Learning Multivariate Distributions by Competitive Assembly of Marginals

Francisco Sánchez-Vega, Jason Eisner, Laurent Younes, and Donald Geman.
Johns Hopkins University, Baltimore, MD, USA

Abstract—We present a new framework for learning high-dimensional multivariate probability distributions from estimated marginals. The approach is motivated by compositional models and Bayesian networks, and designed to adapt to small sample sizes. We start with a large, overlapping set of elementary statistical building blocks, or "primitives", which are low-dimensional marginal distributions learned from data. Each variable may appear in many primitives. Subsets of primitives are combined in a lego-like fashion to construct a probabilistic graphical model; only a small fraction of the primitives will participate in any valid construction. Since primitives can be precomputed, parameter estimation and structure search are separated. Model complexity is controlled by strong biases; we adapt the primitives to the amount of training data and impose rules which restrict the merging of them into allowable compositions. The likelihood of the data decomposes into a sum of local gains, one for each primitive in the final structure. We focus on a specific subclass of networks which are binary forests. Structure optimization corresponds to an integer linear program and the maximizing composition can be computed for reasonably large numbers of variables. Performance is evaluated using both synthetic data and real datasets from natural language processing and computational biology.

Index Terms—graphs and networks, statistical models, machine learning, linear programming

1 Introduction

Probabilistic graphical models provide a powerful tool for discovering and representing the statistical dependency structure of a family of random variables. Generally, these models exploit the duality between conditional independence and separation in a graph in order to describe relatively complex joint distributions using a relatively small number of parameters. In particular, such graded models are potentially well-adapted to small-sample learning, where the bias-variance trade-off makes it necessary to invest in model parameters with the utmost care. Learning models with a very reduced number of samples is no more difficult than with a great many. However, arranging for such models to generalize well to unseen sets of observations, i.e., preventing them from overfitting the training data, remains an open and active area of research in the small-sample domain.

The introduction of carefully chosen topological biases, ideally consistent with prior domain knowledge, can help to guide learning and avoid model overfitting. In practice, this can be accomplished by accepting a restricted set of graph structures as well as by constraining the parameter space to only encode a restricted set of dependence statements. In either case, we are talking about the design of a model class in anticipation of efficient learning.

Our model-building strategy is "compositional" in the sense of a lego-like assembly. We begin with a set of "primitives" — a large pool of low-dimensional, candidate distributions. Each variable may appear in many primitives and only a small fraction of the primitives will participate in any allowable construction. A primitive designates some of its variables as input (α variables) and others as output (ω variables). Primitives can be recursively merged into larger distributions by matching inputs with outputs: in each merge, one primitive is designated the "connector", and the other primitives' α variables must match a subset of the connector's ω variables. Matched variables lose their α and ω designations in the result. The new distribution over the union of variables is motivated by Bayesian networks, being the product of the connector's distribution with the other primitives' distributions conditioned on their α nodes. In fact, each valid construction is uniquely identified with a directed acyclic graph over primitives.

The process is illustrated in Fig. 1 for a set of fourteen simple primitives over twelve variables. This figure shows an example of a valid construction with two connected components using six of the primitives. The form of the corresponding twelve-dimensional probability distribution will be explained in the text. Evidently, many other compositions are possible.

We seek the composition which maximizes the likelihood of the data with respect to the empirical distribution over the training set. Due to the assembly pro-

- *F. Sánchez-Vega, L. Younes and D. Geman are with the Department of Applied Mathematics and Statistics, the Center for Imaging Science and the Institute for Computational Medicine, Johns Hopkins University, 3400 N. Charles St., Baltimore, MD 21218. E-mail: sanchez, laurent.younes, geman@jhu.edu*

- *J. Eisner is with the Department of Computer Science and the Center for Language and Speech Processing, Johns Hopkins University, 3400 N. Charles St., Baltimore, MD 21218. E-mail: jason@cs.jhu.edu*

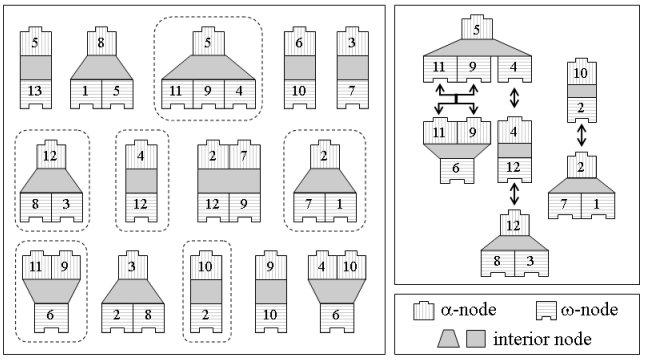

Fig. 1. Simple example of primitives and assemblies.

cess, the global "score" (i.e., expected log-likelihood) of any valid composition decomposes into a sum of local scores, one for each participating primitive; these scores are themselves likelihood ratios corresponding to the gain incurred in fusing the individual variables into the primitive distribution relative to independent variables. This decomposition has several consequences. First, all scores can be precomputed; consequently, parameter estimation (building primitives) and model construction (competitive assembly) are separated. That is, once the primitives are learned the process is data-independent. Second, in many cases, searching over all decompositions for valid compositions can be done by integer linear programming.

The primary intended application is molecular network modeling in systems biology, where it is common to encounter a complex underlying dependency structure among a large number of variables and yet a very small number of samples, at least relative to other fields such as vision and language. DNA microarrays provide simultaneous snap-shot measurements of the levels of expression of thousands of genes inside cells [1], [2], [3]. However, the number of profile measurements per experimental study remains quite small, usually fewer than a few hundreds. Similarly, advances in genotyping microarrays currently make it possible to simultaneously detect single nucleotide polymorphisms (SNPs) over millions of loci practically covering the entire genome of an organism, while the number of individuals in any given study remains orders of magnitude smaller [4], [5], [6]. Thus, any attempt to infer generalizable multivariate distributions from these data, in particular correlation patterns or even higher-dimensional interactions, must deal with well-known trade-offs in computational learning theory between sample size and model complexity [7], and between bias and variance [8].

Our proposals for model-building and complexity control are illustrated with both synthetic and real data. In the former case, experiments include measuring the KL divergence between the optimal composition and the true distribution as a function of the sample size and the number of variables. We compare our graphs with several well-known methods for "reverse engineering" networks, including relevance networks [9], ARACNE [10], CLR [11], which infer graphs from data, and the K2 algorithm [12] for learning Bayesian networks. We present two real-data experiments. One is based on inferring a semantic network from text. The other involves learning dependencies among mutations of the gene *TP53*, which plays a central role in cancer genomics. The substructures in the optimal composition appear biologically reasonable in the sense of aggregating driver mutations and being significantly enriched for certain cell functions. Still, we view our main contribution as methodological, these experiments being largely illustrative.

After discussing related work in Section 2, we will present the general theoretical framework for our models in Section 3, followed by specialization to a specific subclass based on balanced binary trees. In Section 4, we will discuss the choice of a statistically significant set of primitives. These primitives are combined to build the graph structure that maximizes the empirical likelihood of the observed data under a given set of topological constraints. In Section 5 we will show how the corresponding optimization problem can be dealt with using either greedy search or a more efficient integer linear programming formulation. Section 6 discusses the relationship of the foregoing approach to maximum *a posteriori* estimation of graphical model structure and parameters. After this, Section 7 presents some results from synthetic data simulations. In Section 8 we will look at further results obtained using the *20newsgroups* public dataset and the IARC TP53 Database. Finally, we will provide a general discussion and we will sketch some directions for future research.

2 Related Work

Historically, the problem of finding an optimum approximation to a discrete joint probability distribution has been addressed in the literature during the last half century [13]. A seminal paper published by Chow and Liu in the late sixties already proposed the use of information theoretic measures to assess the goodness of the approximation and formulated the structure search as an optimization problem over a weighted graph [14]. Improvements to the original algorithm [15] as well as extensions beyond the original pairwise approach [16] have been proposed. Recently, the popularity of Bayesian networks combined with the need to face small-sample scenarios have led to several works where structural biases are imposed upon the graphs used to approximate the target distribution in order to facilitate learning. Bounded treewidth is an example of such structural constraints. Even though the initial motivation for this approach was to allow for efficient inference [17], [18], [19], there has been work on efficient structure learning [20] and work that uses this kind of bias to avoid model overfitting [21]. Other examples of structural bias aimed

at achieving better generalization properties are the use of L1 regularization to keep the number of edges in the network under control [22], and constraints provided by experts [23]. We will discuss in Section 6 how our method is related to these prior approaches.

Compositional representations of entities as hierarchies of elementary, reusable parts that are combined to form a larger whole constitute an active topic of research in computer vision. Such modeling frameworks are usually built upon a set of composition rules, based on parts and combinations of parts, that progressively define the likelihood of images given the presence of an object at a given pose [24], [25]. A very simple composition rule, based on each part voting for the presence of a relevant object around its location, under the assumption of complex poses, has been proposed in [26]. The hierarchical structures induced by this kind of aggregation procedures provide a convenient tool for hardwiring high-level contextual constraints into the models [27], [28], [29], [30].

Dependency networks, which were proposed in [31] as an alternative to standard Bayesian networks, also provide an interesting example of compositional modeling, since they are built by first learning a set of small graph substructures with their corresponding local conditional probabilities. These "parts" are later combined to define a single joint distribution using the machinery of Gibbs sampling. In any case, the idea of combining compositional modeling and Bayesian networks dates back to the nineties, with the multiply sectioned Bayesian networks (MSBNs) from [32] and the object-oriented Bayesian networks (OOBNs) from [33]. Both approaches, as our work, provide ways to combine a number of elementary Bayesian networks in order to construct a larger model. The final structure can be seen as a hypergraph where hypernodes correspond to those elementary building blocks and hyperlinks are used to represent relations of statistical dependence among them. Hyperlinks are typically associated to so-called "interfaces" between the blocks, which correspond to non-empty pairwise intersections of some of their constituting elements. Even though the actual definition of interface may vary, it usually involves a notion of d-separation of nodes at both of its sides within the network. Later on, the use of a relational structure to guide the learning process [34] and the introduction of structured data types based on hierarchical aggregations [35] (anticipated in [33]) led to novel families of models.

