Learning high-dimensional mixed graphical models with missing values

Inma Tur and Robert Castelo
Universitat Pompeu Fabra, Barcelona Spain
{inma.tur, robert.castelo}@upf.edu

Abstract

Current biomedical instrumentation enables monitoring an increasing amount of mixed discrete and continuous variables. This results in multivariate samples amenable for their analysis using mixed graphical models. The dimension of these data sets, with a number of variables p much larger than the number of observations n, precludes the direct application of classical learning algorithms and specific procedures that work under the $p \gg n$ setting are required for that purpose. Yet, a further obstacle to the wide application of these procedures in the biomedical field arises from the fact that missing observations often occur in clinical and genotype data. The high-dimension of p impedes approaching the problem by simple complete-case analysis and increases substantially the computational burden if we want to use multiply-imputed data sets. Here we show that using limited-order correlations to learn mixed graphical models from data with $p \gg n$ enables a straightforward and effective application of complete-case analysis to the missing data problem. More importantly, because complete-case analysis is only appropriate under the restrictive assumption that data are missing completely at random, we adapt an expectation-maximization algorithm to the limited-order correlation framework and demonstrate its better suitability under the less stringent assumption of data being missing at random.

1 Introduction

Mixed graphical Markov models (GMM) are statistical models representing distributions of discrete and continuous random variables (r.v.) that share a subset of conditional independence restrictions represented by a marked undirected graph G (Lauritzen and Wermuth, 1989). These models constitute a powerful tool to analyze the increasing amount of data recorded by the ever evolving technologies in the biomedical field, which lead to multivariate data sets of p r.v. and n observations where usually $p \gg n$.

When the data set is complete, learning the structure of the graph underlying the data is a well-studied problem. In the classical framework where $n \gg p$, most relevant aspects can be found in the books of Lauritzen (1996) and Edwards (2000). In the opposite setting, where $p \gg n$, Edwards et al. (2010) restrict the learning problem to decomposable mixed GMMs whereas Tur and Castelo (2011) provide an approach without that restriction.

An important problem in the biomedical field is that missing values arise often in clinical and genotype data due to intrinsic properties of biomarkers and molecules that preclude their recording. A straightforward solution, popularly known as complete-case analysis, is to omit observations with missing values. However, the high dimension of p increases the likelihood of having at least one missing value at each observation, and therefore, too few complete multivariate samples may be left to directly apply a graphical model learning algorithm.

Imputation approaches, such as in (Broman et al., 2003) for genotype data from experimental crosses, enable also using existing learning algorithms for complete data. A major challenge when directly applying this approach is that to preserve the sampling properties of the

data, multiply-imputed data sets are required and this can substantially increase the computational cost of learning a graphical model.

In this paper we show that using limited-order correlations in the $p \gg n$ setting (Tur and Castelo, 2011) permits a straightforward and effective application of complete-case analysis. This strategy, however, is only appropriate under the restrictive assumption that data is missing completely at random (MCAR), and it leads to biased estimates of the parameters otherwise. Moreover, even when the MCAR assumption is reasonable, complete-case analysis can yield estimates of higher variability and a loss of statistical power (Little and Rubin, 2002).

For this reason, we extend the learning procedure of Tur and Castelo (2011) to employ an expectation-maximization (EM) algorithm that works under the less stringent assumption of data missing at random (MAR). We show that this provides more satisfying results in both, the MAR and MCAR scenarios.

In Sections 2 and 3 we give an overview of the necessary theory of mixed GMMs and the so-called non-rejection rate, the quantity we employ to learn these graphical models from data with $p \gg n$. In Section 4 we describe the EM algorithm in the context of the specific statistical test employed to calculate the non-rejection rate. In Section 5 we show experimental results under different missing data scenarios. Finally, a short discussion is provided in Section 6.

2 Mixed Graphical Markov Models

Mixed GMMs represent distributions involving discrete r.v. I_δ, $\delta \in \Delta$, and continuous r.v. Y_γ, $\gamma \in \Gamma$, such that an undirected marked graph $G = (V, E)$ is defined with p marked vertices $V = \Delta \cup \Gamma$ and edge set $E \subseteq V \times V$. We assume that the joint distribution of the variables $X = (I, Y)$ is conditional Gaussian (CG-distribution), that is, $Y \sim \mathcal{N}_{|\Gamma|}(\mu(i), \Sigma(i))$ for each joint discrete level $i \in \mathcal{I}$. Considering that discrete genotypes affect mean gene expression values only, we restrict the learning problem to homogeneous mixed GMMs where $\Sigma(i) \equiv \Sigma$ is constant through $i \in \mathcal{I}$.

