Extensions of graphical models with applications in genetics and genomics

Behrouzi, Pariya

Publication date:
2018

Chapter 3

De novo construction of q-ploid linkage maps using discrete graphical models [1]

Abstract

Linkage maps are important tools for genetic research. New sequencing techniques have created opportunities to substantially increase the density of genetic markers. Such revolutionary advances in technology have given rise to new challenges, such as creating high-density linkage maps. Current multiple testing approaches based on pairwise recombination fractions are underpowered in the high-dimensional setting and do not extend easily to polyploid species. We propose to construct linkage maps using graphical models either via a sparse Gaussian copula or a nonparanormal skeptic approach. Linkage groups (LGs), typically chromosomes, and the order of markers in each LG are determined by inferring the conditional independence relationships among large numbers of markers in the genome. Through simulations, we illustrate the utility of our map construction method and compare its performance with other available methods, both when the data are clean and contain no missing observations and when data contain genotyping errors and are incomplete. We apply the proposed method to two genotype datasets: barley and potato from diploid and polypoid populations, respectively. Whereas most tetraploid potato linkage maps until now have been created either from diploid populations or from a subset of marker types, our comprehensive map construction method will be able to deal with

[1] Behrouzi, P., and Wit, E. (2017a). De novo construction of q-ploid linkage maps using discrete graphical models. *arxiv preprint arxiv*:1710.01063v3.

realistic data settings for any biparental diploid and polyploid species containing arbitrary marker types. We have implemented the method in the R package `netgwas` which is freely available at https://CRAN.R-project.org/package=netgwas.

Key words: Linkage mapping; Diploid; Polyploid; Graphical models; Gaussian copula; High-density genotype data.

3.1 Introduction

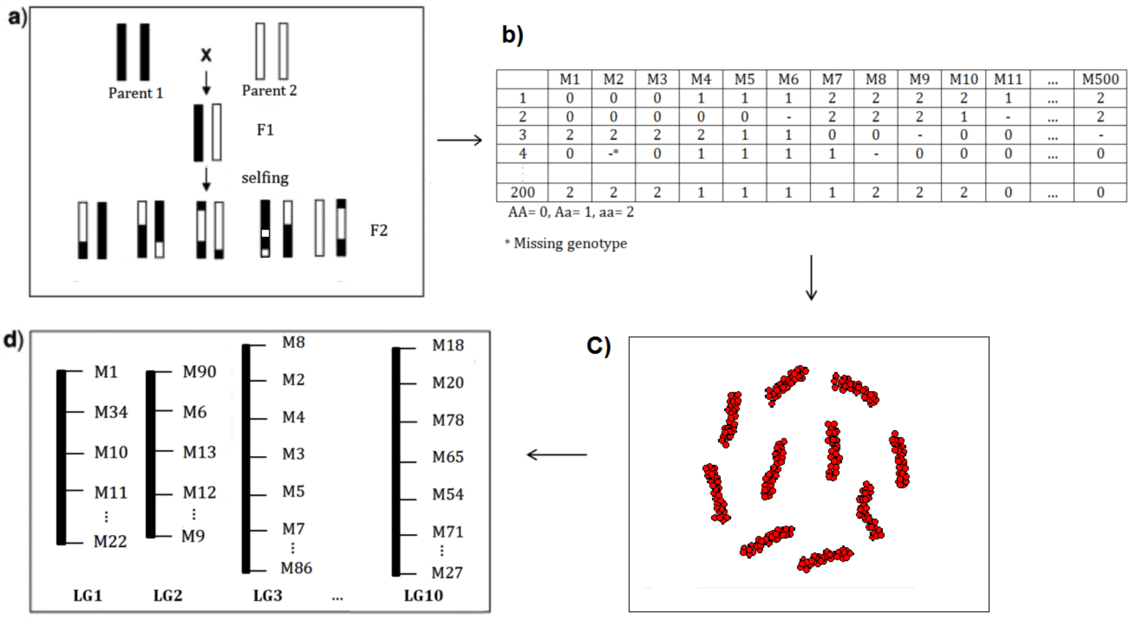
A linkage map provides a fundamental resource to understand the order of markers for the vast majority of species whose genomes are yet to be sequenced. Furthermore, it is an essential ingredient in the often used QTL mapping of genetic diseases, and particularly in identifying genes responsible for heritable or other types of diseases in humans or traits such as disease resistance in plants or animals.

Recent advances in sequencing technology make it possible to comprehensively sequence huge numbers of markers, construct dense maps, and ultimately create a foundation for studying genome structure and genome evolution, identifying quantitative trait loci (QTLs) and understanding the inheritance of multi-factorial traits. Next–generation sequencing (NGS) techniques offer massive and cost–effective sequencing throughput. However, they also bring new challenges for constructing high–quality linkage maps. NGS data can suffer from high rates of genotyping errors, as the observed genotype for an individual is not necessarily identical to its true genotype. Under such circumstances, constructing high–quality linkage maps can be difficult.

Each species is categorized as diploid or polyploid by comparing its chromosome number. Diploids have two copies of each chromosome. For diploid species many algorithms for constructing linkage maps have been proposed. Some of them have been implemented into user-friendly software, such as R/qtl (Broman et al., 2003), JoinMap (Jansen et al., 2001), OneMap (Margarido et al., 2007), and MSTmap (Wu et al., 2008). Among the algorithms for constructing genetic maps, R/qtl estimates genetic maps and identifies genotyping errors in relatively small sets of markers. JoinMap is a commercial software widely used in the scientific genetics community. It uses two methods to construct genetic maps: one is based on regression (Stam, 1993) and the other uses a Monte Carlo multipoint maximum likelihood (Jansen et al., 2001). OneMap has been reported to construct linkage maps in non-inbred populations. However, it is computationally expensive. The MSTMap is a fast genetic map algorithm that determines the order of markers by computing the minimum spanning tree of an associated graph.

Polyploid organisms have more than two chromosome sets. Polyploidy is very common in flowering plants and in different crops such as watermelon, potato, and bread wheat, which contain three (triploid), four (tetraploid), and six (hexaploid) sets of chromosomes, respectively. Despite the importance of polyploid species, statistical tools for construction of their linkage map are underdeveloped. However, Grandke et al. (2017) recently developed a method for this purpose. Their method is based on calculating recombination frequencies between marker pairs, then using hierarchical clustering and an optimal leaf algorithm to detect chromosomes and order markers. Nevertheless, this method can be computationally expensive even for a small numbers of markers. Furthermore, most literature has focused on constructing genetic linkage maps for tetraploids, but these are limited only to autotetraploid species. Only one, TetraploidMap, has been implemented in a software (Preedy and Hackett, 2016), but because it needs manual interaction and visual inspection its implementation is limited. Furthermore, current approaches to polyploid map construction are based mainly on estimation of recombination frequency and LOD scores (Wang et al., 2016), which does not use the full multivariate information in the data.

Different diploid and polyploid map construction methods have made substantial steps toward building better–quality linkage maps. However, the existing methods still suffer from low quality genetic mapping performance, in particular when ratios of genotyping errors and missing observations are high. The main contribution of this chapter is to introduce, for diploid and polyploid species, a novel linkage map algorithm to overcome the difficulties arising routinely in NGS data. With the proposed method we aim to build high–density and high–quality linkage maps using the statistical property called conditional dependence relationships, which reveals direct relations among genetic markers. For diploid scenarios, we evaluated the performance of the proposed method and the other methods in several comprehensive simulation studies, both when the input data were clean and had no missing observations and when the input data were very noisy and incomplete. We measured the performance of the methods in accuracy scores of grouping and ordering. In addition, we studied the performance of our method in constructing linkage maps for simulated polyploids, namely tetraploids and hexaploids. Furthermore, we applied the map construction method in `netgwas` (Behrouzi and Wit, 2017b) to construct maps for two genotype datasets: barley and potato from diploid and tetraploid populations, respectively.



	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	...	M500
1	0	0	0	1	1	1	2	2	2	2	1	...	2
2	0	0	0	0	0	-	2	2	2	1	-	...	2
3	2	2	2	2	1	1	0	0	-	0	0	...	-
4	0	-*	0	1	1	1	1	-	0	0	0	...	0
⋮													
200	2	2	2	1	1	1	1	2	2	2	0	...	0

AA= 0, Aa= 1, aa= 2

* Missing genotype

Fig. 3.1 General view of proposed linkage map estimation process. To illustrate, we use a diploid population containing two copies of each chromosome. (a) Example of mating experiment of an inbred F2 population. (b) Derived genotype data for 200 individuals which have genotyped for 500 markers. (c) Reconstruction of undirected graph between all 500 markers. (d) 10 linkage groups (chromosomes) with markers ordered within each linkage group (LG).

3.2 Genetic background on linkage map

A linkage map is the linear order of genetic markers on a chromosome. Geneticists use it to study the association between genes and traits. In this section we describe the relationship between a linkage map and single nucleotide polymorphism (SNP) markers. For the moment, we assume that each allele can take only one of two values, A or a. This assumption can be relaxed without requiring any methodological adjustments; more will follow in the discussion. Here, we are dealing with markers from high–throughput data such as NGS and SNP arrays.

3.2.1 Linkage map for diploids and polyploids

Diploid organisms contain two sets of chromosomes, one from each parent, whereas polyploids contain more than two sets of chromosomes. In polyploids the number of chromosome sets reflects their level of ploidy: triploids have three sets, tetraploids have four, pentaploids have five, and so forth. Here, we refer to diploids and polyploids as q-ploid $q \geq 2$, where in diploids $q = 2$, triploids $q = 3$, tetraploids $q = 4$, and so on.

The genotype of any q-ploid organism can be homozygous or heterozygous at each single locus on the genome. Different genotype forms of the same gene are called alleles. Alleles can lead to different traits. Alleles are commonly represented by letters; for example, for the gene related to the trait, the allele could be called A and a. In q-ploid individuals there are q copies of allele. If all q allele copies of an organism are identical, the organism is in the homozygous state at that locus; otherwise it is in the heterozygous state. For instance, a tetraploid individual is homozygous for two size alleles, A and a, if all 4 allele copies are either A, or a, which correspond with the genotypes $AAAA$ and $aaaa$, respectively. If a tetraploid individual is heterozygous the following three genotypes would appear: one copy of the A allele and three copies of a (e.g. Aaaa), two copies of A and two copies of a (e.g. AAaa), or three copies of A and one copy of a (e.g. AAAa). Unlike existing methods, our method works not only for diploid organisms but also for all polyploids. Obviously, our method can also be used to analyze simple haploid organisms such as haploid yeast cells.

3.2.2 Mapping population

Mating between two parental lines with recent common biological ancestors is called inbreeding. Mating between parental lines with no common ancestors up to e.g. 4-6 generations is called outcrossing. In both cases, the genomes of the derived progenies are random mosaics of the genomes of the parents. As a consequence of inbreeding parental alleles are attributable to each parental line in the genome of the progeny, whereas in outcrossing this is not the case.

