Gaussian covariance graph models accounting for correlated marker effects in genome-wide prediction

Carlos Alberto Martínez[1], Kshitij Khare[2], Syed Rahman[2], Mauricio A. Elzo[1]

[1]Department of Animal Sciences

[2]Department of Statistics

University of Florida, Gainesville, FL, USA

Correspondence: Carlos Alberto Martínez, Department of Animal Sciences, University of Florida, Gainesville, FL 32611, USA.

Tel: 352-328-1624.

Fax: 352-392-7851.

E-mail: carlosmn@ufl.edu

Abstract

Several statistical models used in genome-wide prediction assume uncorrelated marker allele substitution effects, but it is known that these effects may be correlated. In statistics, graphical models have been identified as a useful tool for covariance estimation in high dimensional problems and it is an area that has recently experienced a great expansion. In Gaussian covariance graph models (GCovGM), the joint distribution of a set of random variables is assumed to be Gaussian and the pattern of zeros of the covariance matrix is encoded in terms of an undirected graph G. In this study, methods adapting the theory of GCovGM to genome-wide prediction were developed (Bayes GCov, Bayes GCov-KR and Bayes GCov-H). In simulated datasets, improvements in correlation between phenotypes and predicted breeding values and accuracies of predicted breeding values were found. Our models account for correlation of marker effects and permit to accommodate general structures as opposed to models proposed in previous studies which consider spatial correlation only. In addition, they allow incorporation of biological information in the prediction process through its use when constructing graph G, and their extension to the multiallelic loci case is straightforward.

Key words Correlated marker effects, genome-enabled prediction, graphical models, high-dimensional covariance estimation.

Introduction

Most of the Bayesian and classical models used in genome-wide prediction (Meuwissen et al., 2001) assume that marker allele substitution effects follow independent distributions which induces a diagonal covariance matrix; however, some biological phenomena point to non-independent effects. On one hand, the existence of linkage disequilibrium (LD) may create a spatial correlation of marker effects (Gianola et al., 2003; Yang and Tempelman, 2012). On the other hand, the complex interactions between regions of the genome and interactions of gene products in the

metabolism also suggest that the assumption of independent effects may not be tenable. Thus, accounting for correlated marker allele substitution effects may increase the predictive performance of statistical models used in genome-wide prediction. Although it has been known that marker effects might be correlated, the problem of accounting for such a correlation has not been widely studied. So far, there have been few studies investigating this interesting problem. Gianola et al., (2003) described a series of frequentist and Bayesian models accounting for within chromosome correlated marker effects. Yang and Tempelman (2012) proposed a Bayesian antedependence model considering a nonstationary correlation structure of SNP effects. The two studies only considered correlations among nearby markers.

Covariance estimation is recognized as a challenging problem in statistics (Stein, 1975), especially in high dimensional problems under the "big p small n" condition where the sample covariance matrix is not of full rank (Rajaratnam et al., 2008). As a consequence, high dimensional covariance estimation using graphical models is a contemporary topic in statistics and machine learning. Regularization methods imposing sparsity on estimators through structural zeros in the covariance or inverse covariance matrix have gain attention during recent years, (Carvalho et al., 2007; Letac and Massan, 2007; Rajaratnam et al., 2008). In these models, the pattern of zeros of the covariance (covariance graph models) or precision matrix (concentration graph models) is defined by means of an undirected graph G. The nodes of this graph represent the underlying random variables, and when the joint distribution of these variables is multivariate Gaussian, pairs of nodes not sharing an edge in G are either, marginally independent (Gaussian covariance graph models) or conditionally (given all other variables) independent (Gaussian concentration graph models). This paper focuses on Gaussian covariance graph models (GCovGM). In statistics, the usefulness of these models in the analysis of high dimensional data exhibiting dependencies is well known (Carvalho et al., 2007; Rajaratnam et al., 2008); consequently, given the need for flexible statistical methods to account for

correlated marker effects in genome-wide prediction, the introduction of GCovGM in this area seems promising. Until now, application of graphical models in quantitative genetics and genomics has entailed miscellaneous problems like pedigree and linkage analysis, detection of QTL (Lauritzen and Sheejan, 2003), causal inference and prediction of genetic values (Rosa et al. 2016), identification of non-informative molecular markers (Scutari et al. 2013) and estimation of linkage disequilibrium networks (Morota et al. 2012). These applications mainly used directed acyclic graph models and none of them addressed the problem of high dimensional covariance estimation.

To our knowledge, this is the first study adapting the theory of GCovGM to account for correlated SNP allele substitution effects in genome-wide prediction. The theory of GCovGM has been developed to estimate the covariance matrix of an observable p-dimensional random vector using N iid observations. In contrast, in genome-wide prediction, the problem involves predicting marker effects, estimating residual variance(s), and estimating the covariance matrix of an unobservable random vector (SNP effects) using one n-dimensional vector with phenotypic information along with genomic information. Thus, the objective of this study was to develop methods that adapt the theory of GCovGM to genome-wide prediction in order to account for correlated marker allele substitution effects.

Materials and methods

This section is split into the following subsections. Firstly, due to the fact that GCovGM theory is not widely known in the realm of quantitative genetics, a brief introduction and details on the challenge encountered when adapting it to genome-wide prediction are presented. Then, statistical methods adapting GCovGM to genome-wide prediction are described along with some approaches to build the graph G. Finally, datasets used to implement our methods are described.

Gaussian Covariance Graph Models

Here, the case of a known graph G is considered. By known G it is meant that the pattern of zeros in the covariance matrix is actually known or that G is defined on the basis of domain-specific knowledge. Some basic concepts in graph theory are provided in supporting information (Appendix A); the reader not familiar with this topic is encouraged to read it before reading the rest of the paper. Hereinafter, the operator $|\cdot|$ represents the determinant when the argument is a matrix and cardinality when the argument is a set. Let $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_N$ be a set of vectors in $\mathbb{R}^p$ identically and independently distributed $MVN(0, \Sigma)$, the target is to estimate Σ. The graph G determines the null entries of Σ as explained above and consequently the parameter space is defined as follows. Let $G = (V, E)$ be an undirected graph with vertex set V and edge set E, then Σ lies in the cone $\mathbb{P}_G = \{A: A \in \mathbb{P}^+ \; and \; A_{ij} = 0 \; whenever \; (i,j) \notin E\}$, where $\mathbb{P}^+$ is the space of positive definite matrices. Thus, $\mathbb{P}_G$ corresponds to the set of all positive definite matrices having null entries whenever the corresponding variables do not share an edge in G. Maximum likelihood estimation is possible only when $N > p$ and because of the constraints that it imposes when adapting GCovGM in genome-wide prediction (see supplementary material, Appendix B) this paper focuses on Bayesian approaches only.

Bayesian estimation

For natural exponential families (as in concentration graph models) a class of conjugate priors corresponding to the Diaconis-Ylvisaker prior (Diaconis and Yilvisaker, 1979) is frequently used. However, covariance graph models correspond to curved exponential families instead of natural exponential families. It is easily checked because $L(\Sigma) \propto \exp\left(-\frac{N}{2} tr(\Sigma^{-1} S) - \frac{N}{2} log|\Sigma|\right), \Sigma \in \mathbb{P}_G,$ where S is the sample covariance matrix, notice that $L(\Sigma)$ does not have the form of a natural exponential family. Silva and Ghahramani (2009) introduced the family of conjugate priors known as inverse G-Wishart ($IGW(U, \delta)$) whose probability density function (pdf) has the

following form: $\pi_{U,\delta}(\Sigma) \propto \exp\left(-\frac{tr(\Sigma^{-1}U)}{2} - \frac{\delta}{2}log|\Sigma|\right), \Sigma \in \mathbb{P}_G$. Let $Y \coloneqq (\boldsymbol{Y}_1, \boldsymbol{Y}_2, \dots, \boldsymbol{Y}_N)$. Under this prior: $\pi_{U,\delta}(\Sigma|Y) \propto L(\Sigma)\pi_{U,\delta}(\Sigma) \propto \exp\left(-\frac{1}{2}tr\left(\Sigma^{-1}(U+NS)\right) - \frac{N+\delta}{2}log|\Sigma|\right), \Sigma \in \mathbb{P}_G$. This corresponds to a $IGW\left(\tilde{U}, \tilde{\delta}\right)$ distribution, $\tilde{U} \coloneqq U + NS, \tilde{\delta} \coloneqq N + \delta$. An important issue that has to be considered now is for which values of matrix U and the shape parameter δ, $\pi_{U,\delta}(\cdot)$ is a valid density. To find sufficient conditions the modified Cholesky decomposition of Σ, $\Sigma = LDL'$, where L is a lower triangular matrix with diagonal entries equal to one and D is a strictly positive diagonal matrix, is used. Then, we have the following transformation (a bijection) $\left\{\Sigma_{ij}\right\}_{i\geq j,(i,j)\in E} \rightarrow \left(\left\{L_{ij}\right\}_{i>j,(i,j)\in E}, D\right)$, which induces the density $\pi_{U,\delta}(L,D) \propto \exp\left(-tr(D^{-1}L^{-1}U(L')^{-1}) - \frac{1}{2}\sum_{j=1}^{p}\left(\delta + 2n_j\right)logD_{jj}\right)$ where $n_j = |\{i: i > j, (i,j) \in E\}| \; \forall \, j = 1,2,\dots,p-1$. From the mathematical point of view, the problem is to find sufficient conditions for the following integral to be finite:

$$\int_{\mathbb{R}^{|E|}} \int_{\mathbb{R}_+^p} \pi_{U,\delta}(L,D)dDdL,$$

after some manipulations, it can be shown that these conditions are the following (Khare and Rajaratnam, 2011). 1) $U \in \mathbb{P}^+$, 2) $\delta - 2n_j > v_k + 2 \; \forall \, j = 1,2,\dots,p-1, \forall \, k = 2,3,\dots,p$, where $v_k = |\{i < k: (i,k) \in E\}|$. Thus, the edge set of G defines the set of values that the shape parameter δ can take because from the second condition above it follows that $\delta > v_* + 2n_* + 2$ where $v_* = \max_{2\leq k\leq p}\{v_k\}, n_* = \max_{1\leq j\leq p-1}\{n_j\}$. For covariance graph models, there is a block Gibbs sampler algorithm to draw samples from the posterior. This sampler is based on partitioning the covariance matrix as: $\Sigma = \begin{bmatrix}\Sigma_{11} & \Sigma'_{.1} \\ \Sigma_{.1} & \Sigma_{-1,-1}\end{bmatrix}$ and it uses the following result. Let $\boldsymbol{\beta}_1 \coloneqq \left(\Sigma_{1j}\right)_{(1,j)\in E}$, i.e., a vector containing the unconstrained (non-null) covariance parameters for variable 1, $\gamma_1 = \Sigma_{11} - \boldsymbol{\Sigma}'_{.1}\Sigma^{-1}_{-1,-1}\boldsymbol{\Sigma}_{.1}$, and $Q_1 =$ a matrix of zeros and ones such that: $\boldsymbol{\Sigma}_{.1} = Q_1\boldsymbol{\beta}_1$, then:

$\boldsymbol{\beta}_1|Q_1,\gamma_1,\Sigma_{-1,-1}\sim MVN\left(A^{-1}Q_1'\Sigma_{-1,-1}^{-1}\boldsymbol{U}_{.1},\gamma_1 A^{-1}\right)$ and $\gamma_1|Q_1,\boldsymbol{\beta}_1,\Sigma_{-1,-1}\sim IG\left(\frac{\delta}{2}-1,\frac{U_{11}-2\boldsymbol{U}_{.1}'\Sigma_{-1,-1}^{-1}Q_1\boldsymbol{\beta}_1+\boldsymbol{\beta}_1'A\boldsymbol{\beta}_1}{2}\right)$, where $A\coloneqq Q_1'\Sigma_{-1,-1}^{-1}U_{-1,-1}\Sigma_{-1,-1}^{-1}Q_1$ and $IG(\cdot,\cdot)$ denotes the Inverse Gamma $(\cdot,\cdot)$ distribution. Using this result and permutations, the partition can be done for the p random variables in every step. Hence, this is not a standard Gibbs sampler because partitions change in every step; however, convergence can be established using results from Asmussen and Glynn (2011). Notice the role of graph G when constructing this block Gibbs sampler, it defines $\boldsymbol{\beta}_j$ and $Q_j, j=1,2\ldots,p$.

The Khare-Rajaratnam family of flexible priors for decomposable graphs

When G is decomposable and its vertices are ordered according to a perfect elimination scheme (Khare and Rajaratnam, 2012), there exists a wider family of more flexible priors developed by Khare and Rajaratnam (2011). The parameter δ of the $IGW(U,\delta)$ family is common for all D_{ii}; however, for decomposable graphs a more flexible prior with pdf of the form $\bar{\pi}_{U,\boldsymbol{\delta}}(\Sigma)\propto \exp\left(-\frac{1}{2}tr(\Sigma^{-1}U)-\sum_{i=1}^{p}\frac{\delta_i}{2}logD_{ii}(\Sigma)\right),\Sigma\in\mathbb{P}_G,U\in\mathbb{P}^+,\boldsymbol{\delta}=(\delta_1,\delta_2,\ldots,\delta_P)$ can be used. In this prior density, every D_{ii} has its own shape parameter δ_i. The price paid for this extra flexibility is that the graph G has to be decomposable. When considering the modified Cholesky decomposition of the covariance matrix, the density in terms of L and D is:

$$\bar{\pi}_{U,\boldsymbol{\delta}}(L,D)\propto \exp\left(-\frac{1}{2}tr((L')^{-1}D^{-1}L^{-1}U)-\sum_{i=1}^{p}\frac{\delta_i-2n_i}{2}logD_{ii}\right),L\in\mathcal{L}_G,D\in\mathcal{D}$$

Sufficient conditions for this to be a proper density are: $U\in\mathbb{P}^+$, $\delta_i>2n_i+v_i+2$ (Khare and Rajaratnam, 2011). This prior is conjugate because the posterior density, given by:

$$\pi(L,D|Y)\propto \exp\left(-\frac{1}{2}tr\left((L')^{-1}D^{-1}L^{-1}(U+NS)\right)-\sum_{i=1}^{p}\frac{N+\delta_i-2n_i}{2}logD_{ii}\right),L\in\mathcal{L}_G,D\in\mathcal{D},$$

is a $\bar{\pi}_{\tilde{U},\tilde{\boldsymbol{\delta}}}(L,D)$ density, where $\tilde{U}$ is as defined above and $\tilde{\boldsymbol{\delta}}_{p\times 1}=\{\tilde{\delta}_i\}=N+\delta_i-2n_i$. Hereinafter, this family of priors will be denoted as GWKR($\boldsymbol{\delta},U$). If in addition to be decomposable the graph is also homogeneous, direct sampling from the posterior can be performed (this case is discussed later), otherwise MCMC methods are used to draw samples from the posterior. Details of a block Gibbs sampler and the proof of its convergence can be found in Khare and Rajaratnam (2011). The full conditional distributions used in Khare and Rajaratnam's Gibbs sampler (Khare and Rajaratnam, 2011) are the following. Let $G=(V,E)$ be a decomposable graph with its vertices ordered according to a perfect elimination scheme (see Appendix A), let LDL' be the modified Cholesky decomposition of the covariance matrix Σ and let $L_{\cdot v}^{G}=(L_{uv})_{u>v,(u,v)\in E}, v=1,2,\dots,p-1$. Then: $L_{\cdot v}^{G}|L\setminus L_{\cdot v}^{G},D,Y\sim N(\mu^{v,G},M^{v,G})\ \forall\ v=1,2,\dots,p-1$, where

$$\mu_u^{v,G}=\mu_u^v+\sum_{u'>v:(u',v)\in E}\ \sum_{\substack{w>v:(w,v)\notin E\\ or\ w<v:\ (L^{-1})_{vw}=0}} M_{uu'}^{v,G}\left(L^{-1}\tilde{U}(L')^{-1}\right)_{vv}\left((LDL')^{-1}\right)_{u'w}\mu_w^v\ \forall\ u>v,(u,v)\in E$$

$$\mu_u^v=\frac{(L^{-1}\tilde{U})_{vu}}{(L^{-1}\tilde{U}(L')^{-1})_{vv}}\ \ \forall\ u\ \text{ such that }\ (L^{-1})_{vu}=0,\ \ ((M^{v,G})^{-1})_{uu'}=\left(L^{-1}\tilde{U}(L')^{-1}\right)_{vv}((LDL')^{-1})_{vv}$$

$\forall\ u,u'>v,(u,v),(u',v)\in E$ and $D_{ii}|L,Y\sim IG\left(\tilde{\delta}_i/2\,,\left(L^{-1}\tilde{U}(L')^{-1}\right)_{ii}/2\right), i=1,2,\dots,p$. In the definition of $\mu^{v,G}$, notation $w:(L^{-1})_{vw}=0$ refers to functional zeros, that is, $(L^{-1})_{vw}$ is zero as a function of the entries of L. Finally, operator "$B\setminus A$" is the relative complement of set A with respect to a set B, also known as the difference of sets A and B; it denotes elements in B but not in A.

Covariance graph models for homogeneous graphs

For covariance graph models, certain properties of the graph $G=(V,E)$ have appealing mathematical consequences on the estimation problem. Covariance graph models take advantage of the fact that homogeneous graphs admit a Hasse ordering of their nodes (see Appendix A). The importance of having a graph with this sort of ordering is summarized in the following theorem (Khare and Rajaratnam, 2011). Let $G=(V,E)$, be a homogeneous graph with a Hasse ordering of its

nodes. Then, $\Sigma = LDL' \in \mathbb{P}_G \Leftrightarrow L \in \mathcal{L}_G \Leftrightarrow L^{-1} \in \mathcal{L}_G$, that is, matrices L and L^{-1} preserve the pattern of zeros in Σ. This theorem is very relevant for the estimation problem we are dealing with because when G is homogeneous, it permits to easily obtain direct samples from the posterior by reparametrization in terms of $T = L^{-1}$. Let $\boldsymbol{x}_i := \{T_{ij}\}_{j<i,(i,j)\in E}$, then, it follows that given D the random vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{m-1}$ are mutually independent and distributed as follows $\boldsymbol{x}_i | D \sim MVN((U^{<i})^{-1}\boldsymbol{U}_{.i}^{<}, D_{ii}(U^{<i})^{-1})$. In addition, $D_{11}, D_{22}, \ldots, D_{pp}$ are also mutually independent with the following marginal distributions $D_{ii} \sim IG\left(\frac{\delta - 2n_i - v_i}{2} - 1, \frac{U_{ii} - (\boldsymbol{U}_{.i}^{<})'(U^{<i})^{-1}\boldsymbol{U}_{.i}^{<}}{2}\right)$, where $v_i = |\{j : j < i, (i,j) \in E\}|$ and $U^{<i}, \boldsymbol{U}_{.i}^{<}$ and U_{ii} correspond to the elements of matrix U when it is partitioned as: $\begin{pmatrix} U^{<i} & \boldsymbol{U}_{.i}^{<} \\ (\boldsymbol{U}_{.i}^{<})' & U_{ii} \end{pmatrix}$, $U^{<i} = \left(U_{jk}\right)_{j,k<i,(i,j),(i,k)\in E}$, and $\boldsymbol{U}_{.i}^{<} = \left(U_{ji}\right)_{j<i,(i,j)\in E}$. From these conditional and marginal distributions, direct sampling can be performed.