All of the above approaches must confront the structure search problem. That is, given a criterion for scoring graphical models of some kind over the observed variables, how do we *computationally* find the single highest-scoring graph, either exactly or approximately? Structure search is itself an approximation to Bayesian model averaging as in [36], but it is widely used because it has the computational advantage of being a combinatorial optimization problem. In the case of Bayesian networks, Spirtes et al. [37, chapter 5] give a good review of earlier techniques, while Jaakkola et al. [38] review more recent alternatives including exact ones. Like many of these techniques (but unlike the module network search procedure in [39]), ours can be regarded as first selecting and scoring a set of plausible building blocks and only then seeking the structure with the best total score [23]. We formalize this latter search as a problem of integer linear programming (ILP), much as in [38], even if our building blocks have, in general, more internal structure. However, in the particular case that we will present in this paper, we search over more restricted assemblies of building blocks, corresponding to trees (generalizing [14]) rather than DAGs. Thus, our ILP formulation owes more to recent work on finding optimal trees, e.g., non-projective syntax trees in computational linguistics [40].

3 Competitive Assembly of Marginals

In this section, we formulate structure search as a combinatorial optimization problem — *Competitive Assembly of Marginals* (CAM) — that is separated from parameter estimation. The family of models that we will consider is partially motivated by this search problem. We also present a specific subclass of model structures based on balanced binary forests.

3.1 General Construction

Our objective is to define a class of multivariate distributions for a family of random variables $\mathbf{X} = (X_i, i \in D)$ for $D = \{1, \ldots, d\}$, where X_i takes values in a finite set Λ. For simplicity, we will assume that all the X_i have the same domain, although in practice each X_i could have a different domain Λ_i. We shall refer to the elements of Λ^D as configurations.

First we mention the possibility of global, structural constraints: for each $S \subset D$, we are given a class $\mathcal{M}_S$ of "admissible" probability distributions over the set of subconfigurations Λ^S (or over S, with some abuse). Our construction below will impose $\mathcal{M}_S$ as a constraint on all joint distributions that we build over S. To omit such constraint, one can let $\mathcal{M}_S$ consist of all probability distributions on S. Let $\mathcal{M}^* = \bigcup_{S \subset D} \mathcal{M}_S$. If $\pi \in \mathcal{M}^*$, we will write $J(\pi)$ for its support, i.e., the uniquely defined subset of D such that $\pi \in \mathcal{M}_{J(\pi)}$.

3.1.1 Primitives as Elementary Building Blocks

The main ingredient in our construction is a family $\mathcal{T}_0$ of relatively simple probability distributions that we call *primitives*. A distribution over $\mathbf{X}$ will be in our class only if it factors into a product of conditional distributions each of which is specified by a primitive. The elements of $\mathcal{T}_0$ are triplets $\phi = (\pi, A, O)$ where $\pi \in \mathcal{M}^*$ and A, O are subsets of $J(\pi)$ that serve as "connection nodes." A set of five primitives is shown in Fig. 2. The variables (or "nodes") in A will be called

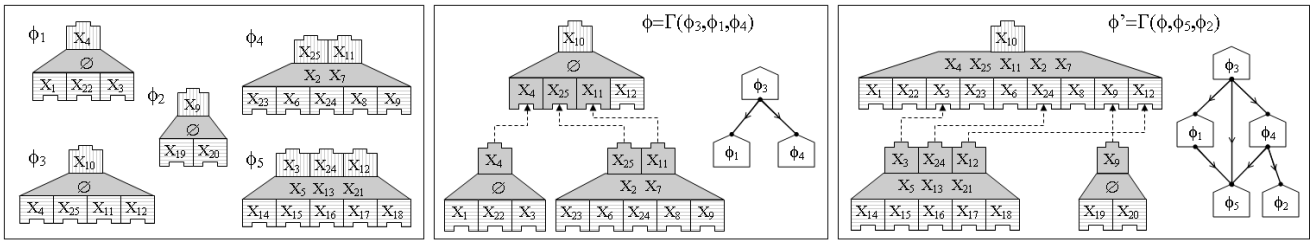

Fig. 2. Example of primitives and merge operations. Left panel shows a set of 5 primitives built from a set of $|D| = 25$ random variables. Center panel illustrates a merge operation where primitives ϕ_1 and ϕ_4 are bound using primitive ϕ_3 as a connector. The resulting new assembly $\phi = \Gamma(\phi_3, \phi_1, \phi_4)$ is shown, as well as its associated primitive DAG (where primitives are drawn as pentagons). Right panel shows the same diagrams for a second merge $\phi' = \Gamma(\phi, \phi_5, \phi_2)$ where the recently created assembly ϕ acts as a connector to bind ϕ_5 and ϕ_2.

α-nodes (A being the *α-set* of primitive ϕ), and the variables in O will be called *ω-nodes* (O being the *ω-set* of ϕ). We require $A, O \neq \emptyset$ and $A \cap O = \emptyset$. Any other nodes, $J(\pi) \setminus (A \cup O)$, are the *interior nodes* of ϕ.

3.1.2 Compositional Modeling by Primitive Merges

One can combine primitives (conditioning on their α-sets as needed) to build more complex distributions, which also have α- and ω-nodes. The merging of three primitives (ϕ_3, ϕ_1, ϕ_4) is illustrated in the second panel of Fig. 2. Here ϕ_3 serves as the binding primitive, or *connector*. The merged assembly ϕ is shown at the top of the third panel and has $|A| = 1, |O| = 9$.

Formally, given a group of primitives $(\phi_0, \phi_1, \ldots, \phi_r)$, where each $\phi_k = (\pi_k, A_k, O_k)$, we define a *merged* distribution π over $S = \bigcup_{k=0}^{r} J(\pi_k)$ as:

$$\pi(x_S) = \pi_0(x_{J(\pi_0)}) \prod_{k=1}^{r} \pi_k(x_{J(\pi_k)} \mid x_{A_k}) \qquad (1)$$

where ϕ_0 serves as the connector. (Here and below, x_I (for $I \subset D$) denotes the restriction of x to I, and we abuse notation by designating joint and conditional probabilities using the same letter as the original probability, the set over which they are defined being implicitly assigned by the variables.)

To ensure that (1) defines a proper distribution, we may only merge $(\phi_0, \phi_1, \ldots, \phi_r)$ when

(M1) $J(\pi_k) \cap J(\pi_0) = A_k \subset O_0$, for all $k = 1, \ldots, r$.
(M2) $J(\pi_k) \cap J(\pi_l) = \emptyset$ for all $k, l = 1, \ldots, r$, $k \neq l$.

(M1) ensures that the α-set of each ϕ_k matches some subset of the connector's ω-set. Together with (M2), it also ensures that, aside from those matchings, primitives $(\phi_0, \phi_1, \ldots, \phi_r)$ are over disjoint sets of variables.

We will say that the group $(\phi_0, \phi_1, \ldots, \phi_r)$ is *mergeable* with ϕ_0 as connector if (M1) and (M2) are satisfied and $\pi \in \mathcal{M}_S$. We then define the resulting merge to be the triplet $\phi = (\pi, A, O)$ with $A = A_0$ and $O = \bigcup_{k=0}^{r} O_k \setminus \bigcup_{k=1}^{r} A_k$. So the α-set of a merge is the α-set of the connector, and its ω-set is the union of the original ω-sets, from which the α-nodes of the non-connector primitives are removed (and become interior nodes in the new structure). This merge or output will be denoted $\phi = \Gamma(\phi_0, \ldots, \phi_r)$.

The merge operation can now be iterated to form probability distributions over increasingly large subsets of variables. This is illustrated in the last panel of Fig. 2, where the merge from the second panel is itself merged with two of the original primitives. For $S \subset D$, we will denote by $\mathcal{T}_S^*$ the set of probability distributions on S that can be obtained by a sequence of merges as defined above. If S is a singleton, we let, by convention, $\mathcal{T}_S^* = \mathcal{M}_S$. Finally, we let $\mathcal{T}^* = \bigcup_{S \subset D} \mathcal{T}_S^*$.

We would define our final model class $\mathcal{F}^*$ (of distributions over D) as $\mathcal{T}_D^*$, except that each distribution in $\mathcal{T}_D^*$ consists of a single connected component. Instead we define $\mathcal{F}^*$ as all product distributions of the form

$$P(x) = \prod_{k=1}^{c} \pi_k(x_{J(\pi_k)}) \qquad (2)$$

where $J(\pi_1), \ldots, J(\pi_c)$ partition D and $\pi_k \in \mathcal{T}_{J(\pi_k)}^*$.

The size and complexity of $\mathcal{F}^*$ are limited by two choices that we made earlier: the set of initial primitives $\mathcal{T}_0$, and the set of admissible distributions $\mathcal{M}^*$. Note that for $S \subsetneq D$, the constraints imposed by $\mathcal{M}_S$ on intermediate merges may be redundant with the final constraints imposed by $\mathcal{M}_D$ (as in Section 3.3 below), or may instead act to further restrict $\mathcal{F}^*$.

The final distribution P is a product of (conditionalized) primitives, whose relationships can be captured by a directed acyclic graph. Indeed, in view of (1), there is an evident connection with Bayesian networks which is amplified in the proposition below.

