Journal of Theoretical Biology 437 (2018) 67–78

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Modeling correlated marker effects in genome-wide prediction via Gaussian concentration graph models

Carlos Alberto Martínez[a,*], Kshitij Khare[b], Syed Rahman[b], Mauricio A. Elzo[a]

[a] Department of Animal Sciences, University of Florida, Gainesville, FL, USA
[b] Department of Statistics, University of Florida, Gainesville, FL, USA

ARTICLE INFO

Article history:
Received 10 November 2016
Revised 25 September 2017
Accepted 15 October 2017
Available online 18 October 2017

Keywords:
Correlated allele substitution effects
Genome-enabled prediction
Graphical models
Sparse covariance estimation

ABSTRACT

In genome-wide prediction, independence of marker allele substitution effects is typically assumed; however, since early stages in the evolution of this technology it has been known that nature points to correlated effects. In statistics, graphical models have been identified as a useful and powerful tool for covariance estimation in high dimensional problems and it is an area that has recently experienced a great expansion. In particular, Gaussian concentration graph models (GCGM) have been widely studied. These are models in which the distribution of a set of random variables, the marker effects in this case, is assumed to be Markov with respect to an undirected graph *G*. In this paper, Bayesian (Bayes G and Bayes G-D) and frequentist (GML-BLUP) methods adapting the theory of GCGM to genome-wide prediction were developed. Different approaches to define the graph *G* based on domain-specific knowledge were proposed, and two propositions and a corollary establishing conditions to find decomposable graphs were proven. These methods were implemented in small simulated and real datasets. In our simulations, scenarios where correlations among allelic substitution effects were expected to arise due to various causes were considered, and graphs were defined on the basis of physical marker positions. Results showed improvements in correlation between phenotypes and predicted additive genetic values and accuracies of predicted additive genetic values when accounting for partially correlated allele substitution effects. Extensions to the multiallelic loci case were described and some possible refinements incorporating more flexible priors in the Bayesian setting were discussed. Our models are promising because they allow incorporation of biological information in the prediction process, and because they are more flexible and general than other models accounting for correlated marker effects that have been proposed previously.

1. Introduction

A feature shared by different statistical models used in genome-wide prediction (Meuwissen et al., 2001) like the family of hierarchical Bayesian regression models known as the Bayesian alphabet (Gianola et al., 2009; Gianola, 2013) and methods based on mixed model equations known as G-BLUP (VanRaden, 2008) and single step G-BLUP (Aguilar et al., 2010), is that marker allele substitution effects are assumed to be mutually independent. Therefore, a diagonal structure for the covariance matrix of these effects is always assumed. However, it is well known that in many populations, especially livestock and plant populations that have undergone selection, linkage disequilibrium (LD) exists. The presence of LD combined with the interactions among genes and complex interactions between gene products taking place in the metabolism point to correlated marker effects. For example, when several SNP are in LD with the same QTL or group of QTL, their effects could be correlated. Another case where SNP effects are expected to be correlated is when a set of SNP are linked to some QTL whose products interact along metabolic pathways.

Thus, there is a need to account for the correlation among marker effects because this is closer to reality and brings advantages like the use of information from correlated marker effects in the prediction of the effect of a given marker, inclusion of biological information in the prediction process and a better knowledge of the covariance structure of marker effects for a particular trait or set of traits. So far, there have been few studies attempting to model a non-diagonal covariance matrix of marker allelic effects in genome-wide prediction. Gianola et al. (2003) proposed Bayesian and frequentist approaches to account for correlated SNP allele substitution effects whereas Yang and Tempelman (2012) developed a hierarchical Bayesian model imposing an autoregressive structure on the covariance matrix of these effects.

* Corresponding author.
E-mail address: carlosmn@ufl.edu (C.A. Martínez).

https://doi.org/10.1016/j.jtbi.2017.10.017
0022-5193/

In statistics, estimation of large covariance matrices when the number of variables is larger than the sample size ("small n, large p") is an important problem of current interest (Khare and Rajaratnam, 2011; Oh et al., 2016). A growing research area to deal with this complex problem is the use of graphical models to find regularized estimators of covariance or precision matrices using both frequentist and Bayesian methods (Carvalho et al., 2007; Letac and Massan, 2007). However, the idea of imposing zeros in the precision matrix is not new; it was proposed by Dempster (1972). Many of these methods share the underlying property of yielding shrinkage estimators (Rajaratnam et al., 2008). Graphical models induce zeros in the covariance or precision matrix, thereby reducing the total number of parameters to be estimated. Models inducing sparsity in the covariance matrix are called covariance graph models, while those inducing sparsity in the precision matrix are called concentration graph models. When the joint distribution of the underlying variables (marker effects in this case) is assumed to be multivariate Gaussian, these zeros translate to appropriate marginal independence assumptions (Gaussian covariance graph models) or conditional independence assumptions (Gaussian concentration graph models).

The patterns of zeros in the covariance or precision matrix can be encoded in terms of an undirected graph G, hence the term "graphical models". (Dempster, 1972; Dawid and Lauritzen, 1993; Silva and Ghahramani, 2006; Khare and Rajaratnam, 2011; Ben-David et al., 2015). The nodes of this undirected graph G represent the underlying random variables. For Gaussian concentration graph models (GCGM), variables not sharing an edge in G are conditionally independent given all other variables (Dawid and Lauritzen, 1993). Equivalently, some authors refer to this property as the undirected Markov property with respect to G (Ben-David et al., 2015). Covariance estimation using graph models is an appealing and very promising topic in statistical genomics because it allows accounting for correlation of marker effects in the prediction of genetic values and phenotypes.

