The conditional independences between variables derived from two independent identically distributed Markov random fields when pairwise order is ignored

ALUN THOMAS[†]

Genetic Epidemiology, University of Utah, 391 Chipeta Way Suite D, Salt Lake City, UT 84108, USA

[Received on 6 January 2009; revised on 2 October 2009; accepted on 5 October 2009]

A result for the equivalence of conditional independence graphs of ordered and unordered vector random variables from first-order Markov models is extended to arbitrary forests. The result is relevant to estimating graphical models for linkage disequilibrium between genetic loci. It explains why, in terms of the conditional independence structure, it sometimes does not matter whether you consider haplotypes or genotypes.

Keywords: graphical modelling; conditional independence graphs; Markov properties; linkage disequilibrium; allelic association.

1. Introduction

In the course of their paper on modelling haplotypes in a population of diploids, [Scheet & Stephens \(2006\)](#) exploit a result that can be stated in a general way as follows.

RESULT 1 *Let $X = (X_1, \dots, X_n)$ and $Y = (Y_1, \dots, Y_n)$ be two independent random realizations from the same joint distribution with the first-order Markov property*

$$P(X = x) = P(X_1 = x_1) \prod_{i=2}^n P(X_i = x_i | X_{i-1} = x_{i-1}).$$

Let $Z(X, Y)$ be the vector obtained by letting each Z_i be the unordered pair $\{X_i, Y_i\}$. Then, Z also has the first-order Markov property, i.e.:

$$P(Z = z) = P(Z_1 = z_1) \prod_{i=2}^n P(Z_i = z_i | Z_{i-1} = z_{i-1}).$$

While at first glance this result seems sensible, those familiar with graphical modelling will find it odd. In a typical application of hidden Markov modelling, X and Y would be the unobservable variables with simple correlation structure introduced in order to explain the more complex structure of the observable Z . Therefore, having Z share the same simple structure negates the point of introducing the hidden variables. Nonetheless, the result is true as is easily shown below. However, it is specific to deriving Z by simply ignoring order and does not extend to higher-order Markov relationships.

[†]Email: alun@genepi.med.utah.edu

The result is particularly relevant in modelling of genetic data. If we assume that X and Y are the states of an individual's alleles at a series of polymorphic loci across a matched pair of paternal and maternal chromosomes, then we cannot typically observe these variables directly, but only the genotypes, i.e. the unordered combination of alleles at each locus. In other words, X and Y are the individual's unobservable 'phase-known haplotypes' and Z is the vector of observable genotypes. The variables X and Y are assumed to be independent, i.e. the locus is said to be in Hardy–Weinberg equilibrium. Although, [Scheet & Stephens \(2006\)](#) use Result 1 in the context of modelling correlations between genetic data, the variables they apply it to are not alleles and genotypes. Instead they apply it to variables indicating what they term 'clusters', meaning the cluster from which an allele or genotype originates.

'Linkage disequilibrium' (LD) is the specific form of 'allelic association', which results from the fact that alleles at genetic loci close to each other tend to be inherited together. Allelic association can also be caused by genetic selection and population admixture. However, LD is the strongest source of association and, very roughly speaking, alleles at loci that are close together are more highly correlated than those more separated in the genome. It is important in genetic studies to account for LD in order to avoid both false-positive results and losing statistical power. Graphical modelling is one approach that can be applied to this problem ([Thomas & Camp, 2004](#); [Thomas, 2005](#); [Verzilli *et al.*, 2006](#); [Thomas, 2007, 2009](#)). The central part of this approach is to estimate a graph that represents the conditional independences between variables given the states of others. In doing so, there is a choice between working directly with the observed genotypes ([Verzilli *et al.*, 2006](#)) or imputing, typically in an iterative way, the unobservable haplotypes and working with these ([Thomas, 2005](#)). The Markov model of Result 1 above is a simple graphical model and the consequence of the result is that the choice between working with observable genotypes and imputed haplotypes is moot in the sense that their correlation structure is represented by the same conditional independence graph.