3.1.3 Atomic Decompositions and Primitive DAGs

Given $\Psi = \{\psi_1 \ldots, \psi_N\} \subset \mathcal{T}_0$, we will let $\mathcal{T}_\Psi$ denote the subset of $\mathcal{T}^*$ obtained by iterations of merge operations involving only elements of Ψ or merges built from them. Equivalently, $\mathcal{T}_\Psi$ is the set of distributions obtained by replacing $\mathcal{T}_0$ by Ψ in the construction above. If $\phi \in \mathcal{T}^*$, we will say that a family of primitives $\Psi = \{\psi_1 \ldots, \psi_N\} \subset \mathcal{T}_0$ is an *atomic decomposition* of ϕ if Ψ is a minimal subset of $\mathcal{T}_0$ such that $\phi \in \mathcal{T}_\Psi$ (i.e., ϕ can be built by merges involving only elements of Ψ and all elements of Ψ are needed in this construction). If $P \in \mathcal{F}^*$ is decomposed as in (2), an atomic decomposition of P is a union of atomic decompositions of each of its independent components. Finally, let $\mathcal{T}_0^*$ (resp. $\mathcal{F}_0^*$) be the set of

atomic decompositions of elements of $\mathcal{T}^*$ (resp. $\mathcal{F}^*$). The set of roots in the atomic decomposition $\Psi \in \mathcal{T}_0^*$ is denoted R_Ψ and defined as the set of indices k such that $A_k \cap J(\pi_l) = \emptyset$ for all $k, l = 1, \ldots, r$, $k \neq l$.

Proposition 1. *Let $\Psi = \{\psi_1, \ldots, \psi_N\} \in \mathcal{T}_0^*$ with $\psi_k = (\pi_k, A_k, O_k)$ and define the directed graph $G(\Psi)$ on $\{1, \ldots, N\}$ by drawing an edge from k to l if and only if $A_l \subset O_k$. Then $G(\Psi)$ is acyclic.*

This proposition is part of a larger one, Proposition S.1, which is stated and proved in Appendix A (see supplemental material). The acyclic graph $G(\Psi)$ is illustrated in Fig. 2 for the merges depicted in the middle and right panels. Notice that the nodes of these graphs are primitives, not individual variables. Consequently, our models are Bayesian networks whose nodes are overlapping *subsets* of our variables $\mathbf{X}$.

3.1.4 Generalization Using Weaker Merging Rules

We remark that the constraints defining merging rules could be relaxed in several ways, resulting in less restricted model families. For example, one could replace (M2) by the weaker condition that supports of non-connectors may intersect over their α-sets, i.e.,

(M2)′ $J(\pi_k) \cap J(\pi_l) \subset A_k \cap A_l$.

Similarly, one can remove the constraint that ω-sets do not intersect α-sets within a primitive, allowing for more flexible connection rules, as defined by (M1) (the ω-set after merging would then simply be the union of all previous ω-sets, without removing the α-sets). Such modifications do not affect the well-posedness of (1). An extreme case is when primitives contain all possible pairs of variables, (M2) is replaced by (M2)′ and the ω-set constraint is relaxed. Then our model class contains all possible Bayesian networks over the variables (i.e., all probability distributions).

3.2 Likelihood

We now switch to a statistical setting. We wish to approximate a target probability distribution P^* on Λ^D by an element of $\mathcal{F}^*$. This will be done by minimizing the Kullback-Leibler divergence between P^* and the model class. Equivalently, we maximize

$$L(P) = E_{P^*}(\log P) = \sum_{x \in \Lambda^d} P^*(x) \log P(x), \qquad (3)$$

where $P \in \mathcal{F}^*$. Typically, P^* is the empirical distribution obtained from a training set, in which case the procedure is just maximum likelihood.

Let each primitive distribution be parametric, $\phi = (\pi(\cdot; \theta), A, O)$, where θ is a parameter defined on a set Θ_ϕ (which can depend on ϕ). From the definition of merge, it is convenient to restrict the distributions in $\mathcal{F}^*$ by separately modeling the joint distribution of the α-nodes and the conditional distribution of the other nodes given the α-nodes. Therefore, we assume $\theta = (\sigma, \tau)$, where the restriction of π to A only depends on σ and the conditional distribution on $J(\pi) \setminus A$ given x_A only depends on τ, i.e.,

$$\pi(x_{J(\pi)}; \theta) = \pi(x_A; \sigma) \pi(x_{J(\pi) \setminus A} \mid x_A; \tau).$$

We assume that single-variable distributions are unconstrained, i.e., there is a parameter $P_j(\lambda)$ for each $\lambda \in \Lambda, j \in D$.

In order to maximize L, we first restrict the problem to distributions $P \in \mathcal{F}^*$ which have an atomic decomposition provided by a fixed family $\Psi = \{\psi_1, \ldots, \psi_N\} \in \mathcal{F}_0^*$. Afterwards, we will maximize the result with respect to Ψ. Let $\theta_k = (\sigma_k, \tau_k) \in \Theta_{\psi_k}, k = 1, \ldots, N$ and let $\ell(\theta_1, ..., \theta_N)$ be the expected log-likelihood (3). Rewriting the maximum of ℓ based on likelihood ratios offers an intuitive interpretation for the "score" of each atomic decomposition in terms of individual likelihood gains relative to an independent model. For any primitive $\phi = (\pi(\cdot; \theta), A, O)$, define

$$\begin{aligned} \rho(\phi) \quad = \quad & \max_\theta E_{P^*} \log \pi(X_{J(\pi)}; \theta) \qquad (4) \\ & - \max_\sigma E_{P^*} \log \pi(X_A; \sigma) + \sum_{j \in J(\pi) \setminus A} H(P_j^*), \end{aligned}$$

where $H(P_j^*) = -E_{P_j^*}(\log P_j^*)$. This is the expected log-likelihood ratio of the estimated parametric model for Ψ and an estimated model in which i) all variables in $J \setminus A$ are independent and independent from variables in A, and ii) the model on A is the original one (parametrized with σ). We think of this as an internal binding energy for primitive ϕ. Similarly, define

$$\mu(\phi) = \max_\sigma E_{P^*} \log \pi(X_A; \sigma) + \sum_{j \in A} H(P_j^*), \qquad (5)$$

the expected likelihood ratio between the (estimated) model on A and the (estimated) model which decouples all variables in A. Then it is rather straightforward to show that the maximum log-likelihood of any atomic decomposition decouples into primitive-specific terms. The proof, which resembles that of the Chow-Liu theorem [14], is provided in Appendix B.

Proposition 2. *Let $\ell(\theta_1, \cdots, \theta_N)$ be the expected log-likelihood of the composition generated by $\Psi = (\psi_1, \ldots, \psi_N) \in \mathcal{F}_0^*$. Then*

$$\max_{\theta_1, \ldots, \theta_N} \ell(\theta_1, \cdots, \theta_N) = \ell^*(\Psi) - \sum_{j \in D} H(P_j^*)$$

where

$$\ell^*(\Psi) = \sum_{k \in R_\Psi} \mu(\psi_k) + \sum_{k=1}^{N} \rho(\psi_k). \qquad (6)$$

Since the sum of entropies does not depend on Ψ, finding an optimal approximation of P^* "reduces" to maximizing ℓ^* over all possible Ψ. More precisely, finding an optimal approximation requires computing

$$\hat{\Psi} = \operatorname*{argmax}_{\Psi \in \mathcal{F}_0^*} \ell^*(\Psi) \qquad (7)$$

(with optimal parameters in (4) and (5)).

The important point is that the values of all ρ's and μ's can be precomputed *for all primitives*. Consequently, due to (6), any valid composition can be scored based only on the contributing primitives. In this way, parameter estimation is separated from finding the optimal Ψ. Obviously, the constraint $\Psi \in \mathcal{F}_0^*$ is far from trivial; it requires at least that Ψ satisfy conditions (i) and (ii) in Proposition S.1 (Appendix A). Moreover, computing $\hat{\Psi}$ typically involves a complex combinatorial optimization problem. We will describe it in detail for the special case that is our focus.

3.3 Balanced Compositional Trees

We now specialize to a particular set of constraints and primitives. In everything that follows, we will assume a binary state space, $\Lambda = \{0, 1\}$, and restrict the admissible distributions $\mathcal{M}^*$ to certain models that can be represented as trees (or forests), for which we introduce some notation. We call this subclass of models *balanced compositional trees*.

If T is a tree, let $J(T) \subset D$ denote the set of nodes in T. For $s \in J(T)$, s^+ denotes the set of children of s and s^- the set of its parents. Because T is a tree, s^- has only one element, unless s is the root of the tree, in which case $s^- = \emptyset$. We will say that T is *binary* if no node has more than two children, i.e., $|s^+| \leq 2$ (we allow for nodes with a single child). If $s^+ = \emptyset$ then s is called a terminal node, or leaf, and we let $\mathcal{L}(T)$ denote the set of all leaves in T. If s has two children, we will arbitrarily label them as left and right, with notation $s.l$ and $s.r$. Finally, T_s will denote the subtree of T rooted at s, i.e., T restricted to the descendants of s (including s). We will say that T is *almost-balanced* if for each $s \in J(T)$ such that $|s^+| = 2$, the number of leaves in $T_{s.l}$ and $T_{s.r}$ differ in at most one unit.

Probability distributions on $\Lambda^{J(T)}$ associated with T are assumed to be of the form:

$$\pi(x_{J(T)}) = p_0(x_{s_0}) \prod_{s \in J(T) \setminus \mathcal{L}(T)} p_s(x_{s^+} \mid x_s) \qquad (8)$$

where s_0 is the root of T, p_0 is a probability distribution and $p_s, s \in J(T) \setminus \mathcal{L}(T)$ are conditional distributions. This definition slightly differs from the usual one for tree models in that children do not have to be conditionally independent given parents.

For $S \subset D$, we will let $\mathcal{M}_S$ denote the set of models provided by (8), in which T is an almost-balanced tree such that $J(T) = S$. The balance constraint is introduced as a desirable inductive bias intended to keep the depth of the trees under control. The set $\mathcal{T}_0$ will consist of primitives $\phi = (\pi, A, O)$ where π is a probability distribution over a subset $J \subset D$ with cardinality two or three; A (the α-set) is always a singleton and we let $O = J \setminus A$. Primitives will be selected based on training data, using a selection process that will be described in the next section.

