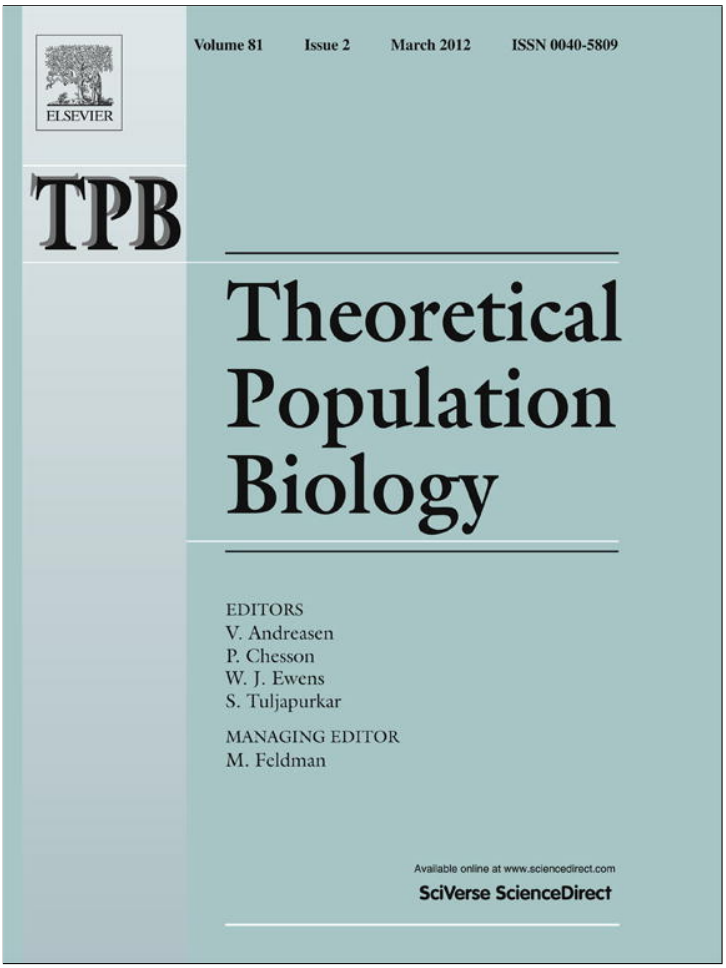

Theoretical Population Biology 81 (2012) 131–143

Contents lists available at SciVerse ScienceDirect

Theoretical Population Biology

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

On the choice of prior density for the Bayesian analysis of pedigree structure

Anthony Almudevar*, Jason LaCombe

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

ARTICLE INFO

Article history:
Received 3 August 2011
Available online 19 December 2011

Keywords:
Pedigree inference
Bayesian inference
Minimum Description Length Principle

ABSTRACT

This article is concerned with the choice of structural prior density for use in a fully Bayesian approach to pedigree inference. It is found that the choice of prior has considerable influence on the accuracy of the estimation. To guide this choice, a *scale invariance* property is introduced. Under a structural prior with this property, the marginal prior distribution of the local properties of a pedigree node (number of parents, offspring, etc.) does not depend on the number of nodes in the pedigree. Such priors are found to arise naturally by an application of the Minimum Description Length (MDL) principle, under which construction of a prior becomes equivalent to the problem of determining the length of a code required to encode a pedigree, using the principles of information theory. The approach is demonstrated using simulated and actual data, and is compared to two well-known applications, CERVUS and COLONY.

1. Introduction

The statistical reconstruction of pedigrees using genotype markers is an important tool in the study of natural and human populations (Garant and Kruuk, 2005; Thomas, 2005; Jones and Wang, 2010b; Jones et al., 2010). The underlying elements of the problem are expressible in terms of classical statistical methods, since markers obey well known probability laws based on Mendelian inheritance. This means that the joint distribution of markers across individuals can be precisely stated for a given pedigree, so that the inverse problem of estimating the pedigree given marker data is statistically well defined.

Suppose we are given marker data for two known generations. Each ordered pair consisting of one individual from each generation forms a putative parent–offspring (PO) pair. One of the central ideas in this field is the *exclusion principle* by which putative PO pairs can be eliminated from consideration based on the genetic incompatibility of the genetic data (extendable to parent pair/offspring triplets). Ideally, the probability of exclusion for false PO pairs is close to one, but still leaves false non-exclusions with a small but nonzero probability, and will occur with a predictable non-zero frequency in large enough samples. Therefore, some method of correcting for false PO pairs is needed. The CERVUS method proposed by Marshall et al. (1998) accomplishes this using a sequence of hypothesis tests which employs a global correction based on the sample. Alternatively, a Bayesian approach which models the number of putative parents in a sample jointly with the total population size from which this sample was drawn was proposed by Nielsen et al. (2001).

A different approach based on graphical modeling has been described in a number of articles (Almudevar, 2003, 2007; Riester et al., 2009, 2010; Cowell, 2009). When a multi-generational pedigree is *complete* (that is, the path of descent between any two individuals in the sample is also represented in the sample), the joint distribution of the marker data is equivalent to a Bayesian network model (Koller and Friedman, 2009). Such models have otherwise been used in a variety of applications ranging from artificial intelligence to the modeling of gene expression data.

This type of model is more flexible, since an arbitrary number of generations can be considered without any adjustments, and it has been shown that pedigrees can be accurately reconstructed even without age or generation assignment data (Almudevar, 2003; Cowell, 2009). The disadvantage is that complexity estimation and control methods natural to the two generation models are not easily extended to arbitrary pedigrees, so a Bayesian methodology which allows variable weighting based on graphical complexity seems necessary, and a number of such approaches have been taken. Prior distributions based on those derived by Nielsen et al. (2001) were incorporated into the general pedigree by Riester et al. (2009), while an approach based on the Minimum Description Length principle was used by Almudevar (2007). In the latter paper, computations suggest that for samples with few missing parents the maximum likelihood works well without any control for complexity. This is because the correct pedigree is close to the maximum complexity permitted given the constraints inherent in pedigrees. The situation changes when many parents are missing, and a well chosen pedigree prior can accurately control the overfitting inevitable with the maximum likelihood approach.

In this article, we develop further the type of pedigree prior density introduced by Almudevar (2007), with a number of consequences. First, we show how a further refinement of the

* Corresponding author.
E-mail address: anthony_almudevar@urmc.rochester.edu (A. Almudevar).

0040-5809/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.tpb.2011.12.003

development can result in inference of much greater accuracy while maintaining the same principle of complexity control. We will also argue that such priors are objective in the sense that they possess a certain scale invariance property which should be regarded as important for any prior in this application.

In Section 2 the Bayesian approach to pedigree inference is discussed, emphasizing the choice of prior density on pedigree structure. A scale invariance property for prior densities is then introduced, which holds that the marginal distribution of any local feature of a pedigree (e.g. number of parents of a given individual) should not depend on the size of the pedigree. An overview of priors proposed in the literature for this type of application is also given.

In Section 3 the construction of prior densities using information theoretic principles is discussed. This is done in the context of the Minimum Description Length principle, a general approach to model selection in which the notion of the goodness of fit of a model, given a set of data, is reformulated as a problem in model assisted data compression. Two coding systems are developed, the *edge code* and the *block code*, each of which may be directly transformed into a prior density. In Section 4 it is then shown that both priors satisfy the scale invariance property, but that important differences remain.

Sections 5 and 6 demonstrate the priors using simulated models, as well as from a data set obtained from 12 large families of Atlantic Salmon. In general, the prior derived from the block code exhibits considerably greater accuracy. The method is also compared to two well known parentage assignment applications, CERVUS 3.0.3 and COLONY 2.0. A concluding section follows. In Appendix A relevant theory from Almudevar (2007) is summarized, while details concerning the calculation of the block code prior density are given in Appendix B.

Software used for this study, incorporated into PEDAPP 2.0, may be downloaded from www.urmc.rochester.edu/biostat/people/faculty/almudevar.cfm.

2. The Bayesian approach to pedigree structure

We introduce a Bayesian network representation of a pedigree model. Consider a set of $\mathcal{U}$ of labeled nodes $1, \ldots, N_I$ (for N_I distinct individual organisms). Let E be a set of directed edges, defined as an ordered pair from $\mathcal{U}$. An edge $j \to i \in E$ implies a PO relationship between j and i. A directed graph is defined by the pair $G = (E, \mathcal{U})$. A graph may also be represented by *parent sets* $S_i \subset \mathcal{U}, i = 1, \ldots, N_I$, where S_i is the set of all parents of i. We adopt the convention that for edge e, $G \cup e$ is the graph $G' = (E \cup \{e\}, \mathcal{U})$. Also, G is an *edgeless* graph when $E = \emptyset$. We will consider N_I to be unbounded, and adopt the following convention for asymptotic order notation. We say $t(N_I) = O(t'(N_I))$ when $\limsup_{N_I \to \infty} |t(N_I)/t'(N_I)| < \infty$ and $t(N_I) = o(t'(N_I))$ when $\limsup_{N_I \to \infty} |t(N_I)/t'(N_I)| = 0$.

If we have a vector of random data $X = (X_1, \ldots, X_{N_I})$, with observation X_i associated with node i, a probabilistic graphical model defines a joint density $f(x)$ for X which is constrained in some intuitive way by a graph $G \subset \mathcal{G}$, for some class of graphs $\mathcal{G}$. An important special case is the Bayesian network (Koller and Friedman, 2009). Let $\mathcal{G}$ be the set of all directed acyclic graphs (DAG). Then $f(x)$ is a Bayesian network if there is some $G = (S_1, \ldots, S_{N_I}) \in \mathcal{G}$ which permits the factorization

$$f(x_1, \ldots, x_{N_I}) = \prod_{i=1}^{N_I} f_i(x_i \mid X_j = x_j,\ j \in S_i). \tag{1}$$

Genotype data X on a complete pedigree satisfies the definition of a Bayesian network for model classes $\mathcal{G}$ of DAGs restricted to graphs with maximum parent set size of 2. To see this, suppose $X = (x_1, \ldots, x_{N_I})$ is the multilocus genotype data for individuals $1, \ldots, N_I$. The pedigree is represented as a DAG G. If the loci are in linkage equilibrium, and the founders of the pedigree are unrelated, then the joint density of the data given G may be written

$$f(x_1, \ldots, x_{N_I}) = \prod_{i=1}^{N_I} p(x_i \mid S_i) \tag{2}$$

where $p(x_i \mid S_i)$ is the probability of genotype x_i given S_i, which is the population genotype frequency of x_i when $S_i = \emptyset$ (usually estimated), and is otherwise directly given by Mendelian inheritance laws.

In developing a full Bayesian approach to the inference of pedigree structure, the density defined in Eq. (1) or (2) may be interpreted as the conditional density $g(X \mid G) = f(x_1, \ldots, x_{N_I})$ so that the posterior density is

$$g(G \mid X) \propto g(X \mid G)\phi(G) \tag{3}$$

where $\phi(G)$ is a prior density on a space of graphs $\mathcal{G}$ (the normalizing constant, which does not depend on G, is omitted). It will be useful to define the following property of a prior.