Inbreeding progenies derive from two homozygous parents. Some inbreeding designs, such as *Backcrossing* (BC), lead to a homozygous population where the derived genotype data include only homozygous genotypes of the parents, namely AA and aa (conveniently coded as 0 and 1). However, some other inbreeding designs such as $F2$ lead to a heterozygous population, where the derived genotype data contain both heterozygous and homozygous genotypes, namely AA, Aa, and aa (conveniently coded as 0, 1 and 2; see Figures 3.1a and 3.1b for an example of a diploid species). Although many other experimental designs are being used in genetic studies, not all existing methods for linkage mapping support all inbreeding experimental designs. However, our proposed algorithm constructs a linkage map for any type of biparental inbreeding experimental designs. In fact, unlike other existing methods, our approach does not require specifying the population type because it is broad and handles any population type that contains at least two distinct genotype states.

Outcrossing or outbred experimental designs, such as full−sib families, derive from two

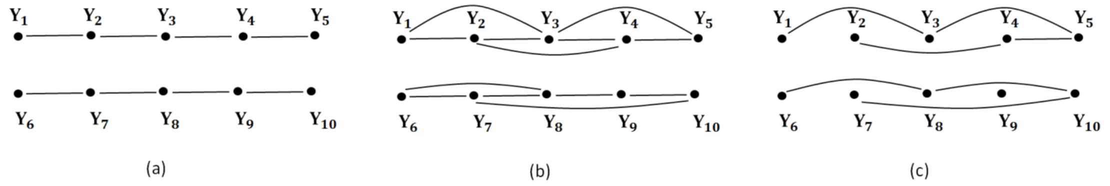


Fig. 3.2 Cartoon example of conditional dependence pattern between neighboring markers in different population types: (a) homozygous, (b) inbred, (c) outcrossing (outbred) populations, where ordered markers $Y_1, \ldots, Y_5$ reside on chromosome 1, and $Y_6, \ldots, Y_{10}$ on chromosome 2.

non–homozygous parents. Thus the genome of the progenies includes a mixed set of many different marker types, including fully informative markers and partially informative markers (e.g. missing markers). Markers are called fully informative when all of the resulting gamete types can be phenotypically distinguished on the basis of their genotypes; they are called partially informative when the gamete types have identical phenotypes.

3.2.3 Meiosis and Markov dependence

During meiosis, chromosomes pair and exchange genetic material (crossover). In diploids, pairing at meiosis occurs between two chromosomes. In polyploids the q chromosome copies may form different types of multivalent pairing. For example, in tetraploids all four chromosome copies may pair at meiosis. Assume a sequence of ordered SNP markers $X_1^c, X_2^c, \ldots, X_d^c$ along chromosome c in a q-ploid species. We describe the Markov dependence structure between markers for different population schemes. (i) During meiosis in inbred populations, genetic material from one of the two parents is copied into the offspring in a sequential fashion, i.e. reading along the genome, until the copying switches in a random fashion to the other parent. Thus, the genome of the offspring is a random but piecewise continuous mosaic of the genomes of its parents. The genotype state at each chromosomal region, or locus, of the offspring is either homozygous maternal, heterozygous, or homozygous paternal. For instance, as a result of genetic linkage and crossover a homozygous maternal genotype will typically be followed by a heterozygous genotype before being able to be followed by a homozygous paternal genotype.

Genetic linkage means that markers located close to one another on a chromosome are linked and tend to be inherited together during meiosis. Another key biological fact is that during meiosis markers on different chromosomes segregate independently; this is called the *independent assortment law.*

For example, in scheme (i) consisting of only a homozygous population, the random variable Y_j which represents the genotype of an individual at location j can be defined as

$$Y_j = \begin{cases} 1 & \text{paternal marker at locus } j \text{ on homologue k,} \\ 0 & \text{otherwise.} \end{cases}$$

This scheme occurs in inbred homozygous populations that include only two genotype states, namely homozygous maternal and homozygous paternal. Mapping populations, such as backcrossing, are included in this scheme. Then, under the assumption of no crossover interference – meaning when a crossover has formed, other crossovers are not prevented from forming – the recombination frequency between the two locations j and $j+1$ is independent of recombination at the other locations on the genome. So, the following holds

$$Pr(Y_{j+1} = y_{j+1} \mid Y_j = y_j, Y_{j-1} = y_{j-1}, \ldots, Y_1 = y_1) = Pr(Y_{j+1} = y_{j+1} \mid Y_j = y_j) \quad (3.1)$$

This equation indicates that the genotype of a marker at location $j+1$ is conditionally independent of genotypes at locations $j-1, j-2, \ldots, 1$ given a genotype at location j. This can be written as

$$Y_{j+1} \perp\!\!\!\perp (Y_1, \ldots, Y_{j-1}) \mid Y_j \quad (3.2)$$

This defines a discrete graphical model $G = (V, E)$ which consists of vertices $V = \{1, \ldots, p\}$ and edge set $E \subseteq V \times V$ with a binary random variable $Y_j \in \{0, 1\}^p$. Given the above property between neighboring markers, we construct linkage maps using conditional (in)dependence models. Figure 3.2a shows a cartoon image of conditional (in)dependencies for this scheme.

Scheme (ii): In inbred populations, one complication arises when in the genotype data we cannot identify each homologue due to heterozygous genotypes. Q-ploid ($q \geq 2$) heterozygous inbred populations, like $F2$, are examples of such cases, where we define X_{jk} as

$$X_{jk} = \begin{cases} 1 & \text{if marker } j \text{ on homologue k is of type } A, \\ 0 & \text{otherwise} \end{cases}$$

where A is one of the two possible alleles at that specific location. Here, X_{jk} represents the allele at homologue k of a chromosome, where the genotype in that location can be written as $X_{j.} = \{X_{j1} \ldots X_{jq}\}$. For example, at marker location j, $X_j = Aaaa$ is one possible genotype for a tetraploid species ($q = 4$); it includes one copy of the desirable allele A where $X_{j1} = 1$, $X_{j2} = 0$, $X_{j3} = 0$, and $X_{j4} = 0$ represent the alleles in the first, second, third and fourth homologues, respectively. The other possible genotypes which include one copy of the desired allele A are $aAaa$, $aaAa$, $aaaA$. Because it is typically impossible to distinguish between genotypes with the same number of copies of a desired allele (e.g. $Aaaa$, $aAaa$, $aaAa$, $aaaA$), we therefore take a random variable Y_j as observed in the number of A alleles at location j:

$$Y_j = \sum_{k=1}^{q} X_{jk}. \tag{3.3}$$

Table 3.1 shows an example of correspondence between Y_j and $X_{j.}$ for a q-ploid species when $q = 4$. We note that a q-ploid species contains $q + 1$ genotype states at location j, as shown in Table 3.1 for a tetraploid species.

Due to *genetic linkage*, the sequence of ordered SNP markers $Y_1, Y_2, \ldots, Y_d$ forms a Markov chain as equation (3.1) with state space S which contains $q + 1$ states. Therefore, the conditional (in)dependence relationship (3.2) between neighboring markers is held. Figure 3.2b presents a cartoon image of the conditional independence graph for this scheme.

Scheme (iii): In outcrossing (outbred) populations, unlike inbred populations, the meaning of “parental" is either unknown or not well defined. In other words, markers in the genome of the progenies can not easily be assigned to their parental homologues. For example, if both non-homozygous parents contain $A_jA_jA_jA_j$ genotype at marker location j, then offspring will also have $A_jA_jA_jA_j$ genotype at marker location j. But we do not know whether that genotype belongs to the paternal or maternal homologue, since both parents have $A_jA_jA_jA_j$ genotype at marker location j. So, in this case we define X_{jk} as follows

Y_j	$X_{j.}$
0	$aaaa$
1	$Aaaa$, $aAaa$, $aaAa$, $aaaA$
2	$AAaa$, $AaAa$, $AaAa$, $AaaA$, $aaAA$
3	$AAAa$, $AaAA$, $AAaA$, $AAAa$, $aAAA$
4	$AAAA$

Table 3.1 Number of copies (dosage) of a reference allele. Relation between different genotypes, $X_{j.}$, and allele dosage, Y_j, for a tetraploid individual, where A is the reference allele.

$$X_{jk} = \begin{cases} 1 & \text{if marker } j \text{ on homologue } k \text{ is of type } A_j, \\ 0 & \text{otherwise} \end{cases}$$

where A_j is one of the possible parental alleles at location j. So, random variable Y_j which represents the dosage of alleles, can be defined as equation (3.3).

Furthermore, in polyploids the *linkage* depends on how a single chromosome pairs during meiosis to generate gametes. In this regard, if both polyploid parents have an A_j allele in all q haploids, then the offspring will also have it, and this will not co-vary with neighboring markers. The possibility of different pairing models during meiosis makes the situation more complex. In diploids, the two homologue chromosomes pair up and form a bivalent, then cross-over before recombinations occur. But polyploid meiosis can occur in various ways; in tetraploids four homologue chromosomes can during meiosis form either two separate bivalents, each of which contributes one haploid, like diploids, or, alternatively, in a more complex situation, the four homologue chromosomes can form quadrivalents, so that cross-over occurs between eight haploids. In both pairing models, bivalent or quadrivalent, crossover events result in recombined haploids that are mosaics of parental chromosomes. Outbred progenies are genetically diverse and highly heterozygous, whereas inbred individuals have little or no genetic variation.

The term (3.1) partially holds for the scheme (iii), where a discrete graphical model can be defined for a multinomial variable $Y_j = \{0, 1, \ldots, q\}$. We use conditional independence to construct linkage maps in outbred populations. However, in this type of population, due to a mixed set of different marker types, the conditional independence relationship between neighboring markers may be more complicated. Many genetic assumptions made in traditional linkage analyses (e.g., known parental linkage phases throughout the genome) do not hold here. For example, when both parents have A_j allele, then their offspring will also have it; however this will not covary with neighboring markers. Figure 3.2c shows a cartoon example of such conditional independence graphs.

To summarize, term (3.1) holds for schemes (i) and (ii), and partially (iii) because transition probability from a genotype at location j to a genotype at location $j + 1$ depends on the recombination frequency between the two locations j and $j + 1$, which is independent of recombination in the other locations. This can be modeled by a discrete Markov process $\{Y_j\}_{j=1,\ldots,d}$ with state space S which contains $q + 1$ genotype states and a transition matrix, which, in case of polyploids ($q \geq 3$), can be calculated with respect to the mode of chromosomal pairing (e.g. bivalent or quadrivalent). The Markov structure of the SNP markers in all three schemes yields a graphical model with as many nodes as markers in a genome. The random variable X_j follows a discrete graphical model whereby the joint distribution $P(X)$ can be factorized as,

$$P(X) = \prod_{c=1}^{C} \prod_{j=1}^{p_c-1} f_{j,j+1}^{(c)}(X_j^{(c)}, X_{j+1}^{(c)}), \tag{3.4}$$

where C defines the number of chromosomes in a genome, and p_c stands for the number of markers in chromosome c (see Section 1.2.1). The outer multiplication of (3.4) shows the *independent assortment law*, and the inner multiplication represents the *genetic linkage* between markers within a chromosome, where the factor $f_{j,j+1}^{(c)}$ indicates the conditional dependence between adjacent markers, given the rest of the markers. Through this probabilistic insight, the inferred conditional (in)dependence relationship between markers provides a high-dimensional space for the construction of a linkage map.

3.3 Algorithm to detect linkage map

We propose to build a linkage map in two steps; first, we reconstruct an undirected graph for all SNP markers on a genome, and second, we determine the correct order of markers in the obtained linkage groups from the first step. We also show how our method handles genotyping errors and missing observations in reconstructing a linkage map.