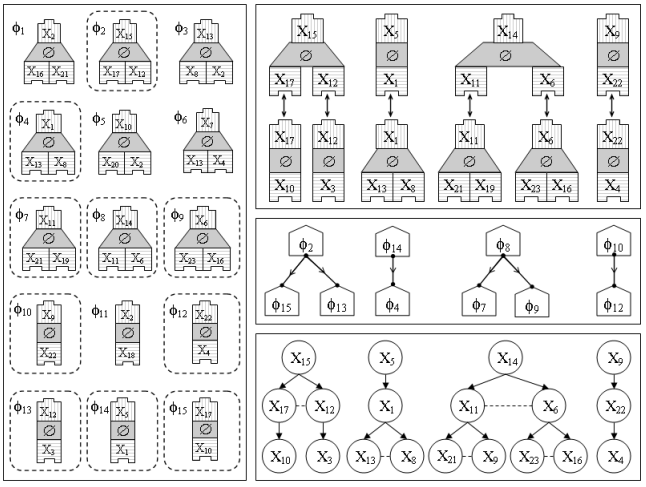

Fig. 3. Illustration of a set of primitives and an atomic decomposition for compositional trees. Left panel shows a pool of 15 primitives built from 23 variables. The encircled ones constitute an atomic decomposition for the four-component model depicted in the top-right panel. The center and bottom right panels show the corresponding DAGs of primitives and variables, respectively. In the last graph, dashed lines are used to link siblings within the same primitive.

Because α-sets have cardinality one, we have $\mu(\phi) = 0$ for all $\phi \in \mathcal{T}_0$ (where μ is defined in (5)), and (6) boils down to maximizing

$$\ell^*(\Psi) = \sum_{k=1}^{N} \rho(\psi_k) \qquad (9)$$

over $\mathcal{F}_0^*$. The description of $\mathcal{F}_0^*$ and of our maximization procedures is given in Section 5.

An example of a set of primitives that can be used to build balanced compositional trees is presented in Fig. 3, together with a sequence of elementary merges.

4 Primitive Selection

From here, we restrict our presentation to the particular case of balanced compositional trees. We discuss selecting an initial set of primitives $\mathcal{T}_0$ and estimating their parameters. We justify the need for a lower bound on the empirical likelihood gain for each accepted primitive and we describe a procedure for choosing this threshold based on controlling the expected number of false positives under the null hypothesis of mutual independence.

4.1 Stepwise Dependency Pursuit

We first specify the allowed representations for primitives, $\phi = (\pi(\cdot; \theta), A, O)$. Primitives are defined over pairs or triplets in D. With pairs, we allow π to be any distribution on $\{0, 1\}^2$. More precisely, letting $A = \{s\}$ and $O = \{t\}$, and using notation from Section 3.2, we let $\sigma = \pi_s(1)$ and $\tau = (\pi_t(1 \mid x_s = 0), \pi_t(1 \mid x_s = 1))$.

For triplets, we introduce several levels of complexity, each adding a single parameter to the joint distribution. Let $A = \{s\}$ and $O = \{u, v\}$. We can make the

joint distribution of (X_s, X_u, X_v) progressively more general with the following steps:

(1)	X_s, X_u, X_v are independent	(3 parameters).
(2)	$X_v \perp (X_s, X_u)$	(4 parameters).
(2′)	$X_u \perp (X_s, X_v)$	(4 parameters).
(3)	$X_u \perp X_v \mid X_s$	(5 parameters).
(4)	$X_u \perp X_v \mid X_s = 0$	(6 parameters).
(4′)	$X_u \perp X_v \mid X_s = 1$	(6 parameters).
(5)	Unconstrained joint	(7 parameters).

Case (1) corresponds to the default singletons, and (2), (2′) involve a pair primitive and a singleton. "True" triplet distributions correspond to (3) through (5). The selection process that we now describe will assign one model to any selected triplet (s, u, v).

If $d = |D|$ is the number of variables, there are $d(d-1)$ possible pairs and $d(d-1)(d-2)/2$ possible triplets. Since we are targeting applications in which d can reach hundreds or more, limiting the pool of primitives is essential to limiting the complexity of both statistical estimation and combinatorial optimization. The selection will be based on a very simple principle: only accept a primitive at a given complexity level when the previous level has been accepted and the expected likelihood increment in passing to the higher-dimensional model is sufficiently large. So, when building a primitive $\phi = (\pi(\cdot;\theta), A, O)$ supported by a set J, with $\theta \in \Theta_\phi$, we will assume a sequence of submodels $\Theta_1 \subset \Theta_2 \subset \cdots \subset \Theta_q$ and let Θ_ϕ be indexed by the largest k such that, for all $l = 1, \ldots, k-1$

$$\max_{\theta \in \Theta_{l+1}} E_{P^*} \log \pi(X_J;\theta) - \max_{\theta \in \Theta_l} E_{P^*} \log \pi(X_J;\theta) \geq \eta$$

where η is a positive constant and P^* is the empirical distribution computed from observations. For example, to form a pair primitive over $J = \{u, v\}$, we compare the joint empirical distribution over J (which estimates three parameters) to the one for which X_u and X_v are independent (which estimates two parameters), and we accept the pair primitive if

$$E_{P^*} \log \frac{P^*(X_u, X_v)}{P^*(X_u)P^*(X_v)} \geq \eta_{(2)}. \tag{10}$$

(For simplicity, we are just writing $P^*(X_u, X_v)$ for the empirical joint distribution of X_u, X_v; in each case the meaning should be clear from context.) In fact, we accept two primitives if this inequality is true: one for which u is the α-node and v the ω-node, and one for which the roles are reversed. Note that selection for pairs reduces to applying a lower bound on the mutual information, the same selection rule as in relevance networks [9].

For triplets, we will apply the analysis to the sequence of models (1), (2)/(2′), (3), … above. For example, to accept a triplet that corresponds to model (3), we first require that model (2) (or (2′)) is preferred to model (1), which implies that either the pair (s, u) or the pair (s, v) is accepted as a primitive using (10). We then demand that model (3) is significantly better than, say, model (2), meaning that

$$E_{P^*} \log \frac{P^*(X_u \mid X_s)P^*(X_v \mid X_s)}{P^*(X_u \mid X_s)P^*(X_v)} \geq \eta_{(3)}. \tag{11}$$

To select model (4), we need model (3) to have been selected first, and then, defining

$$\hat{P}(x_u, x_v \mid x_s) = \begin{cases} P^*(x_u \mid x_s)P^*(x_u \mid x_s) \text{ if } x_s = 0 \\ P^*(x_u, x_v \mid x_s) \text{ if } x_s = 1, \end{cases}$$

we will require

$$E_{P^*} \log \frac{\hat{P}(X_u, X_v \mid X_s)}{P^*(X_u \mid X_s)P^*(X_v \mid X_s)} \geq \eta_{(4)}. \tag{12}$$

Selecting model (5) is done similarly, assuming that either model (4) or (4)′ is selected.

4.2 Determination of the Selection Threshold

The threshold η determines the number of selected primitives and will be based on an estimation of the expected number of false detections. At each step of the primitive selection process, which correspond to the five numbered steps from above, we will assume a null hypothesis consistent with the dependencies accepted up to the previous level and according to which no new dependencies are added to the model. We will define η to ensure that the expected number of detections under this null is smaller than some $\epsilon > 0$, which will be referred to as the *selection threshold*.

We will fix $\eta_{(2)}$ such that the expected number of selected pairs under the assumption that all variables are pairwise independent is smaller than ϵ. Assuming that $m_{(2)}$ pairs have been selected, we will define $\eta_{(3)}$ to ensure that the expected number of selected triplets of type (3) is smaller than ϵ, under the assumption that any triplet of variables must be such that at least one of the three is independent from the others. Similarly, assuming that $m_{(3)}$ triplets are selected, $\eta_{(4)}$ will be chosen to control the number of false alarms under the hypothesis of all candidates following model (3). In some sense, selection at each level is done conditionally to the results obtained at the previous one.

At each step, the expected number of false alarms from model $(k-1)$ to (k) can be estimated by $\hat{\epsilon}$, which is defined as the product of the number of trials and the probability that model (k) is accepted given that model $(k-1)$ is true. Since model (k) is preferred to model $(k-1)$ when the likelihood ratio between the optimal models in each case is larger than $\eta_{(k)}$, $\hat{\epsilon}$ will depend on the distribution of this ratio when the smaller model is true. If the number of observations, n, is large enough, this distribution can be estimated via Wilks' theorem [41] which states that two times the likelihood ratio asymptotically follows a χ^2 distribution, with degrees of freedom given by the number of additional parameters (which is equal to one for each transition).

The number of trials for passing from level (3) to (4) and from level (4) to (5) is the number of selected triplets of the simplest type, i.e., $t_{(3)} = m_{(3)}$ and $t_{(4)} = m_{(4)}$ respectively. Between levels (2) and (3), we make $t_{(2)} = (d-2)m_{(2)}$ trials, and $t_{(1)} = d(d-1)/2$ trials between levels (1) and (2). With this notation, and the fact that each new level involves one additional parameter, we use the following selection process: let $\eta_{(k)}, k = 2, \ldots, 5$ be defined by

$$\eta_{(k)} = \frac{1}{2n} F_{step}^{-1}\left(1 - \frac{\epsilon}{4t_{(k-1)}}\right) \tag{13}$$

where F_{step} is the cumulative distribution function of a χ^2 with 1 d.f. (a standard normal squared) and the factor 4 ensures that the total number of expected false alarms across all levels is no more than ϵ.