Definition 1. A prior density $\phi(G)$ on any space of DAGs $\mathcal{G}$ is of *product form* if it may be written $\phi(G) = \prod_{i=1}^{N_I} \phi_i(S_i)$ for all $G \in \mathcal{G}$.

2.1. Constrained optimization and sampling

The acyclic requirement is a necessity for a pedigree model, otherwise the descendant of an individual could also be its ancestor. However, this constraint accounts for some of the computational difficulties associated with fitting Bayesian network models. In Almudevar (2003) it was noted that if age information is available, the pedigree which maximizes (2) can be determined in polynomial time, if the search is confined to pedigrees conforming to the age constraints. If age information is not available, the greater efficiency resulting from age constraints can be exploited in the following way. Let R be an arbitrary ranking of the individuals, and let $L(R)$ be the maximum value of (2) among all pedigrees in which all parents are of higher rank than offspring. Essentially, R functions as age, so $L(R)$ may be efficiently determined. The problem of maximizing (2) among all possible pedigrees (that is, in the absence of age data) is then equivalent to the problem of maximizing $L(R)$ over the space of all rankings, which may be done much more efficiently than the direct optimization of (2) over the space of pedigrees. Using this method, it was shown in Almudevar (2003) that multigenerational pedigrees could be accurately estimated in the absence of age data, and this result extends to any posterior density (3) when ϕ is of product form.

The problem of sampling from a posterior density on the space of DAGs poses the same difficulties. As in the optimization problem, the introduction of a ranking constraint appears to offer some advantage. It proves to be easy to sample from a posterior density conditional on a rank R, as well as to sample from the space of ranks, provided the prior is of product form. A hierarchical sampler based on this idea was developed in Friedman and Koller (2003). Unfortunately, as acknowledged in that work, the resulting posterior density is not exact, due to the fact that the number of rankings R to which a graph G conforms is not constant on $\mathcal{G}$. A method of correcting for this error was introduced in Ellis and Wong (2008), but relies on the ability to access sampling history which would not be practical in our application.

Unlike many other fields in which Bayesian network models are applied, ranking constraints (age or generational data) are often available for pedigree inference problems, so it will be useful to develop the Bayesian methodology for the rank constrained model,

recognizing the extension to the unconstrained case (for example, partial or no age data) as a distinct problem.

We next develop a formal definition for our constraints. We assume any graph G is selected from class $\mathcal{G}$ of admissible DAGs. If V is a set of constraints on graphs in $\mathcal{G}$ then $\mathcal{G}[V]$ is the subset of $\mathcal{G}$ which conforms to V. The following type of constraint will be of special interest.

Definition 2. Consider subsets $V = (V_1, \ldots, V_{N_I})$ of $\mathcal{U}$, and let $\mathcal{M}^V$ be the set of all directed graphs for which the parent set S_i of i is a subset of V_i for all i. We say V is a *product form* constraint if $\mathcal{M}^V \subset \mathcal{G}$.

A graph G *conforms* to a product form constraint V if the parents of i are in V_i. Let R be a ranking (or permutation) of $(1, \ldots, N_I)$, so that R assigns a unique rank from 1 to N_I, denoted R_i, to node i. We give the following two lemmas.

Lemma 1. *If $V = (V_1, \ldots, V_{N_I})$ is a product form constraint then we must have $V_i = \emptyset$ for at least one i.*

Proof. If there is no $V_i = \emptyset$, then a graph may be selected which contains no founders, which cannot be a DAG. □

Lemma 2. *Consider subsets $V = (V_1, \ldots, V_{N_I})$ of $\mathcal{U}$. If there exists a ranking R such that for each V_i, the elements of V_i (if any) have higher rank than i then V defines a product form constraint.*

Proof. Suppose a selection G under constraint V is not a DAG. It then contains a cycle, which in turn implies that at least one parent has lower rank under R than one of its' offspring. However, this cannot occur under constraint V. □

Product form constraints may take many forms. If each individual is assigned a unique age this is equivalent to a ranking constraint. If ties in age exist, the constraint is coarsened but is still product form. If it is further required that the age difference between parent and offspring exceeds some threshold, the number of constrained graphs is reduced, but the constraint is still product form. The coarsest type of product form constraint, the *cohort constraint*, is defined when $\mathcal{U}$ is partitioned into indexed cohorts, and a parent must be in one or several cohorts of higher index than an offspring.

2.2. Informative and uninformative priors

Although conventions differ, priors are generally classified as informative or uninformative. The purpose of an informative prior is to introduce information about a specific model which assigns relatively higher probabilities to sets of models which contain, or are near, the true model. For example, geographical data in an informative prior may assign higher probabilities to matings between proximate individuals. Sex data may be equally regarded as prior information, or as a constraint on the model space $\mathcal{G}$. In the following development the model space will not incorporate mating constraints. Any sex data will be regarded as prior information, implemented by considering only admissible matings in our sampling algorithms.

The term *uninformative prior* is often used for a prior which is meant to represent a state of uncertainty or indifference regarding the model, which is not our objective. Our approach will be to formulate a property that a prior should arguably satisfy, then to derive such priors. This is similar to the motivation underlying the well known Jeffreys prior (Welsh, 1996), the defining property of which is that it is invariant under reparametrization. In the case of pedigrees, it will be proposed below that a prior density should be scale invariant, in the sense that the marginal distribution of local properties, such as offspring number, should not depend on N_I. Of course, a prior may ultimately contain both an informative and uninformative component. Once a prior is shown to possess the desired property, probabilities may be reweighted to incorporate various forms of prior information, including population size, sampling sparseness, or geographic data (Nielsen et al., 2001; Neff et al., 2001; Hadfield et al., 2006). An example of an informative pedigree prior is proposed in Egeland et al. (2000) of the form $\phi(G) \propto M_I^{b_I} M_P^{b_P} M_G^{b_G}$ where b_I, b_P, b_G are indices of inbreeding, promiscuity and generation span respectively, and M_I, M_P, M_G are nonnegative influence hyperparameters. Inbreeding, for example, may be forbidden by setting $M_I = 0$. Otherwise, the hyperparameters may be less than or greater than one, indicating a role of penalty or reward for the indices, and are subjectively chosen. While interest in this article is in the uninformative prior, both types may be consolidated into a single prior.

A number of priors have been proposed for general Bayesian network models. Most depend on an influence hyperparameter, by which the relative strength of the conditional and prior densities are adjusted. A number of guidelines are proposed for their choice, but the selection must retain some subjectivity.

In Buntine (1991) $\phi(G) \propto \beta^{n_G}(1-\beta)^{\binom{N_I}{2}-n_G}$ is proposed for Bayesian network models, where n_G is the number of edges in G, and β is a hyperparameter. This assumes that under the prior, each possible edge appears at random with probability β, independently of other choices.

In Heckerman et al. (1995) a prior is proposed of the form $\phi(G) \propto \kappa^{\delta}$ where $\kappa \in (0, 1)$ and δ is the size of the symmetric difference of the edge sets of G and a *prior network structure* G_0. It is interesting to note that if G_0 is the edgeless graph then δ becomes the number of edges in G. If κ is set to $1/N_I$ then $\phi(G)$ becomes similar to the prior considered in Almudevar (2007) (see below for further discussion). The use of the edgeless graph in this way is demonstrated in an example in Heckerman et al. (1995) (although other choices for G_0 are demonstrated). In that work it is proposed to set $\kappa = 1/(N' + 1)$, where N' is the *equivalent sample size*, which reflects the user's confidence in the prior network. Methods for assessing N' for the case in which the conditional densities f_i are multinomial with Dirichlet priors are given, but the choice remains subjective, and subject to sensitivity analysis. In general, the rationale for this prior is entirely different from the one proposed here. We note that the form $\phi(G) \propto \kappa^{\delta}$, where δ is the number of edges, is also proposed in Ellis and Wong (2008), where $\kappa \in (0, 1)$ is chosen subjectively.

In Dash and Cooper (2004) it is argued that the uniform prior, particularly restricted to non-forbidden graphs, may be an effective choice (see also Cooper and Herskovits, 1992). We will make the argument below that restricting the maximum indegree to 2 does not change the fact that under the uniform prior the marginal indegree distribution of a node depends profoundly on N_I. A uniform prior for G (along with priors for the parameters of f_i) is also employed in Giudici and Green (1999) although it is acknowledged that it is not neutral with respect to graph complexity, and the proposed methodology admits alternative structural priors.

In Mukherjee and Speed (2008) a prior density for Bayesian networks (easily extended to general graphical models) is proposed of the form $\phi(G) \propto \exp(\lambda \sum_i w_i h_i(G))$, where $h_i(G)$ is a *concordance function* which rewards graphs which conform to prior beliefs (e.g. existence of favored edges, small maximum indegree), w_i are scalar weights, and λ is an influence hyperparameter. While this prior is informative, the prior belief may be of a quite general nature. For example, one example given of a concordance function measures the agreement of the node degree distribution with the power law distribution found in scale-free networks. In Mukherjee and Speed (2008) the hyperparameter λ is chosen using prior

elicitation, as well as the Jeffrey's scale, which relates odds ratios to degrees of belief (see, for example, Kass and Raftery, 1995).

In Sheridan et al. (2010) a prior density is constructed from a random graph model which yields a scale-free network (in particular, the power law distribution for the node degrees). This is applied to the modeling of gene regulatory networks, which are commonly observed to conform to this model. A number of hyperparameters are required, which may be integrated under their own assigned priors. This prior is more objective than most other proposals since it does not make use of any influence hyperparameter.

The reader is also referred to Angelopoulos and Cussens (2008) for a general discussion on the use of informative priors which may be used in complement with uninformative priors.

We finally note the structural prior proposed in Friedman and Koller (2003):

$$\phi(G) \propto \prod_{i=1}^{N_I} \binom{N_I - 1}{|S_i|}^{-1}. \tag{4}$$

The quantity $\binom{N_I-1}{|S_i|}$ equals the number of parent sets equal in size to the one modeled, when the unconstrained space of DAGs is used. In pedigrees, any available age or sex data will restrict feasible parent sets. If we let $n(s, i)$ be the number of feasible parent sets of size s for node i, then (4) may be replaced by

$$\phi_e(G) \propto \prod_{i=1}^{N_I} n(s, i)^{-1}. \tag{5}$$

This is the prior density proposed in Almudevar (2007), and will be discussed further below.

2.3. Definition of scale invariance

To fix ideas, suppose we were to adopt a uniform prior on $\mathcal{G}$. Then consider the number of parents of a fixed node i. As $N_I \to \infty$ we expect the number of parent sets of size k to increase in proportion to N_I^k, so that under the uniform prior density the probability that i has two parents approaches 1 as $N_I \to \infty$. We may then take the point of view that such local structure distributions should not depend on N_I in this way, since they are primarily determined by mating and breeding habits of the species. If any dependence on N_I is expected, possibly relating to sampling sparseness or population density, this should be modeled explicitly, and not depend on any spurious scaling effects of the prior.