It is interesting, therefore, to establish the subclasses of graphical model to which Result 1 can be extended. In what follows we first show that the result itself is correct and then observe that it cannot be extended to higher-order Markov relationships but that it does apply to any graphical model whose conditional independence graph is a forest.

For background information on graphical models and graphs, see [Lauritzen \(1996\)](#) and [Golumbic \(1980\)](#).

2. Results

2.1 Proof of Result 1

Consider the probability distribution of the i th unordered pair Z_i given the states of the $(i-1)$ th ordered pair (X_{i-1}, Y_{i-1}) . First, the case when $Z_i = \{u, v\}$ with $u \neq v$.

$$\begin{aligned}
 & P[Z_i = \{u, v\} | (X_{i-1}, Y_{i-1}) = (x, y)] \\
 &= \sum_a \sum_b P[Z_i = \{u, v\}, (X_i, Y_i) = (a, b) | (X_{i-1}, Y_{i-1}) = (x, y)] \\
 &= P[Z_i = \{u, v\}, (X_i, Y_i) = (u, v) | (X_{i-1}, Y_{i-1}) = (x, y)] \\
 &\quad + P[Z_i = \{u, v\}, (X_i, Y_i) = (v, u) | (X_{i-1}, Y_{i-1}) = (x, y)] \\
 &= P[Z_i = \{u, v\} | (X_i, Y_i) = (u, v)] P[X_i = u | X_{i-1} = x] P[Y_i = v | Y_{i-1} = y] \\
 &\quad + P[Z_i = \{u, v\} | (X_i, Y_i) = (v, u)] P[X_i = v | X_{i-1} = x] P[Y_i = u | Y_{i-1} = y].
 \end{aligned}$$

Clearly, since Z_i is derived from (X_i, Y_i) by ignoring the order

$$P[Z_i = \{u, v\} | (X_i, Y_i) = (u, v)] = P[Z_i = \{u, v\} | (X_i, Y_i) = (v, u)] = 1.$$

Define

$$p_i(u|x) = P[X_i = u | X_{i-1} = x].$$

Then, since X and Y come from the same distribution, we have that

$$P[Z_i = \{u, v\} | (X_{i-1}, Y_{i-1}) = (x, y)] = p_i(u|x)p_i(v|y) + p_i(v|x)p_i(u|y). \quad (1)$$

In a similar way, for the case $u = v$, we have that

$$P[Z_i = \{u, u\} | (X_{i-1}, Y_{i-1}) = (x, y)] = p_i(u|x)p_i(u|y). \quad (2)$$

Since the right-hand sides of both (1) and (2) are symmetrical in x and y , we have that Z_i is independent of the order in which they occur. In other words, Z_i is independent of (X_{i-1}, Y_{i-1}) given Z_{i-1} . Applying this recursively to Z_n, Z_{n-1} and so on, we obtain Result 1.

2.2 Second-order Markov property

We can also follow the above steps for the case of sequences of variables with the second-order Markov property. Instead of $p_i(u|x)$ above, we now define

$$q_i(u|a, c) = P[X_i = u | X_{i-1} = a, X_{i-2} = c].$$

Then, (1) and (2) have the corresponding versions

$$\begin{aligned} P[Z_i = \{u, v\} | (X_{i-1}, Y_{i-1}) = (a, b), (X_{i-2}, Y_{i-2}) = (c, d)] \\ = q_i(u|a, c)q_i(v|b, d) + q_i(u|b, d)q_i(v|a, c) \end{aligned}$$

and

$$P[Z_i = \{u, u\} | (X_{i-1}, Y_{i-1}) = (a, b), (X_{i-2}, Y_{i-2}) = (c, d)] = q_i(u|a, c)q_i(u|b, d).$$

From these, we can see that Z_i is again independent of the order of (a, b) and (c, d) taken separately; however, it depends on the order of (a, b) and (c, d) relative to each other. That is, a and c must occur successively in one sequence and b and d in the other. As this information is not derivable from the unordered Z_{i-1} and Z_{i-2} , Result 1 does not extend beyond the case of the first-order Markov property.