For small values of n, the approximation based on Wilks' theorem is in principle not valid. Based on Monte-Carlo simulations, however, we observed that it can be considered as reasonably accurate for $n \geq 20$, which covers most practical cases. When $n < 20$, we propose to choose $\eta_{(k)}$ using Monte-Carlo (for very large values of d, the number of required Monte Carlo replicates may become prohibitively large, but learning distributions for extremely large d and $n < 20$ may be a hopeless task to begin with).

5 Structure Search Algorithm

The procedure defined in the previous section yields the collection, $\mathcal{T}_0$, of building blocks that will be composed in the final model. Each of these blocks, say ψ, comes with their internal binding energy, $\rho(\psi)$, which can be precomputed. The structure search problem, as described in (9), consists in maximizing

$$\ell^*(\Psi) = \sum_{k=1}^{N} \rho(\psi_k)$$

over all groups $\Psi = \{\psi_1, \ldots, \psi_N\} \in \mathcal{F}_0^*$, i.e., all groups of primitives that lead to a distribution on D that can be obtained as a sequence of legal merges on Ψ.

We start by describing $\mathcal{F}_0^*$. Recall that each family $\Psi = \{\psi_1, \ldots, \psi_N\} \subset \mathcal{T}_0$ defines an oriented graph $G(\Psi)$ on D, by inheriting the edges associated to each of the ψ_k's. We have the following fact (the proof is provided in Appendix C, as supplemental material).

Proposition 3. *A family of primitives $\Psi \subset \mathcal{T}_0$ belongs to $\mathcal{F}_0^*$ if and only if*

(i) *The α-nodes of the primitives are distinct.*
(ii) *The primitives do not share edges*
(iii) *$G(\Psi)$ is an almost-balanced binary forest.*

These conditions can be checked without seeking a particular sequence of admissible merges that yields $G(\Psi)$. That is, the structure search problem reduces to maximizing $\ell^*(\Psi)$ over all $\Psi = \{\psi_1, \ldots, \psi_N\}$ such that $G(\Psi)$ is an almost-balanced forest. This is still hard: when the true underlying distribution is rich in dependencies (yielding a large set $\mathcal{T}_0$), the number of possible Ψ's explodes combinatorially as the number of variables increases. Because of this, the exhaustive enumeration of all possible forests is not feasible. We propose two alternatives: a greedy search heuristic and a reformulation of the search as an ILP optimization problem, which can be solved using publicly available software (we worked with the Gurobi optimizer).

5.1 Greedy Search Solution

We begin with an edgeless graph where all variables are treated as singletons, i.e. $\Psi_0 = \emptyset$. The search operates by progressively adding new elements to Ψ until no such option exists. At step k of the procedure, with a current solution denoted Ψ_k, we define the next solution to be $\Psi^{(k+1)} = \Psi_k \cup \{\psi_{k+1}\}$ where ψ_{k+1} is chosen as the primitive for which ρ is maximized over all primitives that complete Ψ_k into a legal structure (and the procedure stops if no such primitive exists). At the end of each step, the set $\mathcal{T}_0$ can be progressively pruned out from all primitives that will never be used to complete the current Ψ_k, i.e., primitives that share an edge, or an α-node, with already selected ψ_j's, or primitives with ω-nodes coinciding with already allocated α-nodes. Of course, this strategy risks getting trapped in local maxima and is not guaranteed to find the optimal global solution.

5.2 Integer Linear Programming Solution

Exact maximization of $\ell^*(\Psi)$ is an ILP problem. Let V be the set of vertices and let $\mathcal{E}$ be the set of (oriented) edges present in $\mathcal{T}_0$. Here, whenever we speak of an edge we refer to edges in the graph structure associated to each primitive, where each node corresponds to a variable (as opposed to hyperedges in the higher level hypergraph where each node corresponds to a different primitive). The graph structure for pair primitives consists of an oriented edge from the α-node to the ω-node. The graph for triplet primitives consists of two oriented edges from the α-node to each of the ω-nodes (as shown in Fig. 3).

Introduce binary selector variables $x_\psi, \psi \in \mathcal{T}_0$ and $y_e, e \in \mathcal{E}$. For $e \in \mathcal{E}$, let $\mathcal{T}_e$ be the set of $\psi \in \mathcal{T}_0$ that contain e. We want to rephrase the conditions in Proposition 3 using linear equalities on the x's, y's and other auxiliary variables. (The meaning of the notation x, y, is different, in this section only, of what it is in the rest of the paper, in which it is used to denote realizations of random variables.)

We formulate the ILP here only in the specific setting of balanced compositional trees (Section 3.3), although the approach generalizes to other cases where $G(\Psi)$ is restricted to be a forest. If we wished to allow $G(\Psi)$ to be any DAG, we would modify the ILP problem to rule out only *directed* cycles [38], [42].

The first constraint is, for all $e \in \mathcal{E}$,

$$\sum_{t \in \mathcal{T}_e} x_t = y_e,$$

which ensures that every selected edge belongs to one and only one selected primitive.

We also need all edges in each selected primitive to be accounted for, which gives, for all $\psi \in \mathcal{T}_0$,

$$(|\psi| - 1)x_\psi \leq \sum_{e \in \psi} y_e$$

where $|\psi|$ is the number of vertices in ψ (two or three).

For every directed edge $e = (v, v')$ with $v, v' \in V$, let its reversal be $\bar{e} = (v', v)$. Our next constraint imposes $y_e + y_{\bar{e}} \leq 1$. Note that this constraint is made redundant by the acyclicity constraint. Still, it may be useful to speed up the solver.

Vertices must have no more than one parent and no more than two children, which gives, for all $v \in V$,

$$\sum_{(v',v) \in \mathcal{E}} y_{(v',v)} \leq 1 \quad \text{and} \quad \sum_{(v,v') \in \mathcal{E}} y_{(v,v')} \leq 2.$$

We also ensure that no vertex is the α-node of two distinct selected binary primitives. For $v \in V$, let Ψ_v denote the subset of $\mathcal{T}_0$ containing binary primitives with $\{v\}$ as an α-node. Then we want, for all $v \in V$

$$\sum_{\psi \in \Psi_v} x_\psi \leq 1.$$

The remaining conditions are more complex and require auxiliary variables. We first ensure that the graph associated to the selected ψ's is acyclic. This can be done by introducing auxiliary flow variables $f_e, e \in \mathcal{E}$ with the constraint

$$\begin{cases} -C(1 - y_e) + y_e + \sum_{e' \to e} f_{e'} \leq f_e \leq y_e + \sum_{e' \to e} f_{e'} \\ 0 \leq f_e \leq C y_e \end{cases}$$

where C is large enough (e.g., $C = |\mathcal{E}|$) and $e' \to e$ means that the child in edge e' coincides with the parent in edge e. (If $y_e = 1$, this condition implies $f_e = 1 + \sum_{e' \to e} f_{e'}$ which is impossible unless $f_e = \infty$ if the graph has a loop.)

The last condition is for balance. Introduce variables $g_e, e \in \mathcal{E}$ and $h_e, e \in \mathcal{E}$ with constraints

$$\begin{cases} 0 \leq h_e \leq y_e \\ h_e \leq 1 - y_{e'} \text{ if } e \to e' \\ h_e \geq 1 - \sum_{e \to e'} y_{e'} - C(1 - y_e) \\ -C(1 - y_e) + \sum_{e \to e'} g_{e'} \leq g_e \leq h_e + \sum_{e \to e'} g_{e'} \\ y_e \leq g_e \leq C y_e \\ \text{for all triplets } \psi,\ |g_{e(\psi)} - g_{e'(\psi)}| \leq 1 + C(1 - x_\psi) \end{cases}$$

where $e(\psi)$ and $e'(\psi)$ denote the two edges in triplet ψ. The variable h_e equals 1 if and only if e is a terminal edge. The variable g_e counts the number of leaves (or terminal edges). We have $g_e = 0$ if $y_e = 0$. If e is terminal and $y_e = 1$, then the sum over descendants vanishes and the constraints imply $g_e = 1$. Otherwise ($h_e = 0$ and $y_e = 1$), we have $g_e = \sum_{e \to e'} g_{e'}$. The last inequality ensures that the trees are almost-balanced.

The original problem can now be solved by maximizing $\sum_{\psi \in \mathcal{T}_0} \rho(\psi) x_\psi$ subject to these constraints, the resulting solution being $\hat{\Psi} = \{\psi : x_\psi = 1\}$.

ILP is a general language for formulating NP-complete problems. Although the worst-case runtime of ILP solvers grows exponentially with the problem size, some problem instances are much easier than others, and modern solvers are reasonably effective at solving many practical instances of moderate size. We show empirical runtimes in Section 2 of the supplemental material, together with an analysis of the size of the ILP encoding. Note that one can improve upon greedy search even without running the ILP solver to convergence, since the solver produces a series of increasingly good suboptimal solutions en route to the global optimum. Also, when the number of variables and constraints in the ILP problem becomes computationally prohibitive, we can adopt a hybrid search strategy: start by running a greedy search (which typically leads to a forest with several independent components) and then solve multiple ILP problems as the one described above, each restricted to $\psi's$ that are supported by the set of variables involved in each of those components. Even though the solution may still not be globally optimal, this coarse-to-fine approach may lead to improved performances over the use of greedy search and ILP alone.

6 Connection with MAP Learning of Bayesian Network Structure

To situate our statistical approach with respect to the prior work of Section 2, we now discuss how it relates to MAP estimation of Bayesian networks.

6.1 Primitives as Bayesian Networks

The approach that we propose in this paper consists in constructing small-dimensional primitives, possibly having complex parametrizations (if allowed by the data-driven selection process), and assembling them globally into a model covering all variables subject to complexity contraints. Note that, even though the global relationship among primitives is organized as a Bayesian network, as described in Section 3.1, the distribution specified by each primitive can be modeled in an arbitrary way. We conceive of these primitives as small modeling units, and the parametric representation introduced in Section 3.2 can be based on any appropriate model (one can use, for example, a Markov random field built by progressive maximum entropy extension [43], selected similarly to Section 4).