It will be useful to distinguish between two types of empirical quantities derived from G. *Scale quantities* $H(G)$ are proportional to N_I. Examples include cohort size, number of edges, number of offspring with exactly one parent, and so on. *Local quantities* $W(G)$ do not depend on N_I. These are typically local characteristics of a specific node (number of offspring, number of siblings), or rates of scale quantities (i.e. $H(G)/N_I$, where $H(G)$ is a scale quantity).

We will now formally define the notion of *scale invariance* for a prior density $\phi(G)$. For a given graph G the *neighborhood* of node i is denoted $N_i(G) = (U^o, U^p)$ where U^o and U^p are the sets the offspring and parents, respectively, of i.

Definition 3. A prior density ϕ on graph space $\mathcal{G}[V]$ is *scale invariant* if for any node i the marginal prior distribution of $N_i(G)$ does not depend on N_I, or if this dependence vanishes as $N_I \to \infty$.

It will be of some value to devise more precise, and testable, statements of Definition 3, which we term *scale invariance properties*.

Definition 4 (*SI-1*)**.** A prior density ϕ on graph space $\mathcal{G}[V]$ has *scale invariance property* SI − 1 if for any node i, and k for which $P(|S_i| = k) > 0$, the ratio $P(|S_i| = k+1)/P(|S_i| = k)$ does not depend on N_I, or if this dependence vanishes as $N_I \to \infty$.

We do not require that $P(|S_i| = k)$ define a uniform distribution, since the prior density could be easily adjusted to make it so. It is also important to note that $P(|S_i| = k)$ itself may depend on N_I, as long as the relative values do not.

Next, consider a fixed graph G and let $\mathcal{G}_i^+[G, V]$ be the set of graphs in $\mathcal{G}[V]$ which may be formed by adding an edge $j \to i$ to G. In an alternative formulation of the scale invariance property it is assumed that the probability mass of all such graphs is of the same order of magnitude as $\phi(G)$.

Definition 5 (*SI-2*)**.** A prior density ϕ on graph space $\mathcal{G}[V]$ has *scale invariance property* SI − 2 if for any graph $G \in \mathcal{G}[V]$ with $\phi(G) > 0$ the ratio $P(\mathcal{G}_i^+[G, V])/\phi(G)$ does not depend on N_I, or if this dependence vanishes as $N_I \to \infty$.

3. Information theoretic prior densities—definition

The Minimum Description Length (MDL) principle, due to Rissanen (1978) proposes as a model score the length of an encoding of X aided by model G. In order to be decoded, the model itself must be included in the code. We expect the code length of a model to increase with its complexity, hence the effect is to penalize overfitting. It is important to note that this methodology is given in several forms, representing several levels of refinement (Grünwald, 2007). The object of the so-called *crude* or *two-part code* MDL is a code length of form

$$B(G \mid X) = B(X \mid G) + B(G), \tag{6}$$

where $B(G)$ is the code length of the model, and $B(X \mid G)$ is the (model-assisted) code length of the data. Coding theory (Hamming, 1986) predicts that for data with density $g(x \mid G)$ a code for x exists which approaches the minimum possible expected code length of $-\log g(x \mid G)$, where the log base is the number of characters in the coding alphabet (which may be taken to be e without loss of generality). By a reciprocal exponential transformation, we may express (6) as

$$g(G \mid x) \propto g(x \mid G) \exp(-B(G)). \tag{7}$$

We next consider the problem of coding pedigrees, then we will explore the implications of accepting $\phi(G) \propto \exp(-B(G))$ as a prior density.

3.1. Prior densities derived from coding theory

We start with the approach taken in Almudevar (2007), summarized briefly here. Let $N_E(G)$ be the number of edges in G. It will be convenient to consider the adjacency matrix adj(G) of a graph G, which is a 0–1 matrix in which element (i, j) equals 1 if and only if j is a parent of i. This matrix can be represented directly using N_I^2 bits, but this is inefficient when the number of edges is significantly less than this number. To see this, we first note that one of M objects may be coded using $\log_2 M$ bits using a *fixed length* code (we can achieve our purpose without assuming M is a power of 2, so that $\log_2 M$ need not be an integer). Then graph G can be directly coded using

$$B_N(G) = 2N_E(G) \log_2 N_I \tag{8}$$

bits by listing the 2 nodes defining each edge (a node can be coded with $\log_2 N_I$ bits using a fixed length code).

Suppose instead we code an ordered list of the parent sets. We first code the size of the parent set, letting $b(k)$ be the number

of bits required to code integer k. Then suppose there are $n(k, i)$ parent sets of size k available to node i. Using a fixed length code (separately for each parent set size), G may be coded with length

$$B_E(G) = \sum_{i=1}^{N_I} b(|S_i|) + \log_2 n(|S_i|, i). \tag{9}$$

Asymptotic theory developed in Almudevar (2007) (see Appendix A) confirms that under general conditions the code $B_E(G)$ scales with N_I according to

$$B_E(G) = N_E(G) \log_2 N_I + o(N_E(G) \log_2 N_I). \tag{10}$$

This leads to the edge code prior ϕ_e defined in (5), using instead a natural logarithmic scale for the code length. It is interesting to note that $B_E(G)$ and $B_N(G)$ can be compared directly because they are, at least asymptotically, multiples of $\log_2 N_I$, the length of a node label. On this basis we conclude that the *edge code* (9) is more efficient than the *node code* (8).

Now suppose we are given the following adjacency matrix for an order $n = 7$ graph G:

$$\text{adj}(G) = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \mathbf{1} & \mathbf{1} & 0 & \mathbf{1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \mathbf{1} & 1 & 0 & 1 & 0 & 0 & 0 \\ \mathbf{1} & 1 & 0 & 1 & 0 & 0 & 1 \\ \mathbf{1} & 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}. \tag{11}$$

If we use an edge code to code G, the code length will be approximately $15 \log(n)$. Notice, however, that G contains prominent block structure, defining a *block* as any submatrix consisting entirely of ones. In particular, G may be partitioned into three blocks, defined by ordered pairs of row and column labels, namely $F_1 = \{3, 5, 6, 7\} \times \{1, 2, 4\}$, $F_2 = \{6, 7\} \times \{7\}$ $F_3 = \{1\} \times \{4\}$. By convention, the pair of sets defining a block will be referred to as *offspring* and *parent* sets.

Denote the size of the sets defining a block F_i to be u_i and v_i. In our example, $(u_1, v_1) = (4, 3)$, $(u_2, v_2) = (2, 1)$ and $(u_3, v_3) = (1, 1)$. A block F_i may be completely specified by its *sides*, the left-most $u_i \times 1$ submatrix of F_i, and the top-most $1 \times v_i$ submatrix of F_i. The sides of F_1 for G are indicated in bold type in (11). The sides then represent a subset of the edges defining F_i, totaling in number $u_i + v_i - 1$ (the 'corner' need not be counted twice).

We then say an unordered list of disjoint blocks $\tilde{F} = (F_1, \ldots, F_k)$ *represents* graph G if $\text{adj}(G)$ contains all blocks in $\tilde{F}$, and every 1-element of $\text{adj}(G)$ is contained in some block in $\tilde{F}$. Such a representation may always be constructed, if necessary by using a 1×1 block for each edge. We define the *block number* of $\tilde{F}$ to be

$$N_B(\tilde{F}) = \sum_{i=1}^{k} u_i + v_i - 1.$$

The total number of edges in block F_i will be $u_i v_i$, then note inequalities

$$u_i + v_i - 1 \leq u_i v_i; \qquad \min\{u_i, v_i\} \geq 1,$$
$$u_i + v_i - 1 < u_i v_i; \qquad \min\{u_i, v_i\} \geq 2.$$

This implies

$$N_B(\tilde{F}) \leq N_E(G),$$
$$N_B(\tilde{F}) < N_E(G) \text{ if } \max_i \min\{u_i, v_i\} \geq 2.$$

A block representation $\tilde{F}$ of G will generally not be unique, but $N_B(\tilde{F})$ will never be larger than $N_E(G)$, and if a 2×2 block exists in $\text{adj}(G)$ then we may always construct a block representation $\tilde{F}$ for which $N_B(\tilde{F})$ is strictly smaller than $N_E(G)$. In the above example we have $N_B(\tilde{F}) = 9$ and $N_E(G) = 15$, with the difference entirely attributable to block F_1.

This suggests a coding scheme which may be more efficient than an edge code in an asymptotic sense. Suppose, given a block representation $\tilde{F}$ of graph G, we coded the subgraph of G consisting of the sides of each block in $\tilde{F}$, using an edge code. This would result in a code length of

$$B_B(G) = N_B(\tilde{F}) \log_2 N_I + o(N_B(\tilde{F}) \log_2 N_I), \tag{12}$$

provided we could code information giving the block membership of each coded edge in an efficient manner, that is, using code length $o(N_B(\tilde{F}) \log_2 N_I)$. That (12) can be achieved under general conditions was proven in Almudevar (2007) (see Appendix A). We refer to such a code as a *block code*.

In addition to its greater efficiency, the block code has an interpretation which is directly relevant to the problem of pedigree estimation. Consider the two pedigrees G_A and G_B in Fig. B.1. Both have 4 edges, so $N_E(G_A) = N_E(G_B) = 4$. On the other hand, G_B has a 2×2 block, so there is a block representation $\tilde{F}_B$ for which $N_B(\tilde{F}_B) = 3$. Examining G_A, it can be seen that for any block representation $\tilde{F}_A$ we must have $N_B(\tilde{F}_A) = 4$. The block code, under the MDL principle, thus favors G_B over G_A. We will show below that this comparison has a quite precise probabilistic interpretation (Section 4.3), with important consequences for pedigree inference.

3.2. Canonical representation of block code

A block representation of G is generally not unique. Trivially, each edge may form a separate block, but this would be clearly inefficient. A more instructive example is given by the following adjacency matrix:

$$\text{adj}(G) = \begin{bmatrix} 0 & 0 & 0 & 0 & \mathbf{1} & \mathbf{1} \\ 0 & 0 & 0 & 0 & \mathbf{1} & \mathbf{1} \\ 0 & 0 & 0 & 0 & 0 & \mathbf{1} \\ 0 & 0 & 0 & 0 & 0 & \mathbf{1} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{13}$$

Two reasonable choices for block structure would be $F_1 = \{1, 2\} \times \{5\}$ and $F_2 = \{1, 2, 3, 4\} \times \{6\}$ on the one hand, and $F_1' = \{1, 2\} \times \{5, 6\}$ and $F_2' = \{3, 4\} \times \{6\}$ on the other. Clearly, there is a need to devise a canonical block structure. We do this by first enumerating all possible parent sets. For each, a block is formed with the set of all nodes which are offspring of all members of the parent set, assuming that set is nonempty. In this way, all edges are included in exactly one block. For (13) the canonical block structure would consist of F_1', F_2'. As it happens, this block structure has a smaller block number than the given alternative (5 versus 6). This suggests that the problem of determining the optimal block structure would be important for this methodology, but a solution is not known to the authors.