We want to close this section by emphasizing that the existence of a more flexible family of prior distributions for decomposable GCovGM and the simplification of the estimation problem in the case of homogeneous GCovGM are examples of the benefits of encoding the covariance structure in terms of a graph because by taking advantage of some of its properties, the estimation problem can be modified in such a way that appealing features (e.g., more generality, simplification of algorithms, closed form expressions) emerge.

Adapting GCovGM to genome-wide prediction

In this section, we explain why GCovGM theory cannot be directly applied to genome-wide prediction, and we describe the challenges that have to be overtaken when adapting this theory to account for correlated marker effects. The key point is that the estimation problem is not the same as the one described in the previous sections; hereinafter, this estimation problem will be referred to as the standard problem. Now, we describe the statistical problem found in genome-wide prediction. The model considered here is the following multiple linear regression model:

$$\boldsymbol{y} = W\boldsymbol{g} + \boldsymbol{e} \quad (1)$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is an observable random vector containing response variables (e.g., corrected phenotypes or deregressed BV), $\boldsymbol{g} \in \mathbb{R}^m$ is an unknown vector of marker allele substitution effects, $\boldsymbol{e} \in \mathbb{R}^n$ is a vector of residuals, $W_{n\times m}$ is a matrix whose entries correspond to one to one mappings from the set of genotypes to a subset of the integers for every individual at every locus $W = \{w_{ij}\} = \begin{cases} 1, if\ genotype = BB \\ 0, if\ genotype = BA \\ -1, if\ genotype = AA \end{cases}$, where w_{ij} is map corresponding to the genotype of the i^{th} individual for the j^{th} marker. The distributional assumptions are: $\boldsymbol{g}|\Sigma \sim MVN(0, \Sigma)$ and $\boldsymbol{e}|\sigma^2 \sim MVN(0, \sigma^2 I)$ which implies $\boldsymbol{y}|\boldsymbol{g}, W, \sigma^2 \sim MVN(W\boldsymbol{g}, \sigma^2 I)$. Recall that in the standard problem, the target is to estimate the covariance matrix of an observable vector-valued random variable under the assumption of multivariate normality and to this end, a sample of size $N > 1$ of independent and identically distributed random vectors is used. On the other hand, the problem being addressed in this study is to predict the allelic effects of molecular markers accounting for correlation among these random variables using phenotypic ($\boldsymbol{y}$) and genomic (W) data. This requires estimating the covariance matrix of marker allele substitution effects and the residual variance. Typically, phenotypic data correspond to a single n-dimensional vector. Because marker allele substitution effects are unknown, from the statistical point of view, the target is estimating the covariance matrix (Σ) of an unobservable m-dimensional random variable ($\boldsymbol{g}$) as well as the residual variance (σ^2), and to predict $\boldsymbol{g}$ using a single n-dimensional vector of phenotypes and the genomic information contained in W. Hence, the problem considered in this study is quite different to the standard problem and consequently, GCovGM theory cannot be applied directly to genome-wide prediction. Thus, when considering the theory of GCovGM as a means to model correlated marker effects, statistical methods adapting it to the genome-wide prediction problem have to be developed.

Finally, it is worth mentioning that Zhang et al. (2013) proposed methods to estimate covariance matrices corresponding to the sum of a low rank symmetric matrix and a sparse matrix, which is the case of the phenotypic covariance matrix $Var[y] \coloneqq V = W\Sigma W' + \sigma^2 I$, but these methods require a sample size larger than one and do not estimate Σ directly.

A hierarchical Bayes formulation

The flexibility of Hierarchical Bayesian modeling permits to cope with the problem of adapting GCovGM to genome wide prediction; it provides a simple and principled solution. Basically, the approach involves modification of the joint conditional prior of marker effects, that is, the conditional prior of $\boldsymbol{g}$. The parametric Bayesian linear regression models conventionally used in genome-wide prediction share the same sampling distribution and differ in the priors posed over marker effects. Due to this fact, this family of models is known as the "Bayesian alphabet" (Gianola et al., 2009). All these models specify the joint conditional prior distribution of marker effects as the product of the conditional priors of each marker, that is, joint priors are built under the assumption of conditional independence. These priors are typically Gaussian, finite mixtures of Gaussian distributions or finite mixtures of point mass at zero and a Gaussian distribution (the so-called spike and slab priors). Thus, they depend on unknown variance components associated with the Gaussian distributions and the finite mixture priors also involve parameters corresponding to mixing probabilities. Here, in order to take into account correlation between marker effects we formulate a hierarchical model where the joint conditional prior of $\boldsymbol{g}$ corresponds to a multivariate Gaussian distribution with a non-diagonal covariance matrix whose structure is dictated by a known undirected graph G. Thus, G reflects marginal independence assumptions made about marker effects. We consider the two families of distributions presented above the IGW and the GWKR. Consequently, our models can be thought of as new members of the Bayesian alphabet because they also assume a multivariate Gaussian sampling distribution and differ from existing models in the specification of

the prior distribution of marker effects. The residual variance is given the following conjugate prior: $\sigma^2 \sim IG\left(\frac{a}{2}, \frac{b}{2}\right)$. Regarding the covariance matrix of marker effects, under the conventional GCovGM problem, Khare and Rajaratnam (2011) provided recursive equations to find the posterior mean in closed form for homogeneous graphs. However, as explained in the previous section, the target here is estimating the covariance matrix of an unobservable random vector which leads to a different problem; therefore, even for this sort of graphs sampling from the joint posterior distribution is required. To this end, the following simple but useful property permits the use of a Gibbs sampler. Notice that under model 1 it follows that the full conditional distribution of Σ satisfies $\pi(\Sigma|Else) = \pi(\Sigma|\boldsymbol{g}, G)$. This property, and the conjugacy of the priors considered here (IGW and GWKR), imply that the full conditional of Σ pertains to the same family of the prior. Therefore, because it is possible to obtain samples from these families and all other full conditionals are standard distributions, a Gibbs sampler can be implemented (Robert and Casella, 2010). Under the model termed Bayes GCov: $\Sigma|G \sim IGW(\delta, U)$ which can be used for general graphs (i.e., non-decomposable graphs). Then, the joint posterior has the following form:

$$\pi(\boldsymbol{g}, \sigma^2, \Sigma|\boldsymbol{y}, G) \propto (\sigma^2)^{-\frac{n}{2}} \exp\left(\frac{-1}{2\sigma^2}(\boldsymbol{y} - W\boldsymbol{g})'(\boldsymbol{y} - W\boldsymbol{g})\right) |\Sigma|^{-1/2} \exp\left(\frac{-1}{2}\boldsymbol{g}'\Sigma^{-1}\boldsymbol{g}\right)$$

$$\times \exp\left(-\frac{tr(\Sigma^{-1}U)}{2} - \frac{\delta}{2} log|\Sigma|\right) (\sigma^2)^{-\left(\frac{b}{2}+1\right)} \exp\left(\frac{-a}{2\sigma^2}\right)$$

and $\Sigma|Else \sim IGW(\delta^*, U^*), U^* \coloneqq U + \boldsymbol{g}\boldsymbol{g}', \delta^* \coloneqq \delta + 1$. If G is decomposable and the conditional prior for Σ is a GWKR($\boldsymbol{\delta}, U$) distribution, then this variation of the model is referred to as Bayes GCov-KR. In this case, the full conditional distribution of Σ is GWKR($\boldsymbol{\delta}^*, U^*$), where ${\boldsymbol{\delta}^*}_{m\times 1} \coloneqq \{\delta_i^*\} = 1 + \delta_i - 2n_i$. Finally, under the conditional prior $IGW(\delta, U)$, if the graph is homogeneous, the model is denoted as Bayes GCov-H just to stress the fact that this is a homogenous GCovGM and therefore the model is reparametrized in terms of the modified Cholesky decomposition of Σ. In this

case, the Gibbs sampler is more efficient due to the fact that direct samples from the full conditional distribution of Σ can be drawn.

We close this section by emphasizing the following connection between our models and the model known as Bayes A (Meuwissen et al., 2001) which assumes independent marker effects. This model independently assigns the same scaled inverse chi-squared distribution with parameters (v, S) to the variances of marker effects. Notice that under the assumption of uncorrelated marker effects, that is, assuming $E = \emptyset$, which implies $\Sigma = diag(\sigma_1^2, \ldots, \sigma_m^2)$, the $IGW(\delta, U)$ prior reduces to

$$\pi(\Sigma|G) \propto \prod_{j=1}^{m} \exp\left(-\frac{U_{jj}}{2\sigma_j^2}\right)\left(\sigma_j^2\right)^{-\frac{\delta}{2}}.$$

Now, recall the restriction $\delta - 2n_j > v_j + 2$ which implies $\delta > 2$; therefore, $\pi(\Sigma|G)$ is the product of m inverse gamma densities with rate parameter $U_{jj}/2$ and shape parameter $\delta/2 - 1, 1 \leq j \leq m$. In particular, if $U = I_m$ (the identity matrix of dimension $m \times m$), then, this prior density corresponds to the product of m scaled inverse chi-squared densities with parameter $(\delta - 2, 1/(\delta - 2))$. Similarly, in this situation the GWKR($\boldsymbol{\delta}, U$) prior amounts to posing independent scaled inverse chi-squared priors with parameters $\left(\delta_j - 2, 1/\left(\delta_j - 2\right)\right), 1 \leq j \leq m$. Thus, when assuming uncorrelated marker effects and $U = I_m$, Bayes GCov (and therefore Bayes GCov-H) reduces to Bayes A assuming $\sigma_1^2, \ldots, \sigma_m^2 \overset{iid}{\sim}$ scaled inverse chi-squared $(\delta - 2, 1/(\delta - 2))$. Finally, Bayes GCov-KR reduces to a model that could be seen as a variation of Bayes A where the prior distributions of the variances of maker effects have different parameters, i.e., $\sigma_1^2, \ldots, \sigma_m^2 \overset{ind}{\sim}$ scaled inverse chi-squared $\left(\delta_j - 2, 1/\left(\delta_j - 2\right)\right), 1 \leq j \leq m$.