2.1 Decomposable Mixed GMMs

An important subclass of mixed GMMs is defined by decomposable marked graphs.

Definition 1. A triple (A, B, C) of disjoint subsets of V forms a decomposition of an undirected marked graph G if $V = A \cup B \cup C$ and the following three conditions hold: 1. C is a complete subset of V; 2. C separates A from B; and 3. $C \subseteq \Delta$ or $B \subseteq \Gamma$.

Thus, an undirected marked graph G is decomposable if there exists a proper decomposition (A, B, C) such that the subgraphs $G_{A\cup C}$ and $G_{B\cup C}$ are decomposable. When this holds, we say that $C_1 = \{A \cup C\}$ and $C_2 = \{B \cup C\}$ are cliques and, $S = \{C\}$ is a separator of G.

2.2 Maximum Likelihood Estimates of Mixed GMMs

Let $\mathcal{X} = \{x^{(\nu)}\} = \{(i^{(\nu)}, y^{(\nu)})\}$ be a sample of $\nu = 1, \ldots, n$ i.i.d. observations from a homogeneous CG-distribution. For an arbitrary subset $A \subseteq V$, we abbreviate to $i_A = i_{A\cap\Delta}$, $\mathcal{I}_A = \mathcal{I}_{A\cap\Delta}$ and $y_A = y_{A\cap\Gamma}$ and the following sampling statistics are defined:

$$n(i) = \#\left\{\nu : i^{(\nu)} = i\right\}, \tag{1}$$

$$s(i) = \sum_{\nu:i^{(\nu)}=i} y^{(\nu)}, \tag{2}$$

$$\bar{y}(i) = s(i)/n(i), \tag{3}$$

$$ss(i) = \sum_{\nu:i^{(\nu)}=i} y^{(\nu)}(y^{(\nu)})^T, \tag{4}$$

$$ssd(i) = ss(i) - s(i)s(i)^T/n(i), \tag{5}$$

$$ssd(A) = \sum_{i_A\in\mathcal{I}_A} ssd(i_A), \tag{6}$$

$$ssd_A(A) = \sum_{i_A\in\mathcal{I}_A} ssd_{A\cap\Gamma}(i_A), \tag{7}$$

$$ssd = ssd(V) \text{ and } ssd_A = ssd_A(V). \tag{8}$$

Lauritzen (1996, Prop. 6.10) shows that if $n \geq |\Gamma| + |\mathcal{I}|$ then the likelihood function attains its maximum almost surely if and only if $n(i) > 0$ for all $i \in \mathcal{I}$. Then, the maximum likelihood estimate (MLE) is given as having moment characteristics equal to the empirical moments, i.e.,

$$\hat{p}(i) = n(i)/n,\ \hat{\mu}(i) = \bar{y}(i),\ \hat{\Sigma} = ssd/n. \tag{9}$$

Analogously, decomposable mixed GMMs admit explicit MLEs: in the homogeneous case, (Lauritzen, 1996, Prop. 6.21) shows that the MLE exists almost surely if and only if $n(i_C) \geq |C \cap \Gamma| + |\mathcal{I}_C|$ for all cliques C of G and $i_C \in \mathcal{I}_C$. In that case, it is given with the following canonical parameters (Lauritzen, 1996, pg. 189):

$$\hat{p}(i) = \prod_{j=1}^{k} \frac{n(i_{C_j})}{n(i_{S_j})}, \tag{10}$$

$$\hat{h}(i) = n\left\{\sum_{j=1}^{k} \left[ssd_{C_j}(C_j)^{-1}\bar{y}_{C_j}(i_{C_j})\right]^{|\Gamma|} - \left[ssd_{S_j}(S_j)^{-1}\bar{y}_{S_j}(i_{S_j})\right]^{|\Gamma|}\right\}, \tag{11}$$

$$\hat{K} = n\left\{\sum_{j=1}^{k}\left[ssd_{C_j}(C_j)^{-1}\right]^{|\Gamma|} - \left[ssd_{S_j}(S_j)^{-1}\right]^{|\Gamma|}\right\}, \tag{12}$$

where $S_1 = \emptyset$ and in Eq.(12) $[M]^{|\Gamma|}$ is a $|\Gamma| \times |\Gamma|$ matrix obtained from a $|A| \times |A|$ matrix $M = \{m_{\gamma\eta}\}_{|A|\times|A|}$ with $A \subseteq \Gamma$ such that $[M]^{|\Gamma|}_{\gamma\eta} = m_{\gamma\eta}$ if $\gamma, \eta \in A$ and $[M]^{|\Gamma|}_{\gamma\eta} = 0$ otherwise. Analogously, in Eq. (11) $[M]^{|\Gamma|}$ is a $|\Gamma|$-length vector obtained from a $|A|$-length vector M.