3.3. Block code prior density

In this section we develop a prior density based on the canonical block structure, used throughout this section. Assume there are b blocks in the canonical block representation of graph G, and let f_i, g_i be the number of offspring and parent block sets of size i, $i = 1, \ldots, N_I - 1$. It will also be useful to let h_{ij} be the number of $i \times j$ blocks. The block code prior is then

$$\phi_b(G) \propto \left(b! K_c K_p\right)^{-1}, \quad \text{where}$$

$$K_c = \prod_{i=1}^{N_I-1} \left(\binom{N_I}{i} \right), \qquad K_p = \prod_{i=1}^{N_I-1} \left(\binom{N_I}{i} \right). \tag{14}$$

The construction of the block code is described in detail in Almudevar (2007). Under the canonical block partition, j may belong to more than one block as a parent, but to one block only as an offspring. We may therefore define $F_i(G)$ to be the unique block in which i is an offspring. If i has no parent then $F_i(G) = \emptyset$, otherwise we set $F_i(G) = S_o \times S_p$, where S_o, S_p are the nonempty offspring and parent sets defining the block. Clearly, $i \in S_o$.

Given a fixed node i and graph G we consider constructing $G' = G \cup \{j \to i\}$, assuming $\{j \to i\}$ is not already in G. This can change the block structure in a number of ways. While we may expect an expansion of $F_i(G)$, it is also possible that the net result will be the incorporation of $F_i(G)$ into another block. Additionally, the expansion of $F_i(G)$ may also force the contraction or elimination of another block. For example, if we accept the block partition F_1', F_2' for adj(G) in (13), then add edge $5 \to 3$, then F_1' is expanded from $\{1, 2\} \times \{5, 6\}$ to $\{1, 2, 3\} \times \{5, 6\}$, but F_2' is then contracted from $\{3, 4\} \times \{6\}$ to $\{4\} \times \{6\}$. An *independent block expansion* will refer to an expansion which forces no changes to any other block under the canonical partition.

There will be special interest in the ratios $\phi_b(G')/\phi_b(G)$. We first define rates

$$\begin{aligned} \pi_i^f &= f_i/N_I, \\ \pi_i^g &= g_i/N_I, \quad i = 1, \ldots, N_I - 1, \\ \pi_{ij}^h &= h_{ij}/N_I, \quad i, j = 1, \ldots, N_I - 1, \\ \mu_b &= b/N_I. \end{aligned}$$

Throughout, we assume that f_i, g_i, h_{ij} and b are scale quantities, which implies that $\pi_i^f, \pi_i^g, \pi_{ij}^h$ and μ_b are local quantities. Then define the following coefficients:

$$\lambda_k^f = \begin{cases} \dfrac{\pi_1^f}{1-\pi_1^f}; & k = 0 \\ 2\dfrac{\pi_2^f(1-\pi_1^f)}{\pi_1^f}; & k = 1 \\ (k+1)\dfrac{\pi_{k+1}^f}{\pi_k^f}; & k \geq 1 \end{cases}$$

and

$$\lambda_k^g = \begin{cases} \dfrac{\pi_1^g}{1-\pi_1^g}; & k = 0 \\ 2\dfrac{\pi_2^g(1-\pi_1^g)}{\pi_1^g}; & k = 1 \\ (k+1)\dfrac{\pi_{k+1}^g}{\pi_k^g}; & k \geq 1 \end{cases}. \tag{15}$$

The ratio $\phi_b(G')/\phi_b(G)$ depends only on the coefficients f_k, g_k and b. The strategy taken is therefore to express the transition from G to G' as a sequence of transitions of these coefficients. The methodology is explained in Appendix B. To summarize briefly, Table B.2 lists a number of elementary transitions, along with an associated coefficient. If a transition sequence can be expressed as a sequence of these elementary transitions, then the resulting ratio $\phi_b(G')/\phi_b(G)$ is approximated (with relative error of order $O(1/N_I)$) as the product of their associated coefficients.

4. Information theoretic prior densities—scale invariance

In this section, the scale invariance properties of the edge code and block code prior are verified. We will assume that there are no other constraints other than those implied by the constraint V defined in Section 2.1, which will generally be assumed to be a product form constraint. Of course, sex data is often available, and could be used to prohibit inadmissible matings in a straightforward way.

4.1. Scale invariance of the edge code

In Friedman and Koller (2003) the prior defined in (4) is characterized as generating a marginal uniform prior on the size of the parent set for a fixed node i, while the same characterization is made of (5) in Almudevar (2007). In fact, this is not exactly true for the unrestricted DAG space. To see this, we need only consider the case $N_I = 2$. In this case there are 3 DAGs, two each consisting of the single edge $2 \to 1$ or $1 \to 2$, and the edgeless graph. Then prior density (4) assigns the same value to each, therefore the marginal prior density for the parent set size for each node is $(2/3, 1/3)$ for values $(0, 1)$. The uniformity property (a special case of property SI-1) will hold under various forms of order restrictions, which we now show.

Theorem 1. *Under a product form constraint V and prior density (5) defined on $\mathcal{G}[V]$, the marginal prior distribution of the parent set size is uniform (and therefore satisfies scale invariance property* SI − 1*).*

Proof. Under the product form constraint, the distributions of parent sets $S_1, \ldots, S_{N_I}$ are independent on $\mathcal{G}[V]$. Therefore,

$$\begin{aligned} P(|S_i| = k) &= \sum_{|S_i|=k} Cn(s, i)^{-1} \\ &= n(s, i) Cn(s, i)^{-1} \\ &= C, \end{aligned}$$

where C is a constant of normalization which does not depend on k. This completes the proof. □

Note that on the unconstrained model space $\mathcal{G}$ the parent set distributions are not independent under (5). For example, if $S_1 = \{2\}$, then S_2 cannot contain 1.

We next show that the edge code also satisfies property SI-2.

Theorem 2. *Under a product form constraint V and prior density (5) defined on $\mathcal{G}[V]$, the marginal prior distribution satisfies scale invariance property* SI − 2*.*

Proof. For fixed node i set $N_i = |V_i|$. Assume $N_i \geq 2$. For the edge code prior (5) we have $n(0, i) = 1, n(1, i) = N_i, n(2, i) = N_i(N_i - 1)/2$. Then, for fixed graph G and $G' \in \mathcal{G}_i^+[G, V]$ we have

$$\frac{\phi_e(G')}{\phi_e(G)} = \frac{1}{N_i}, \quad |S_i| = 0 \quad \text{and} \quad \frac{\phi_e(G')}{\phi_e(G)} = \frac{2}{N_i - 1}, \quad |S_i| = 1.$$

If we then compare the probability of a graph G with the probability of the set of graphs obtained by adding one parent to that graph we have

$$\begin{aligned} \frac{\sum_{G' \in \mathcal{G}_i^+[G,V]} \phi_e(G')}{\phi_e(G)} &= 1, \quad |S_i| = 0 \quad \text{and} \\ \frac{\sum_{G' \in \mathcal{G}_i^+[G,V]} \phi_e(G')}{\phi_e(G)} &= 2, \quad |S_i| = 1, \end{aligned} \tag{16}$$

that is, this ratio does not depend on N_i for all $N_i \geq 2$, which completes the proof. □

4.2. Scale invariance of the block code

The block code density is more complex than the edge code, so we consider only property SI-2. We further assume that a cohort constraint V holds, and that the cohort sizes are scale quantities (as defined Section 2.3). We may therefore set $q_i = |V_i|/N_I$, where q_i is a local quantity.

Theorem 3. *Suppose a block code density is defined on $\mathcal{G}[V]$ where V is a cohort constraint. Then the block code density satisfies scale invariance property* SI − 2.

Proof. For fixed node i set $N_i = |V_i|$. Assume $N_i \geq 1$. We consider two cases defined by $|S_i(G)|$.

Case 1: $|S_i(G)| = 0$. Let $V_i^1 \subset V_i$ be the set of nodes which form blocks as single parents. Let k_j be the number of offspring of any $j \in V_i^1$. Then the addition of $j \to i$ for $j \in V_i^1$ results in an independent expansion of a block of dimension $k_j \times 1$, forming new graph G' from transitions $f_{k_j} \to f_{k_j} - 1$ and $f_{k_j+1} \to f_{k_j+1} + 1$. From Case 5 of Table B.2 we have

$$\frac{\phi_b(G')}{\phi_b(G)} = \frac{\lambda_{k_j}^f}{N_I}(1 + O(1/N_I)). \tag{17}$$

Then, if $j \notin V^1$, we must have an independent creation of a new 1×1 block, so that combining cases 1, 2 and 3 of Table B.2 gives

$$\frac{\phi_b(G')}{\phi_b(G)} = \frac{\mu_b^{-1}}{N_I}\lambda_0^f\lambda_0^g(1 + O(1/N_I)). \tag{18}$$

All members of V_i may be used to construct $G' \in \mathcal{G}_i^+[G, V]$, therefore the sum $\sum_{G' \in \mathcal{G}_i^+[G,V]} \phi_b(G')/\phi_b(G)$ consists of $q_i N_I$ terms of the form (17) or (18), each of which consists of local quantities multiplied by factor N_I^{-1} up to relative error $O(1/N_I)$, therefore ϕ_b satisfies property SI-2 for this case.

Case 2: $|S_i(G)| = 1$. It is given that i has one parent, say j, and $F_i(G)$ is a $k_j \times 1$ block. A new graph G' is formed from G by adding edge $j' \to i$ for any $j' \in V_i - \{j\}$. Let V_i^b be the set of all nodes $j' \in V_i$ which form a $k_{j'} \times 2$ block $S_o \times \{j, j'\}$, and let $\mathcal{G}_i^b[G, V]$ be the set of all such graphs $G' = G \cup \{j' \to i\}$ where $j' \in V_i^b$. Then define $\mathcal{G}_i^{+/b}[G, V] = \mathcal{G}_i^+[G, V] - \mathcal{G}_i^b[G, V]$. Whether or not $G' \in \mathcal{G}_i^b[G, V]$ has important implications, so we take the two cases separately.