Some criteria to define $\boldsymbol{G}$

One of the first steps to carry out analyses with our models is defining the graph G, that is, defining the marginal covariance structure of allelic effects. To this end, some approaches based on genetic criteria are presented in this section. The first one is based on the idea of spatial correlation (Gianola et al. 2003, Yang and Tempelman 2012). Using a physical or linkage map, a window is defined based on a given map distance, or a given number of markers and it is slid across each chromosome. The order of markers is dictated by the physical or linkage map. This strategy induces a differentially banded or a banded covariance matrix. A second approach is based on the use of biological information. Using tools such as gene annotation information, markers can be clustered based on their function using approximations like those presented in Do et al. (2015), Peñagaricano et al. (2015), Abdollahi-Arpanahi et al. (2016), and Mcleaod et al. (2016). This creates groups or sets of loci and there are two options: permit correlations among effects of markers in different blocks or not. Finally, linkage disequilibrium between loci can be used. In this case, one of the metrics used to assess LD is chosen and those pairs of loci having a metric greater than a predefined threshold will be neighbors in G.

Simulation study

One of the main issues related to GCovGM is their computational burden in certain cases, e.g., when dealing with general graphs. In this paper, our main objective was to develop the covariance graph methodology for genome-wide prediction. Scaling the computational efficiency of the proposed approach for large scale implementation is an object of future research. Consequently, to ensure computational tractability, two small datasets were simulated in order to implement the proposed models. A single genome formed by 5 chromosomes of 10 cM length each, with 1605 biallleic markers and 1000 biallelic QTL was simulated. This genome was created via a forward-in-time approach using software QMSim (Sargolzaei and Schenkel, 2013). To create the population, 4000 historical generations of size 1000 (500 males and 500 females) were simulated in

order to reach mutation-drift equilibrium and to induce LD (Sargolzaei and Schenkel, 2013). Subsequently, 65 founders (20 males and 45 females) were randomly chosen and three generations of random mating were simulated. The total number of individuals was 200. Using this population, phenotypic records were created as the sum of the additive genetic value and an error term using two different approaches to simulated QTL effects which created two datasets that hereinafter will be referred to as dataset 1 and dataset 2. For dataset 1, QTL effects were drawn from independent zero mean Gaussian distributions and were scaled such that the additive genetic variance was equal to 50. On the other hand, for dataset 2, QTL allele substitution effects were simultaneously drawn from a multivariate Gaussian distribution with null vector mean and a banded covariance matrix with bandwidth of size 10. These effects were then scaled in order to have an additive genetic variance of 50. In both datasets, residuals were drawn from independent Gaussian distributions with null mean and variance equal to 50, consequently, heritability was 0.5. In dataset 1, models considered only marker effects, that is, QTL genotypes were not used and only markers with a minor allele frequency larger than 0.08 were considered in the analyses. In contrast, the analyses carried out using dataset 2 considered only QTL effects (i.e., SNP genotypes were not used); like in scenario 1, QTL with minor allele frequency smaller or equal than 0.08 were discarded. Ten replicates of each dataset were simulated. The graph G was based on windows defined by a fixed number of marker loci (6, dataset1) or QTL (6, dataset 2), which induces a decomposable-non-homogeneous graph; therefore, models Bayes GCov-KR and Bayes GCov were fitted. Bayes A, a Bayesian model assuming uncorrelated effects, which is frequently used in genome-wide prediction, was also fitted. Training sets were formed by individuals from generations zero and one, and validation sets were comprised of individuals from generation 2.

Predictive performance was assessed using the following criteria. Pearson correlation of phenotypes and predicted additive genetic values in the validation set (predictive ability) and the

Pearson correlation between true and predicted additive genetic values (accuracy) in training and validation sets. In each analysis, 15000 MCMC samples (first 5000 were considered burn in) were obtained using the Gibbs samplers described above. Analyses were performed using in-house R scripts (R Core Team, 2015).

Results

The average (across replicates) number of SNP and QTL considered in the analyses (i.e., having a minor allele frequency larger than 0.08) was 1487.4 and 927 respectively. Table 1 summarizes the performance of the models fitted to datasets 1 and 2. According to average predictive ability (APA), average accuracy in the training set (AAT) and average accuracy in the validation set (AAV), our models clearly outperformed Bayes A in the two simulated datasets, differences being more marked in the case of independent QTL effects (dataset 1). In these datasets, the flexibility of the GWKR priors yielded a better predictive performance. Also, the performance of our methods tended to be less variable; Bayes A showed a smaller variation only for APA in dataset 2.

Discussion

General comments about the models

In this study, the theory of GCovGM was adapted to genome-wide prediction through hierarchical Bayesian modeling. This development permits to account for correlated marker allele substitution effects in a flexible way. This flexibility is due to the ability of our models to accommodate covariance structures arising from biological considerations such as information from metabolic pathways and not only from the assumption of spatial correlation as has been done in previous studies (Gianola et al., 2003; Yang and Tempelman, 2012). Thus, covariances between effects of markers which are not physically linked are permitted. Furthermore, the possibility of

defining the graph G using tools such as gene annotation provides a way to incorporate biological information in the prediction process. The use of biological information (e.g., genome annotation) in genome-wide prediction has been used in previous studies (Do et al.,2015; Abdollahi-Arpanahi et al., 2016; Mcleaod et al., 2016), but they did not use this information to account for correlation among marker effects. These studies reported modest, moderate or null increments in predictive performance when incorporating biological information in the prediction problem.

Several approaches to define the graph based on biological principles were presented. These approaches involve the assumption of spatial correlation and the aforementioned use of existing bioinformatics tools to create "functional" sets of SNP whose effects are correlated. In general, the second strategy would induce graphs with no special properties. However, due to the theoretical and numerical advantages of decomposable graphs discussed previously, it is convenient to work with this sort of graphs whenever possible. To this end, in a submitted paper (Martínez et al., 2016), we have proven two propositions and a corollary that provide conditions on the edges set and the ordering of markers, such that the induced graph is decomposable. For the sake of completeness, these propositions and the corollary are presented in Appendix B. Proposition 1 in Appendix B is the most general, but when G is defined using biological information and subsets of different "functional" SNP sets are allowed to be correlated, its conditions are more difficult to satisfy. On the other hand, proposition 2 in Appendix B and its corollary are more restrictive in terms of the covariance structure, but they provide easier ways to order markers and define the edge set, that guarantee decomposability. Once the "functional" sets have been defined, if these conditions are not satisfied, these theoretical results provide a basis to find a decomposable super-graph containing the original graph G, an idea that has been used in graphical models (Lauritzen, 1996). Such a super-graph has been referred to as the cover of G (Khare and Rajaratnam, 2012).

In GCovGM, the family of homogeneous graphs is the one with more attractive properties. This is why the implementation of Bayes GCov-H is easier and faster because direct sampling of Σ is feasible. However, finding this kind of graphs is, in general, not an easy problem. An example of a homogeneous graph is a rooted tree where all nodes are children of a single parent (the root). Thus, under the approach of using biological information to define the graph G, a homogeneous graph can easily be found as follows: The tree structure mentioned above is imposed to each "functional" set and no correlations between effects of markers in different sets are allowed. It also holds when each "functional" set is assumed to be a complete. All the strategies mentioned above might appear restrictive, but notice that assuming independent marker effects amounts to imposing a covariance structure as well. In fact it is a special case of our approach when the edge set is the empty set.

Here, the focus was on Bayesian models because under the GCovGM framework, they can deal with the "big p small n" setting. However, in Appendix C, a frequentist approach to find the empirical BLUP of $\boldsymbol{g}$ is presented. This formulation is based on the EM algorithm (Dempster et al., 1977) combined with GCovGM theory and it permits obtaining estimators of dispersion parameters Σ and σ^2 which are used to build the mixed model equations corresponding to model 1 whose solution yields the empirical BLUP of $\boldsymbol{g}$ (Henderson,1963). This formulation involves a partition of data induced by the assumption that different groups (e.g., half-sib families) have different sets of marker effects. Such an assumption has also been considered by other authors like Gianola et al. (2003).

Even with the aid of bioinformatics, biochemistry and physiology to construct the graph G, it may not reflect the actual underlying covariance structure, but important correlations might be captured resulting in an improvement of the accuracy of genome-wide prediction. Covariance model selection involves finding the pattern of zeros and estimating the non-zero elements of either the precision or the covariance matrix (Bickel and Levina 2008; Khare et al., 2013). Model selection in GCovGM has not been as well studied as its counterpart in Gaussian concentration graph models.

There exist some frequentist methods that induce sparsity based on penalized likelihood approaches (Bien and Tibshirani, 2011) and others based on the idea of inducing sparsity in the parameter L of the modified Cholesky decomposition of Σ (Rothman et al., 2010). From the Bayesian perspective, some methods based on the Bayesian Lasso have been proposed, e.g., Wang (2012), but their main limitation is the computational burden. In order to overcome this problem, Wang (2015) proposed a method to perform covariance model selection with improved computational efficiency. On the other hand, Silva and Kalaitzes (2015) developed an approach to improve the efficiency of MCMC algorithms used to perform Bayesian inference and showed its application in covariance model selection, and Silva (2013) proposed a method based on acyclic directed mixed graphs (a generalization of directed acyclic graphs) that can be used to estimate the covariance matrix when the pattern of zeros is unknown. Some of these methods could be implemented in genome-wide prediction following approaches similar to those presented in this study.