3 The Non-rejection Rate

In order to address the problem of learning mixed GMMs from data with $p \gg n$ and missing values we will use the limited-order correlation approach introduced by Tur and Castelo (2011) and based on the *non-rejection rate* (Castelo and Roverato, 2006). This approach requires testing conditional independences $X_\alpha \perp\!\!\!\perp X_\beta | X_Q$ using a likelihood ratio test between two decomposable models.

Without loss of generality, assume that $V = \{\alpha, \beta, Q\}$ with $V = \Delta \cup \Gamma$. Let (γ, η) denote a pair of continuous variables $(\gamma, \eta \in \Gamma)$, and (δ, γ) a pair of mixed variables $(\delta \in \Delta, \gamma \in \Gamma)$, so that either $Q = V\backslash\{\gamma, \eta\}$ or $Q = V\backslash\{\delta, \gamma\}$ as the conditioning subset. In the pure continuous case, the likelihood ratio statistic raised to the power $2/n$ corresponding to $\gamma \perp\!\!\!\perp \eta | Q$ is (Lauritzen, 1996, pg. 192):

$$\Lambda_{\gamma\eta.Q} = \frac{|ssd_\Gamma||ssd_{\Gamma\backslash\{\gamma,\eta\}}|}{|ssd_{\Gamma\backslash\{\gamma\}}||ssd_{\Gamma\backslash\{\eta\}}|}. \tag{13}$$

In the mixed case, the likelihood ratio statistic raised to the power $2/n$ corresponding to $\delta \perp\!\!\!\perp \gamma | Q$ is (Lauritzen, 1996, pg. 194):

$$\Lambda_{\delta\gamma.Q} = \frac{|ssd_\Gamma||ssd_{\Gamma^*}(\Delta^*)|}{|ssd_{\Gamma^*}||ssd_\Gamma(\Delta^*)|}, \tag{14}$$

where $\Gamma^* = \Gamma\backslash\{\gamma\}$ and $\Delta^* = \Delta\backslash\{\delta\}$. Tur and Castelo (2011) show that for decomposable mixed GMMs, likelihood ratios in (13) and (14) follow exactly the beta distributions:

$$\Lambda_{\gamma\eta.Q} \sim \mathcal{B}\left(\frac{n - |\Gamma| - |\mathcal{I}| + 1}{2}, \frac{1}{2}\right), \text{ and}$$

$$\Lambda_{\delta\gamma.Q} \sim \mathcal{B}\left(\frac{n - |\Gamma| - |\mathcal{I}| + 1}{2}, \frac{|\mathcal{I}_{\Delta^*}|(|\mathcal{I}_\delta| - 1)}{2}\right), \tag{15}$$

respectively. This provides a proper control of the Type-I error when $q = |Q|$ grows and allows one to define a quantity called the non-rejection rate (NRR) for mixed GMMs (Tur and Castelo, 2011) corresponding to a linear measure of association over all marginal distributions of size $q < (n-2)$. The NRR is calculated for every pair of variables and is based on the outcome of a moderate number of these conditional independence tests for different Q subsets of size q sampled uniformly at random; see (Castelo and Roverato, 2006) for further details.

Using marginal distributions of size $(q+2) < n$ facilitates the use of complete-case analysis by replacing $\mathcal{I}_A$ in Eq. (7) by $\mathcal{I}_A^{obs}$ such that $\mathcal{I}_A^{obs} \subseteq \mathcal{I}_A$ contains only the combined levels from $A \cap \Delta$ that are fully observed. Analogously, for missing continuous values, we replace $y^{(\nu)}$ by their observed components $y^{(\nu)}_{obs}$ in Eqs. (2), (3) and, (4).