First, assume $G' \in \mathcal{G}_i^b[G, V]$. Then there are two types of transitions. If $k_j = 1$ then a 1×1 block is removed, and a $k_{j'} \times 2$ block is increased by 1 offspring. This gives transitions $b \to b - 1, f_1 \to f_1 - 1, g_1 \to g_1 - 1, f_{k_{j'}} \to f_{k_{j'}} - 1$ and $f_{k_{j'}+1} \to f_{k_{j'}+1} + 1$. Alternatively, if $k_j > 1$ then a $k_j \times 1$ block is decreased by 1 offspring, and a $k_{j'} \times 2$ block is increased by 1 offspring. This gives transitions $f_{k_j} \to f_{k_j} - 1, f_{k_j-1} \to f_{k_j-1} + 1, f_{k_{j'}} \to f_{k_{j'}} - 1$ and $f_{k_{j'}+1} \to f_{k_{j'}+1} + 1$. Combining the appropriate cases from Table B.2 gives

$$\frac{\phi_b(G')}{\phi_b(G)} = \begin{cases} \mu_b \dfrac{\lambda_{k_{j'}}^f}{\lambda_0^f\lambda_0^g}(1 + O(1/N_I)); & k_j = 1 \\ \dfrac{\lambda_{k_{j'}}^f}{\lambda_{k_j-1}^f}(1 + O(1/N_I)); & k_j > 1. \end{cases} \tag{19}$$

Next, we consider $G' \notin \mathcal{G}_i^b[G, V]$. If $k_j = 1$, this results in an independent expansion of $F_i(G)$ from 1×1 to 1×2, otherwise, this results in a reduction of $F_i(G)$ and the creation of a new 1×2 block. For $k_j = 1$ this results in transitions $g_1 \to g_1 - 1$ and $g_2 \to g_2 + 1$. For $k_j > 1$ this results in transitions $b \to b + 1, f_{k_j} \to f_{k_j} - 1, f_{k_j-1} \to f_{k_j-1} + 1, f_1 \to f_1 + 1$ and $g_2 \to g_2 + 1$. Combining the appropriate cases from Table B.2 gives

$$\frac{\phi_b(G')}{\phi_b(G)} = \begin{cases} \dfrac{\lambda_1^g}{N_I}(1 + O(1/N_I)); & k_j = 1 \\ \dfrac{2\mu_b^{-1}}{N_I}\dfrac{\lambda_0^f\pi_2^g}{\lambda_{k_j-1}^f}(1 + O(1/N_I)); & k_j > 1. \end{cases} \tag{20}$$

We then note that we expect $|V_i^b|$ to be a local quantity, since it is no larger than the number of offspring of the parent of i, so that $|\mathcal{G}_i^{+/b}[G, V]| = q_i N_I(1 + O(1/N_I))$, so that, directly from (20)

$$\frac{\sum_{G' \in \mathcal{G}_i^{+/b}[G,V]} \phi_b(G')}{\phi_b(G)} = \begin{cases} q_i\lambda_1^g(1 + O(1/N_I)); & k_j = 1 \\ q_i \dfrac{2\mu_b^{-1}\lambda_0^f\pi_2^g}{\lambda_{k_j-1}^f}(1 + O(1/N_I)); & k_j > 1. \end{cases} \tag{21}$$

The theorem holds by noting that $\sum_{G' \in \mathcal{G}_i^+[G,V]} \phi_b(G')/\phi_b(G)$ is the sum of (21) and $|V_i^b|$ terms of the form (19), each of which is asymptotically independent of N_I. □

4.3. A comparison of edge code and block code priors

While both the edge code and block code priors ϕ_e and ϕ_b satisfy a scale invariance property, important differences remain. We will illustrate this using a simple example. Suppose we are given genetic data for two sets of individuals, partitioned into known parental and offspring cohorts. Referring to Fig. B.1, suppose we fix 2 individuals d and e of the offspring generation. Then suppose the data permits us to consider two hypothetical pedigrees, graphs A and B, involving individuals a, b and c of the parental generation. The choice reduces to deciding whether a or c is the second parent of e. We will examine the role of the prior densities in this choice. Suppose there are n parents from which to choose. For the first graph A, d and e are half-siblings with one mutual parent and two distinct parents, so there are $n(n-1)(n-2)/2$ ways to choose such a parental configuration. For graph B, d and e are full siblings. There are $n(n-1)/2$ ways to choose such a parental configuration. Thus, the number of half-sibling models is approximately n times the number of full-sibling models, so that the compatibility of the full-sibling model with the data should be regarded as a more unlikely event, and therefore assigned greater significance.

Now suppose we estimate the ratio of the prior probabilities that would be assigned to graphs A and B by an edge code and a block code, making use of the approximations given in (10) and (12). We have already noted that $N_E(A) = N_E(B) = 4$, so for an edge code based prior the ratio would be $O(1)$. Then we have $N_B(A) = 4, N_B(B) = 3$, interpreting B as a 2×2 block. The ratio of prior probabilities for graphs A to B would be $O(1/n)$ under a block code based prior, the reciprocal of the respective frequencies of the isomorphisms of A and B for two fixed offspring as previously discussed. Thus ϕ_b assigns greater evidentiary weight to graph B, in a manner consistent with its relative rarity among randomly chosen graph types.

The effect can be seen more precisely from Theorem 3. Recall that when G' is constructed from G by adding a new parent j' to node i the order of the density ratio is $\phi_b(G')/\phi_b(G) = O(1/N)$ if $|S_i(G)| = 0$, or if $|S_i(G)| = 1$ and $j' \notin V_i^b$, but is $\phi_b(G')/\phi_b(G) = O(1)$ if $|S_i(G)| = 1$ and $j' \in V_i^b$. In the latter case, adding a parent $j' \in V_i^b$ forces i to join an already existing two-parent family. Thus, ϕ_b behaves like ϕ_e when a new edge is added, except for this latter case, to which the block code prior assigns greater evidentiary weight by a factor of n, as predicted by the approximation formulas.

5. Simulation results

In this section we examine the impact of the choice of graph code by performing simulations analogous to those in Almudevar (2007). The focus of the inference is on correctly estimating the total descendant statistic for founder nodes. The statistic is clearly dependent on the structure of the graph, G, and the choice of node, i, and so we let $D_i(G)$ denote the total descendant number for node i given graph G. As the estimation of the graph G from genotype data is subject to various types of error, it is natural to consider the posterior density of $D_i(G)$.

Genotype data is obtained from known test pedigrees, which are generated from replications of the base pedigree B, defined as the union of two subpedigrees, $B1$ and $B2$, shown in Fig. B.2, with founder descendant numbers 39 and 14 respectively. The subpedigrees may therefore be interpreted as different *fitness classes*. We assume that generation data is provided, and that parents are always of the previous generation. Furthermore we suppose a locus has eight alleles, with allele frequencies conforming to a Zipf distribution ($p_i \propto 1/i, i = 1, \ldots, 8$). The three posterior densities considered in the simulation employ the uniform prior (i.e. the likelihood), the edge code prior and the block code prior. Our simulations are based on 5, 10, 15, and 20 replications of the base pedigree, respectively.

The posterior density defined by the uniform and edge code priors is of product form, and so may be sampled using independent sampling of parent sets (see Almudevar, 2007). This is not true of the block code prior, so we resort to the Hastings–Metropolis MCMC method to sample from the posterior density. In our implementation the state-space for the chain is the set of all DAG's conforming to generational order constraints. Given a current graph G, the algorithm selects a child node in the graph with uniform probability, and selects a new pair of parents again with uniform probability. Exchanging the current parent pair with the new parent pair creates a proposal graph G^T.

The estimated posterior densities for $D_i(G)$ of the founders are shown in Fig. B.3, with the true values indicted by a horizontal line. First, the reduction in bias for the edge code in comparison to the likelihood conforms to that reported in Almudevar (2007). It is immediately apparent, however, that the block code prior produces a significantly narrower, but accurate, posterior density than those produced by the likelihood or the edge code prior. This is clearly attributable to the greater evidentiary weight given to the assignment of common parent sets of size two to multiple offspring, and holds for all examples regardless of pedigree size.

6. Example—Atlantic Salmon

As a further demonstration of the method, we use microsatellite data from a sample of 781 Atlantic Salmon (Herbinger et al., 1999). This sample contains 12 large full sibling groups ranging is size from 8 to 140, comprising a total of 760 individuals. For each sibling group the maximum likelihood parental genotype pair was determined. A parent cohort was created using the resulting 24 genetic parents (in effect, the true parents), and by adding M additional spurious parents by sampling genotypes using a population genotype distribution estimated from the true parents. Four sampling models were created by setting $M = 50, 100$, and removing 6 of the true parents for each value (including two complete parent pairs). Suppose $X = (x_1, \ldots, x_{N_I})$ is the multilocus genotype data for individuals $1, \ldots, N_I$, and let G be a pedigree with parent sets S_i. In the posterior density (3) the conditional density $g(X \mid G)$, is set to (2), with prior density ϕ set as uniform, or determined by the edge code or block code.

To explore a number of issues, two other well known parentage assignment applications were also applied to the data, in particular CERVUS (version 3.0.3) (Slate et al., 2000) and COLONY (version 2.0) (Jones and Wang, 2010a). Comparisons must take into account some differences in assumptions, models and application domains. We first note that CERVUS accepts as input distinct offspring and parent samples, with the inference consisting entirely of parentage assignments. On the other hand, COLONY is designed to estimate sibling structure in the absence of one or both parents. In our comparison, therefore, we do not estimate sibling relationships not implied by parent–offspring links, since this is provided only by COLONY. We also note that COLONY requires sex data for parents, so this information was used in all estimates. COLONY and CERVUS both require estimates of the probability that a parent is included in the sample, in which case the known proportion was used. All applications make use of allele frequencies. Since varying methodologies are used, a single set of allele frequency estimates obtained from the complete candidate parent sample was used in all estimates.

CERVUS is based on likelihood ratios for parent–offspring pair or triplet relationships against unrelatedness (or LOD scores). These are considered separately for each offspring, and large LOD scores will imply parentage assignment based on an empirical *delta score* distribution (the difference in LOD scores between the most likely and second most likely parent or parent pair). This means that the procedure is not a true pedigree estimate, but rather a set of parentage assignments possessing high statistical confidence. In our example, we use the 95% critical value of the delta score distribution to accept a parent–offspring pair or triplet, and a pedigree estimate is constructed from these assignments.

COLONY is quite different in that it computes a likelihood score for family clusters, which are sets of offspring in some defined sibling or half sibling relationship, along with members of the candidate parent sample assigned as parents. The likelihood of the pedigree is then the product of the cluster likelihoods. If one or more parents are missing, the likelihood is calculated by conditioning on a genotype distribution of any missing parent, so that COLONY will estimate sibling structure in the absence of parental information, uniquely among the methods considered here.

This leads to an important caveat to be made in our comparison. The model defined in (2) will be strictly correct for Cases 1–2, but not Cases 3–4. This is because in the case of missing parents siblings become founders in the resulting pedigree, but are clearly related. A similar effect holds when only one parent is removed. The consequence of this is that the model, like CERVUS, does not exploit information regarding sibling structure that would be informative for parentage assignment. Of course, model (2) can be modified to model missing parents. In fact, a number of authors (Coombs et al., 2010; Walling et al., 2010) have recently demonstrated that parentage assignment is more accurate when incorporating sibling inference, providing more argument for a true pedigree level inference. However, this issue is not directly relevant to an evaluation of the properties of the priors.

