Journal of Machine Learning Research 23 (2022) 1-55 Submitted 7/19; Published 2/22

The AIM and EM Algorithms for Learning from Coarse Data

Manfred Jaeger JAEGER@CS.AAU.DK
Department for Computer Science
Aalborg University
Selma Lagerlöfs Vej 300, 9220 Aalborg, Denmark

Editor: XuanLong Nguyen

Abstract

Statistical learning from incomplete data is typically performed under an assumption of ignorability for the mechanism that causes missing values. Notably, the expectation maximization (EM) algorithm is based on the assumption that values are missing at random. Most approaches that tackle non-ignorable mechanisms are based on specific modeling assumptions for these mechanisms. The adaptive imputation and maximization (AIM) algorithm has been introduced in earlier work as a general paradigm for learning from incomplete data without any assumptions on the process that causes observations to be incomplete. In this paper we give a thorough analysis of the theoretical properties of the AIM algorithm, and its relationship with EM. We identify conditions under which EM and AIM are in fact equivalent, and show that when these conditions are not met, then AIM can produce consistent estimates in non-ignorable incomplete data scenarios where EM becomes inconsistent. Convergence results for AIM are obtained that closely mirror the available convergence guarantees for EM. We develop the general theory of the AIM algorithm for discrete data settings, and then develop a general discretization approach that allows to apply the method also to incomplete continuous data. We demonstrate the practical usability of the AIM algorithm by prototype implementations for parameter learning from continuous Gaussian data, and from discrete Bayesian network data. Extensive experiments show that the theoretical differences between AIM and EM can be observed in practice, and that a combination of the two methods leads to robust performance for both ignorable and non-ignorable mechanisms.

Keywords: incomplete data, missing at random, coarsened at random, expectation maximization, Bayesian networks

1. Introduction

1.1 Learning from Coarse Data

Learning from incomplete data is a fundamental problem for machine learning. The most prevalent form of incomplete data is data with missing values. However, data can be incomplete also in other ways: for example, numeric data values may be given in binned form, where instead of a precise value x, only an interval $[a, b]$ that x falls into is recorded. A special case of this is *right-censored* data, where values exceeding a certain threshold t are only reported as "$> t$". Similarly, grouped observations of categorical attributes can also occur, for example when in a personal data form the field 'citizenship' is filled with 'European' rather than a specific nationality. In the particular case where a class label

is subject to such set-valued observations, this situation has been considered under the names of *learning from partially labeled* data (Jin and Ghahramani, 2003; Cour et al., 2011) and *superset learning* (Hüllermeier and Cheng, 2015). For finite sample spaces W, the *coarse data model* (Heitjan and Rubin, 1991; Couso et al., 2017) provides a simple and fully general framework for dealing with all forms of incompleteness. According to this model, an incomplete ("coarse") observation can be given by any subset of W. For example, if W is the set of all countries, then a coarse observation 'European' is the subset of European countries. A personal data record (*nationality*=Belgian,*gender*=?,*education*=university) with a missing value for the *gender* attribute corresponds to the set of complete data records {(Belgian, male, university),(Belgian, female, university),(Belgian, non-binary, university)}.

1.2 Imputation: a Brief Survey

Almost all methods that have been developed to deal with incomplete data can be understood as variations on the theme of *imputation*: the incomplete data is turned into a complete data set, and then learning is performed based on this (hypothetical) complete data. In many approaches it is assumed that incompleteness only occurs in the form of missing values.

In the simplest type of imputation procedure, data item with missing values is turned into a complete item by filling in the missing values. This can be done by using average or mode values as default fill-ins, or in the manner of more sophisticated *hot-deck* imputations, where the imputed values are chosen based on the observed values in similar ("donor") data instances (Andridge and Little, 2010). Broadly speaking, many versions of hot-deck imputation can be understood as filling in missing values by values obtained from nearest-neighbor predictions. Other classification or regression methods can also be used for imputing missing values. An empirical comparison of a variety of such methods is given by Jerez et al. (2010). Single imputation methods are computationally simple, and quite popular in practice, because they result in a complete data set to which all learning methods and analysis techniques can be applied. However, theoretical justifications for this approach can only be given under strong modeling assumptions for the incomplete data, and for restricted types of statistical inferences on the imputed completion (Andridge and Little, 2010).