Another set of relevant problems that create the need for extending the models proposed here are the following. Sparse estimation of the covariance matrix via graphical models when priors for marker effects correspond to finite mixtures like the so-called spike and slab priors, cases where the assumption of Gaussian distribution of $\boldsymbol{y}$ is not suitable (e.g., binary variables, count data), and the implementation of Gaussian copula graphical models (Dobra and Lenkoski, 2011) which could be used to formulate a hierarchical model that permits getting rid of making assumptions about the parametric representation of the prior distribution of marker allele substitution effects. However, it is important to notice that until now, these models have been developed to estimate the precision matrix; to our knowledge, estimation of the covariance matrix using this kind of models has not been investigated.

Extension to multiallelic loci

Here, biallelic loci were considered, but in some cases multiallelic loci have to be dealt with. In the future, models could be fit using genotypes for actual genes instead of molecular markers. In such a case, there could be more than two alleles per locus. A similar situation occurs when fitting effects of haplotypes built from two or more consecutive markers (Meuwiseen et al., 2001; Calus et al., 2008). The methods developed here can be easily extended to the multiallelic case. If there are a_k alleles at locus k, then the corresponding columns of the design matrix are formed by defining $a_k - 1$ variables as follows:

$$W^k = \{w_{ij}^k\} = \begin{cases} 1, if\ genotype = A_j A_j \\ 0, if\ genotype = A_j - \ , j = 1,2, \dots, a_k - 1 \\ -1, if\ genotype = - - \end{cases},$$

where w_{ij}^k is the genotype of the i^{th} individual for the j^{th} allele of locus k and " $-$ " represents an allele different from A_j. The graph G can be built based on the ideas discussed above, with extra considerations at the intra-locus level. For example, it could be assumed that effects of alleles of the same locus are all correlated.

Data analyses

In general, Bayes GCov and Bayes GCov-KR outperformed Bayes A. Differences between our models and Bayes A were more marked when QTL effects were independent and models considered SNP effects (dataset 1). In this scenario, independent QTL effects were simulated, but models were fitted in terms of SNP effects; consequently, allelic effects of markers in high LD with the same QTL or set of QTL's could be correlated. This correlation may be the reason behind the superior performance of our models when compared to Bayes A. On the other hand, in the ideal scenario where the model considers the causal variants (the QTL) instead of markers (i.e., models fitted to dataset 2) the benefit of accounting for marginal correlation was smaller as suggested by a smaller difference in the three criteria used to assess predictive performance. This behavior may suggest that when considering the causal variants instead of proxies like the SNP, models assuming

independent effects yield an acceptable predictive performance even when the true covariance matrix is non-diagonal. Hopefully, this ideal scenario where the causal variants determining a phenotype, or at least most of them, are known will be reached in the near future. The largest difference between the method Bayes A and our methods (15.7%) was observed for AAV in dataset 1, while the smallest one (0.5%) was observed for AAT in dataset 2, in both cases, when comparing it with Bayes GCov-KR. Although Bayes GCov-KR had higher APA, AVA and ATA values in these simulated datasets, notice that the differences compared to Bayes G-Cov were small, being slightly larger in dataset 2; therefore, in these simulations the gain in fitting a more complex model which considers as many shape parameters as markers did not yield a remarkable gain in accuracy or predictive ability. The gains in accuracy in the validation set observed in dataset 1 are larger than those found by Yang and Tempelman (2012) when comparing their antedependence models with their independent marker effects counterparts Bayes A and Bayes B, whereas gains in accuracy observed in dataset 2 were comparable (they found increments in accuracy of breeding values in the testing population up to 3%). The simulated data in Yang and Tempelman (2012) were similar to dataset 1, where it is expected that correlation among SNP effects arises from physical proximity to the same causal variants. They also considered a heritability value of 0.5. However, Yang and Tempelman (2012) considered a much smaller number of QTL (30). In addition, they considered models fitting different subsets of SNP and they found cases where Bayes B (a model assuming independent effects) outperformed their model ante-Bayes A (which accounts for correlated marker effects). They attributed these results to the small number of simulated QTL because in cases where the number of QTL controlling the phenotype is relatively small, models posing spike and slab priors over marker allele substitution effects, like Bayes B, tend to perform better. In a mice population, Yang and Tempelman (2012) also found that Bayes B outperformed ante-Bayes A in terms of predictive ability. Finally, the study of Gianola et al. (2003) did not consider data analysis.

Final remarks

This paper introduces the theory of GCovGM in the context of genome-wide prediction which permits to account for correlated marker effects in a very flexible way in terms of the marginal covariance structure. Models developed here also allow incorporating biological information in the prediction process through its use when building graph G.

Acknowledgements

C.A. Martínez thanks Fulbright Colombia and "Departamento Adiministrativo de Ciencia, Tecnología e Innovación" COLCIENCIAS for supporting his PhD and Master programs at the University of Florida through a scholarship, and Bibiana Coy for her love, support and constant encouragement.

References

Asmussen, S., & Glynn, P.W. (2011). A new proof of convergence of MCMC via the ergodic theorem. *Statistics and Probability Letters, 81*, 1482-1485.

Bickel, P.J., & Levina, E. (2008). Covariance regularization by thresholding. *The Annals of Statistics, 36(6)*, 2577-2604.

Bien, J. & Tibshirani, R.J.. (2011). Sparse estimation of a covariance matrix. *Biometrika,98(4)*, 807-820.

Calus M.P.L., Meuwissen, T.H.E., de Roos, A.P.W., & Veerkamp, R.F. (2008). Accuracy of genomic selection using different methods to define haplotypes. *Genetics, 178*, 553-561.

Carvalho, C.M., Massam, H., & West, M. (2007). Simulation of hyper-inverse- Wishart distributions in graphical models. *Biometrika, 94*, 647-659.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B, 39(1)*, 1-38.

Diaconis, P., & Ylvisaker, D. (1979). Conjugate priors for exponential families. *The Annals of Statistics, 7(2)*, 269-281.

Gianola, D., Perez-Enciso, M., & Toro, M.A. (2003). On marker-assisted prediction of genetic value: Beyond the Ridge. *Genetics, 163*, 347-365.

Gianola, D., de los Campos, G., Hill, W.G., Manfredi, E., & Fernando, R.L. (2009). Additive genetic variability and the bayesian alphabet. *Genetics, 183*, 347-363.

Henderson, C.R. (1963). Selection index and expected genetic advance. In W.D. Hanson & H.F. Robinson (Eds.), *National Academy of Sciences-National Research Council,* publication 982.

Illumina, Inc., 2011. GoldenGate Bovine3K Genotyping BeadChip. Illumina Data Sheet. San Diego, CA. www.illumina.com/Documents//products/datasheets/datasheet_bovine3k.pdf.

Khare, K., & Rajaratnam, B. (2011). Wishart distributions for decomposable covariance graph models. *The Annals of Statistics, 39(1)*, 514-555.

Khare, K., & Rajaratnam, B. (2012). Sparse matrix decompositions and graph characterizations. *Linear Algebra and its Applications, 437*, 932-947.

Khare, K., Oh, S., & Rajaratnam, B. (2015). A convex pseudo-likelihood framework for high dimensional partial correlation estimation with convergence guarantees. *Journal of the Royal Statistical Society Series B, 77(4)*, 803-825.

Lauritzen, S.L. (1996). Graphical models. The Clarendon Press, Oxford University Press, New York, NY.

Lauritzen, S.L., & Sheehan, N.A. (2003). Graphical models for genetic analyses. *Statistical Science, 18(4)*, 489-514.

Letac, G., & Massan, H. (2007). Wishart distributions for decomposable graphs. *The Annals of Statistics, 35(3)*, 1278-1323.

Martínez, C.A., Khare, K., Rahman, S., & Elzo, M.A. (2016). Modelling correlated marker effects in genome-wide prediction via Gaussian concentration graph models. *arXiv preprint*, arXiv:1611.03361.

Meuwissen, T.H.E., Hayes B.J., & Goddard, M.E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*, 1819-1829.

Morota, G., Valente, B.D., Rosa, G.J.M., Weigel, K.A., & Gianola, D. (2012). An assessment of linkage disequilibrium in Holstein cattle using a Bayesian network. *Journal of Animal Breeding and Genetics, 129*, 474-487.

Peñagaricano, F., Weigel, K.A., Rosa, G.J.M., & Kathib, H. (2013). Inferring quantitative trait pathways associated with bull fertility from a genome-wide association study. *Frontiers in Genetics, 3*, 307.

R Core Team (2015). R: A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. *URL https://www.R-project.org/.*

Rajaratnam, B., Massam, H., & Carvalho, C. (2008). Flexible covariance estimation in graphical Gaussian models. *The Annals of Statistics, 36(6)*, 2818-2849.

Rosa, G.J.M., Felipe, V.P.S., & Peñagaricano, F. (2016). Applications of graphical models in quantitative genetics and genomics. In: Kadarmideen, H.N. (eds.), Systems biology in animal production and health Vol. 1, pp 95-116. Springer, NY.

Rothman, A., Levina, E., & Zhu, J. (2010). A new approach to Cholesky-based covariance regularization in high dimensions. *Biometrika, 97(3)*, 539-550.

Sargolzaei, M., & Schenkel, F.S. (2013). *QMSim User's Guide Version 1.10*. Centre for Genetic Improvement of Livestock, Department of Animal and Poultry Science, University of Guelph, Guelph, Canada.

Scutari, M., Mackay, I., & Balding, D. (2013). Improving the efficiency of genomic selection. *Statistical Applications in Genetics and Molecular Biology, 12(4)*, 517-527.

Silva, R., & Ghahramani, Z. (2009). The Hidden Life of Latent Variables: Bayesian Learning with Mixed Graph Models. *Journal of Machine Learning Research, 10,* 1187-1238.