To evaluate the pedigree estimates, it is noted that there are $2 \times 760 = 1520$ separate parent assignments which define the pedigree. An offspring may have less than two parent assignments. A false positive (FP) is an assigned parent of an offspring not truly that offsprings' parent. A false negative (FN) is a true parent of an offspring not assigned in an estimate. Also included in the estimate is the total number of assignments actually made. For CERVUS and COLONY a single estimate is returned. For the Bayesian methods, the average of these quantities within the posterior sample is reported, as well as the quantities associated with the maximum posterior density pedigree.

The results are summarized in Table B.1. In the CERVUS estimates, the FP rate is quite good (1.25% or less) but the FN rate is very high. This reflects the essentially conservative nature of the

approach which, as a series of hypothesis tests, controls for the FP rate only. Of the Bayesian methods, those based on the uniform and edge code priors do not offer an accuracy comparable to the remaining methods. This leaves two methodologies, COLONY and the block code prior method, as yielding satisfactory compromises between FP and FN rates.

We find the relative strengths of the block code prior method and COLONY reflected in the results of Table B.1. The maximum Bayesian estimate is exactly correct for Cases 1–2, with errors (FP, FN) = (30, 30) and (36, 36) respectively for COLONY. For Cases 3–4, COLONY provides the more accurate estimates, with (FP, FN) = (30, 86) and (23, 72) respectively, compared to (FP, FN) = (89, 100) and (45, 100) for the block code prior method.

This naturally raises a question as to whether or not the maximum likelihood criterion is sufficient for this problem, or whether or not some further model selection criterion is needed. It has often been noted that maximum likelihood estimates have a bias toward more complex models, so we will present a brief discussion of the likelihood model used by CERVUS. The cluster likelihood differs from the pedigree likelihood (2) in that it incorporates a likelihood contribution only from parents assigned to some cluster, which penalizes solutions with larger numbers of clusters. A similar effect is present when conditioning on missing parents, and this case was analyzed in Almudevar and Anderson (2011). By comparing the most significant likelihood contributions of competing models, it was found that this form of complexity control can be predicted to work well when $\dot{p}^{4d} \ll 0.5^n$, where $\dot{p}$ is a representative allele frequency, d is the number of loci and n is the largest sibling group size. When this condition fails, the likelihood criterion can be predicted to spuriously split large true clusters. Since only the right side of the inequality depends on the size of the pedigree, it can be concluded that the clustering structure of the likelihood used by COLONY does not entirely solve the scaling problem.

A somewhat more conjectural issue regarding COLONY was also discussed in Almudevar and Anderson (2011). It proved to be quite easy to modify the pedigree estimate calculated by COLONY so as to increase the likelihood by splitting larger true clusters. More generally, whenever several pedigree estimates were available for a common data set, it was found that the one with the *highest* likelihood was the *least* accurate. While this pertains to single generation sibling reconstruction, we find a similar effect in the examples considered here.

COLONY will archive, where feasible, the 1000 highest likelihood estimates. These were available for Cases 1 and 3. The total error FP + FN was calculated for each, as well as a *parent complexity* measure, defined as the number of unique parent sets in the estimate (the 2 missing parents case is considered to be one set, with single and two parent sets considered distinct). Pairwise plots for these measures are shown in Fig. B.4. In both Cases 1 and 3 we find a positive association between log likelihood value and error, to the extent that, at least in these cases, selecting from among the *lowest* log likelihood estimates would yield the *highest* accuracy (it should be stressed that the archive appears to have been confined to the very latest stages of the algorithm's transition history). A strong relationship between complexity and error is also indicated. The lack of a likelihood/complexity gradient in either example suggests that the algorithm has converged to the neighborhood of a local maximum, at which point the likelihood exhibits little sensitivity to complexity. Overall, the relationships between the three measures strongly suggest that a careful application of a suitable complexity or scaling model would improve accuracy (we have already proposed that the higher accuracy of COLONY in Cases 3–4 may be due to the modeling of sibling structure in the absence of parents, which is not included in (2)). In addition, it may be the case that the reported accuracy of COLONY is due not to the determination of the maximum likelihood estimate, but to the inability of its simulated annealing optimization method to conduct a search sufficient in scope to find it.

The same analysis was applied to the block code posterior sample (Fig. B.5). The situation differs considerably, in that higher posterior density values are now associated with greater accuracy. For Case 1, for which the maximum posterior estimate has zero error, there is little variation in the parent complexity within the posterior sample. For Case 3, the same relationship between complexity and accuracy is observed, but the posterior density possesses a clear sensitivity to complexity.

Finally, we note that the running times for the block code based posterior sample were 28.2, 29.1, 27.7 and 29.3 min for Cases 1–4 respectively, using 2×10^9 transitions, and keeping 1000 equally spaced transitions for the sample after a burn-in period of 10^9 transitions. The running times for COLONY were 7, 7, 9 and 14 h for Cases 1–4 respectively.

7. Conclusion

This article has considered the problem of choosing a prior density for pedigree structure for use in a fully Bayesian approach to multigenerational pedigree inference. Focus was on uninformative priors, which could then be used in combination with other informative priors proposed in the literature. It was found that the choice of prior has considerable influence, and must therefore be made with care.

To guide the choice of prior, the property of *scale invariance* was defined. When a structural prior has this property, the marginal prior distribution of the local properties of a pedigree node (number of parents, offspring, etc.) does not depend on the number of nodes in the pedigree. Such priors are found to arise naturally by an application of the Minimum Description Length principle, under which construction of a prior becomes equivalent to the problem of determining the length of a code required to encode a pedigree using the principles of information theory. Two coding systems leading to scale invariant priors are presented. Interestingly, one of them (the block code) yields smaller code lengths in an asymptotic sense, and appears to perform much better as an inference tool than the other (the edge code, also discussed in Almudevar (2007)). In a two generation parentage assignment example, the Bayesian method, using the block code, provided significantly greater overall accuracy than CERVUS 3.0.3, and an accuracy comparable to COLONY 2.0.

This suggests the intriguing possibility that accuracy in inference can be directly related to the problem of determining the most efficient coding method.

Acknowledgments

The author wishes to thank Christophe Herbinger (Dalhousie University) and Patrick O'Reilly (Bedford Institute of Oceanography, Fisheries and Oceans Canada) for the use of the Atlantic Salmon data set. This work was supported by NIH grant HG004648.

Appendix A. Asymptotic code lengths

Let $\mathcal{P}_n$ be the class of probability distributions on $\mathcal{I}_n = \{0, 1, \ldots, n\}$, and let $\mathcal{Y}_n$ be the class of real-valued functions on $\mathcal{I}_n$. For a distribution $P = (p_0, p_1, \ldots, p_n) \in \mathcal{P}_n$ and function $g \in \mathcal{Y}_n$ we denote the expectation operator

$$E_P[g(\omega)] = \sum_{i=0}^{n} g(i)p_i.$$

Suppose we are given a sequence $P^n \in \mathcal{P}_n, n \geq 1$. Define the triangle function $t_n(i) = \min(i, n - i) \in \mathcal{Y}_n$, and let $b_1^n = E_{P^n}[t_n(\omega)]$. We will make use of the following assumptions on the sequence P^n.

(B1) $\lim_{n\to\infty} b_1^n$ exists and is greater than 0 (possibly ∞),

Table B.1
Summary of pedigree estimates for Cases 1–4 based on the Atlantic Salmon data. There are 12 sibling groups with 24 distinct parents, with $M = 50$, 100 spurious parents. For Cases 1–2 all parents are in the sample, for Cases 3–4 6 parents are removed. Quantities represented are denoted FP (false positives), FN (false negatives), N (true number of parents assigned in pedigree), $\hat{N}$ (estimated parents assigned in pedigrees). For the three posterior densities demonstrated, the average values of FP, FN and $\hat{N}$ within the posterior samples are given, as well as the value associated with the maximum posterior density pedigree estimate. For COLONY (COL) and CERVUS (CER) the quantity listed is that associated with a single pedigree estimate returned.

		Bayesian average			Bayesian maximum			COL	CER
		Unif	Edge	Block	Unif	Edge	Block		
Case 1	FP	454.3	338.4	0.28	464	352	0	30	19
$M = 100$	FN	467.2	751.9	12.4	488	822	0	30	880
24 parents	N	1520	1520	1520	1520	1520	1520	1520	1520
	$\hat{N}$	1507.1	1106.6	1507.9	1496	1050	1520	1520	659
Case 2	FP	245.0	197.3	0.29	264	185	0	36	27
$M = 50$	FN	260.4	515.1	11.8	296	548	0	36	456
24 parents	N	1520	1520	1520	1520	1520	1520	1520	1520
	$\hat{N}$	1504.6	1202.2	1508.5	1488	1157	1520	1520	1091
Case 3	FP	704.6	421.1	86.7	694	444	89	30	1
$M = 100$	FN	406.3	564.6	112.5	408	575	100	86	883
18 parents	N	975	975	975	975	975	975	975	975
	$\hat{N}$	1273.3	831.5	949.2	1261	844	964	919	93
Case 4	FP	472.2	268.5	50.1	477	294	45	23	9
$M = 50$	FN	254.5	422.8	111.9	270	431	100	72	720
18 parents	N	975	975	975	975	975	975	975	975
	$\hat{N}$	1192.7	820.8	913.2	1182	838	920	926	264

(B2) $\lim_{n\to\infty} \frac{E_{P^n}[\log(\max(t_n(\omega),1))]}{\log n} = 0$,
(B3) For any $\epsilon > 0$ there exists t_ϵ such that

$$\limsup_{n\to\infty} E_{P^n}\left[\frac{t_n(\omega)}{b_1^n} I\left\{\frac{t_n(\omega)}{b_1^n} < t_\epsilon\right\}\right] \le \epsilon.$$

Let $\tilde{G} = \{G_n : n \ge 1\}$ be a sequence of graphs where G_n is of order n. Let $\tilde{d}_n = (d_1, \ldots, d_n)$ be the list of parent set sizes in G_n. Then, for canonical block representation $\tilde{F}_n$ of G_n, consisting of k_n blocks, let $\tilde{u}_n = (u_1, \ldots, u_{k_n})$, $\tilde{v}_n = (v_1, \ldots, v_{k_n})$ be the dimensions of the k_n blocks, so that the ith block has dimension $u_i \times v_i$. Let P_d^n, P_u^n, P_v^n be the empirical distributions of $\tilde{d}_n, \tilde{u}_n, \tilde{v}_n$, respectively. We make a further assumption that for all large enough n, we have $\max \tilde{d}_n < n/2$, $\max \tilde{u}_n < n/2$ and $\max \tilde{v}_n < n/2$. Then by Theorem 1 of Almudevar (2007), if P_d^n otherwise satisfies conditions (B1)–(B3) then

$$B_E(G_n) = N_E(G_n)\log n + o\left(N_E(G_n)\log n\right). \tag{A.1}$$

Similarly, if we assume that P_u^n, P_v^n each satisfy (B1)–(B3), and further assume that $0 < \liminf_n n^{-1}k_n \le \limsup n^{-1}k_n < \infty$, then by Theorem 2 of Almudevar (2007) we must have

$$B_B(\tilde{F}_n) = N_B(\tilde{F}_n)\log n + o\left(N_B(\tilde{F}_n)\log n\right). \tag{A.2}$$

In summary, approximations (A.1) and (A.2) hold under conditions which describe the distribution of quantities such as parent set size or offspring set size found in graphs G_n, as $n \to \infty$, which essentially state that the means approach a finite nonzero constant (or approach ∞ at a slow enough rate), and that the tail probabilities are not too large.