4 An EM algorithm for the Exact Conditional Independence Test

The expectation-maximization (EM) algorithm (Dempster et al., 1977) is a method to find the MLEs in statistical models when these models depend on unobserved latent variables. In the context of GMMs with $n \gg p$, Didelez and Pigeot (1998) use the EM-algorithm to provide MLEs of a mixed GMM

with missing values under the MAR assumption. Later, Geng et al. (2000) developed a more efficient EM algorithm (PIEM) for decomposable mixed GMMs.

Here we deal with observations that follow a homogeneous CG-distribution where discrete and/or continuous r.v. are allowed to have missing values. We denote by $x_{obs} = (i_{obs}, y_{obs})$ the non-missing components of an observation from such distribution.

Our learning approach is based on calculating the NRR for each pair of variables (α, β). This requires performing conditional independence tests by computing a likelihood ratio between a saturated and a constrained decomposable model. Such a ratio only involves the *ssd* matrices derived from the decomposition of both models; see Eqs. (13), (14). Therefore, our procedure applies the EM algorithm to the saturated model to obtain the *ssd* matrix estimate corresponding to the r.v. in $\{\alpha, \beta, Q\}$.

Under the constrained model, however, if at each observation where β is non-missing, the values in the separator are also observed, Geng et al. (2000) show that the decomposition $\{\alpha, \beta, Q\}$ is lossless. In that case, we can apply the EM algorithm separately to the clique $C_1 = \{\alpha, Q\}$ using all the observations of the data set, denoted by $\mathcal{X}$, and to the clique $C_2 = \{\beta, Q\}$ and the separator $S = \{Q\}$ using only those observations where β is observed, denoted by $\mathcal{X}^\beta$, to get the corresponding *ssd* matrix estimates. If the constrained model cannot be decomposed losslessly, then we first compute the expectation of the sufficient statistics in the separator S using only those observations required to obtain the lossless decomposition. The detailed algorithm is as follows.

Initialization: First, the moment parameters of the model $\{p_0(i), \mu_0(i), \Sigma_0\}$ are initialized. We use $p_0(i) = |\mathcal{I}|/n$ for each $i \in \mathcal{I}$, $\mu_0(i) = 0$ for each $\gamma \in \Gamma$ and, $\Sigma_0 = I_{|\Gamma|}$ (identity matrix). However, in our experience, this particular algorithm is not very sensitive to other choices of the initial parameters.

E-step: The E-step computes the expectation of the sufficient statistics (1), (2) and, (4) given the observed data and the current estimation of the parameters for each model. A lossless decomposition is obtained using the strategy described before and we perform the E-step to the clique C_2 using the observations $\mathcal{X}^\beta$ and to the C_1 using all the observations $\mathcal{X}$:

$$\begin{aligned} E\{n(i_d)|x_{obs}\} &= \sum_{\nu=1}^{n} pr(I_d = i_d|x_{obs}^{(\nu)}) = \\ &\sum_{\nu=1}^{n} \sum_{i' \in \mathcal{I}: i'_d = i_d} pr(I = i'|x_{obs}^{(\nu)}) \end{aligned}$$

where

$$pr(I = i'|x_{obs}^{(\nu)}) = \frac{\exp k(i')}{\sum_{s \in \mathcal{S}} \exp k(s)} \qquad (16)$$

and

$$\begin{aligned} k(i) &= y_{obs}^T \Sigma_{\{obs,obs\}}^{-1} \mu(i)_{obs} - \frac{1}{2}[y_{obs}^T \Sigma_{\{obs,obs\}}^{-1} y + \\ &\mu(i)_{obs}^T \Sigma_{\{obs,obs\}}^{-1} \mu_{obs}(i)] + \log p(i). \end{aligned}$$

The set $\mathcal{S} = \{(i_{obs}, i_{mis})|i_{mis} \in \mathcal{I}_{mis}\}$ in (16) is the set of all combinations of discrete levels given the observed ones.