In the case in which these primitives are also modeled as Bayesian networks, the global distribution of our model is obviously also represented as such.

This case includes the compositional trees introduced in Section 3.3, for which deriving a Bayesian network representation in cases (2)–(5) is straightforward.

In such a case, an alternative characterization of our method is that we perform a structure search over Bayesian networks that can be partitioned into previously selected primitives. In this regard, it can be compared with other inductive biases, including the well-studied case of restricting the tree-width of the graph, which leads to a maximum-likelihood structure search problem that is equivalent to finding a maximum-weight *hyperforest* [44], [45]. Our global constraints $\mathcal{M}^*$ could be used to impose such a tree-width restriction on the graph *over primitives*, during greedy or global search. In particular, the compositional trees of Section 3.3 restrict this graph to tree-width 1, yielding our simpler combinatorial problem of finding a maximum-weight *forest* whose nodes are (possibly complex) primitives. Relaxing (M2) to (M2)′ in Section 3.1, we remark that if our primitives consist of all Bayesian networks on subsets of $\leq w+1$ variables, then assembling them under the global compositional-tree constraint gives a subset of Bayesian networks of tree-width w, while dropping the global constraint gives the superset $CPCP^w$ [46].

6.2 Primitive Selection vs. MAP Estimation

Suppose we omit the initial step of primitive filtering. Then purely maximum-likelihood estimation could be done using our global structure search algorithms from Section 5. Naturally, however, maximum-likelihood estimation will tend to overfit the data by choosing models with many parameters. (Indeed, this is the motivation for our approach.) A common remedy is instead to maximize some form of penalized log-likelihood. For many penalization techniques, this can be accomplished by the same global structure search algorithm over assemblies of primitives. One modifies the definition of each binding energy $\rho(\psi_k)$ in our maximization objective (9) to add a constant penalty that is more negative for more complex primitives ψ_k [23]. Before the search, it is safe to discard any primitive ψ such that the penalized binding energy $\rho(\psi) < \rho(\psi')$ for some ψ' with $(J(\psi), A(\psi)), O(\psi)) = (J(\psi'), A(\psi'), O(\psi'))$, since then ψ cannot appear in the globally optimal structure [23]. The filtering strategy in 4.1 can be regarded as a slightly more aggressive version of this, if the penalties are set to 0 for triplets of type (1), $-\eta_{(2)}$ for those of type (2)/(2′), $-(\eta_{(2)}+\eta_{(3)})$ for those of type (3), and so on.

One can regard the total penalty of a structure as the log of its prior probability (up to a constant). The resulting maximization problem can be seen as MAP estimation under a structural prior. To interpret our η penalties in this way would be to say that a random model, *a priori*, is $\exp \eta_{(3)}$ times less likely to use a given triplet of type (3) than one of type (2). However, our actual approach differs from the above MAP story in two ways. First, it is not fully Bayesian, since in Section 4.2, we set the η parameters of the prior from our training data (cf. empirical Bayes or jackknifing). Second, a MAP estimator would include the η penalties in the global optimization problem—but we use these penalties only for primitive selection.

Why the difference? While our approach is indeed somewhat similar to MAP estimation with the above prior, that is not our motivation. We do not actually subscribe to a prior belief that simple structures are more common than complex structures in our application domains (Section 8). Furthermore, our goal for structure estimation is not to minimize the posterior risk under 0-1 loss, which is what MAP estimation does. Rather, we seek an estimator of structure that bounds the risk of a model under a loss function defined as the number of locally useless correlational parameters. Our structure selection procedure conservatively enforces such a bound ϵ (by keeping the false discovery rate low even within $\mathcal{T}_0$ as a whole, and *a fortiori* within any model built from a subset of $\mathcal{T}_0$). Subject to this procedure, we optimize likelihood, which minimizes the posterior risk under 0-1 loss for a *uniform* prior over structures and parameters.

We caution that ϵ does not bound the number of incorrect edges relative to a true model. $\mathcal{T}_0$ includes all correlations that are valid within a primitive, even if they would vanish when conditioning also on variables outside the primitive (cf. [47]). To distinguish direct from indirect correlations, our method uses only global likelihood as in [14]. In the small-sample regime, the resulting models (as with MAP) can have structural errors but at least remain predictive without overfitting—as we now show. Bounding the number of incorrect edges would have to *underfit*, rejecting all edges, even if true or useful, that might be indirect.

7 Simulation Study

We assessed the performance of our learned models using synthetic data. Here we present an overview of our results. The full description of our simulations is provided in Section 1 of the supplemental material.

We first run several experiments to measure the effect of sample size, number of variables, selection threshold and search strategy upon learning performance when the true model belongs to our model class $\mathcal{F}^*$. We evaluated the quality of the estimation by computing the KL divergence between the known, ground truth distribution and the distribution learned using our models. We evaluated network reconstruction accuracy by building ROC and precision-recall (PR) curves in terms of true-positive and false-positive edges. The actual curves and further details can be found in Section 1.1 of the supplemental material. Besides the obvious fact that both quality of estimation and network reconstruction accuracy improved with

increasing sample sizes, our results showed that i) the choice of an excessively large selection threshold leads to model overfitting and ii) for very small samples, the distribution learned using CAM can be better (in terms of KL divergence to the ground truth) than the distribution learned by using the true generating graph and estimating parameters from data.

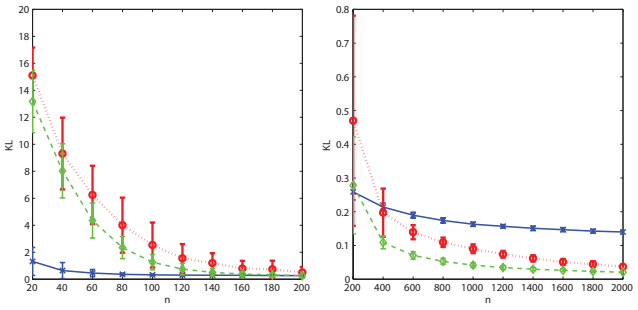

Fig. 4. KL divergence between the Bayesian network ground truth distribution ($d = 14$) and the distributions learned using CAM (solid, blue, cross), K2 (dotted, red, circle) and the true generating graph with MLE parameters (dashed, green, diamond). Results are shown for a fixed choice of parameters and averaged over 100 random replicates (see Section 1.2 of the supplemental material for details).

We compared CAM to other methods from the literature, namely Bayesian networks [48] (represented by the K2 algorithm [12]), relevance networks (RN) [9], ARACNE [10] and CLR [11]. Full details are included as Section 1.2 of the supplemental material. First, we sampled from a balanced binary forest like the ones described in previous sections. We found that CAM outperformed all the alternatives, both in terms of KL to the true distribution and network reconstruction accuracy. This was particularly evident for small samples. Of course, these results were not surprising because we were in the favorable case where the ground truth belonged to our constrained family of models. Next, we considered the unfavorable case where we sampled from a generic Bayesian network with a more complex dependency structure than those allowed within our model class. Fig. 4 shows an example of the type of curves that we observed (curves for other choices of parameters are shown in Fig. S.7 of supplemental material). CAM offered the best performance for small samples by favoring bias over variance. In fact, for small enough sample sizes CAM performed better than using the true graph and only estimating parameters, as we had remarked before. For larger sample sizes, CAM performed worse than the alternatives. This was expected: when data are plentiful and the dependency structure is rich, learning models more complex than ours becomes feasible. Precision-recall curves (Fig. S.8) showed that, for same levels of recall, K2 achieves the least precision. In general, RN, CLR and particularly ARACNE offer the best performances, although CAM is comparable for small samples. For larger sample sizes (over 100) CAM has lower precision than the others (except K2).

8 Real-Data Experiments

Our first experiment involves a semantic network learning task for which the results are reasonably easy to interpret. In the second one, we learn a network of statistical dependencies among somatic mutations in gene *TP53*, which plays an important role in cancer.

8.1 Learning Semantic Networks from Text

The *20newsgroups* dataset [49] is a collection of approximately 20,000 documents that were obtained from 20 different Usenet newsgroups. We worked with a reduced version made publicly available by the late Sam Roweis through his website at the New York University. The data are summarized in a matrix with binary occurrence values for $d = 100$ words across $n = 16,242$ newsgroup postings, these values indicating presence or absence of the word in a specific document. We discarded some words with general or ambiguous meanings, leaving 66 words that were clearly associated to six well differentiated semantic categories (*computers*, *space*, *health*, *religion*, *sports* and *cars*). Intuitively, we would expect the occurrences of words such as *dog* and *boat* to be approximately independent, whereas not so for say *hockey* and *nhl*.

We first measured the effect of the observed sample size. We evaluated edgewise network reconstruction accuracy using a hand-crafted ground-truth network where words that are semantically related were linked by an undirected edge. We chose random subsets of documents of different sizes and we learned a network for each of them using CAM, RN, CLR and ARACNE (K2 was not used because the causal ordering of the variables is unknown). The net result (not shown) is that all methods provide roughly comparable ROC and PR curves. We then compared the predictive performance of CAM versus K2 and a baseline edgeless Bayesian network as a function of sample size, by computing average log-likelihood on different sets of hold-out samples. We arbitrarily chose an entropy ordering for K2 and, in order to provide a fair comparison, we enforced the same ordering constraint upon the set of candidate CAM structures. Our results show that CAM outperforms the alternatives when sample sizes are small (details can be found in Section 3.1 of supplemental material).