Silva, R. (2013). A MCMC approach for learning the structure of Gaussian acyclic directed mixed graphs (2013). In: Giudici, P., Ingrassia, S., Vichi, M. (eds.), Statistical Models for Data Analysis, pp 343-352. Springer, NY.

Silva, R., & Kalaitzis, A. (2015). Bayesian inference via projections. *Statistics and Computing, 25*, 739-753.

Stein, C. (1975). Estimation of a covariance matrix. In *Reitz lecture*. 39th annual meeting. IMS Atlanta, GA

Wang, H. (2012). Bayesian Graphical Lasso models and efficient posterior computation. *Bayesian Analysis, 7(4),* 867-866.

Wang, H. (2015). Scaling it up: Stochastic search structure learning in graphical models. *Bayesian Analysis, 10,* 351-377.

Yang, W., & Tempelman, R.J. (2012). A Bayesian Antedependence Model for Whole Genome Prediction. *Genetics, 190,* 1491-1501.

Zhang, L., Sarkar, A., & Mallick, B.K. (2013). Bayesian low rank and sparse covariance matrix decomposition. *arXiv preprint,* arXiv:1310.4195.

Table 1 Average (over 10 replicates) predictive abilities (APA), accuracies in training (AAT) and validation (AAV) sets for simulated datasets 1 and 2 (standard deviations in brackets).

Model	Dataset 1			Dataset 2		
	APA	**AAT**	**AAV**	**APA**	**AAT**	**AAV**
Bayes GCov	0.432 (0.075)	0.739 (0.056)	0.573 (0.078)	0.432 (0.128)	0.716 (0.038)	0.557 (0.099)
Bayes GCov-KR	0.441 (0.071)	0.740 (0.058)	0.573 (0.075)	0.444 (0.094)	0.743 (0.056)	0.566 (0.089)
Bayes A	0.352 (0.161)	0.684 (0.064)	0.417 (0.081)	0.404 (0.048)	0.711 (0.051)	0.526 (0.123)

Appendix A: Basic Concepts in Graph Theory

Undirected graph. An undirected graph G is defined as a collection of two objects $G = (V, E)$ where V is the set of vertices (finite) and $E \subseteq V \times V$ is the set of edges satisfying:

$$(u, v) \in E \Leftrightarrow (v, u) \in E.$$

Neighbor vertices. Let $G = (V, E)$ be an undirected graph. The vertices $u, v \in V$ are said to be neighbors if $(u, v) \in E$.

P-path. A p-path is a collection of p distinct vertices $u_1, u_2, \dots, u_p$ such that $(u_i, u_{i+1}) \in E, i = 1,2, \dots, p-1$, that is, (u_i, u_{i+1}) are neighbors for $i = 1,2, \dots, p-1$.

P-cycle. A p-cycle is a collection of p distinct vertices $u_1, u_2, \dots, u_p$ such that $(u_i, u_{i+1}) \in E, i = 1,2, \dots, p-1$ and $(u_p, u_1) \in E$

Clique. A subset $V_0 \subset V$ is a clique if $(u, v) \in E \; \forall \, u, v \in V_0$.

Maximal clique. A subset $V_0 \subset V$ is defined to be a maximal clique if V_0 is a clique and there does not exist a clique $\bar{V}$ such that $V_0 \subset \bar{V} \subseteq V$.

Ordered graphs. Let $G = (V, E)$ and let σ be an ordering of V, that is, a bijection from V to $\{1,2, \dots, |V|\}$. Then, the ordered graph $G_\sigma = (\{1,2, \dots, |V|\}, E_\sigma)$ is defined as follows: $(i, j) \in E_\sigma \; iff \; (\sigma^{-1}(i), \sigma^{-1}, (j)) \in E$.

Perfect elimination ordering. An ordering σ of a graph $G = (V, E)$ is defined to be a perfect elimination ordering if a triplet $\{i, j, k\}$ with $i > j > k$ such that $(i, j) \notin E_\sigma$ and $(i, k), (j, k) \in E_\sigma$ does not exist.

Subgraph. The graph $G' = (V', E')$ is said to be a subgraph of graph $G = (V, E)$ if $V' \subseteq V$ and $E' \subseteq E$.

Induced subgraph. Consider the graph $G = (V, E)$ and a subset $A \subseteq V$. Define $E_A = (A \times A) \cap E$. The subgraph $G_A = (A, E_A)$ is defined to be a subgraph of G induced by A. **Decomposable graph**. An undirected graph $G = (V, E)$ is a decomposable graph if it does not contain a cycle of length greater than or equal to four as an induced subgraph. It turns out that decomposable graphs are

characterized by the existence of a perfect elimination ordering of their vertices; therefore, a graph $G = (V, E)$ is decomposable iff its vertices admit a perfect elimination ordering.

Connected graph. A graph G is said to be connected if any pair of distinct vertices in G are connected, that is, there exists a path between them.

Directed edges. An edge is said to be directed if $(u, v) \notin E$ whenever $(v, u) \in E$. If (v, u) is a directed edge then v is said to be a *parent* of u and u is said to be a *child* of v.

Directed graph. A graph $\mathcal{D} = (V, E)$ such that its edges are directed is defined as a directed graph.

Directed acyclic graph. A directed acyclic graph (DAG) is a directed graph with no cycles.

Tree. A tree is a connected graph with no cycle of length greater or equal than 3.

Rooted tree. A rooted tree is a tree in which a particular node is distinguished from the others and designated the root of the tree. This node is the ancestor of all other nodes in the tree. An ancestor of a node u in a rooted tree with root node r is any node in the path from r to u.

Homogeneous graph. An undirected graph $G = (V, E)$ is defined to be homogeneous if for all $(u, v) \in E$, either:

$$\{i: i = u \; or \; (i, u) \in E\} \subseteq \{i: i = v \; or \; (i, v) \in E\}$$

or

$$\{i: i = v \; or \; (i, v) \in E\} \subseteq \{i: i = u \; or \; (i, u) \in E\}.$$

An equivalent definition is the following. A graph $G = (V, E)$ is said to be homogeneous if it is decomposable and it does not have a 4-path as an induced subgraph. Homogeneous graphs have an equivalent representation in terms of directed rooted trees called Hasse diagrams.

Hasse diagram. A Hasse diagram is built as follows. For $i \in V$, let $\mathcal{N}(u) \coloneqq \{i: i = u \; or \; (i, u) \in E\}$. Whenever $\mathcal{N}(u) \subseteq \mathcal{N}(v)$ we write $v \to u$. If $u \to v$ and $v \to u$ it is said that there is a equivalence relation between u and v. Using this relation, equivalence classes are created. For example, if $\mathcal{N}(u) = \mathcal{N}(v)$, then u and v are in the same equivalence class. The equivalence classes are the nodes of the Hasse diagram, formally, if $\bar{u}$ denotes the equivalence class containing node u, then the Hasse diagram of G is a directed graph with node set $V_H \coloneqq \{\bar{u}: u \in E\}$. The edge set E_H is defined as follows. If $\bar{u} \neq \bar{v}$, $u \to v$, and $\nexists \, k$ such that $u \to k \to v$ then put a directed edge from u to v.

Hasse perfect vertex elimination scheme or Hasse ordering. Once the Hasse diagram of G has being built, the nodes of G are ordered in the following way. The ordering is descending starting from the root of the tree; therefore, nodes pertaining to equivalence classes on the top of the Hasse diagram are assigned the largest levels. Within every equivalence class with more than one node, the ordering is arbitrary. Hence, the ordering is not unique. Any ordering that gives an ancestor a higher level than any of its descendants in the Hasse diagram of G is defined to be a Hasse perfect vertex elimination scheme or simply a Hasse ordering of the nodes of G.

Appendix B: Maximum likelihood estimation in covariance graph models

Maximum likelihood estimation of Σ for general graphs, standard problem

If the sample size N is larger than p, then maximum likelihood estimation of Σ is feasible. After removing constant terms from the negative of the log-likelihood the following is the objective function to be minimized: $l^*(\Sigma) = tr(\Sigma^{-1}S) + log|\Sigma|, \Sigma \in \mathbb{P}_G$, where S is the sample covariance

matrix. Notice that the objective involves Σ^{-1} instead of Σ. This objective function is not convex, which makes this minimization more difficult than the minimization problem for concentration graph models. One important feature of covariance graph models is that they correspond to curved exponential families instead of the well-studied exponential families as is the case of concentration graph models (Khare and Rajaratnam 2011), it poses a more challenging problem.

An iterative conditional fitting (ICF) algorithm to minimize $l^*(\Sigma)$ was developed by Chaudhuri et al. (2007); however, because we are dealing with a non-convex optimization problem, convergence to a global or even a local minimum is not guaranteed.