$$E\{s(i_d)_\gamma|x_{obs}\} = \sum_{\nu=1}^{n} pr(I_d = i_d|x_{obs}^{(\nu)}) E(Y_\gamma|y_{obs}^{(\nu)}, i_d)$$

$$E\{ss(i_d)_\gamma|x_{obs}\} = \sum_{\nu=1}^{n} pr(I_d = i_d|x_{obs}^{(\nu)}) \times \left[E(Y_\gamma|y_{obs}^{(\nu)}, i_d)^2 + c_{\gamma\gamma}\right]$$

$$E\{ss(i_d)_{\gamma,\eta}|x_{obs}\} = \sum_{\nu=1}^{n} pr(I_d = i_d|x_{obs}^{(\nu)}) \times \left\{E(Y_\gamma|y_{obs}^{(\nu)}, i_d) E(Y_\eta|y_{obs}^{(\nu)}, i_d) + c_{\gamma\eta}\right\}$$

where

$$E(Y_\gamma|y_{obs}^{(\nu)}, i_d) = \mu(i)_\gamma - \Sigma_{\{\gamma,obs\}} \Sigma_{\{obs,obs\}}^{-1} \{y_{obs} - \mu(i)_{obs}\}$$

and

$$c_{\gamma\eta} = cov(Y_{\gamma,\eta}|y_{obs}, i) = \Sigma_{\{(\gamma,\eta),(\gamma,\eta)\}} - \Sigma_{\{(\gamma,\eta),obs\}} \Sigma_{\{obs,obs\}}^{-1} \Sigma_{\{obs,(\gamma,\eta)\}}$$

M-step: This step updates the new estimates maximizing the conditional expectations of the sufficient statistics found in the E-step. For the saturated model we compute the parameters of Eq. (9) whereas for the constrained model we only have to compute Eqs. (10), (11) and, (12).

Convergence criteria: For the constrained model, we first have to transform the canonical parameters to moment parameters:

$$\hat{\mu}(i) = \hat{K}^{-1}\hat{h}(i) \quad \text{and} \quad \hat{\Sigma} = \hat{K}^{-1}.$$

Then, in order to check the convergence of each model, the E-step and M-step are iterated until the maximum of

$$\left\{ \frac{|\mathcal{D}\left(\hat{n}(i)\right)|}{\sqrt{(\hat{n}(i)+1)}}, \frac{|\mathcal{D}\left(\hat{\mu}(i)_\gamma\right)|}{\sqrt{\hat{\sigma}_{\gamma\gamma}}}, \frac{|\mathcal{D}\left(\hat{\sigma}_{\gamma\eta}\right)|}{\sqrt{\hat{\sigma}_{\gamma\gamma}\hat{\sigma}_{\eta\eta} + (\hat{\sigma}_{\gamma\eta})^2}} \right\}$$

for $i \in \mathcal{I}$ and $\gamma, \eta \in \Gamma$ is smaller than a tolerance value ϵ, for instance $\epsilon = 0.01$. In the numerators, $\mathcal{D}$ denotes the difference operator between moment parameters of two consecutive iterations.

Once the convergence criterion is achieved, what we are interested in, is the last update of the *ssd* matrices we calculate at each iteration to perform the M-step.

5 Results

In this section we show experimental results focused on the case of mixed interactions with discrete missing values. This is, in fact, the more common situation when inferring associations between continuous gene expression profiles and discrete genotypes.

5.1 Synthetic Homogeneous Mixed GMM data

We build synthetic homogeneous mixed GMMs as it is done in (Tur and Castelo, 2011). First, in order to bound the size of any minimal subset separating every pair of vertices we sample an undirected d-regular graph (Harary, 1969) structure $G = (V, E)$ with $V = \Delta \cup \Gamma$ and excluding edges $(i, j) \in \Delta \times \Delta$. Then, we generate random parameters according to the conditional independences encoded in G and rendering all pairs $(i, j) \in \Delta \times \Delta$ marginally independent.

In fact, we generate two sets of parameters: one in which the interactions between a discrete and a continuous r.v. (mixed linear interactions) are strong and another in which these interactions are weak. To generate the first model, joint levels of discrete r.v. are assigned with a uniform distribution, a nominal mean Pearson correlation $\rho = 0.5$ is employed to generate the random covariance matrix Σ, and a standard deviation value $\sigma = 3$ is set for sampling mixed linear interaction parameters $h(i)$ with a strong effect such that $\mu(i) = \Sigma \cdot h(i)$. To build the model with weak mixed linear interactions, joint levels of discrete r.v. and Σ are generated as in the previous case but now the parameter $h(i)$ is generated using a smaller standard deviation $\sigma = 1.5$. Finally, from each set of parameters we generate n observations to build the corresponding data set $\mathcal{X} = \{x^{(\nu)}\} = \{(i^{(\nu)}, y^{(\nu)})\}$, $\nu = 1, \ldots, n$, by first sampling, for every observation, a joint level $i \in \mathcal{I}$ according to their (uniform) probability distribution, and, secondly, by sampling a multivariate normal observation from $\mathcal{N}_{|\Gamma|}(\mu(i), \Sigma)$ for the continuous r.v. Y.