Appendix B. Ratios of block code densities

It will be convenient to regard the block code density ϕ_b as a restriction of the following function of $2N - 1$ non-negative integers

$$h(f_1, \ldots, f_{N-1}, g_1, \ldots, g_{N-1}, b) = \left(b!K_cK_p\right)^{-1}, \quad \text{where}$$

$$K_c = \prod_{i=1}^{N-1} \binom{\binom{N}{i}}{f_i},$$

$$K_p = \prod_{i=1}^{N-1} \binom{\binom{N}{i}}{g_i},$$

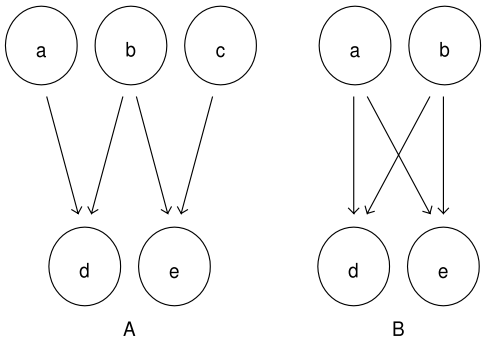

Fig. B.1. Example pedigrees used in Section 4.3.

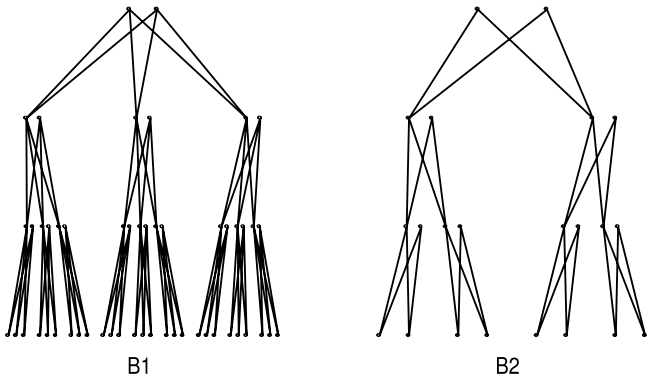

Fig. B.2. Example pedigrees for simulation study, Section 5.

and where the restriction is defined by constraint $\sum_i f_i = \sum_j g_j = b$. Suppose a, d, k are integers with $k \ge 1$ and $a > |d|$. We assume $|d|$ and k are bounded above, and we set $\alpha = a/N$. Then we have approximation

$$\frac{\binom{\binom{N}{k}}{a}}{\binom{\binom{N}{k}}{a+d}} = \tau(\alpha, d, k, N)(1 + O(1/N)), \quad \text{where}$$

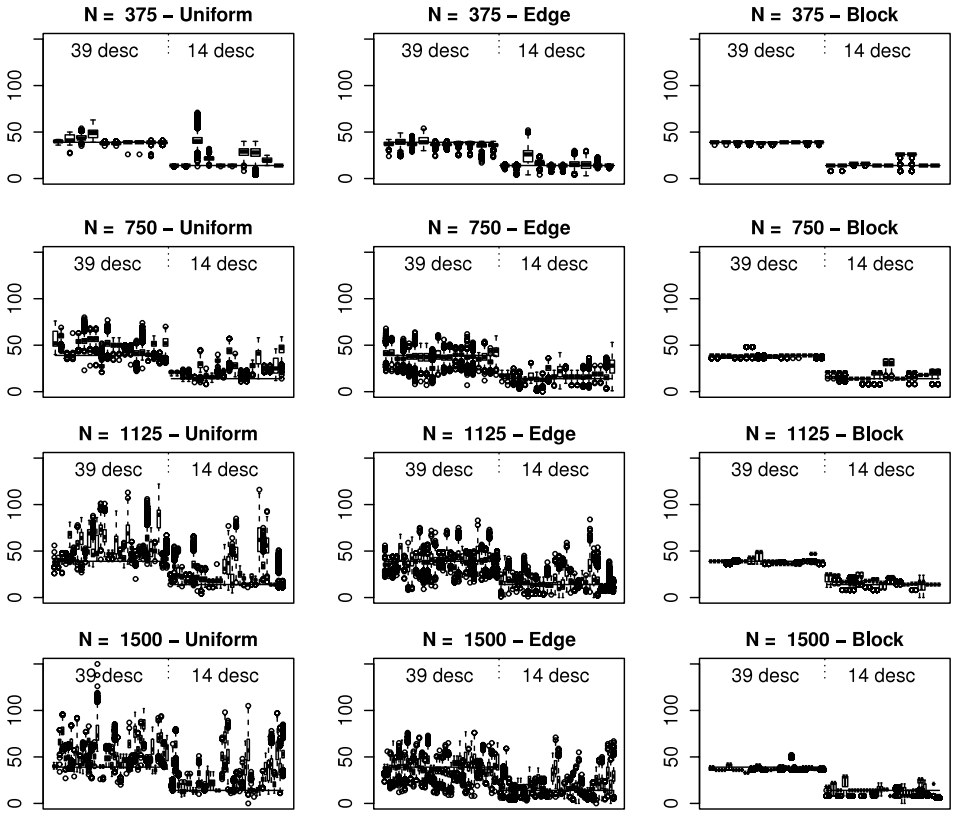

Fig. B.3. Estimated posterior densities of $D_i(G)$ for a varying number of pedigree replications. Two founders are selected from each base pedigree replication, and each score is separately represented. The solid horizontal lines represent the true total descendant number.

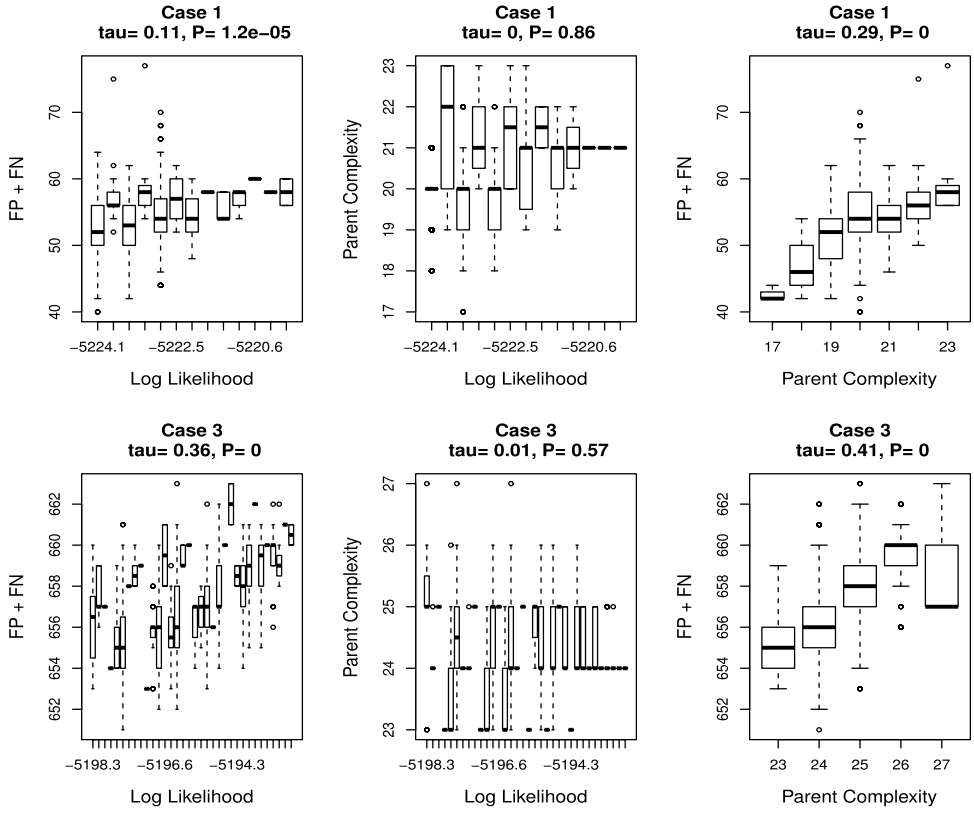

Fig. B.4. *COLONY* 2.0 *analysis*. Pairwise plots of (a) log likelihood; (b) total error FP + FN; (c) parental complexity (number of unique parent sets); obtained from archive of 1000 highest likelihood pedigrees using COLONY 2.0. Cases 1 and 3 are defined in Table B.1. Due to repeated values on the horizontal axis, measurements on the vertical axis are represented as boxplots. Also given is Kendall's tau, which measures pairwise concordance of the axes measurements (−1, 1, 0 represent perfect negative, positive and zero concordance. *P*-values signify rejection of zero concordance). The log likelihood values were rounded for the plots, but not for the concordance estimates.