3.3.1 Estimating marker-marker network

To reconstruct an undirected graph between SNP markers in a q-ploid species we propose two methods: the sparse ordinal glasso approach (Behrouzi and Wit, 2017a) and the nonparanormal skeptic approach (Liu et al., 2012) (the latter discussed under Supplementary Materials). The former method can deal with missing values, whereas the latter is computationally faster.

An undirected graphical model for the joint distribution (3.4) of a random vector $Y = (Y_1, \ldots, Y_p)$ is associated with a graph $G = (V, E)$, where each vertex j corresponds to a variable Y_j. The pair (j, l) is an element of the edge set E if and only if Y_j is dependent of Y_l, given the rest of the variables. In the graph estimation problem, we have n samples of the random vector Y, and it is our aim to estimate the edge set E. Depending on how various mapping populations are produced, Y represents either binary variables $Y = \{0, 1\}$, as in homozygous populations, or multinomial variables $Y = \{0, 1, \ldots, q + 1\}$ where q is the ploidy level. For example in diploids q is 2 and in tetraploids 4.

Sparse ordinal glasso A relatively straightforward approach to discover the conditional (in)dependence relation among markers is to assume underlying continuous variables $Z_1, \dots, Z_p$ for markers $Y_1, \dots, Y_p$, which can not be observed directly. In our modeling framework, Y_j and Z_j define observed rank and true latent value, respectively, where each latent variable corresponds to one observed variable. The relationship between Y_j and Z_j is expressed by a set of cut-points $(-\infty, C_1^{(j)}], (C_1^{(j)}, C_2^{(j)}] \dots, (C_q^{(j)}, \infty)$, which is obtained by partitioning the range of Z_j into $q_j - 1$ disjoint intervals. Thus, $y_j^{(i)}$, which represents the genotype of the i-th sample for the j-th marker, can be written as follows

$$y_j^{(i)} = \sum_{k=1}^{q} k \times 1_{\{C_{q-1}^{(j)} < z_j^{(i)} \leq C_q^{(j)}\}} \qquad i = 1, 2, \dots, n, \tag{3.5}$$

where we define $\mathcal{D} = \{z_j^{(i)} \in \mathbb{R} \mid C_{q-1}^{(j)} < z_j^{(i)} \leq C_q^{(j)}\}$. We use a high dimensional Gaussian copula with discrete marginals. We assume

$$Z \sim N_p(0, \Sigma)$$

where the $p \times p$ precision matrix $\Theta = \Sigma^{-1}$ contains all the conditional independence relationships between the latent variables. Given our parameter of interest Θ, we nonparametrically estimate the cut-points for each $j = 1, \dots, p$ as follows

$$\widehat{C}_q^{(j)} = \begin{cases} -\infty & \text{if } q = 0\ ; \\ \Phi^{-1}(\sum_{i=1}^{n} I(y_j^{(i)} \leq q)/n) & \text{if } q = 1, \dots, q_j - 1; \\ +\infty & \text{if } q = q_j. \end{cases}$$

Penalized EM algorithm In genotype datasets we commonly encounter situations where the number of genetic markers p exceeds the number of samples n. To solve this dimensionality problem we propose to impose an l_1 norm penalty on the likelihood consisting of the absolute value of the elements of the precision matrix Θ. Furthermore, to be able to deal with commonly occurring missing values in genotype data we implement an EM algorithm (McLachlan and Krishnan, 2007), which iteratively finds the penalized maximum likelihood estimate $\widehat{\Theta}_\lambda$. This algorithm proceeds by iteratively computing the conditional expectation of complete log-likelihood and optimizing it. In the E-step we compute the conditional expectation in the penalized log-likelihood

$$Q_\lambda(\boldsymbol{\Theta} \mid \widehat{\boldsymbol{\Theta}}^{(m)}) = \frac{n}{2}\Bigg[\log|\boldsymbol{\Theta}| - tr(\frac{1}{n}\sum_{i=1}^{n} E_{Z^{(i)}}(Z^{(i)}Z^{(i)t}|y^{(i)}, \widehat{\boldsymbol{\Theta}}^{(m)}, \widehat{\mathcal{D}})\boldsymbol{\Theta}) - p\log(2\pi)\Bigg] \\ - \lambda||\Theta||_1 \tag{3.6}$$

where λ is a nonnegative tuning parameter. To calculate the conditional expectation $\bar{R} = \frac{1}{n}\sum_{i=1}^{n} E_{Z^{(i)}}(Z^{(i)}Z^{(i)t}|y^{(i)}, \widehat{\Theta}^{(m)}, \widehat{\mathcal{D}})$ we propose two different approaches, namely Gibbs sampling and an approximation method (Behrouzi and Wit, 2017a). Further details on the calculation of the conditional expectation are provided in the Supplementary Materials. The M-step is a maximization problem which can be solved efficiently using either graphical lasso (Friedman et al., 2008)

$$\widehat{\boldsymbol{\Theta}}^{(m+1)}_{glasso} = \arg\max_{\boldsymbol{\Theta}} \Big\{ \log|\boldsymbol{\Theta}| - tr(\bar{R}\boldsymbol{\Theta}) - \lambda||\boldsymbol{\Theta}||_1 \Big\}$$

or the CLIME estimator (Cai et al., 2011)

$$\widehat{\Theta}^{(m+1)}_{\text{CLIME}} = \arg\min_{\Theta} ||\Theta||_1 \qquad \text{subject to} \qquad ||\bar{R}\Theta - I_p||_\infty \leq \lambda,$$

where I_p is a p-dimensional identity matrix.

In large-scale genotyping studies, it is common to have missing genotype data. Before determining the number of linkage groups and ordering markers, we handle the missing data within the E-step of the EM algorithm, where we calculate the conditional expectation of true latent variables given the observed ranks. If an observed value, $y_j^{(i)}$ is missing, we take the unconditional expectation of the corresponding latent variable. In the EM framework we can easily handle high ratios of missingness in the data.

3.3.2 Determining linkage groups

A group of loci that are correlated defines a linkage group (LG). Depending on the density and proximity of the underlying markers each LG corresponds to a chromosome or part of a chromosome. The number of discovered linkage groups is controlled by the tuning parameter λ (section 3.3.1). We use the extended Bayesian criterion (eBIC), which has successfully been applied by Yin and Li (2011) in selecting sparse Gaussian graphical models for genomic data to determine the number of linkage groups. The eBIC is defined as

$$eBIC(\lambda) = -2\ell(\widehat{\Theta}_\lambda) + (\log n + 4\gamma \log p)\mathrm{df}(\lambda), \tag{3.7}$$

where $\ell(\widehat{\Theta}_\lambda)$ is the non-penalized likelihood and $\gamma \in [0,1]$ is an additional parameter. And $df(\lambda) = \sum_{1 \leq i < j \leq p} I(\widehat{\theta}_{ij,\lambda} \neq 0)$ where $\widehat{\theta}_{ij,\lambda}$ is (i,j)th entry of the estimated precision matrix $\widehat{\Theta}_\lambda$ and I is the indicator function. In case of $\gamma = 0$ the classical BIC is obtained. Typical values for γ are $1/2$ and 1. We select the value of λ that minimizes (3.7) for $\gamma = \frac{1}{2}$. It is notable that in existing map construction methods the construction of linkage groups is usually done by manually specifying a threshold for pairwise recombination frequencies; this, however, influences the output map, whereas our method detects LGs automatically in a data–driven way.

Figure 3.1(c) shows an example of an estimated conditional independence graph between markers. This graph includes 10 distinct sub–graphs, each of which corresponds to a linkage group. In this graph, given all markers on a genome, markers within the linkage groups are conditionally dependent, due to *genetic linkage*, and markers between linkage groups are conditionally independent, due to the *independent assortment law*.

Some genotype studies suffer from low numbers of samples or they contain signatures of epistatic selection (Behrouzi and Wit, 2017a), which may cause bias in determining the linkage groups. To address this problem, besides the model selection step, we use the fast-greedy algorithm to detect the linkage groups in the inferred graph. This community detection algorithm reflects the two biological concepts of genetic linkage and independent assortment in a sense that it defines communities which are highly connected within, and have few links between communities.

3.3.3 Ordering markers

Assume that a set of d markers has been assigned to the same linkage group. Let $G(V^{(d)}, E^{(d)})$ be a sub–graph on the set of unordered d markers, where $V^{(d)} = \{1, \ldots, d\}$, $d \leq p$ and the edge set $E^{(d)}$ represents the estimated edges among d markers where $E^{(d)} \subseteq E$. We remark that the precision matrix $\widehat{\Theta}_\lambda^{(d)}$, a submatrix of $\widehat{\Theta}_\lambda$, contains all conditional dependence relations between the set of d markers. Depending on the type of mating between the parental lines we introduce two methods to order markers, one based on dimensionality reduction and another based on bandwidth reduction. Both methods result in a one-dimensional map.

Inbred In inbred populations, markers in the genome of the progenies can be assigned to their parental homologues, resulting in a simpler conditional independence pattern between neighboring markers. In the case of inbreeding, we use multidimensional scaling (MDS) to represent the original high-dimensional space in a one-dimensional map while attempting to maintain pairwise distances. We define the distance matrix D which is a $d \times d$ symmetric matrix where $D_{ii} = 0$ and $D_{ij} = -\log(\rho_{ij})$ for $i \neq j$. Here, the matrix ρ represents the conditional correlation among d objects which can be obtained as $\rho_{ij} = -\frac{\theta_{ij}}{\sqrt{\theta_{ii}}\sqrt{\theta_{jj}}}$, where θ_{ij} is the ij-th element of the precision matrix Θ.
We aim to construct a configuration of d data points in a one–dimensional Euclidean space by using information about the distances between the d nodes. Given the distance matrix D, we define a linear ordering L of d elements such that the distance $\widehat{D}$ between them is similar to D. We consider a metric MDS, which minimizes $\widehat{L} = \arg\min_L \sum_{i=1}^{d} \sum_{j=1}^{d} (D_{ij} - \widehat{D}_{ij})^2$ across all linear orderings.

Outbred An outbred population derived from mating two non-homozygous parents results in markers in the genome of progenies that can not easily be assigned to their parental homologues. Neighboring markers that vary only on different haploids will appear as independent, therefore requiring a different ordering algorithm [see Figure 3.2c]. In that case, to order markers we use the reverse Cuthill-McKee (RCM) algorithm (Cuthill and McKee, 1969). This algorithm is based on graph models. It reduces the bandwidth of the associated adjacency matrix, $A_{d \times d}$, for the sparse matrix $\widehat{\Theta}_{\lambda}^{(d)}$. The bandwidth of the matrix A is defined by $\beta = \max_{\theta_{ij} \neq 0} |i - j|$. The RCM algorithm produces a permutation matrix P such that PAP^T has a smaller bandwidth than does A. The bandwidth is decreased by moving the non-zero elements of the matrix A closer to the main diagonal. The way to move the non-zero elements is determined by relabeling the nodes in graph $G(V_d, E_d)$ in consecutive order. Moreover, all of the nonzero elements are clustered near the main diagonal.