5.2 Synthetic Missing data

Missing values are generated under two different assumptions. We simulate, in one hand, data under the MCAR mechanism by which the occurrence of a missing value does not depend on other observed or unobserved values. On the other hand, we simulate missing data under the less stringent, and more realistic, MAR assumption. In this case, whether a value is missing or not depends on other observed values but not on unobserved ones; see (Little and Rubin, 2002).

Given a complete data set $\mathcal{X}$ sampled as described in the previous subsection, we create a missing indicator (MI) variable M_δ for each $\delta \in \Delta$. When $M_{\delta k} = \text{TRUE}$, then we remove the discrete value $i^{(k)}$ from variable δ in observation k of data set $\mathcal{X}$. Values for MI variables are sampled using the following logistic model (Greenland and Finkle, 1995):

$$\Pr\left(M_\delta = \text{TRUE} | Y = y\right) = expit(a + by) \quad (17)$$

where $expit(x) = exp(x)/\left(1 + exp(x)\right)$. The in-

tercept is set to $a \sim \mathcal{N}(\Phi^{-1}(\tau), 1)$, where Φ is the cumulative distribution function of a standard normal distribution, so that the number of missing values for each discrete r.v. equals τn where τ is the desired fraction of missing values. In this paper we consider a maximum $\tau = 0.2$ as genotypes with higher missing rates are generally discarded (The International HapMap Consortium, 2007). The coefficient b determines the dependence of M_δ on Y such that with $b = \ln(1)$ we obtain missing values under the MCAR assumption while $b = \ln(3)$ produces a strong MAR effect with a 3-fold increase in the odds of missing data for each unit increase in Y.

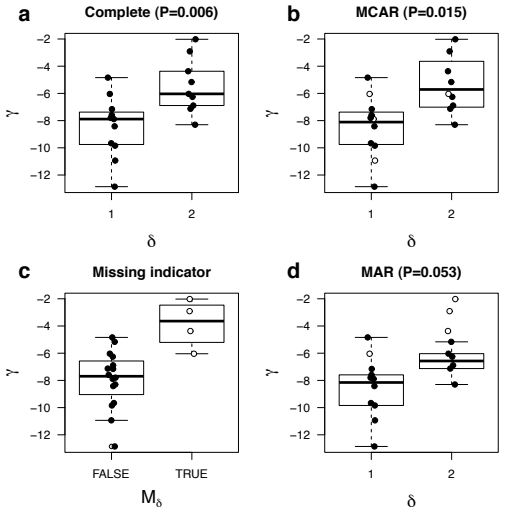

Figure 1: (a) Complete data without missing values. (b) MCAR data. (c) M_δ depends on γ with $b = \ln(2)$. (d) MAR data using M_δ from (c). White dots correspond to observations with missing discrete values while black dots to non-missing observations. Boxplots in (a, b, d) are calculated from non-missing observations (black dots). P indicates p-value for $H_0 : \delta \perp\!\!\!\perp \gamma | \emptyset$.

We have simulated $n = 20$ observations from a correlated pair of discrete and continuous r.v. (δ, γ) obtaining the association shown in Figure 1a, with a fraction $\tau = 0.2$ of missing values under the MCAR (b) and MAR (c, d) mechanisms. Figure 1 clearly shows how the MAR effect in (c) eliminates the association between δ and γ in (d) at a significance level $\alpha = 0.05$.

5.3 Analysis of Statistical Power

In this section we assess how the MCAR and MAR assumptions affect the statistical power to detect strong and weak mixed linear associations using complete-case analysis and the EM-algorithm in the exact conditional independence test.

For this purpose, we simulate 2 mixed GMMs where $|\Delta| = |\Gamma| = 1$ and δ and γ are connected so that the association between them in one model is strong and in the other, the association is weak. We sample 10,000 data sets $\{\mathcal{X}^s, \mathcal{X}^w\}$ of $n = 100$ observations from the strong and the weak model. From each previous data set, we simulate a MCAR data set and a MAR data set with strong effect $b = \ln(3)$ according to missing values rates $\tau = \{0.2, 0.15, 0.1, 0.05\}$. We perform conditional independence tests between δ and γ from the original complete data sets and from the data sets with missing values using the EM-algorithm and complete-case analysis. Finally, we calculate the statistical power as $1 - \beta$ where β is the type-II error using a theoretical quantile $\alpha = 0.05$.