Applications of graphical models in genome-wide prediction have been focused on directed acyclic graph models (AKA Bayesian networks). Some studies applying this kind of graphical models to prediction problems in quantitative genomics are Malovini et al. (2009), Chang and McGeachie (2011), Han et al. (2012), Scutari et al. (2013), and Scutari et al. (2014). To our knowledge, we are the first to adapt GCGM to account for correlated SNP allele substitution effects in genome-wide prediction. A challenge encountered in this process is the following. The theory of GCGM is developed to estimate the precision matrix of an observable p-dimensional random vector using a sample of size n. However, estimation of dispersion parameters in genome-wide prediction involve the problem of estimating residual variance(s) and the more challenging problem of estimating the covariance or precision matrix of an unobservable random vector containing SNP effects using a single n-dimensional vector of phenotypes as well as genotypes. Thus, the objectives of this study were to introduce the theory of GCGM in genome-wide prediction by developing methods to adapt it to perform sparse estimation of the precision matrix of allele substitution effects, and to evaluate its impact on the accuracy of genome-wide prediction.

2. Methods

This section is organized as follows. Because the theory of GCGM is not widely known by quantitative geneticists, it is briefly presented in Section 2.1 along with the challenges encountered when adapting it to genome-wide prediction. Section 2.2 presents the Bayesian approach to solve the problem (Bayes G and Bayes G-D), whereas an EM algorithm is developed in Section 2.3 to provide a frequentist solution (GML-BLUP). In Section 2.4, several approaches to build the graph G, some of them leading to decomposable graphs, are presented. Finally, in Section 2.5, simulated and real datasets used to implement our methods are described.

2.1. Gaussian concentration graph models

In this section, frequentist and Bayesian approaches for Gaussian concentration graph models are presented. The off-diagonal entries of the inverse covariance matrix correspond to the conditional covariance between pairs of variables given all other variables; therefore, under a GCGM, assumptions about the structure of this matrix are made and, as explained in the introduction, this structure is represented using a graph G (Dawid and Lauritzen, 1993; Letac and Massam, 2007).

When the pattern of zeros is unknown, i.e., the graph is unknown; the interest is to find the null entries and to estimate the non-null entries of the inverse covariance matrix. This problem is known as model selection (Bickel and Levina 2008; Rajaratnam et al., 2008; Khare et al., 2015). Here, the case of a known graph G is considered. The scenario of known G encompasses situations where either the pattern of zeros of the precision matrix is actually known or domain-specific knowledge permits the definition of G. Before discussing concentration graph models, the reader not familiar with basic concepts in graph theory is referred to Appendix A.

2.1.1. The estimation problem

The statistical problem is the following. Suppose that $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n$ is a set of vectors in $\mathbb{R}^p$ identically and independently distributed $MVN(0, \Sigma)$. The inverse of the covariance matrix Σ is usually denoted by Ω. In a GCGM, the target is to estimate Ω instead of Σ. Let $G = (V, E)$ be a graph with vertices set V and edges set E, notice that $|V| = p$. As mentioned before, the graph G defines the null entries in Ω and hence it defines the sparsity pattern. The vertices of G represent the set of random variables we are dealing with. Formally, the parameter space is the following cone (Dawid and Lauritzen, 1993): $\Omega \in \mathbb{P}_G = \{A : A \in \mathbb{P}^+ \text{ and } A_{ij} = 0 \text{ whenever } (i, j) \notin E\}$, where $\mathbb{P}^+$ is the space of positive definite matrices.

2.1.1.1. Maximum likelihood estimation. The negative of the log-likelihood function has the form:

$$l(\Omega) = c + \frac{n}{2} tr(\Omega S) - \frac{n}{2} log|\Omega|, \ \ \Omega \in \mathbb{P}_G,$$

where c involves all terms not depending on Ω, and S is the sample covariance matrix defined as:

$$S = \frac{1}{n}\sum_{i=1}^{n} \left(\boldsymbol{Y_i} - \bar{\boldsymbol{Y}}\right)\left(\boldsymbol{Y_i} - \bar{\boldsymbol{Y}}\right)', \ \ \bar{\boldsymbol{Y}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{Y_i}.$$

The negative of the log-likelihood can be slightly modified to obtain the following objective function whose minimization is equivalent to the minimization of $l(\Omega)$:

$$l^*(\Omega) = tr(\Omega S) - log|\Omega|, \ \ \Omega \in \mathbb{P}_G \tag{2.1}$$

In general, there is no closed form solution. An iterative proportional fitting algorithm developed by Speed and Kiiveri (1986) allows minimizing $l^*(\Omega)$. It is based on a partition of the covariance matrix according to the maximal cliques of G.

The properties of the graph have mathematical and statistical consequences on the estimation problem (Letac and Massan, 2007; Khare and Rajaratnam, 2011; Khare and Rajaratnam, 2012). Therefore, it is important to assess certain properties of G because their impact on the estimation problem can be advantageous. In this case, it turns out that when n is greater than the size of the largest clique in G, $l(\Omega)$ is strictly convex and consequently it has a unique