Theoretical Population Biology 107 (2016) 52–64

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

An information theoretic approach to pedigree reconstruction

Anthony Almudevar

Department of Biostatistics and Computational Biology, University of Rochester, United States

ARTICLE INFO

Article history:
Received 1 April 2015
Available online 8 October 2015

Keywords:
Pedigree reconstruction
Graphical models
Minimum Description Length principle
Bayesian inference

ABSTRACT

Network structure is a dominant feature of many biological systems, both at the cellular level and within natural populations. Advances in genotype and gene expression screening made over the last few decades have permitted the reconstruction of these networks. However, resolution to a single model estimate will generally not be possible, leaving open the question of the appropriate method of formal statistical inference. The nonstandard structure of the problem precludes most traditional statistical methodologies. Alternatively, a Bayesian approach provides a natural methodology for formal inference. Construction of a posterior density on the space of network structures allows formal inference regarding features of network structure using specific marginal posterior distributions.

An information theoretic approach to this problem will be described, based on the Minimum Description Length principle. This leads to a Bayesian inference model based on the information content of data rather than on more commonly used probabilistic models. The approach is applied to the problem of pedigree reconstruction based on genotypic data. Using this application, it is shown how the MDL approach is able to provide a truly objective control for model complexity.

A two-cohort model is used for a simulation study. The MDL approach is compared to COLONY-2, a well known pedigree reconstruction application. The study highlights the problem of genotyping error modeling. COLONY-2 requires prior error rate estimates, and its accuracy proves to be highly sensitive to these estimates. In contrast, the MDL approach does not require prior error rate estimates, and is able to accurately adjust for genotyping error across the range of models considered.

1. Introduction

The inference of network structure has assumed increasing importance in the life sciences with the advent of high-dimensional molecular data. That discernible forms of dependence in such data can be used to infer network structure has been confirmed by mathematical theory and numerous applications ranging from gene regulatory networks to pedigrees (Rissanen et al., 2007; Lee and Tzou, 2009; Vignes et al., 2011; Marbach et al., 2012). However, the nonstandard form of the inference, in which the 'parameter' is a graph or similar object, precludes classical statistical methods, leaving open the problem of controlling for false positives and determining confidence levels.

In this article we review a general approach to this problem, based on the following three principles:

1. A full Bayesian solution permits formal inference that is accurate and computationally efficient.
2. The correct choice of the prior density on network structure is crucial. Reasonable principles of invariance exist with which to guide the choice of uninformative prior, to which informative prior information can be appended.
3. Information theory, as proposed under the *Minimum Description Length* (MDL) principle, provides the mathematical basis for Bayesian models with predictable and intuitive properties. Probabilistic models, which often rely on untestable assumptions, are not needed. The problem is formulated as a data compression problem, with *models* interpreted as forms of regularity which may be exploited for greater efficiency.

This approach will be applied to the problem of pedigree reconstruction (PR), a seminal problem in population biology involving the inference of joint kinship forms using genotypic data (Pemberton, 2008; Jones et al., 2010; Harrison et al., 2013).

This article will cover the following topics. A brief introduction to graphical models, in particular the Bayesian network, will be given in Section 2. The relationship between this model and the problem of pedigree inference will be reviewed, giving conditions under which a pedigree may be modeled as a Bayesian network, and the implications for cases in which this does not hold.

Section 3 will review methodologies associated with the inference of graph structure. The basis for a Bayesian methodology

E-mail address: anthony_almudevar@urmc.rochester.edu.

http://dx.doi.org/10.1016/j.tpb.2015.09.006
0040-5809/

will be given, with a discussion of both the choice of prior distribution for graph structure, and of related computational issues.

Section 4 introduces the Minimum Description Length (MDL) principle, and its relationship to Bayesian inference. Section 4.1 provides a brief introduction to basic coding theory, on which most MDL analysis is based.

Section 5 elaborates on the discussion on prior distributions of Section 3.1 in the context of the MDL method. It is shown how a rigorous solution to an inference problem can be derived by a purely objective (and apparently unrelated) criterion, in particular, optimal data compression.

In Section 6 the MDL principle is applied to PR. A general approach is outlined, based on the efficient coding of genotype data assisted by pedigree models. Issues such as genotyping errors, missing data and linkage are discussed.

In Section 7 the approach of Section 6 is applied to a pedigree model based on two generational cohorts. The MDL method is compared to the widely used PR application COLONY 2 (Jones and Wang, 2010; Wang, 2013; Wang and Scribner, 2014). The emphasis is on the respective abilities of the methods to accurately adjust for genotyping error. Section 8 summarizes the results in a conclusion.

2. Graphical models, Bayesian networks and pedigrees

We first introduce some terminology. A graph $G = (V, E)$ is a collection of nodes V and edges E, which are either ordered (directed) or unordered (undirected) pairs of nodes. Given the directed edge $a \to b$ we say that a is a *parent* of b, or that b is a *child* of a. A *directed path* from nodes a_1 to a_m is any sequence of nodes $a_1, \ldots, a_m$ such that G contains directed edges $a_i \to a_{i+1}$, $i = 1, \ldots, m-1$. If a directed path exists from a to c, then c is a *descendant* of a, and a is an *ancestor* of c. A *cycle* is a directed path with a common start and end node. A *directed acyclic graph* (DAG) is a directed graph which contains no cycles. A directed graph is completely defined by specifying the *parent sets* S_i for each node i. The number of parents of a node (equal to $|S_i|$) is referred to as its *indegree*. In a DAG, a node without parents is a *founder* (a DAG must contain at least one). A *subgraph* of G is a graph for which nodes and edges are subsets of V and E, respectively.