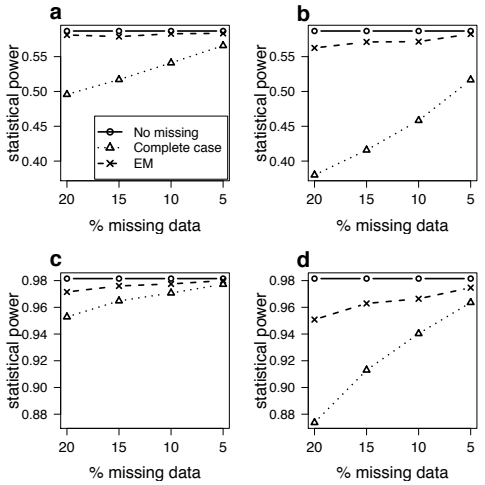

Figure 2: Statistical power to detect a mixed linear association (δ, γ). (a-b) Weak association, (c-d) strong association; (a-c) obtained under MCAR and (b-d) under a MAR effect.

In Figure 2 we assess the statistical power in each of the previously described scenarios. We

observe that EM algorithm yields higher power in front of weak and strong associations, than complete-case analysis even when missing data is sampled according to the MCAR mechanism.

5.4 Accuracy of Likelihood Ratio Estimates

We want to assess the accuracy of the estimation of the likelihood ratio when it is calculated with both, the EM-algorithm and complete-case analysis.

To that end, we simulate 1,000 data sets of $n = 30$ observations with strong mixed linear associations $\mathcal{X}^s$ from a common graph G with $d = 3$, $|\Delta| = 3$, and $|\Gamma| = 47$. For each data set, we select a pair of mixed discrete and continuous r.v. (δ, γ) uniformly at random between present and absent edges. We remove 20% of values of δ under the MCAR ($b = \ln(1)$) and MAR ($b = \ln(3)$) mechanisms.

For each complete and missing data sets we calculate the likelihood ratio in Eq. (14) for $\delta \perp\!\!\!\perp \gamma | Q$ where Q is formed by variables indexed by vertices adjacent to γ in G.

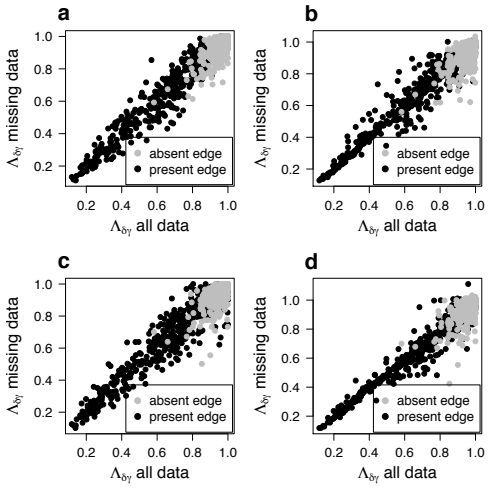

Figure 3: Comparison of likelihood ratio from a complete and a missing data set. (a-b) MCAR, (c-d) MAR; (a-c) complete-case, (b-d) EM-algorithm.

In Figure 3 we observe that estimates of likelihood ratios for present edges with complete-case analysis are significantly more variable than those obtained with the EM-algorithm under both, MCAR and MAR assumptions (one-sided F-test of equality of variances, p-value < 0.05).

5.5 Analysis of Non-rejection Rate

To conclude this section, we want to assess the performance of using different strategies to learn a mixed GMM from data with $p \gg n$ and missing values by means of estimating the NRR. We sample a single data set with $n = 40$ observations from a d-regular graph with strong mixed linear interactions on $p = 400$ and $d = 20$, and where $|\Delta| = 10$, $|\Gamma| = 390$. We estimate NRR values for each pair of mixed vertices using conditioning sets of size $q = 21$.

In this case, for each discrete variable $\delta \in \Delta$ we generate M_δ according to the following model:

$$\Pr\left(M_\delta = \text{TRUE} | Y = (y_1, \ldots, y_{|\Gamma|}\right) = expit\left(a + b \sum_{i=1}^{|\Gamma|} \frac{y}{|\Gamma|}\right)$$

with $a \sim \mathcal{N}(\Phi^{-1}(0.2), 1)$ and $b = \ln(3)$.