The algorithm is based on the following partition of Σ:

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma'_{.1} \\ \Sigma_{.1} & \Sigma_{-1,-1} \end{bmatrix} \qquad (B.1)$$

where Σ_{11} is the 1,1 entry of Σ, $\boldsymbol{\Sigma_{.1}}$ is the first column of Σ without the first entry and $\Sigma_{-1,-1}$ is the submatrix of Σ resulting from deleting its first row and column. Using the standard rules for inversion by partitioning:

$$\Sigma^{-1} = \begin{bmatrix} \dfrac{1}{\gamma_1} & \dfrac{-\boldsymbol{\Sigma}'_{.1}\Sigma^{-1}_{-1,-1}}{\gamma_1} \\ \dfrac{-\Sigma^{-1}_{-1,-1}\boldsymbol{\Sigma}'_{.1}}{\gamma_1} & \dfrac{\Sigma^{-1}_{-1,-1} + \Sigma^{-1}_{-1,-1}\boldsymbol{\Sigma}_{.1}\boldsymbol{\Sigma}'_{.1}\Sigma^{-1}_{-1,-1}}{\gamma_1} \end{bmatrix}$$

where $\gamma_1 = \Sigma_{11} - \boldsymbol{\Sigma}'_{.1}\Sigma^{-1}_{-1,-1}\boldsymbol{\Sigma}_{.1}$. Notice that knowing Σ, we can get $\left(\boldsymbol{\Sigma}_{.1}, \Sigma_{-1,-1}, \gamma_1\right)$ and vice versa; consequently, we have a one to one transformation. By using permutations, the same partition can be performed for every one of the p random variables represented in graph G. The high level of the algorithm is the following:

1) Partition Σ as $\left(\boldsymbol{\Sigma}_{.1}, \Sigma_{-1,-1}, \gamma_1\right)$, 2) minimize $l^*(\Sigma)$ with respect to $\boldsymbol{\Sigma}_{.1}$ treating as fixed the current values of $\Sigma_{-1,-1}$ and γ_1 and 3) minimize $l^*(\Sigma)$ with respect to γ_1 fixing the current values of $\Sigma_{.1}$ and $\Sigma_{-1,-1}$. The same is repeated for the p variables and it corresponds to one iteration of the algorithm. The minimization problem is solved by minimizing the following quadratic form with respect to $\boldsymbol{\beta}_1$ (Chaudhuri et al.,2007):

$$\frac{-1}{\gamma_1}\left(2\boldsymbol{\beta}'_1 Q'_1\Sigma^{-1}_{-1,-1}\boldsymbol{S}_{.1} - \boldsymbol{\beta}'_1 Q'_1\Sigma^{-1}_{-1,-1}S_{-1,-1}\Sigma^{-1}_{-1,-1}Q_1\boldsymbol{\beta}_1\right)$$

where $\boldsymbol{\beta}_1 := \left(\Sigma_{1j}\right)_{(1,j)\in E}$, $\boldsymbol{S}_{.1}$ and $S_{-1,-1}$ are elements obtained after partitioning S as Σ was partitioned in $(B.1)$ and Q_1 is a matrix of zeros such that: $\boldsymbol{\Sigma_{.1}} = Q_1\boldsymbol{\beta_1}$. This is a standard problem and its solution is $\widehat{\boldsymbol{\beta}}_1 = \left(Q'_1\Sigma^{-1}_{-1,-1}S_{-1,-1}\Sigma^{-1}_{-1,-1}Q_1\right)^{-1}Q'_1\Sigma^{-1}_{-1,-1}\boldsymbol{S}_{.1}$. On the other hand, the solution to the second minimization is: $\hat{\gamma}_1 = S_{11} - 2\boldsymbol{\Sigma}'_{.1}\Sigma^{-1}_{-1,-1}\boldsymbol{S}_{.1} + \Sigma'_{.1}\Sigma^{-1}_{-1,-1}S_{-1,-1}\Sigma^{-1}_{-1,-1}\boldsymbol{\Sigma}_{.1}$. Kauermann (1996) proposed to modify the objective function in order to make it a function of Σ, which makes the problem convex. The new objective function has the form: $\tilde{l}(\Sigma) = tr(\Sigma S^{-1}) + log|\Sigma|, \Sigma \in \mathbb{P}_G$. A trick based on using the maximal cliques of G is applied to solve this problem and the solution is known as the Kauermann's dual estimator. Under certain conditions, convergence of the ICF algorithm at least to a local stationary point can be proved (Drton et al., 2006).

Maximum likelihood estimation of Σ for homogenous graphs, standard problem

Recall that for homogeneous graphs, $\Sigma = LDL' \in \mathbb{P}_G \Leftrightarrow L \in \mathcal{L}_G \Leftrightarrow L^{-1} \in \mathcal{L}_G$; therefore, the objective function is written in terms of (L, D) instead of Σ. Also, recall that after removing constant terms from the negative log-likelihood we get: $l^*(\Sigma) = tr(\Sigma^{-1}S) + log|\Sigma|, \Sigma \in \mathbb{P}_G$. The bijection from $\mathbb{P}_G$ to $\mathcal{L}_G \times \mathcal{D}$ induces:

$$l^*(L, D) = tr((L')^{-1}D^{-1}L^{-1}S) + log|D|, L \in \mathcal{L}_G, D \in \mathcal{D}$$

reparameterization in terms of $T = L^{-1}$ yields:

$$l^*(L, D) = tr(T'D^{-1}TS) + log|D|, T \in \mathcal{L}_G, D \in \mathcal{D}$$

$$= \sum_{i=1}^{p} \frac{1}{D_{ii}} (\boldsymbol{T}_{i.} S \boldsymbol{T}'_{i.}) + log\, D_{ii} \qquad (B.2)$$

where $\boldsymbol{T}_{i.}$ Is the i^{th} row of T.

To obtain the MLE of Σ, every summand in $(B.2)$ is minimized with respect to D_{ii} and $\boldsymbol{T}_{i.}$. Define $\boldsymbol{x}_i := \{T_{ij}\}_{j<i,(i,j)\in E}$; $N^<(i) := \{j : j < i, (i,j) \in E\}$ and construct the following matrix from the sample covariance matrix:

$$S_i = \begin{pmatrix} S^{<i} & S^{<}_{.i} \\ (S^{<}_{.i})' & S_{ii} \end{pmatrix} \qquad (B.3)$$

where $\boldsymbol{S}^{<}_{.i} = (S_{ki})_{k<i,(i,k)\in E}, S^{<i} = (S_{kl})_{k,l\in N^<(i)}$. Then, the MLE are:

$$\widehat{\boldsymbol{x}}_i = (S^{<i})^{-1}\boldsymbol{S}^{<}_{.i}$$

$$\widehat{D}_{ii} = S_{ii} - (\boldsymbol{S}^{<}_{.i})'(S^{<i})^{-1}\boldsymbol{S}^{<}_{.i}$$

Combining all $\widehat{D}_{ii}$ and $\widehat{\boldsymbol{x}}_i$ we can build $\widehat{D}$ and $\widehat{T}$ and using them we have $\widehat{\Sigma} = \widehat{L}\widehat{D}\widehat{L}'$.

Maximum likelihood estimation of Σ in genome-wide prediction

Unlike the Bayesian approach, envisaging a frequentist solution to the problem of adapting GCGM to genome-wide prediction under the model presented in the manuscript is not straightforward and we could not find a direct and principled method to cope with this problem. Therefore, some *ad hoc* extra assumptions were done in order to provide a frequentist formulation. The method proposed here involves two steps. The first one combines the EM algorithm (Dempster et al., 1977) with GCovGM to estimate covariance components. The second one involves plugging these estimates into mixed model equations corresponding to model 1 in order to obtain the empirical BLUP of $\boldsymbol{g}$ (Henderson, 1963).

According to the rationale of the EM-algorithm, we define $\boldsymbol{g}$ as the augmented or missing data, then we find the maximizers of the complete likelihood as if $\boldsymbol{g}$ were observable and finally we compute their expected values with respect to the distribution of the missing or augmented data given the observed data. As mentioned in the manuscript, maximum likelihood estimation of Σ is only possible if $N > m$. In model 1, we have a single n-dimensional vector $\boldsymbol{y}$ and the target is to estimate the residual variance and the covariance matrix of the m-dimensional vector $\boldsymbol{g}$; therefore, in terms of the standard problem: N=1.Thus, an *ad hoc* solution is to assume that data can be split into $f > m$ groups such that each group has a different vector of marker effects, that is, $\boldsymbol{y}_i = W_i\boldsymbol{g}_i + \boldsymbol{e}_i, \forall\, i = 1,2,\ldots,f$. Currently, as more and more animals are genotyped, for SNP panels of moderate density (e.g., 50K) the case $n > m$ can be found. For example, this is the case of the Holstein population in the US. However, this is not the most common situation and it is important to notice that it does not

imply that $f > m$ which is the necessary condition to carry out maximum likelihood estimation of Σ. One of the simplest ways to split a population into f groups is by considering families (e.g., half-sibs or full-sibs) as in Gianola et al. (2003). Currently, the requirement $f > m$ will be met by very few populations when considering a relatively small number of markers and this is the reason for not considering the frequentist approach in the manuscript. Notwithstanding, in this appendix we provide an approach to carry out maximum likelihood estimation of the dispersion parameter $\boldsymbol{\theta} \coloneqq (\Sigma, \sigma^2)$ in a genome-wide prediction model based on multiple linear regression which later permits to obtain the empirical BLUP of $\boldsymbol{g}$.