3.4 Simulation study

In this section, we study the performance of the proposed method for different diploids and polyploids. In section 3.4.1 we perform a comprehensive simulation study to compare the performance of the proposed algorithm with other available tools in diploid map constructions, namely JoinMap (Jansen et al., 2001) and MSTMap (Wu et al., 2008). The former is based on Monte Carlo maximum likelihood and the latter uses a minimum spanning tree of a graph.

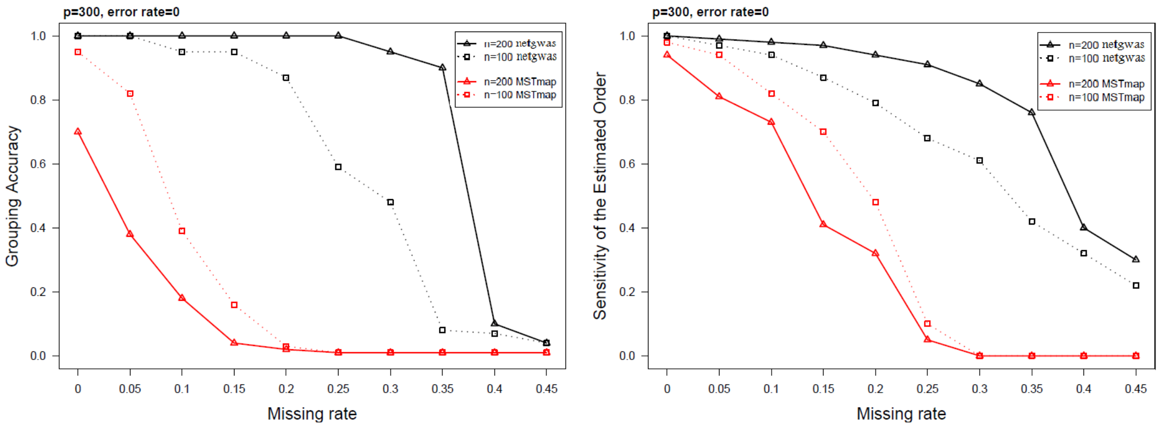
In section 3.4.2 we perform a simulation study to examine the algorithm performance on polyploids. At this moment the proposed method is the only one that constructs linkage maps for polyploid species automatically without any manual adjustment. Thus, in this case we can not compare the proposed method with other methods.

3.4.1 Diploid species

We simulate genotype data from an inbred $F2$ population. This population type generates discrete random variables with values $Y = \{0, 1, 2\}$ associated with the three distinct genotype states, AA, Aa, and aa at each marker. The procedure in generating genotype data is as follows: first, two homozygous parental lines are simulated with genotypes AA and aa at each locus. A given number of markers, p, are spaced along the predefined chromosomes. Then, two parental lines are crossed to give an $F1$ population with all heterozygous genotypes Aa at each marker location. Finally, a desired number of individuals, n, are simulated from the gametes produced by the $F1$ population.

A genotyping error means that the observed genotype for an individual is not identical to its true genotype, for example, observing genotype AA when Aa is the true genotype. Genotyping errors can distort the final genetic map, especially by incorrectly ordering markers and inflating map length. Therefore, to order markers that contain genotyping errors is an essential task in constructing high-quality linkage maps. To investigate this, we create genotyping errors in the simulated datasets [see Supplementary Materials] by randomly flipping the heterozygous loci along the chromosomes to either one of the homozygous allele. We inserted missing observations randomly along chromosomes simply by deleting genotypes.

For each simulated data, we compare the performance of the map construction in `netgwas` with two other models: JoinMap, and MSTMap. We compute two criteria: grouping accuracy (GA) and ordering accuracy (OA), to assess the performance of the above mentioned tools in estimating the correct map. The former measures the closeness of the estimated number of linkage groups to the correct number, and the latter calculates the ratio of markers that are correctly ordered. We define the grouping accuracy as follows: $GA = \frac{1}{1+(LG-\widehat{LG})^2}$, where LG stands for actual number of linkage groups and $\widehat{LG}$ is the estimated number of linkage groups. The GA criterion is a positive value with a maximum of 1. A high value of GA indicates good performance in determining the correct number of linkage groups. To compute ordering accuracy, we calculate the Jaccard distance, d_J,



(a) (b)

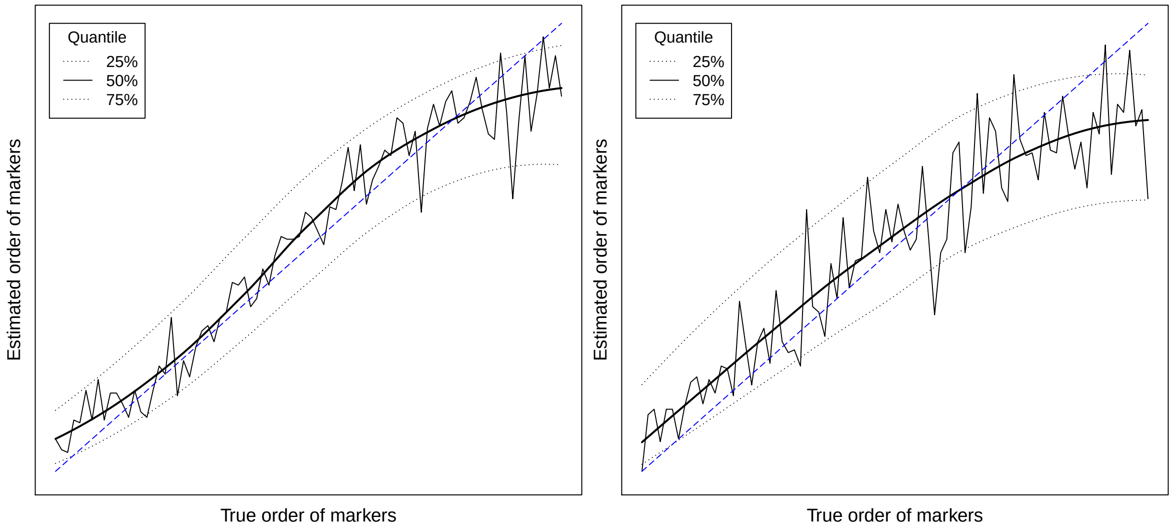
Fig. 3.3 Comparison of performance between map construction in `netgwas` and MSTMAP for different missingness rates with no genotyping errors. Variables p and n represent numbers of markers and individuals in simulated diploid genotype datasets. (a) Reports grouping, and (b) shows ordering accuracy scores for 50 independent runs

which measures mismatches between the estimated order and the true order. We define the ordering accuracy of the estimated map as $OA = \frac{1}{1+d_J}$. This measurement lies between 0 and 1, where 1 and 0 stand for a perfect and a poor ordering, respectively.

In terms of computation, `netgwas` runs in parallel. In the performed simulations, we ran the map construction functions, both in `netgwas` and the MSTMAP on a Linux machine with 24 2.5 GHz Intel Xeon processors and 128 GB memory. JOINMAP runs only on Windows. We ran it on a Windows machine with 3.20 GHz Intel Xeon processors and 8 GB RAM memory.

Evaluation of estimated maps in presence of missing genotypes We studied the effect of different ratios of missingness in the accuracy of the estimated linkage maps using two methods: `netgwas` and MSTMAP. The simulated data contained 300 markers for both $n = 100$ and $n = 200$ individuals where the missingness rates ranged from 0 to 0.45. In these sets of simulations we assumed no genotyping error [More simulation sets are performed in Supplementary Materials].

Figure 3.3 evaluates the accuracy of estimated maps in terms of grouping (Figure 3.3a) and ordering accuracies (Figure 3.3b). In general, this figure shows that `netgwas` constructed significantly better maps than MSTMAP across the full range of missingness rates. More specifically, for a moderate number of individuals, $(n = 200)$, Figure 3.3a shows that `netgwas` correctly estimated the actual number of linkage groups for missingness rates up to 0.25; when rates were between 0.25 and 0.35, `netgwas` estimated with high accu-



(a) (b)

Fig. 3.4 Performance of `netgwas` on different polypoid simulated datasets. Median, lower quartile, and upper quartile of estimated order versus true order for (a) tetraploids ($q = 4$), and (b) hexaploid simulated datasets ($q = 6$). Solid lines indicate median and smoothed median. Blue dashed line indicates ideal ordering.

racy (≥ 0.90) the actual number of linkage groups. Only when the missing rate was higher than 35% did `netgwas` begin to estimate the actual number of linkage groups poorly. For $n = 100$ `netgwas` correctly estimated the actual number of linkage groups up to 0.05 missingness, and very accurately (≥ 0.9) estimated the number of linkage groups for missingness rates between 0.05 and 0.2. With more than 20% missingness the accuracy diminished. MSTMAP always made significantly poorer estimates of the actual number of linkage groups than did `netgwas`; its performance immediately began to drop as soon as there was some level of missingness. Surprisingly, it estimated the number of linkage groups better when $n = 100$ than $n = 200$, but this may have been a fluke.

Figure 3.3b shows the ordering accuracy within each correctly estimated linkage group. Ordering quality in `netgwas` was significantly better than MSTMAP for both $n = 100$ and $n = 200$. More specifically, when $n = 100$ and the missing rate equaled zero, `netgwas` ordered markers perfectly (100% accuracy) and MSTMAP orders markers with a high accuracy (95%). In addition, with increased missingness rates, the map construction function in `netgwas` outperformed that of the MSTMAP in ordering markers within each LG. Surprisingly, when the number of individuals increased, MSTMAP performed more poorly in ordering markers.

3.4.2 Polyploid species

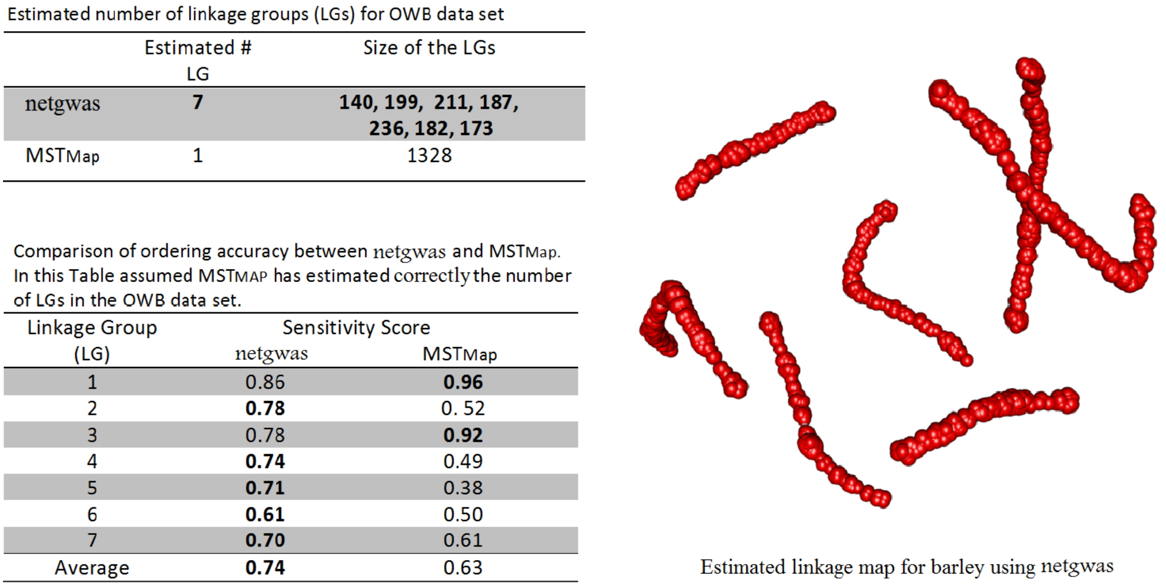
We also applied `netgwas` to simulated outbred polyploid genotype datasets. We used PedigreeSim (Voorrips and Maliepaard, 2012) to simulate $F1$ mapping populations in tetraploids ($q = 4$) and hexaploids ($q = 6$) with $n = 200$ individuals. PedigreeSim simulates polyploid genotypes with different configurations, such as chromosomal pairing modes during meiosis. The simulated tetraploids ($q = 4$) are motivated by autotetraploid potato where $Y = \{0, 1, 2, 3, 4\}$ corresponds to the five biallelic tetraploid genotype states ($aaaa$, $Aaaa$, $AAaa$, $AAAa$, $AAAA$), which are created across 12 chromosomes. The simulated hexaploids ($q = 6$) are motivated by allohexaploid peanut, a polyploid species that contains 10 chromosomes, where $Y = \{0, 1, 2, 3, 4, 5, 6\}$ corresponds to the seven genotype states ($aaaaaa$, $Aaaaaa$, $AAaaaa$, $AAAaaa$, $AAAAaa$, $AAAAAa$, $AAAAAA$) across its genome. In total, 50 populations, each consisting of $p = 1000$ markers, were simulated for each scenario.

