.

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.