2.3 Conditional independence graphs

The central feature of graphical models is the ‘conditional independence graph’ that encodes the relationships between variables. If the joint distribution of a set of variables $X = \{X_1, \dots, X_n\}$ factorizes as

$$P(X_1, \dots, X_n) = \prod_i f_i(T_i), \quad \text{where } T_i \subset \{X_1, \dots, X_n\}, \quad (3)$$

where the f_i are positive functions of their arguments T_i , we can define a conditional independence graph by associating each variable with a vertex and having undirected edges between any two variables that appear as arguments in the same factor.

2.4 Trees and forests

An undirected ‘tree’ is any connected undirected graph with no cycles. A joint distribution of a set of variables has a conditional independence graph that is a tree if and only if the variables can be ordered so that

$$P(X_1, \dots, X_n) = P(X_1) \prod_{i=2}^n P(X_i | X_{\alpha(i)}), \quad \text{where } \alpha(i) < i, \quad \forall i.$$

Choosing X_1 arbitrarily and arranging the others in the order the corresponding vertices are reached in a breadth first, or depth first, search out from the starting point is one way of doing this. For each i , the vertex $\alpha(i)$ will be the first vertex visited in the unique path in the tree from vertex i back to the starting point.

If we now define

$$p'_i(u|x) = P[X_i = u | X_{\alpha(i)} = x],$$

we have results similar to (1) and (2) above

$$P[Z_i = \{u, v\} | (X_{\alpha(i)}, Y_{\alpha(i)}) = (x, y)] = p'_i(u|x)p'_i(v|y) + p'_i(v|x)p'_i(u|y)$$

and

$$P[Z_i = \{u, u\} | (X_{\alpha(i)}, Y_{\alpha(i)}) = (x, y)] = p'_i(u|x)p'_i(u|y).$$

As before, these are symmetrical in x and y so Z_i is independent of $(X_{\alpha(i)}, Y_{\alpha(i)})$ given $Z_{\alpha(i)}$. Also, as before, we can apply this result recursively from n down to 1. In this case, we are effectively recursing from the leaf nodes of the tree back to the arbitrary root. Hence, we get a generalization of Result 1 for any conditional independence graph that is a tree.

An undirected ‘forest’ is any undirected graph with no cycles. Any forest can be considered as a collection of disjoint trees. Since the trees are unconnected, we can apply the above to each tree independently and so Result 1 applies also to forests. Since a tree is a forest with a single component, we can state these generalizations as follows.

RESULT 2 *Let $X = (X_1, \dots, X_n)$ and $Y = (Y_1, \dots, Y_n)$ be two independent random realizations from the same joint distribution with a conditional independence graph that is a forest F . Let the variables $X_1, \dots, X_n$ be associated with vertices $1, \dots, n$ of F , respectively. Let $Z(X, Y)$ be the vector obtained by letting each Z_i be the unordered pair $\{X_i, Y_i\}$. Then, the conditional independence graph of Z is also F with $Z_1, \dots, Z_n$ associated with vertices $1, \dots, n$, respectively.*

3. Discussion

Thomas (2007) showed that low-order Markov models do not model LD appropriately and that richer classes of graphical models such as those of Scheet & Stephens (2006), Thomas (2007) and Verzilli *et al.* (2006) are needed. However, general graphical models, or even the more tractable subclass of decomposable models, are not easy to search in an efficient fashion. Thomas (2009) showed that graphical models with ‘interval’ conditional independence graphs are a rich enough class to model LD well and have a convenient characterization that allows efficient searching of the space. Trees and forests are a similarly well characterized and easy to manipulate class of graphs. The consequence of Result 2 is that

if we stay within this class, we can estimate conditional independence graphs either directly from genotype data or from imputed haplotypes. The graphs estimated from genotypes and haplotypes should, given enough data, be the same. Whether this subclass of models is sufficiently flexible to give good results will depend on the context and even specific data being analysed. Note, however, that much of the initial development of graphical models was done exclusively with trees and forests, and substantial progress was made under this restriction (Kelly & Barclay, 1973; Pearl, 1982, 1986; Kim & Pearl, 1983).