We used the mean square error (MSE) as a measure for evaluating the performance of the proposed method on detecting the true number of chromosomes. In the tetraploid simulation the mean of MSE was 0.52, and for the hexaploid simulation it was 0.15. Figure 3.4 shows the performance of the proposed method in ordering markers for tetraploids (Figure 3.4a) and hexaploids (Figure 3.4b). The solid line shows the median of estimated order of each marker across a chromosome versus the true order, and the lower (25%) and upper quartiles (75%) of the estimated marker order is shown as dashed lines. This figure shows that, although ordering markers in outcrossing families is challenging [see section 3.2.3], the proposed method orders markers reasonably well.

3.5 Construction of linkage map for diploid barley

In the literature a barley genotyping dataset is used to compare different map construction methods for real-world diploid data. This genotyping dataset is generated from a doubled haploid population, which results in homozygous individual plants, $Y_{ij} \in \{0, 1\}$. Barley genotype data are the result of crossing Oregon Wolfe Barley Dominant with Oregon Wolfe Barley Recessive (see http://wheat.pw.usda.gov/ggpages/maps/OWB). The Oregon Wolfe Barley (OWB) data include $p = 1328$ markers that were genotyped on $n = 175$ individuals of which 0.02% genotypes are missing. The barley dataset is expected to yield 7 linkage groups, one for each of the 7 barley chromosomes.

As shown in Figure 3.5, through estimating $\widehat{\Theta}_\lambda$, which contains conditional (in)dependence relationships between barley markers, we were able to correctly detect the 7 barley chro-



Estimated number of linkage groups (LGs) for OWB data set

	Estimated # LG	Size of the LGs
netgwas	**7**	**140, 199, 211, 187, 236, 182, 173**
MSTMap	1	1328

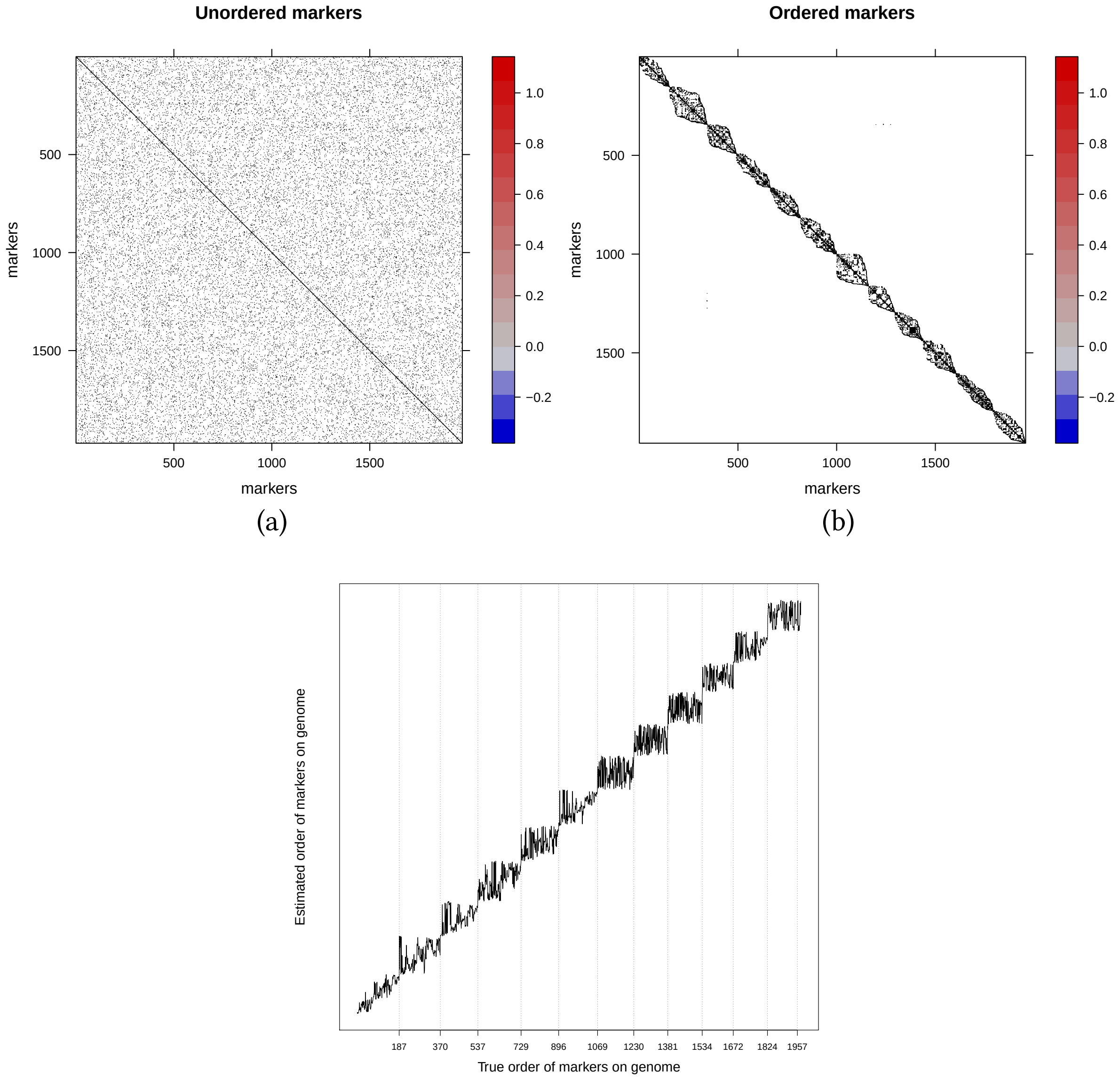
Comparison of ordering accuracy between netgwas and MSTMap. In this Table assumed MSTMAP has estimated correctly the number of LGs in the OWB data set.

Linkage Group	Sensitivity Score	
(LG)	netgwas	MSTMap
1	0.86	**0.96**
2	**0.78**	0. 52
3	0.78	**0.92**
4	**0.74**	0.49
5	**0.71**	0.38
6	**0.61**	0.50
7	**0.70**	0.61
Average	**0.74**	0.63

Estimated linkage map for barley using netgwas

Fig. 3.5 Summary of comparison between netgwas and MSTMAP in barley data. Table summarizes estimated number of LGs (chromosomes) and size of markers within each LG. Below, average ordering accuracy scores for the two methods. Right figure estimated undirected graph in netgwas for the barley data. This consists of 7 sub–graphs, each showing a chromosome.

mosomes as sub–graphs in the estimated undirected graph. Furthermore, using the conditional correlation matrix as distance in the multi-dimensional scaling approach helped us to order markers with high accuracy. In addition, Figure 3.5 reports the result of applying the two methods: netgwas and MSTMAP, to construct a linkage map for the barley data. The top part of Figure 3.5 shows that our method correctly estimated the true number of chromosomes. Also, the size of markers within each chromosome is consistent with the number of markers that reported in Cistué et al. (2011). MSTMAP was not able to estimate the true number of chromosomes and grouped all 1328 markers as one linkage group. The bottom of Figure 3.5 shows the accuracy of estimated marker order in 7 barley chromosomes. To be able to compare marker order in both methods we used the actual map to cluster markers in the map resulting from MSTMAP. Thus, at the bottom of Figure 3.5 it is assumed that the MSTMAP has estimated the correct number of chromosomes. Average ordering of accuracy scores across the linkage groups in netgwas is higher than those in MSTMAP except with chromosomes 1 and 3.



(a)

(b)

Fig. 3.6 Construction linkage map in potato. (a) Estimated precision matrix for unordered genotype data of tetraploid potato. (b) Estimated precision matrix after ordering markers. (c) True order of markers across potato genome, versus estimated order. Each dashed line represents a chromosome. All potato chromosomes detected correctly.

3.6 Construction of linkage map for tetraploid potato

World-wide, the potato is the third most important food crop (Bradshaw and Bonierbale, 2010). However, the complex genetic structure of tetraploid potatoe's (Solanum tuberosum L.) makes it difficult to improve important traits such as disease resistance in this crop. Thus there is a great interest in constructing linkage maps in the potato to identify markers related to disease resistance genes.

The full-sib mapping population MSL603 consists of 156 F1 plants resulting from a cross between female parent "Jacqueline Lee" and male parent "MSG227-2". The obtained genotype data contain 1972 SNP markers (Massa et al., 2015) with five allele dosages which are associated with the random variables $Y_j \in \{0, 1, \ldots, 4\}$ for $j = 1, \ldots, 1971$.

Figure 3.6 represents the result of applying the proposed map construction method to the unordered potato genotype data. Figure 3.6a shows the estimated sparse precision matrix for the unordered genotype data. Figure 3.6b represents the estimated precision matrix after ordering markers; it reveals the number of potato chromosomes as blocks across the diagonal. The potato genome contains 12 chromosomes. The proposed method correctly identifies the 12 chromosomes. The estimated linkage map contains 1957 markers. Figure 3.6c compares the estimated order versus the true order of markers. Each dashed line shows the estimated linkage groups. The markers ordered with reasonable precision, given that the ordering of markers has always been a challenging task in linkage map constructions, and in particular for polyploid species.

3.7 Conclusion

Construction of linkage maps is a fundamental and necessary step for detailed genetic study of diseases and traits. A high-quality linkage map provides opportunities for greater throughput gene manipulation and phenotype improvement. Here we have introduced a novel method for constructing linkage maps from high-throughput genotype data where the number of genetic markers exceeds the number of individuals. The proposed method constructs a linkage map for any biparental diploid or polyploid population. We proposed to build linkage maps in two steps: (i) inferring conditional independence relationships between markers on the genome; (ii) ordering markers in each linkage group, typically a chromosome. In the first step of the proposed method we used the Markov properties of adjacent markers: the genotype of an individual haploid at marker Y_j given its genotype at Y_{j-1} or Y_{j+1} is conditionally independent of the genotype at any other marker location. This property defines a graphical model for discrete random variables.