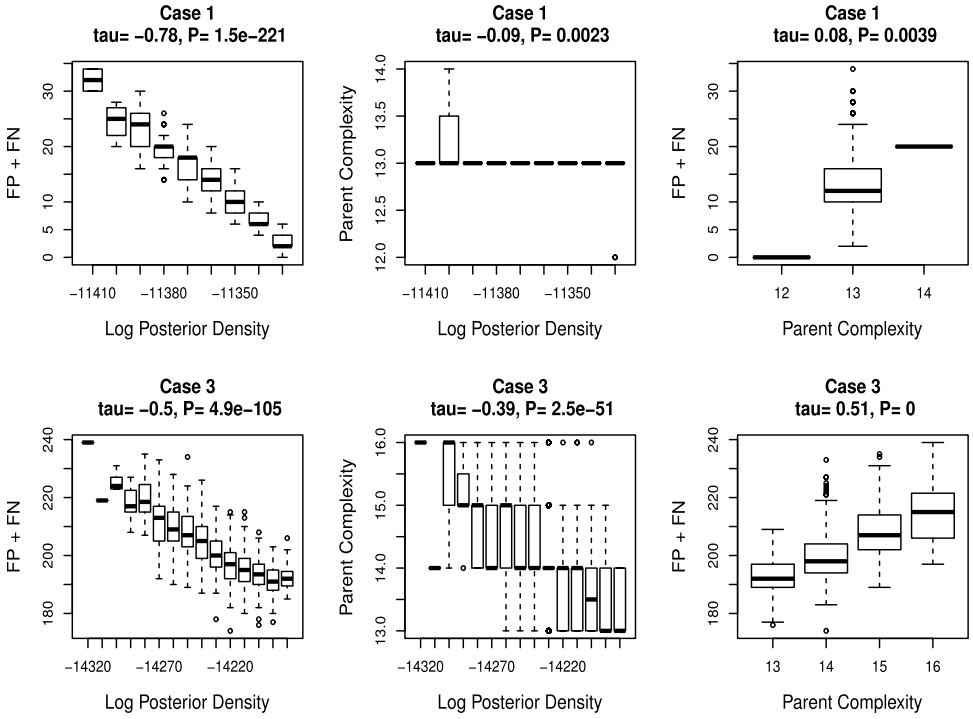

Fig. B.5. *Block code posterior analysis.* Pairwise plots of (a) log posterior density; (b) total error FP + FN; (c) parental complexity (number of unique parent sets); obtained from block code posterior sample. Cases 1 and 3 are defined in Table B.1. Due to repeated values on the horizontal axis, measurements on the vertical axis are represented as boxplots. Also given is Kendall's tau, which measures pairwise concordance of the axes measurements (−1, 1, 0 represent perfect negative, perfect positive and zero concordance. *P*-values signify rejection of zero concordance). The log posterior density values were rounded for the plots, but not for the concordance estimates.

$$\tau(\alpha, d, k, N) = \begin{cases} \left(\dfrac{\alpha}{1-\alpha}\right)^d; & k = 1 \\ \left(\dfrac{\alpha k!}{N^{k-1}}\right)^d; & k > 1. \end{cases} \tag{B.1}$$

Denote vector $\tilde{G} = (f_1, \ldots, f_N - 1, g_1, \ldots, g_N - 1, b)$. We note that h is a product of functions of the components of $\tilde{G}$ taken separately. Based on (B.1) we may estimate the ratio $h(\tilde{G}')/h(\tilde{G})$, where $\tilde{G}'$ is obtained from $\tilde{G}$ through specific transitions. For example, for transition $f_k \to f_k + 1$ we have, for $k \geq 1$,

$$\begin{aligned} \frac{h(\tilde{G}')}{h(\tilde{G})} &= \frac{\binom{N}{f_k}}{\binom{N}{f_k+1}} \\ &= \tau(\pi_k^f, 1, k, N)(1 + O(1/N)) \\ &= \begin{cases} \dfrac{\pi_1^f}{1-\pi_1^f}; & k = 1 \\ \dfrac{\pi_k^f k!}{N^{k-1}}; & k > 1. \end{cases} \end{aligned}$$

Furthermore, the ratio $h(\tilde{G}')/h(\tilde{G})$ for $\tilde{G}'$ obtained from multiple transitions can be obtained in the same way. An important case is $f_k \to f_k - 1, f_{k+1} \to f_{k+1} + 1$ which can be calculated in the same way as

$$\begin{aligned} \frac{h(\tilde{G}')}{h(\tilde{G})} &= \frac{\binom{N}{f_k}\binom{N}{f_{k+1}}}{\binom{N}{f_k-1}\binom{N}{f_{k+1}+1}} \\ &= \tau(\pi_k^f, -1, k, N)\tau(\pi_{k+1}^f, 1, k+1, N)(1 + O(1/N)) \\ &= N^{-1}\lambda_k^f(1 + O(1/N)), \end{aligned}$$

Table B.2
Elementary transitions with associated coefficients, giving the dominant term for ratio $h(\tilde{G}')/h(\tilde{G})$.

Case	Transitions	Dominant term
1	$b \to b + 1$	$N^{-1}\mu_b^{-1}$
2	$f_1 \to f_1 + 1$	λ_0^f
3	$g_1 \to g_1 + 1$	λ_0^g
4	$g_2 \to g_2 + 1$	$N^{-1}2\pi_2^g$
5	$f_k \to f_k - 1,\ f_{k+1} \to f_{k+1} + 1$	$N^{-1}\lambda_k^f$
6	$g_k \to g_k - 1,\ g_{k+1} \to g_{k+1} + 1$	$N^{-1}\lambda_k^g$

using the λ coefficients defined in Eq. (15). Table B.2 gives ratios for several important types of elementary transitions, along with their associated coefficients. In general, for multiple transitions constructed from these elementary transitions, density ratios may be approximated as the product of their associated coefficients. For example, if we wish to calculate the density ratio of graphs G and G', where G' is obtained from G by adding a single 1×1 block, which no other affect on the block structure. This results in transitions $b \to b + 1, f_1 \to f_1 + 1, g_1 \to g_1 + 1$, so by taking the product of the appropriate terms from Table B.2 we have

$$\frac{\phi_b(G')}{\phi_b(G)} = \frac{\mu_b^{-1}}{N}\lambda_0^f\lambda_0^g(1 + O(1/n)).$$

References

Almudevar, A., 2003. A simulated annealing algorithm for maximum likelihood pedigree reconstruction. Theoretical Population Biology 63, 63–75.
Almudevar, A., 2007. A graphical approach to relatedness inference. Theoretical Population Biology 71, 213–229.
Almudevar, A., Efficient coding of labelled graphs. In: Information Theory Workshop, 2007, ITW'07, IEEE, 2–6, 2007, pp. 523–528.
Almudevar, A., Anderson, E.C., 2011. A new version of PRT software for sibling groups reconstruction with comments regarding several issues in the sibling reconstruction problem. Molecular Ecology Resources (early view).

Angelopoulos, N., Cussens, J., 2008. Bayesian learning of Bayesian networks with informative priors. Annals of Mathematics and Artificial Intelligence 54, 53–98.
Buntine, W.L., 1991. Theory refinement in Bayesian networks. In: Uncertainty in Artificial Intelligence: Proceedings of the Seventh Conference. Morgan Kaufmann, San Mateo, CA, pp. 52–60.
Coombs, J., Letcher, B., Nislow, K., 2010. Pedagree: software to quantify error and assess accuracy and congruence for genetically reconstructed pedigree relationships. Conservation Genetics Resources 2, 147–150. doi:10.1007/s12686-010-9202-9.
Cooper, G.F., Herskovits, E., 1992. A Bayesian method for the induction of probabilistic networks from data. Machine Learning 9, 309–347.
Cowell, R.G., 2009. Efficient maximum likelihood pedigree reconstruction. Theoretical Population Biology 76 (4), 285–291.
Dash, D., Cooper, G.F., 2004. Model averaging for prediction with discrete Bayesian networks. Journal of Machine Learning Research 5, 1177–1203.
Egeland, T., Mostad, P.F., Mevåg, B., Stenersend, M., 2000. Beyond traditional paternity and identification cases: selecting the most probable pedigree. Forensic Science International 110, 47–59.
Ellis, B., Wong, W.H., 2008. Learning causal Bayesian network structures from experimental data. Journal of the American Statistical Association 103 (482), 778–789.
Friedman, N., Koller, D., 2003. Being Bayesian about network structure: a Bayesian approach to structure discovery in Bayesian networks. Machine Learning 50, 95–125.
Garant, D., Kruuk, L.E.B., 2005. How to use molecular marker data to measure evolutionary parameters in wild populations. Molecular Ecology 14, 1843–1859.
Giudici, P., Green, P.J., 1999. Decomposable graphical Gaussian model determination. Biometrika 86 (4), 785–801.
Grünwald, P.D., 2007. The Minimum Description Length Principle. MIT Press, Cambridge, Massachusetts.
Hadfield, J.D., Richardson, D.S., Burke, T., 2006. Towards unbiased parentage assignment: combining genetic, behavioural and spatial data in a Bayesian framework. Molecular Ecology 15, 3715–3730.
Hamming, R.W., 1986. Coding and Information Theory. Prentice-Hall, Englewood Cliffs, New Jersey.
Heckerman, D., Geiger, D., Chickering, D.M., 1995. Learning Bayesian networks: the combination of knowledge and statistical data. Machine Learning 20, 197–243.
Herbinger, C., O'Reilly, P.T., Doyle, R.W., Wright, J.M., O'Flynn, F., 1999. Early growth performance of Atlantic Salmon full-sib familes reared in single family tanks or in mixed family tanks. Aquaculture 173, 105–116.
Jones, A.G., Small, C.M., Paczolt, K.A., Ratterman, N.L., 2010. A practical guide to methods of parentage analysis. Molecular Ecology Resources 10, 6–30.
Jones, O.R., Wang, J., 2010a. Colony: a program for parentage and sibship inference from multilocus genotype data. Molecular Ecology Resources 10 (3), 551–555.
Jones, O.R., Wang, J., 2010b. Molecular marker-based pedigrees for animal conservation biologists. Animal Conservation 13, 26–34.
Kass, R.E., Raftery, A.E., 1995. Bayes factors. Journal of the American Statistical Association 90 (430), 773–795.
Koller, D., Friedman, N., 2009. Probabilistic Graphical Models: Principles and Techniques. MIT Press, Cambridge, MA.
Marshall, T.C., Slate, J., Kruuk, L.E.B., Pemberton, J.M., 1998. Statistical confidence for likelihood-based paternity inference in natural populations. Molecular Ecology 7, 639–655.
Mukherjee, S., Speed, T.P., 2008. Network inference using informative priors. Proceedings of the National Academy of Sciences 105, 14313–14318.
Neff, B.D., Repka, J., Gross, M.R., 2001. A Bayesian framework for parentage analysis: the value of genetic and other biological data. Theoretical Population Biology 59, 315–331.
Nielsen, R., Mattila, D.K., Clapham, P.J., Palsbøll, P.J., 2001. Statistical approaches to paternity analysis in natural populations and applications to the North Atlantic humpback whale. Genetics 157, 1673–1682.
Riester, M., Stadler, P.F., Klemm, K., 2009. Franz: reconstruction of wild multgeneration pedigrees. Bioinformatics 25, 2134–2139.
Riester, M., Stadler, P.F., Klemm, K., 2010. Reconstruction of pedigrees in clonal plant populations. Theoretical Population Biology 78, 109–117.
Rissanen, J., 1978. Modeling by the shortest data description. Automatica 14, 465–471.
Sheridan, P., Kamimura, T., Shimodaira, H., 2010. A scale-free structure prior for graphical models with applications in functional genomics. PLoS One 5 (11), e13580.
Slate, J., Marshall, T., Pemberton, J., 2000. A retrospective assessment of the accuracy of the paternity inference program cervus. Molecular Ecology 9 (6), 801–808.
Thomas, S.C., 2005. The estimation of genetic relationships using molecular markers and their efficiency in estimating heritability in natural populations. Philosophical Transactions of the Royal Society, Series B 360, 1457–1467.
Walling, C.A., Pemperton, J.M., Hadfield, J.D., Kruuk, L.E.B., 2010. Comparing parentage inference software: reanalysis of a red deer pedigree. Molecular Ecology 19 (9), 1914–1928.
Welsh, A.H., 1996. Aspects of Statistical Inference. Wiley and Sons, New York, NY.