Next, we learned the network shown in Fig. 5 using the full dataset. The network is componentwise optimal. It is not guaranteed to be globally optimal because the dual gap for the global optimization problem was non-negligible; still, the result appeared to be stable after extensive computation (see Section 2.3 of supplemental material). We observe a very good correspondence between network components and semantic categories. In fact, there is only one merge between components that may seem questionable: the *space* and *religion* components end up being connected by the word *earth*. This is a consequence of the local

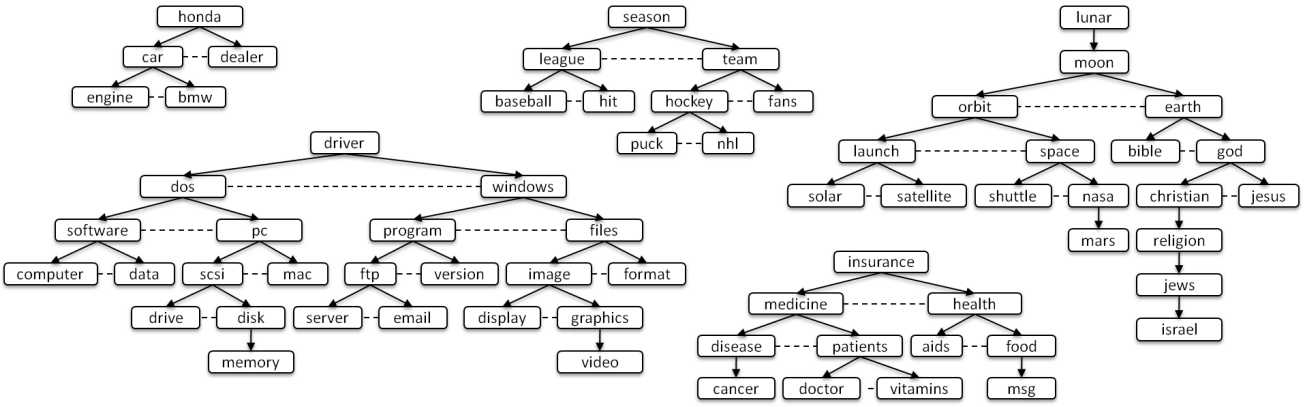

Fig. 5. Final network learned using a subset of words from the *20newsgroups* dataset and the full sample size. This result was obtained using the ILP search procedure. Dashed lines are used to identify siblings that belong to the same primitive within our model. Words belonging to the same semantic categories (*computers*, *space*, *health*, *religion*, *sports* and *cars*) tend to form proximity clusters within the network.

scope of our primitives and the fact that this word is frequently used within both semantic contexts. For comparison purposes, we learned two graphs using RN (see Section 3.2 of supplemental material). In the first case, the threshold for mutual information was the same as the one used for the CAM binary primitives in Fig. 5; in the RN case all the variables wound up in the same connected component and the large number of learned interactions made it somewhat difficult to interpret the result. In the second case, we equalized the number of learned edges, leading to the isolation of twenty variables as singletons (and thus to the failure to learn some important connections).

8.2 Learning Statistical Dependencies among Somatic Mutations in Gene *TP53*

The *TP53* gene is located on the short arm of chromosome 17 and encodes the p53 tumor suppressor protein, which plays a fundamental role in many human cancers [50], [51]. The p53 protein is activated when cells are stressed or damaged and it blocks their multiplication, providing an important mechanism to prevent tumor proliferation. Mutations in the *TP53* gene are primarily of the missense type. They are known to cause direct inactivation of the p53 protein in about half of the cancers where this protein fails to function correctly and indirect inactivation in many other cases [50]. Because of this, understanding the effect of these mutations can provide very valuable insight into the mechanisms of cancer [52].

Most of the somatic mutations reported in the literature are compiled in the International Agency for Research on Cancer (IARC) TP53 Database [53]. We used version R15 (updated in Nov. 2010) and worked with the somatic mutations dataset, which mainly consists of missense mutations detected by DNA sequencing in tumor samples and mutations within exons 5-8. The original dataset contains measurements for $d = 4{,}356$ different mutations and $n = 25,101$ different tissue samples. (Discarding mutations appearing in fewer than two samples leaves $d = 2{,}000$ and $n = 23{,}141$.) We can represent each sample as a d-dimensional binary vector indicating which mutations were observed in a tissue extracted from a patient. Since patients can have more than one cancer, multiple samples may come from the same individual. Still, we treat the vectors of observations as independent samples from an underlying multivariate distribution which characterizes the dependency structure among mutations in a cancer population.

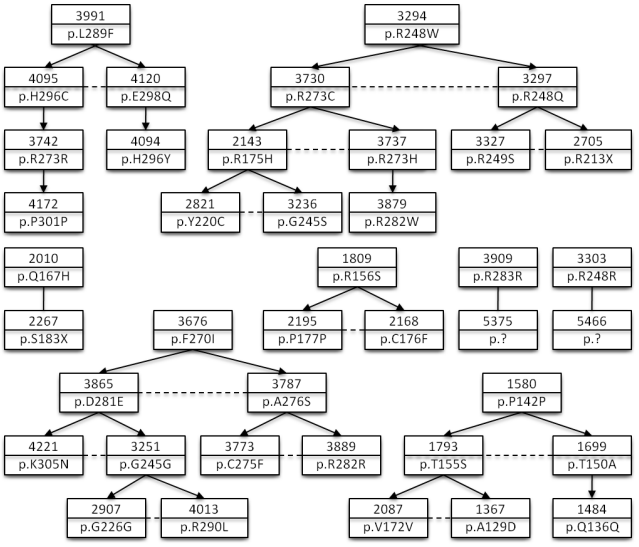

Fig. 6. Some components of the TP53 somatic mutations network learned using CAM (full network is provided in Fig. S.12). Dashed lines link siblings within the same primitive. For each node, we provide the unique mutation identifier in the IARC Database and the (standard) mutation description at the protein level, where *p.XzY* means substitution of amino acid X by amino acid Y at codon z; e.g., $p.R248W$ represents substitution of *Arg* by *Trp* at codon 248.

Selecting $\epsilon = 1$ leads to 760 candidate primitives. Using CAM, we learned a network that contained 68 different mutations (out of the original set of 2,000 candidates). Fig. 6 shows some of the components in this network (the full network graph is provided in Fig. S.12, Section 4 of supplemental material).

Each somatic mutation in IARC is annotated with biochemical details about the actual nucleotide variation, as well as clinical data for the patient and tissue where the mutation was observed. Enrichment of shared annotations is statistically significant for several subsets of mutations within our network, both at the pairwise and the component levels, which suggests that these mutations might be functionally related. Furthermore, based on a hypergeometric null, the mutations in our network were significantly enriched for several biological indicators such as presence in CpG sites and associated gain of function (GoF) at the protein level. The same type of test shows that our network is significantly enriched for "deleterious" or "driver" mutations (which are known to have a negative impact on the phenotype, as opposed to "neutral" or "passenger" ones). A detailed explanation of our statistical analysis, including some comments on the biological interpretation of our results, is provided in Section 4 of the supplemental material.

9 Conclusions and Future Work

We have introduced a new modeling framework that combines local model selection (designing and estimating primitives) with a compositional step that merges primitives into valid graphical arrangements and multivariate distributions. This construction makes it possible to adjust model complexity to sample size by controlling the dimension of the parameter space. The introduction of structural biases can be used to decrease variance and avoid model overfitting, which is critical in small sample regimes.

Our approach has been validated using both synthetic and real data. For simulated data, our method outperforms general Bayesian networks in approximating the true generating distribution for small sample sizes. Our approach is also comparable with methods designed for recovering graphs rather than distributions. Finally, experiments with real data, particularly genetic mutations in the *TP53* gene, demonstrate that the CAM algorithm can cluster mutations into biologically plausible groups.

Even though in this paper we have only discussed the case of discrete random variables, the CAM framework generalizes to the continuous case where discrete distributions are replaced by probability density functions. Also, we have focused on balanced binary trees, which simplifies both primitive learning and the combinatorial optimization problem for competitive assembling. These constraints yield networks which are easy to interpret, since the limits on topological complexity often lead to a final graph with several components of moderate size. These limits include strong restrictions to the set of allowable values for incoming and outgoing vertex degrees. However, our entire framework applies more generally to any family of primitives, such as those depicted in Fig. 1, allowing us to move beyond the strong structural constraints imposed by trees in the context of moderate to large sample sizes. In particular, such extensions might allow for learning networks with "hubs" and scale-free properties provided near-optimal assemblies can be identified.

Acknowledgment

Francisco Sánchez-Vega thanks "la Caixa" Foundation and Caja Madrid Foundation for support through graduate fellowships. The work of Laurent Younes is partially supported by NSF CCF-0625687. The work of Donald Geman is partially supported by NIH-NCRR Grant UL1 RR 025005 and NSF CCF-0625687.

References

[1] D. J. Duggan, M. Bittner, Y. Chen, P. Meltzer, and J. M. Trent, "Expression profiling using cDNA microarrays," *Nat. Gen. Suppl.*, vol. 21, pp. 10–14, 1999.
[2] R. J. Lipshutz, S. P. Fodor, T. R. Gingeras, and D. J. Lockhart, "High-density synthetic oligonucleotide arrays," *Nat. Gen.*, vol. 21, pp. 20–24, 1999.
[3] Y. Wang, D. J. Miller, and R. Clarke, "Approaches to working in high-dimensional data spaces: Gene expression microarrays," *British J. Cancer*, February 2008.
[4] J. H. Moore and M. D. Ritchie, "The challenges of whole-genome approaches to common diseases," *J. Amer. Med. Assoc.*, vol. 291, no. 13, pp. 1642–1643, 2004.
[5] M. Morley, C. Molony, T. Weber, J. Devlin, K. Ewens, R. Spielman, and V. Cheung, "Genetic analysis of genome-wide variation in human gene expression," *Nature*, 2004.
[6] M. I. McCarthy, G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. A. Ioannidis, and J. N. Hirschhorn, "Genome-wide association studies for complex traits: Consensus, uncertainty and challenges," *Nat. Rev. Gen.*, vol. 9, no. 5, pp. 356–369, 2004.
[7] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, 2001.
[8] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," *Neural Comput.*, vol. 4, no. 1, pp. 1–58, 1992.
[9] A. J. Butte and I. S. Kohane, "Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements," *Pac. Symp. Biocomput.*, pp. 418–29, 2000.
[10] A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Favera, and A. Califano, "ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context," *BMC Bioinformatics*, vol. 7, p. S7, 2006.
[11] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner, "Large-scale mapping and validation of E. coli transcriptional regulation from a compendium of expression profiles," *PLoS Biol.*, vol. 5, no. 1, p. e8, Jan 2007.
[12] G. F. Cooper and T. Dietterich, "A Bayesian method for the induction of probabilistic networks from data," in *Machine Learning*, 1992, pp. 309–347.
[13] D. T. Brown, "A note on approximations to discrete probability distributions," *Info. and Control*, vol. 2, no. 4, pp. 386–392, 1959.
[14] C. I. Chow and C. N. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Trans. Inf. Theory*, vol. 14, pp. 462–467, 1968.
[15] M. Meila, "An accelerated Chow and Liu algorithm: Fitting tree distributions to high-dimensional sparse data," in *Proc. ICML*. San Francisco: Morgan Kaufmann, 1999, pp. 249–257.
[16] T. Szántai and E. Kovács, "Hypergraphs as a means of discovering the dependence structure of a discrete multivariate probability distribution," *Ann. of Operat. Res.*, pp. 1–20, 2010.
[17] F. R. Bach and M. I. Jordan, "Thin junction trees," in *NIPS 14*, 2001, pp. 569–576.