We employed a Gaussian copula graphical model combined with a penalized EM algorithm to estimate a sparse precision matrix $\widehat{\Theta}_\lambda$. This method iteratively computes the conditional expectation of the complete penalized log-likelihood, and optimizes it to estimate $\widehat{\Theta}_\lambda$. The method can also deal with missing values, which are very common in genotype datasets. The nonparanormal skeptic is an alternative approach that is computationally faster but can not deal with missing genotypes. Depending on the type of mapping population, inbred or outbred, in step 2 of the proposed linkage map construction we use either a multidimensional scaling approach or the Cuthill-McKee algorithm, respectively. Both ordering algorithms result in a one-dimensional map. We noted that in outcrossing populations it is difficult to order markers because a clear definition of the parental genotype is lacking.

We performed several simulation studies to compare the performance of the proposed method with other commonly used diploid map construction tools. To address the challenges in the construction of a linkage map from genotype data, we studied the performance of the proposed method on simulated data with high ratios of missingness and genotyping error. As shown in our simulation studies, our method, called `netgwas`, outperformed the commonly available linkage map tools, both when the input data were clean with no missing observations and when the input data were noisy and incomplete.

As outlined in Cervantes-Flores et al. (2008), constructing linkage maps in polyploids, with outcrossing behavior, is a challenging task. So far, based on our experience, no method has been developed to construct polyploid linkage maps for a large number of different marker types without any manual adjustment and/or visual inspection. Based on the simulated polyploids with outcrossing behavior, the proposed method detected the true number of linkage groups with high accuracy, and ordered markers with reasonable precision.

We applied the proposed method to two genotype studies involving barley and potato. In the barley map construction, we correctly detected its 7 chromosomes, whereas other method grouped all markers in one linkage group. The `netgwas` method ordered markers with higher accuracy in most of the chromosomes. The method detected all the potato chromosomes, although it identified chromosome 10 as two linkage groups. Its ordering of markers within each chromosome was a substantial improvement of what has been possible up until now. We remark that the proposed map construction method uses all possible marker types, unlike the other map construction methods, which use a subset of markers (Grandke et al., 2017).

We point out that `netgwas` also works for multi-allelic loci, which are locations in a genome that contain three or more observed alleles. For example, assume that A, T, and G are three possible alleles at location j on a genome, unlike the most usual cases whereby only two alleles can be observed at a location (e.g. A and G). We propose to analyze either separately or jointly a dataset containing multi-allelic loci. In the former case, observed alleles count once as reference, and therefore allow for one separate dataset. In the above example three datasets will be generated: the first dataset counts the number of A alleles as a reference, the second dataset counts the number of T alleles as a reference, and the third dataset counts the G allele as a reference. Each dataset can be analyzed separately;

to control similarity between the estimated precision matrices the fused graphical lasso can be used. The final map can be obtained through ordering markers in an estimated precision matrix. In the latter case, in the example above, we combine all three datasets as one dataset in such a way that it creates three replicates of $n \times p$ dimension. Moreover, we analyze the obtained dataset and construct the final linkage map.

3.8 Supporting information

Computing conditional expectation

We calculate $\bar{R}$ in equation (6) of this chapter as

$$E\left[Z^{(i)}Z^{(i)t}|Y^{(i)},\widehat{\Theta}^{(m)}\right] = E\left[Z^{(i)}|Y^{(i)},\widehat{\Theta}^{(m)}\right]E\left[Z^{(i)}|Y^{(i)},\widehat{\Theta}^{(m)}\right]^t + cov\left[Z^{(i)}|Y^{(i)},\widehat{\Theta}^{(m)}\right] \tag{3.8}$$

The conditional random variable $Z|Y$ follows a truncated p-variate normal distribution. Wilhelm and Manjunath, (2010) provided the analytical solution to compute moments of truncated multivariate normal distribution. However, their approach is feasible for only very few variables. Here, we propose instead to simulate a large number of samples from the truncated p-variate normal distribution and compute the sample conditional covariance matrix and sample conditional mean to estimate $E\big[Z^{(i)}Z^{(i)t}|Y^{(i)},\widehat{\Theta}^{(m)}\big]$ using the equation (3.8).

Alternatively, we use an efficient approximate estimation algorithm, which is implemented in Behrouzi and Wit, (2017). The variance elements in the conditional expectation matrix can be calculated through the second moment of the conditional $Z_j^{(i)} \mid Y^{(i)}$, and the rest of the elements in this matrix can be approximated through $E(Z_j^{(i)}Z_{j'}^{(i)} \mid y^{(i)};\widehat{\Theta},\widehat{\mathcal{D}}) \approx E(Z_j^{(i)} \mid y^{(i)};\widehat{\Theta},\widehat{\mathcal{D}})\; E(Z_{j'}^{(i)} \mid y^{(i)};\widehat{\Theta},\widehat{\mathcal{D}})$ using mean field theory. The first and second moment of $z_j^{(i)}|y^{(i)}$ can be written as

$$E(Z_j^{(i)} \mid y^{(i)},\widehat{\Theta},\widehat{\mathcal{D}}) = E[E(Z_j^{(i)} \mid z_{-j}^{(i)},y_j^{(i)},\widehat{\Theta},\widehat{\mathcal{D}}) \mid y^{(i)},\widehat{\Theta},\widehat{\mathcal{D}}], \tag{3.9}$$

$$E((Z_j^{(i)})^2 \mid y^{(i)},\widehat{\Theta},\widehat{\mathcal{D}}) = E[E((Z_j^{(i)})^2 \mid z_{-j}^{(i)},y_j^{(i)},\widehat{\Theta},\widehat{\mathcal{D}}) \mid y^{(i)},\widehat{\Theta},\widehat{\mathcal{D}}], \tag{3.10}$$

where $z_{-j}^{(i)} = (z_1^{(i)},\ldots,z_{j-1}^{(i)},z_{j+1}^{(i)},\ldots,z_p^{(i)})$. The inner expectations in (3.9) and (3.10) are relatively straightforward to calculate. $z_j^{(i)} \mid z_{-j}^{(i)}, y_j^{(i)}$ follows a truncated Gaussian distribution on the interval $[c^{(j)}_{y^{(i)}_j}, c^{(j)}_{y^{(i)}_j+1}]$ with parameters $\mu_{i,j}$ and $\sigma^2_{i,j}$ given by

$$\mu_{ij} = \widehat{\mathbf{\Sigma}}_{j,-j}\widehat{\mathbf{\Sigma}}^{-1}_{-j,-j}z^{(i)t}_{-j},$$

$$\sigma^2_{i,j} = 1 - \widehat{\mathbf{\Sigma}}_{j,-j}\widehat{\mathbf{\Sigma}}^{-1}_{-j,-j}\widehat{\mathbf{\Sigma}}_{-j,-j}.$$

Let $r_{k,l} = \frac{1}{n}\sum_{i=1}^{n} E(Z^{(i)}_k Z^{(i)}_l \mid y^{(i)}, \widehat{\mathbf{\Theta}}, \widehat{\mathcal{D}})$ be the (k,l)-th element of empirical correlation matrix $\bar{R}$, then to obtain the $\bar{R}$ two simplifications are required.

$$\begin{aligned} &E(Z^{(i)}_k Z^{(i)t}_l \mid y^{(i)}, \widehat{\mathbf{\Theta}}, \widehat{\mathcal{D}}) \approx E(Z^{(i)}_k \mid y^{(i)}, \widehat{\mathbf{\Theta}}, \widehat{\mathcal{D}})E(Z^{(i)}_l \mid y^{(i)}, \widehat{\mathbf{\Theta}}, \widehat{\mathcal{D}}) \quad \text{if } 1 \le k \ne l \le p, \\ &E(Z^{(i)}_k Z^{(i)t}_l \mid y^{(i)}, \widehat{\mathbf{\Theta}}, \widehat{\mathcal{D}}) = E((Z^{(i)}_k)^2 \mid y^{(i)}, \widehat{\mathbf{\Theta}}, \widehat{\mathcal{D}}) \qquad\qquad\qquad \text{if } k = l. \end{aligned}$$

Applying the results in the appendix to the conditional $z^{(i)}_j \mid z^{(i)}_{-j}, y^{(i)}_j$ we obtain

$$E(Z^{(i)}_j \mid y^{(i)}; \widehat{\mathbf{\Theta}}, \widehat{\mathcal{D}}) = \widehat{\mathbf{\Sigma}}_{j,-j}\widehat{\mathbf{\Sigma}}^{-1}_{-j,-j}E(Z^{(i)^t}_{-j} \mid y^{(i)}; \widehat{\mathbf{\Theta}}, \widehat{\mathcal{D}}) + \frac{\phi(\widehat{\delta}^{(i)}_{j,y^{(i)}_j} - \phi(\tilde{\delta}^{(i)}_{j,y^{(i)}_j+1})}{\Phi(\tilde{\delta}^{(i)}_{j,y^{(i)}_j+1}) - \Phi(\tilde{\delta}^{(i)}_{j,y^{(i)}_j})}\tilde{\sigma}^{(i)}_j, \tag{3.11}$$

$$\begin{aligned} E((Z^{(i)}_j)^2 \mid y^{(i)}; \widehat{\mathbf{\Theta}}, \widehat{\mathcal{D}}) = {} & \widehat{\mathbf{\Sigma}}_{j,-j}\widehat{\mathbf{\Sigma}}^{-1}_{-j,-j}E(Z^{(i)^t}_{-j} Z^{(i)}_{-j} \mid y^{(i)}; \widehat{\mathbf{\Theta}}, \widehat{\mathcal{D}})\widehat{\mathbf{\Sigma}}^{-1}_{-j,-j}\widehat{\mathbf{\Sigma}}^{t}_{j,-j} + (\tilde{\sigma}^{(i)}_j)^2 \\ & + 2\frac{\phi(\tilde{\delta}^{(i)}_{j,y^{(i)}_j}) - \phi(\tilde{\delta}^{(i)}_{j,y^{(i)}_j+1})}{\Phi(\tilde{\delta}^{(i)}_{j,y^{(i)}_j+1}) - \Phi(\tilde{\delta}^{(i)}_{j,y^{(i)}_j})}[\widehat{\mathbf{\Sigma}}_{j,-j}\widehat{\mathbf{\Sigma}}^{-1}_{-j,-j}E(Z^{(i)^t}_{-j} \mid y^{(i)}; \widehat{\mathbf{\Theta}}, \widehat{\mathcal{D}})]\tilde{\sigma}^{(i)}_j \\ & + \frac{\delta^{(i)}_{j,y^{(i)}_j}\phi(\tilde{\delta}^{(i)}_{j,y^{(i)}_j}) - \tilde{\delta}^{(i)}_{j,y^{(i)}_j+1}\phi(\tilde{\delta}^{(i)}_{j,y^{(i)}_j+1})}{\Phi(\tilde{\delta}^{(i)}_{j,y^{(i)}_j+1}) - \Phi(\tilde{\delta}^{(i)}_{j,y^{(i)}_j})}(\tilde{\sigma}^{(i)}_j)^2, \end{aligned} \tag{3.12}$$

where $Z^{(i)}_{-j} = (Z^{(i)}_1, \ldots, Z^{(i)}_{j-1}, Z^{(i)}_{j+1}, \ldots, Z^{(i)}_p)$ and $\tilde{\delta}^{(i)}_{j,y^{(i)}_j} = [c^{(i)}_j - E(\tilde{\mu}_{ij} \mid y^{(i)}; \widehat{\mathbf{\Theta}}, \widehat{\mathcal{D}})]/\tilde{\sigma}_{ij}$. In this way, an approximation for $\bar{R}$ is obtained as follows:

$$\tilde{r}_{kl} = \left\{ \begin{array}{ll} \frac{1}{n}\sum_{i=1}^{i=n} E(Z^{(i)}_k \mid y^{(i)}, \widehat{\mathbf{\Theta}}^{(m)}, \widehat{\mathcal{D}})E(Z^{(i)}_l \mid y^{(i)}, \widehat{\mathbf{\Theta}}^{(m)}, \widehat{\mathcal{D}}) & \text{if } 1 \le k \ne l \le p \\ \frac{1}{n}\sum_{i=1}^{i=n} E((Z^{(i)}_k)^2 \mid y^{(i)}, \widehat{\mathbf{\Theta}}^{(m)}, \widehat{\mathcal{D}}) & \text{if } k = l. \end{array} \right.$$

The latent graphical model discussed in this chapter, though it is a natural approach, is computationally expensive for a large number of variables ($p > 2000$). We therefore describe here an alternative method to construct high–dimensional undirected graphical models.