We estimate NRR values using complete-case analyis and the EM-algorithm with $\epsilon = 0.001$. Moreover, we also use the multiple imputation method of Broman et al. (2003), implemented through the `sim.geno()` function from the `R/qtl` package, that uses a hidden Markov model to impute missing genotypes in experimental crosses. We use it here to create 10 imputed data sets from the simulated data set with missing values. Then, the NRR is calculated from each of these imputed data sets and averaged at each pair of vertices.

In Figure 4, we show precision-recall curves using the NRR values from all possible mixed pairs (δ, γ) with $\delta \in \Delta$ and $\gamma \in \Gamma$ for each method. We can observe that NRRs calculated from multiply-imputed data sets perform worse than using complete-case analysis and the EM algorithm. However, between these two latter approaches, even though EM is slightly better, there are no substantial differences in performance, as opposed to what we previously observed when we carefully assessed statistical power and accuracy.

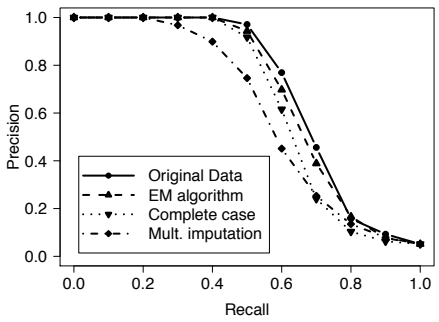

Figure 4: Precision recall curves from NRR values estimated with the original and missing data. In the latter case, the EM algorithm, complete-case analysis and multiple-imputation methods were used.

6 Discussion

In this paper we have introduced a statistical procedure to learn mixed GMMs from data with $p \gg n$ and missing values, based on complete-case analysis and the EM algorithm. Our experimental results show that, for the purpose of a conditional independence test, the EM-algorithm yields more accurate estimates of the likelihood ratio for the presence of a mixed interaction (Fig. 3), and higher statistical power to detect it (Fig. 2), compared to complete-case analysis under both, the MCAR and MAR assumptions for missing data.

Intringuinly, both methods perform similarly when using the NRR to learn the structure of a mixed GMM (Fig. 4) and more research will be necessary to elucidate the circumstances under which these methods perform differently in the context of structure learning with missing data.

Availability: The described algorithms are implemented in the `qpCItest()` function from the Bioconductor project package `qpgraph`.

Acknowledgments

This work is supported by a project grant [TIN2011-22826] and part of it was developed while the first author, supported by a FPI predoctoral fellowship [BES-2009-024901], was visiting the Dept. of Statistics at Oxford University, funded by a short-term visit fellowship [EEBB-2011-43932], all funded by the Spanish Ministerio de Economía y Competitividad.

References

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

R. Castelo and A. Roverato. 2006. A robust procedure for gaussian graphical model search from microarray data with p larger than n. *J Mach Learn Res*, 7:2621–2650.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc B*, 39(1):1–38.

V. Didelez and I. Pigeot. 1998. Maximum likelihood estimation in graphical models with missing values. *Biometrika*, 85(4):960–966.

D. Edwards, G. de Abreu, and R. Labouriau. 2010. Selecting high-dimensional mixed graphical models using minimal AIC or BIC forests. *BMC Bioinformatics*, 11(1):18.

D. Edwards. 2000. *Introduction to graphical modelling*. Springer.

Z. Geng, K. Wan, and F. Tao. 2000. Mixed graphical models with missing data and the partial imputation EM algorithm. *Scand J Stat*, 27(3):433–444.

S. Greenland and W. D. Finkle. 1995. A critical look at methods for handling missing covariates in epidemiologic regression analyses. *Am J Epidemiol*, 142(12):1255–1264.

F. Harary. 1969. *Graph theory*. Addison-Wesley.

S. Lauritzen and N. Wermuth. 1989. Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann Stat*, 17(1):31–57.

S. Lauritzen. 1996. *Graphical Models*. Oxford University Press.

R. J. A. Little and D. B. Rubin. 2002. *Statistical Analysis With Missing Data*. Probability and Statistics. Wiley, second edition.

The International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449:851–861.

I. Tur and R. Castelo. 2011. Learning mixed graphical models from data with p larger than n. In *Proc. 27th Conf. Uncertainty in Artificial Intelligence (UAI)*, pages 689–697, Barcelona, Spain.