A probabilistic graphical model (PGM) generally consists of a graph G with random variables $X = (X_1, \ldots, X_N)$ associated with nodes labeled $V = \{1, \ldots, N\}$ which possesses joint density $f(x) = f(x_1, \ldots, x_N)$. If $S \subset V$ then $X[S]$ denotes the vector of components X_i associated with nodes $i \in S$. In PGMs such as *Bayesian networks* or *Markov networks* (Koller and Friedman, 2009) the density f and graph G are related in the sense that f satisfies a collection of conditional independence constraints implied by G (formally, only the density f is needed to completely define the model). PGMs may differ in classes of graphs used (DAGs for Bayesian networks and undirected graphs for Markov networks). The PGM most relevant to the problem of PR is the Bayesian network, which we discuss next.

2.1. Bayesian networks

The *Bayesian network* (BN) is a type of graphical model which has been used in a number diverse fields ranging from artificial intelligence to the modeling of gene regulatory networks (Pearl, 1988; Koller and Friedman, 2009; Scutari, 2010). Suppose we are given a random vector $X = (X_1, \ldots, X_N)$, with joint density $f(x) = f(x_1, \ldots, x_N)$. Then f is a BN if there is at least one DAG G on nodes $\{1, \ldots, N\}$ such that

$$f(x) = \prod_{i=1}^{N} f(x_i \mid x[S_i]), \tag{1}$$

where S_i are the parent sets defining G and $f(x_i \mid x[S_i])$ is the density of X_i conditional on $X[S_i]$. If $S_i = \emptyset$ then $f(x_i \mid x[S_i])$ is equal to the marginal density of X_i. Note that (1) implies that founder components of X are independent. Conditions under which (1) holds are well known, and take several forms. For our purposes we cite the following (Koller and Friedman, 2009).

Definition 1. A random vector X satisfies the *local Markov property* (LMP) for some DAG G if each X_i is independent of its nondescendants when conditioned on $X[S_i]$ (this implies that values of X_i associated with founders of G are mutually independent).

BNs are alternatively characterized in the following way.

Definition 2. A random vector X is *complete* for some DAG G if its distribution can be factorized according to (1) for G.

A central result in the theory of BNs (Pearl, 1988) is that a multivariate density f may be decomposed according to (1) when the LMP of Definition 1 holds, so that Definitions 1 and 2 are in this sense equivalent, and both define a BN.

Intuitively, Definition 2 implies that there are no missing edges in G, that is, edges that would be needed to permit the factorization (1). As an example, consider a BN model for data $X = (X_1, X_2, X_3)$, based on DAG G with edges $1 \to 2$ and $1 \to 3$. This means that the distribution of X can be factorized as

$$f(x) = f(x_1) f(x_2 \mid x_1) f(x_3 \mid x_1),$$

so that X is complete. Furthermore, the subvectors (X_1, X_2) and (X_1, X_3) are also complete with respect to G (or the relevant subgraph of G), but this is not true of subvector (X_2, X_3).

The factorization (1) makes the BN quite tractable, both analytically and computationally. To see this, suppose a BN is used to model a multivariate normal density. In most cases, the indegree within G possesses an upper bound significantly smaller than N. This will mean that the number of parameters defining the model is of order $O(N)$. On the other hand, a general multivariate normal density requires order $O(N^2)$ parameters. Thus, a significant reduction in modeling complexity follows from Definition 1.

2.2. Pedigree reconstruction and graphical models

There is a clear relationship between graphical models and pedigree reconstruction that has been noted in the literature (Almudevar, 2003, 2007b; Riester, 2009; Cowell, 2009; Almudevar and LaCombe, 2012; Sheehan et al., 2014). The DAG is a natural representation for a pedigree, constructed using directed edges from parent to offspring. Clearly, a pedigree interpreted as a directed graph cannot contain a cycle. Furthermore, the laws of Mendelian inheritance, at least under certain linkage assumptions, imply the type of conditional independence assumptions which define a Bayesian network. We next consider the question of when a PR problem can be modeled as a BN.

Suppose we are given a set of labeled individuals $V = \{1, \ldots, N\}$, to be interpreted as nodes in a graph. We may refer to the larger population $\mathcal{V}$ from which V is sampled. We also have data $X = (X_1, \ldots, X_N)$ where X_i represents a genotype observation from L loci.

A *pedigree graph* (PG) $G(V)$ may be defined as an exhaustive specification of all parent–offspring (PO) dyads within V (as a directed edge from parent to child). If $|S_i| = 0$ then i is a *founder*, and if $|S_i| = 1$ then we say i is a *semifounder* (the parent not included in V is a *hidden parent*). If $S_i \subset V$ are the two parents of i, then the conditional distribution $f(x_i \mid x[S_i])$ of the offspring genotypes X_i given the parental genotypes is well known, following from Mendelian probability laws. If i is a founder, then $f(x_i \mid x[S_i]) = f(x_i)$ is the marginal distribution of X_i, equivalent to