Nonparanormal SKEPTIC

As alternative, we use the nonparanormal skeptic approach (Liu et al., 2012) to estimate the penalized concentration matrix Θ. In this approach, instead of using the transformed data to estimate precision matrix Θ, a sample correlation matrix Γ can be computed from pairwise rank correlations, such as Kendall's tau and Spearman's rho which measure the strength of association between two ranked variables. For the random vector $y_j^{(1)}, \ldots, y_j^{(n)}$ the Kendall's tau and Spearman's rho are given, respectively, by

$$\widehat{\tau}_{jl} = \frac{2}{n(n-1)} \sum_{i,\acute{i}=1}^{n} \text{sign}(y_j^{(i)} - y_j^{(\acute{i})})(y_l^{(i)} - y_l^{(\acute{i})})$$

and

$$\widehat{\rho}_{jl} = \frac{\sum_{i=1}^{n}(r_j^i - \bar{r}_j)(r_l^i - \bar{r}_l)}{\sqrt{\sum_{i=1}^{n}(r_j^i - \bar{r}_j)^2 . \sum_{i=1}^{n}(r_l^i - \bar{r}_l)^2}}$$

$$\widehat{\Gamma}_{jl} = \begin{cases} \sin(\frac{\pi}{2}\widehat{\tau}_{jl}) & j \neq l \\ 1 & j = l \end{cases} \qquad ; \qquad \widehat{\Gamma}_{jl} = \begin{cases} 2\sin(\frac{\pi}{6}\widehat{\rho}_{jl}) & j \neq l \\ 1 & j = l. \end{cases}$$

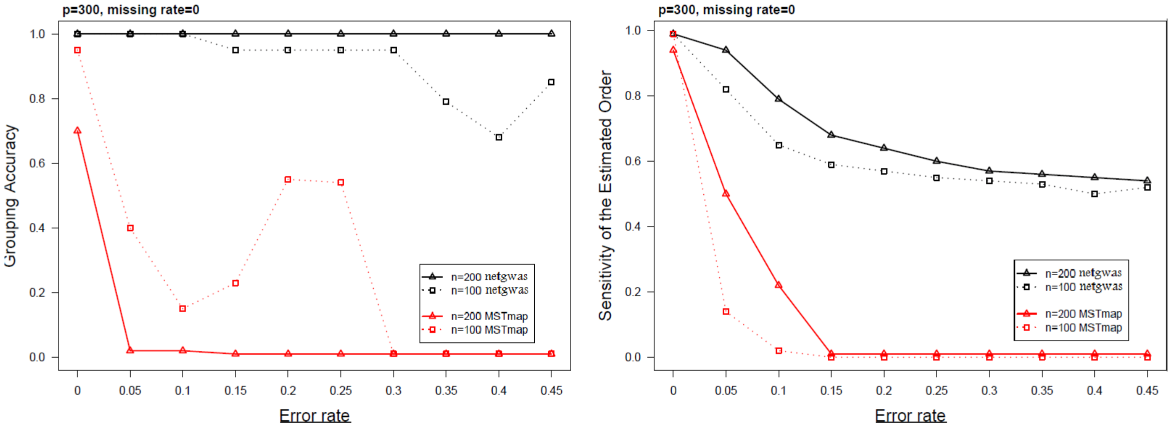
To estimate the sparse precision matrix and the graph, one can use either the graphical lasso

$$\widehat{\Theta}_{\text{glasso}} = \arg\max_{\boldsymbol{\Theta}} \left\{ \log|\boldsymbol{\Theta}| - tr(\Gamma\boldsymbol{\Theta}) - \lambda||\boldsymbol{\Theta}||_1 \right\} \tag{3.13}$$

or CLIME estimator, with $\widehat{\Gamma}$ as input

$$\widehat{\Theta}_{\text{CLIME}} = \arg\min_{\Theta} ||\Theta||_1 \qquad \textbf{subject to} \qquad ||\widehat{\Gamma}\Theta - I_p||_\infty \leq \lambda, \tag{3.14}$$

Although both methods involve convex optimization problems, these can be efficiently solved.



(a) (b)

Fig. 3.7 Genotyping errors randomly distributed over genetic markers: comparison between map construction in `netgwas` and MSTMap in presence of different choices of error rates and no missing data. (a) Reports grouping, and (b) shows ordering accuracy scores for 50 independent runs.

Simulation study

Evaluation of estimated maps in presence of genotyping errors

We studied the accuracy of the estimated linkage maps where genotyping errors are randomly distributed across the genetic markers. The simulated data contain a ratio of “bad markers", ranging from 0 up to 0.45 genotyping errors. We activated the error-detection feature in MSTMap. Figure 3.7a shows that when datasets contain genotyping errors, netgwas perfectly estimates the correct number of LGs, in particular when the sample size is sufficient, $n > 100$. In addition, the quality of the estimated linkage maps – in terms of estimating the actual number of LGs and ordering of markers – is significantly better in the netgwas than those in the MSTMap, even with activation of its error-detection feature.

Based on our simulations, we remark that with both netgwas and MSTMap erroneous markers remain in the estimated linkage map. However, netgwas orders them in the correct LG (see Figure 3.7), whereas MSTMap performs poorly in detecting LGs as well as in correctly ordering markers.

In general, for moderate numbers of individuals, when data contain genotyping errors the netgwas constructs a linkage map that is very close to the actual map in the accuracy of both the estimation and the ordering of linkage groups. This is because conditional independence is an effective way to recover relationships among genetic markers.