Result 2 addresses only one aspect of a graphical model: the conditional independence graph. The other element of the model is the set of parameters of the marginal or conditional distributions describing how variables depend on each other. These will, of course, be different, in particular, the number of parameters will differ. For example, if we consider the conditional probabilities of one 2-allele locus given another, we have two free parameters to estimate: $P(X_i = 0|X_j = 0)$ and $P(X_i = 0|X_j = 1)$. Since a locus with two alleles has three genotypes, the corresponding genotype relationship will involve six free parameters. In general, it will depend on the strength of correlations between loci, and hence, the ability to reconstruct the haplotypes well, whether it will be better to address the six parameter known data problem or the two parameter imputed data problem.

The ultimate goal of modelling the joint distribution of alleles is usually to infer the complete phase-known haplotypes from the unphased genotypes. Recent applications often observe genotypes at only a subset of the loci at which haplotypes are required (Li *et al.*, 2006; Marchini *et al.*, 2007; Browning & Browning, 2008) and rely on LD modelling to impute the missing data. Such missing data problems are very naturally addressed using graphical models, for instance, the author's programs for estimating LD models based on observed genotypes apply such an imputation iteratively (Thomas, 2005). Under the forest restriction, the structure of such models could be estimated directly from genotype data and then be applied, with suitable parameter re-estimation, to haplotypes. In fact, although this is an informal observation rather than the result of systematic investigation, it is often the case that graphs for models estimated directly from genotypes and from haplotypes imputed from the genotypes have similar structure, even when the graphs are not trees.

Funding

National Institutes of Health (R01 GM81417) and U.S. Department of Defense (W81XWH-07-01-0483) to A.T.

REFERENCES

- BROWNING, S. R. & BROWNING, B. L. (2008) Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.*, **81**, 1084–1097.
- GOLUMBIC, M. C. (1980) *Algorithmic Graph Theory and Perfect Graphs*. New York: Academic Press.
- KELLY, C. W. & BARCLAY, S. (1973) A general Bayesian model for hierarchical inference. *Organ. Behav. Hum. Perform.*, **10**, 388–403.
- KIM, J. H. & PEARL, J. (1983) A computational model for causal and diagnostic reasoning in inference systems. *Proceeding of the 8th International Joint Conference on Artificial Intelligence*. Karlsruhe, Germany, pp. 190–193.
- LAURITZEN, S. L. (1996) *Graphical Models*. Oxford: Clarendon Press.
- LI, Y., DING, J. & ABECASIS, G. R. (2006) Mach 1.0: rapid haplotype reconstruction and missing genotype inference. *Am. J. Hum. Genet.*, **79**, S2290.

- MARCHINI, J., HOWIE, B., MYERS, S., MCVEAN, G. & DONNELLY, P. (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.*, **39**, 906–913.
- PEARL, J. (1982) Reverend Bayes on inference engines: a distributed hierarchical approach. *Proceedings of the AAAI National Conference on Artificial Intelligence*. Pittsburgh, PA, pp. 133–136.
- PEARL, J. (1986) Fusion, propagation and structuring in belief networks. *Artif. Intell.*, **29**, 241–288.
- SCHEET, P. & STEPHENS, M. (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.*, **78**, 629–644.
- THOMAS, A. (2005) Characterizing allelic associations from unphased diploid data by graphical modeling. *Genet. Epidemiol.*, **29**, 23–35.
- THOMAS, A. (2007) Towards linkage analysis with markers in linkage disequilibrium. *Hum. Hered.*, **64**, 16–26.
- THOMAS, A. (2009) Estimation of graphical models whose conditional independence graphs are interval graphs and its application to modeling linkage disequilibrium. *Comput. Stat. Data. Anal.*, **53**, 1818–1828.
- THOMAS, A. & CAMP, N. J. (2004) Graphical modeling of the joint distribution of alleles at associated loci. *Am. J. Hum. Genet.*, **74**, 1088–1101.
- VERZILLI, C. J., STALLARD, N. & WHITTAKER, J. C. (2006) Bayesian graphical models for genomewide association studies. *Am. J. Hum. Genet.*, **79**, 100–112.