[18] A. Chechetka and C. Guestrin, "Efficient principled learning of thin junction trees," in *NIPS*, Vancouver, Canada, 2007.
[19] D. Shahaf and C. Guestrin, "Learning thin junction trees via graph cuts," *J. Mach. Learning Res.— Proc. Track*, vol. 5, pp. 113–120, 2009.
[20] M. Narasimhan and J. Bilmes, "PAC-learning bounded tree-width graphical models," in *Proc. UAI*, 2004, pp. 410–417.
[21] G. Elidan and S. Gould, "Learning bounded treewidth Bayesian networks," in *NIPS*, 2008, pp. 417–424.
[22] S. I. Lee, V. Ganapathi, and D. Koller, "Efficient structure learning of Markov networks using L1-regularization," in *Proc. NIPS*, Cambridge, MA, 2007, pp. 817–824.
[23] C. P. de Campos, Z. Zeng, and Q. Ji, "Structure learning of Bayesian networks using constraints," in *Proc. ICML*, 2009.
[24] S. Geman, D. F. Potter, and Z. Chi, "Composition systems," *Quart. Appl. Math.*, vol. 60, no. 4, pp. 707–736, 2002.
[25] S. C. Zhu and D. Mumford, "A stochastic grammar of images," *Found. Trends Comp. Graph. Vision*, vol. 2, no. 4, pp. 259–362, 2006.
[26] Y. Amit and A. Trouvé, "POP: Patchwork of parts models for object recognition," *Int. J. of Comput. Vision*, vol. 75, no. 2, pp. 267–282, Nov. 2007.
[27] A. Dobra, C. Hans, B. Jones, J. R. Nevins, G. Yao, and M. West, "Sparse graphical models for exploring gene expression data," *J. Multivar. Anal.*, vol. 90, pp. 196–212, July 2004.
[28] J. Utans, "Learning in compositional hierarchies: Inducing the structure of objects from data," in *NIPS 6*, 1994, pp. 285–292.
[29] R. D. Rimey and C. M. Brown, "Control of selective perception using Bayes nets and decision theory," *Int. J. Comput. Vision*, vol. 12, pp. 173–207, April 1994.
[30] B. Neumann and K. Terzic, "Context-based probabilistic scene interpretation," in *Artificial Intell. in Theory and Practice III*, ser. IFIP Adv. in Inform. and Commun. Tech. Springer Boston, 2010, vol. 331, pp. 155–164.
[31] D. Heckerman, D. M. Chickering, C. Meek, R. Rounthwaite, and C. Kadie, "Dependency networks for inference, collaborative filtering, and data visualization," *J. Mach. Learning Res.*, pp. 49–75, 2000.
[32] Y. Xiang, F. V. Jensen, and X. Chen, "Multiply sectioned Bayesian networks and junction forests for large knowledge-based systems," *Comp. Intell.*, vol. 9, pp. 680–687, 1993.
[33] D. Koller and A. Pfeffer, "Object-oriented Bayesian networks," in *Proc. UAI*, 1997, pp. 302–313.
[34] N. Friedman, L. Getoor, D. Koller, and A. Pfeffer, "Learning probabilistic relational models," in *Proc. IJCAI*, 1999, pp. 1300–1309.
[35] E. Gyftodimos and P. A. Flach, "Hierarchical Bayesian networks: An approach to classification and learning for structured data," in *Methods and Applications of Artificial Intelligence*, ser. Lecture Notes in Computer Science, G. A. Vouros and T. Panayiotopoulos, Eds., 2004, vol. 3025, pp. 291–300.
[36] D. Pe'er, "Bayesian network analysis of signaling networks: A primer," *Sci. STKE*, vol. 2005, no. 281, p. pl4, 2005.
[37] P. Spirtes, C. Glymour, and R. Scheins, *Causation, Prediction, and Search*, 2nd ed. MIT Press, 2001.
[38] T. Jaakkola, D. Sontag, A. Globerson, and M. Meila, "Learning Bayesian network structure using LP relaxations," in *Proc. AISTATS*, vol. 9, 2010, pp. 358–365.
[39] E. Segal, D. Koller, N. Friedman, and T. Jaakkola, "Learning module networks," in *J. Mach. Learning Res.*, 2005, pp. 525–534.
[40] A. Martins, N. Smith, and E. Xing, "Concise integer linear programming formulations for dependency parsing," in *Proc. ACL-IJCNLP*, 2009, pp. 342–350.
[41] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *Ann. Math. Statist.*, no. 9, pp. 60–62, 1938.
[42] C. E. Miller, A. W. Tucker, and R. A. Zemlin, "Integer programming formulation and traveling salesman problems," *J. Assoc. for Computing Machinery*, vol. 7, pp. 326–329, 1960.
[43] S. Della Pietra, V. Della Pietra, and J. Lafferty, "Inducing features of random fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 4, pp. 380–393, Apr. 1997.
[44] D. Karger and N. Srebro, "Learning Markov networks: Maximum bounded tree-width graphs," in *Proc. 12th ACM-SIAM Symp. on Discrete Algorithms*, 2001.
[45] N. Srebro, "Maximum likelihood bounded tree-width Markov networks," *Artificial Intell.*, vol. 143, pp. 123–138, 2003.
[46] K.-U. Höffgen, "Learning and robust learning of product distributions," in *Proc. COLT*, 1993, pp. 77–83.
[47] J. Schäfer and K. Strimmer, "An empirical Bayes approach to inferring large-scale gene association networks," *Bioinformatics*, vol. 21, no. 6, pp. 754–764, 2005.
[48] D. Heckerman, "A tutorial on learning Bayesian networks," Microsoft Research, Tech. Rep. MSR-TR-95-06, March 1995.
[49] K. Lang, "Newsweeder: Learning to filter netnews," in *Proc. ICML*, 1995, pp. 331–339.
[50] B. Vogelstein, D. Lane, and A. J. Levine, "Surfing the p53 network," *Nature*, no. 408, 2000.
[51] A. J. Levine, C. A. Finlay, and P. W. Hinds, "P53 is a tumor suppressor gene," *Cell*, vol. 116, pp. S67—S70, 2004.
[52] M. S. Greenblatt, W. P. Bennett, M. Hollstein, and C. C. Harris, "Mutations in the p53 tumor suppressor gene: Clues to cancer etiology and molecular pathogenesis," *Cancer Res.*, vol. 54, no. 18, pp. 4855–4878, 1994.
[53] A. Petitjean, E. Mathe, S. Kato, C. Ishioka, S. V. Tavtigian, P. Hainaut, and M. Olivier, "Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: Lessons from recent developments in the IARC TP53 database," *Human Mut.*, vol. 28, no. 6, pp. 622–629, 2007.

Francisco Sánchez-Vega graduated in Telecommunications Engineering in 2005 from ETSIT Madrid and ENST Paris. Also in 2005, he was awarded a M.Res. in Applied Mathematics for Computer Vision and Machine Learning from ENS Cachan. He arrived at Johns Hopkins University in 2006 and received a M.Sc.Eng. in Applied Mathematics and Statistics in 2008. He is currently a Ph.D. candidate at this same department and a member of the Center for Imaging Science and the Institute for Computational Medicine at JHU.

Jason Eisner holds an A.B. in Psychology from Harvard University, a B.A./M.A. in Mathematics from the University of Cambridge, and a Ph.D. in Computer Science from the University of Pennsylvania. Since his Ph.D. in 2001, he has been at Johns Hopkins University, where he is Associate Professor of Computer Science and a core member of the Center for Language and Speech Processing. Much of his research concerns the prediction and induction of complex latent structure in human language.

Laurent Younes Former student of the Ecole Normale Superieure in Paris, Laurent Younes was awarded the Ph.D. from the University Paris Sud in 1989, and the thesis advisor certification from the same university in 1995. He was a junior, then senior researcher at CNRS (French National Research Center) from 1991 to 2003. He is now professor in the Department of Applied Mathematics and Statistics at Johns Hopkins University (that he joined in 2003). He is a core faculty member of the Center for Imaging Science and of the Institute for Computational Medicine at JHU.

Donald Geman received the B.A. degree in Literature from the University of Illinois and the Ph.D. degree in Mathematics from Northwestern University. He was a Distinguished Professor at the University of Massachusetts until 2001, when he joined the Department of Applied Mathematics and Statistics at Johns Hopkins University, where he is a member of the Center for Imaging Science and the Institute for Computational Medicine. His main areas of research are statistical learning, computer vision and computational biology. He is a fellow of the IMS and SIAM.