It is also assumed that: $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_f$ are iid $MVN(0, \Sigma)$, $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_f$ are independent $MVN\left(0, \sigma^2 I_{n_i}\right)$ random variables and $Cov(g_i, e_{i'}) = 0, \forall\ 1 \leq i, i' \leq f$, where n_i is the number of observations in group i; therefore, $\sum_{i=1}^{f} n_i = n$. Under these assumptions, the complete log-likelihood can be written as:

$$l(\sigma^2, \Sigma) = constants - \frac{n}{2} log\sigma^2 + \frac{f}{2}\left(-log|\Sigma| - tr\left(\Sigma^{-1} S_g\right)\right) - \frac{\|\boldsymbol{y} - W^* \boldsymbol{g}^*\|_2^2}{2\sigma^2} \quad (B.4)$$

$$S_g = \frac{1}{f}\sum_{i=1}^{f} \boldsymbol{g}_i \boldsymbol{g}_i{}', \boldsymbol{g}^* \coloneqq \left(\boldsymbol{g}_1{}' \cdots \boldsymbol{g}_f{}'\right)', W^* = Block\ Diag.\{W_i\}_{i=1}^{n}.$$

The expected values of sufficient statistics for the covariance parameters taken with respect to the conditional distribution of the missing data given the observed data have to be found. The sufficient statistic for $\boldsymbol{\theta}$ is $\left(S_g, \boldsymbol{e}^{*\prime}\boldsymbol{e}^*\right), \boldsymbol{e}^* = \boldsymbol{y} - W^*\boldsymbol{g}^*$. Also, given $\boldsymbol{y}$, $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_f$ are independent with the following distributions: $\boldsymbol{g}_i|\boldsymbol{y}_i \sim MVN\left(K_i^{-1}\frac{W_i'y_i}{\sigma^2}, K_i^{-1}\right)$, where $K_i \coloneqq \frac{W_i'W_i}{\sigma^2} + \Sigma^{-1}$. Similarly, it follows that $\boldsymbol{e}^*|\boldsymbol{y} \sim MVN\left(\sigma^2 V^{-1}\boldsymbol{y}, \sigma^2(I - \sigma^2 V^{-1})\right)$, where $V = W^{*\prime} I_f \otimes \Sigma W^* + R$. Hence,

$$E\left[S_g|\boldsymbol{y}\right] = \frac{1}{f}\sum_{i=1}^{f} K_i^{-1}\left[I_m + \frac{1}{(\sigma^2)^2} W_i' y_i y_i' W_i K_i^{-1}\right] \quad (B.5)$$

$$E[\boldsymbol{e}^{*\prime}\boldsymbol{e}^*|\boldsymbol{y}] = \sigma^2(n - \sigma^2 tr(V^{-1}) + \sigma^2 \boldsymbol{y}' V^{-1} V^{-1}\boldsymbol{y}) \quad (B.6)$$

Applying the Woodbury's identity, $E\left[S_g|\boldsymbol{y}\right]$ can be alternatively expressed as:

$$E\left[S_g|\boldsymbol{y}\right] = \frac{1}{f}\Sigma\left\{fI_m - \left[\sum_{i=1}^{f} W_i' V_i^{-1}\left(I_{n_i} - \boldsymbol{y}_i \boldsymbol{y}_i' V_i^{-1}\right) W_i\right]\Sigma\right\} \quad (B.7)$$

where $V_i \coloneqq W_i \Sigma W_i' + \sigma^2 I_{n_i}$. It does not require inversion of Σ, it requires inverting f $n_i \times n_i$ matrices. The expectation step of this EM algorithm consists of using either B.5 or B.7 to compute $E\left[S_g|\boldsymbol{y}\right]$ and B.6 to compute $E[\boldsymbol{e}^{*\prime}\boldsymbol{e}^*|\boldsymbol{y}]$, the maximization step is the one involving GCovGM. At iteration t, the maximization step involves the following computations:

$$(\hat{\sigma}^2)^{(t+1)} = \frac{\widehat{\boldsymbol{q}}^{(t)}}{n}, \widehat{\boldsymbol{q}}^{(t)} \coloneqq E[\boldsymbol{e}^{*\prime}\boldsymbol{e}^*|\boldsymbol{y}]\Big|_{\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}}$$

$$\hat{\Sigma}^{(t+1)} = h\left(\hat{S}_g{}^{(t)}\right), \hat{S}_g{}^{(t)} \coloneqq E\left[S_g|\boldsymbol{y}\right]\Big|_{\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}}$$

where $\hat{\Sigma}^{(t+1)}$ is computed using methods explained before. For homogeneous graphs, function $h(\cdot)$ has closed forms after reparametrizing the objective function in terms of (T, D) as shown previously in this section. Once the algorithm converges and the maximum likelihood estimates of Σ and σ^2 are

obtained, these are plugged in the mixed model equations corresponding to model 1 to obtain the empirical BLUP of $\boldsymbol{g}$ (Henderson, 1963):

$$\widehat{\boldsymbol{g}} = \left(W'W + \hat{\sigma}^2\hat{\Sigma}^{-1}\right)^{-1}W'\boldsymbol{y}.$$

References (only those not included in the manuscript are presented here)

Chaudhuri, S., Drton, M., & Richardson, T.M. (2007). Estimation of a Covariance Matrix with Zeros. *Biometrika, 94(1),* 199-216.

Drton, M.,Eichler, M., & Richardson, T.S. (2006). Computing Maximum Likelihood Estimates in Recursive Linear Models with correlated Errors. *ArXiv preprint, arXiv,* 0601631.

Kauermann, G. (1996). On a dualization of graphical Gaussian models. *Scandinavian Journal of Statistics, 23(1),* 105-116.

Appendix C: Conditions to find decomposable graphs

The following proposition establishes which approaches will induce decomposable graphs. Hereinafter, the "functional blocks" mentioned in approach the approach considering the use of gene annotation will be referred to as blocks. In this approach, when effects of markers in different blocks are not allowed to be correlated, the corresponding strategy will be referred to as approach F1. On the other hand, when the effects of subsets or markers in different blocks are assumed to be correlated, the corresponding strategy will be referred to as approach F2.

If a block contains a subset of markers with effects correlated with the effects of a subset of markers in another block, these blocks are said to be linked. Let B be the total number of blocks and $\mathcal{L}$ be the set of pairs of linked blocks. Let Ψ be the set of blocks linked with at least two other blocks, $\forall\, l \in \Psi$ let Γ_l be the set of blocks linked to block l and $\forall a \in \Gamma_l$, let C_{l_a}be the subset of markers in block l whose effects are correlated with effects of a subset of markers in block $a, 1 \leq a \leq B, a \neq l$.

Proposition 1

The graphs induced under approaches considering correlation of groups of nearby markers and approach F1, are decomposable. In addition the graph induced under the approach F2 is decomposable if there exists an ordering of markers σ' that along with the edge set satisfy the following conditions.

Condition 1.1 For all possible triplets of linked blocks $\{l, l', l''\}$ such that $C_{l_{l'}} \neq C_{l_{l''}}$, $C_{l'_l} \neq C_{l'_{l''}}$, $C_{l''_l} \neq C_{l''_{l'}}$, and the sets $I_l \coloneqq C_{l_{l'}} \cap C_{l_{l''}}$, $I_{l'} \coloneqq C_{l'_l} \cap C_{l'_{l''}}$ and $I_{l''} \coloneqq C_{l''_l} \cap C_{l''_{l'}}$, are all non-empty, the following never happens: $\sigma'(i) > \sigma'(j) > \sigma'(k)$, $i \in C_{l_{l''}} \cap I_l^c, j \in C_{l'_l}$ or $i \in C_{l_{l''}}, j \in C_{l'_l} \cap I_{l'}^c$, and $k \in I_{l''}$; if there are triplets of linked blocks $\{l, l', l''\}$ such that exactly one of the three sets $\{I_l, I_{l'}, I_{l''}\}$, say I_l is empty, then: $\min\{\sigma'(k), \sigma'(i), \sigma'(j)\} = \sigma'(k), \forall\, k \in C_{l_{l'}} \cup C_{l_{l''}}\ \forall\, j \in I_{l'}\ \forall\, i \in I_{l''}$ and if exactly two of these sets, say $\{I_l, I_{l'}\}$ are empty, then for

either l or l', say l, $\sigma'(k) < \sigma'(i)\ \forall k \in C_{l_{l'}} \cup C_{l_{l''}}, \forall i \in I_{l''}$. Superindex C indicates the complement with respect to the index set of the corresponding block.

Condition 1.2 For every possible triplet of blocks $\{l, l', l''\}$ the following does not happen: $\sigma'(k) < \sigma'(j) < \sigma'(i), k \in I_l, j \in C_{l'_l}, i \in C_{l''_l}, C_{l'_{l''}} = \emptyset$.

Condition 1.3 For every duplet of linked blocks $\{l, l'\}$ the following does not hold: $\exists\ i \in l, \{j, k\} \in l'$ such that $\sigma'(i) > \sigma'(j) > \sigma'(k), i \in C_{l_{l'}}, j \in C^C_{l'_l}, k \in C_{l'_l}$.

Condition 1.4 For each pair of linked blocks (l, l'), $C_{l_{l'}} \times C_{l'_l} \in E_\sigma$, that is, the effect of each marker in $C_{l_{l'}}$ is correlated with the effects of all marker in $C_{l'_l}$.

Moreover, conditions 1.1, 1.2 and 1.3 are necessary whereas condition 1.4 is not.

This proposition involves all possible orderings of markers. However, if markers are ordered in such a way that markers in the same block are given consecutive indices, the number of possible orderings is reduced. Thus, in order to provide a simpler way to order markers, the following proposition only requires the existence of an ordering of the blocks and a structure on the edges set satisfying certain conditions that permit to find a perfect elimination ordering of markers.

Proposition 2

If there exists an ordering ρ of the blocks which coupled with the structure of the edges set satisfy condition 1.4 plus the following conditions:

Condition 2.1 $C_{l_a} = \cdots = C_{l_m} \coloneqq C_l\ \forall\ l \in \Psi$

Condition 2.2 For every possible triplet of blocks $\{l, l', l''\}$ the following does not happen: $(l, l'), (l, l'') \in L, (l', l'') \notin L, \rho(l) < \rho(l') < \rho(l'')$.

Then the following ordering strategy (denoted by σ) of marker loci is a perfect elimination ordering: once blocks have been ordered according to ρ, markers are ordered in such a way that the smaller the index of a block the smaller the indices of the markers pertaining to that block. The ordering inside each block is done as follows: markers in C_l are given the largest indices in block l. In addition, under this ordering strategy, condition 2.2 is also necessary for σ to be a perfect elimination ordering whereas condition 2.1 is not.

Corollary to Proposition 2

Consider the "super graph" formed by regarding the blocks as super nodes and $\mathcal{L}$ as a "super vertices set". Then, under conditions 2.1 and 1.4, if the "super vertices set" admits a perfect elimination ordering, the ordering defined in proposition 2 corresponds to a perfect elimination scheme.