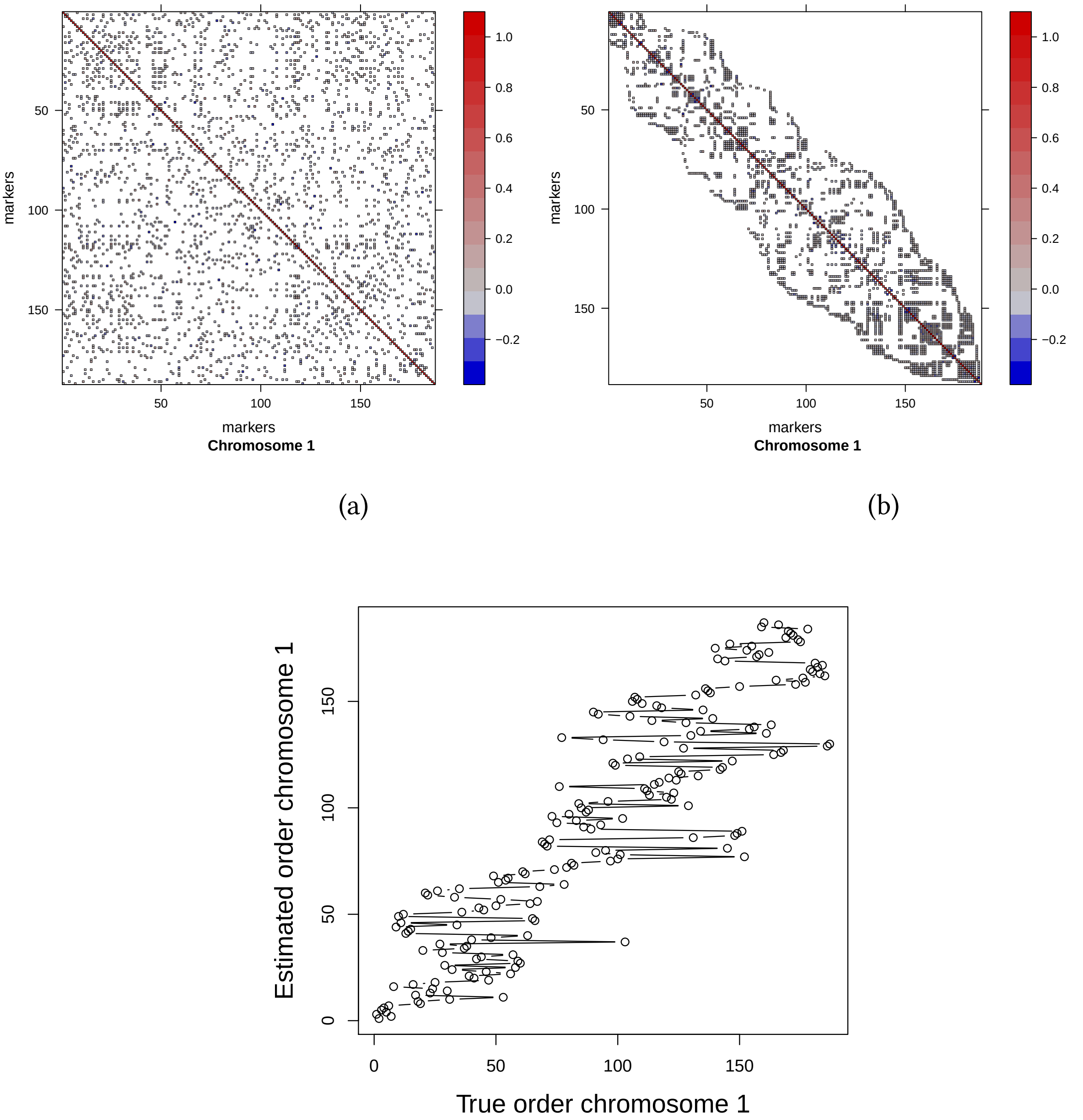
		Grouping Accuracy			Ordering Accuracy		
Missing rate	Error rate	netgwas	MSTMap	JoinMap	netgwas	MSTMap	JoinMap
p=300 & n=200							
0	0	**1.00** (0.00)	0.70 (0.12)	0.00 (0.00)	**1.00** (0.00)	0.94 (0.00)	0.00 (0.00)
0.05	0.05	**1.00** (0.00)	0.30 (0.19)	0.00 (0.00)	**0.91** (0.03)	0.77 (0.11)	0.00 (0.00)
0.10	0.10	**1.00** (0.00)	0.04 (0.03)	0.00 (0.00)	**0.73** (0.03)	0.46 (0.26)	0.00 (0.00)
0.15	0.15	**1.00** (0.01)	0.01 (0.00)	0.00 (0.00)	**0.65** (0.04)	0.00 (0.00)	0.00 (0.00)
0.20	0.20	**1.00** (0.00)	0.01 (0.00)	0.00 (0.00)	**0.59** (0.02)	0.00 (0.00)	0.00 (0.00)
0.25	0.25	**0.95** (0.16)	0.01 (0.00)	0.00 (0.00)	**0.53** (0.03)	0.00 (0.00)	0.00 (0.00)
p=500 & n=200							
0	0	**1.00** (0.00)	0.55 (0.34)	0.00 (0.00)	**1.00** (0.00)	0.90 (0.09)	0.00 (0.00)
0.05	0.05	**1.00** (0.00)	0.10 (0.07)	0.00 (0.00)	**0.77** (0.04)	0.61 (0.12)	0.00 (0.00)
0.10	0.10	**1.00** (0.00)	0.01 (0.00)	0.00 (0.00)	**0.60** (0.03)	0.18 (0.23)	0.00 (0.00)
0.15	0.15	**1.00** (0.00)	0.01 (0.00)	0.00 (0.00)	**0.56** (0.01)	0.00 (0.00)	0.00 (0.00)
0.20	0.20	**0.95** (0.16)	0.01 (0.00)	0.00 (0.00)	**0.54** (0.01)	0.00 (0.00)	0.00 (0.00)
0.25	0.25	**0.90** (0.21)	0.01 (0.00)	0.00 (0.00)	**0.51** (0.03)	0.00 (0.00)	0.00 (0.00)
p=1000 & n=200							
0	0	**1.00** (0.00)	0.61 (0.36)	0.00 (0.00)	**1.00** (0.00)	0.91 (0.06)	0.00 (0.00)
0.05	0.05	**1.00** (0.00)	0.04 (0.03)	0.00 (0.00)	**0.56** (0.00)	0.51 (0.09)	0.00 (0.00)
0.10	0.10	**1.00** (0.00)	0.44 (0.16)	0.00 (0.00)	0.52 (0.00)	**0.78**(0.02)	0.00 (0.00)
0.15	0.15	**1.00** (0.01)	0.05 (0.00)	0.00 (0.00)	**0.52** (0.00)	0.60 (0.13)	0.00 (0.00)
0.20	0.20	**0.95** (0.14)	0.01 (0.00)	0.00 (0.00)	**0.51** (0.00)	0.00 (0.00)	0.00 (0.00)
0.25	0.25	**0.95** (0.14)	0.01 (0.00)	0.00 (0.00)	**0.51** (0.00)	0.00 (0.00)	0.00 (0.00)
p=300 & n=100							
0	0	**1.00** (0.00)	0.70 (0.12)	0.00 (0.00)	**1.00** (0.00)	0.94 (0.00)	0.00 (0.00)
0.05	0.05	**0.95** (0.16)	0.82 (0.30)	0.00 (0.00)	0.76(0.06)	**0.94**(0.09)	0.00 (0.00)
0.10	0.10	**1.00** (0.00)	0.31 (0.31)	0.00 (0.00)	**0.64** (0.05)	**0.64** (0.07)	0.00 (0.00)
0.15	0.15	**0.90** (0.21)	0.02 (0.01)	0.00 (0.00)	**0.56** (0.07)	0.24 (0.18)	0.00 (0.00)
0.20	0.20	**0.40** (0.44)	0.01 (0.00)	0.00 (0.00)	**0.45** (0.10)	0.00 (0.00)	0.00 (0.00)
0.25	0.25	**0.40** (0.35)	0.01 (0.00)	0.00 (0.00)	**0.38** (0.11)	0.00 (0.00)	0.00 (0.00)
p=500 & n=100							
0	0	**1.00** (0.00)	0.80 (0.26)	0.00 (0.00)	**1.00** (0.00)	0.93 (0.05)	0.00 (0.00)
0.05	0.05	**1.00** (0.00)	0.34 (0.29)	0.00 (0.00)	0.62 (0.01)	**0.67** (0.10)	0.00 (0.00)
0.10	0.10	**1.00** (0.00)	0.08 (0.07)	0.00 (0.00)	**0.55** (0.02)	0.41 (0.08)	0.00 (0.00)
0.15	0.15	**0.87** (0.28)	0.05 (0.00)	0.00 (0.00)	0.50 (0.10)	**0.60** (0.13)	0.00 (0.00)
0.20	0.20	**0.51** (0.37)	0.01 (0.00)	0.00 (0.00)	**0.50** (0.07)	0.00 (0.00)	0.00 (0.00)
0.25	0.25	**0.21** (0.31)	0.01 (0.00)	0.00 (0.00)	**0.46** (0.16)	0.00 (0.00)	0.00 (0.00)
p=1000 & n=100							
0	0	**1.00** (0.00)	0.74 (0.35)	0.00 (0.00)	**1.00** (0.00)	0.82 (0.08)	0.00 (0.00)
0.05	0.05	**1.00** (0.00)	0.13 (0.07)	0.00 (0.00)	**0.53** (0.01)	0.50 (0.04)	0.00 (0.00)
0.10	0.10	**0.95** (0.16)	0.01 (0.00)	0.00 (0.00)	**0.52** (0.01)	0.13 (0.16)	0.00 (0.00)
0.15	0.15	**0.95** (0.15)	0.00 (0.00)	0.00 (0.00)	**0.49** (0.04)	0.00 (0.00)	0.00 (0.00)
0.20	0.20	**0.85** (0.24)	0.00 (0.00)	0.00 (0.00)	**0.46** (0.07)	0.00 (0.00)	0.00 (0.00)
0.25	0.25	**0.82** (0.30)	0.00 (0.00)	0.00 (0.00)	**0.44** (0.05)	0.00 (0.00)	0.00 (0.00)

Table 3.2 Summary of performance measures of linkage map construction in simulated F2 populations for netgwas, MSTMap and JoinMap at different rates of missingness and genotyping errors. The table presents the grouping and ordering accuracy scores for 50 independent runs and the standard deviation in parentheses. Best scores are boldfaced.

Evaluation of estimated maps for incomplete and noisy data

The ordering accuracy scores in Table 3.2 should be interpreted carefully, as inverting the order of flanking markers reduces the number of correct orderings and ultimately decreases ordering accuracy scores. In all scenarios, the `netgwas` more accurately detected linkage groups. Furthermore, this method performed well in ordering markers. Overall, except in a few cases where MSTMAP performed better, given the various ratios of noisy and incomplete data, `netgwas` estimated genetic linkage with greater accuracy than the other two methods.

Finally, we remark that determining linkage groups in JOINMAP requires the user to specify an input parameter, thereby influencing its output. However, our proposed method does not depend on any manual threshold or manual determination of the linkage groups.



(a)

(b)

(c)

Fig. 3.8 Visualizing the bandwidth reduction ordering algorithm in chromosome 1 of potato. (a) Shows estimated concentration sub−matrix for chromosome 1; (b) Represents result of performing reverse Cuthill-McKee algorithm on (a); (c) Evaluates estimated order resulting from (b) versus true order for chromosome 1 in potato.

References

Behrouzi, P. and Wit, E. (2017a). Detecting epistatic selection with partially observed genotype data using copula graphical models. *arXiv preprint arXiv:1710.00894* .

Behrouzi, P. and Wit, E. C. (2017b). netgwas: An r package for network-based genome-wide association studies. *arXiv preprint arXiv:1710.01236* .

Bradshaw, J. E. and Bonierbale, M. (2010). Potatoes. In *Root and tuber crops*, pages 1–52. Springer.

Broman, K. W., Wu, H., Sen, Ś., and Churchill, G. A. (2003). R/qtl: Qtl mapping in experimental crosses. *Bioinformatics* **19,** 889–890.

Cai, T., Liu, W., and Luo, X. (2011). A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* **106,** 594–607.

Cervantes-Flores, J. C., Yencho, G. C., Kriegner, A., Pecota, K. V., Faulk, M. A., Mwanga, R. O., and Sosinski, B. R. (2008). Development of a genetic linkage map and identification of homologous linkage groups in sweetpotato using multiple-dose aflp markers. *Molecular Breeding* **21,** 511–532.

Cistué, L., Cuesta-Marcos, A., Chao, S., Echávarri, B., Chutimanitsakun, Y., Corey, A., Filichkina, T., Garcia-Mariño, N., Romagosa, I., and Hayes, P. M. (2011). Comparative mapping of the oregon wolfe barley using doubled haploid lines derived from female and male gametes. *Theoretical and applied genetics* **122,** 1399–1410.

Cuthill, E. and McKee, J. (1969). Reducing the bandwidth of sparse symmetric matrices. In *Proceedings of the 1969 24th national conference*, pages 157–172. ACM.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9,** 432–441.

Grandke, F., Ranganathan, S., van Bers, N., de Haan, J. R., and Metzler, D. (2017). Pergola: fast and deterministic linkage mapping of polyploids. *BMC Bioinformatics* **18,** 12.

Jansen, J., De Jong, A., and Van Ooijen, J. (2001). Constructing dense genetic linkage maps. *Theoretical and Applied Genetics* **102,** 1113–1122.

Liu, H., Han, F., Yuan, M., Lafferty, J., Wasserman, L., et al. (2012). High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics* **40,** 2293–2326.

Margarido, G., Souza, A., and Garcia, A. (2007). Onemap: software for genetic mapping in outcrossing species. *Hereditas* **144,** 78–79.

Massa, A. N., Manrique-Carpintero, N. C., Coombs, J. J., Zarka, D. G., Boone, A. E., Kirk, W. W., Hackett, C. A., Bryan, G. J., and Douches, D. S. (2015). Genetic linkage mapping of economically important traits in cultivated tetraploid potato (solanum tuberosum l.). *G3: Genes| Genomes| Genetics* **5,** 2357–2364.

McLachlan, G. and Krishnan, T. (2007). *The EM algorithm and extensions*, volume 382. John Wiley & Sons.

Preedy, K. and Hackett, C. (2016). A rapid marker ordering approach for high-density genetic linkage maps in experimental autotetraploid populations using multidimensional scaling. *Theoretical and Applied Genetics* **129,** 2117–2132.

Stam, P. (1993). Construction of integrated genetic linkage maps by means of a new computer package: Join map. *The Plant Journal* **3,** 739–744.

Voorrips, R. E. and Maliepaard, C. A. (2012). The simulation of meiosis in diploid and tetraploid organisms using various genetic models. *BMC bioinformatics* **13,** 248.

Wang, H., van Eeuwijk, F. A., and Jansen, J. (2016). The potential of probabilistic graphical models in linkage map construction. *Theoretical and Applied Genetics* pages 1–12.

Wu, Y., Bhat, P. R., Close, T. J., and Lonardi, S. (2008). Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. *PLoS genetics* **4,** e1000212.

Yin, J. and Li, H. (2011). A sparse conditional gaussian graphical model for analysis of genetical genomics data. *The annals of applied statistics* **5,** 2630.