In *multiple imputation* (Rubin, 1978, 1996) several completions are constructed by sampling missing values from the conditional distribution of the unobserved variables, given the observed values, and a prior Bayesian model for both the data and the missing-data mechanism (Rubin, 1996). Inferences about quantities of interest are obtained by averaging estimates obtained separately from each imputed complete data set. The variance in the sample of estimates can further be used as a basis to analyze the sensitivity of the inferences to the missing values, or the choice of the prior model. Furthermore, point estimates can be replaced by interval estimates reflecting the uncertainty due to incompleteness. The theoretical justification of multiple imputation essentially rests on the assumption that the prior model is correct (however, an analysis in more frequentist terms can also be given (Rubin, 1996)).

Several authors suggest that in the absence of well-justified assumptions on the missingness mechanism one should consider all possible data completions, and only determine

interval-valued estimates that represent all the estimates that would be obtained from any of the completions. We refer to this *all imputations* approach as *conservative inference* (Horowitz and Manski, 2001). Conservative inference approaches have received some attention for parameter learning in Bayesian networks, where it also has been suggested to refine the conservative interval estimates to point estimates by maximizing entropy (Cowell, 1999), or by taking convex combinations of extremal solutions (Ramoni and Sebastiani, 1998). A robust version of the Naive Bayesian classifier based on conservative interval estimates is developed by Corani and Zaffalon (2008). A related type of approach that also leads to point estimates is to maximize lower or upper bounds on the likelihood functions that are induced by different data completions (Hüllermeier, 2014; Guillaume and Dubois, 2015).

All approaches mentioned so far can be described in terms of finite collections of imputed data completions. Going one step further, one can consider probability distributions over possible completions. We will refer to such probability distributions as *fractional completions*. They are the basis of the EM algorithm (Dempster et al., 1977), where in the expectation step one (implicitly) considers the expected distribution over completions, given a current complete data model.

1.3 Missing and Coarsened at Random

As is well known, the EM algorithm is based on the assumption that the missingness mechanism is *ignorable* in the sense that values are *missing at random* (*MAR*) (Rubin, 1976). Exact definitions of *MAR* that are found in the literature may exhibit some subtle differences (Jaeger, 2005a; Seaman et al., 2013; Mealli and Rubin, 2015). A rough (but not fully accurate) intuition is that data is *MAR*, if whether or not an attribute value is missing does not depend on the actual value. For the more general coarse data model, the *MAR* assumption has been generalized to the *coarsened at random (*CAR*)* assumption (Heitjan and Rubin, 1991). In spite of its greater generality, the *CAR* assumption is actually conceptually simpler and more transparent than the original *MAR* assumption. Roughly speaking, data is *CAR*, if the coarse observation of the subset $U \subseteq W$ is equally likely to happen for each of the possible underlying complete data points $w \in U$. An in-depth analysis and characterization of different versions of *MAR* and *CAR* is given by Jaeger (2005a).

The *MAR* or *CAR* assumptions are quite restrictive, and notoriously difficult to validate (Cator, 2004; Manski, 2005; Jaeger, 2006a; Mohan and Pearl, 2014). When these assumptions appear unrealistic, one would prefer methods that do not rely on them. One such approach is to include in the data analysis an explicit model for the mechanism that causes values to be missing (Little and Rubin, 1987, Chapter 11). This approach has been rigorously pursued by Mohan et al. (Mohan et al., 2013; Van den Broeck et al., 2015; Mohan et al., 2018), who use graphical models to represent the joint distribution of the underlying complete and the observed incomplete data. Using a graph-based concept of *MAR* that is somewhat stronger and simpler than the original one by Rubin (1976), the authors develop techniques for efficient estimation under *MAR* (Van den Broeck et al., 2015), and for consistent estimation in some non-*MAR* scenarios (Mohan et al., 2018). Specialized techniques for fitting an explicit model for the missingness mechanism jointly with a complete data model have also been introduced in the context of collaborative filtering (Steck, 2010;

Hernández-Lobato et al., 2014) and learning from positive and unlabeled data (Bekker and Davis, 2018).

A method for dealing with non-*CAR* data without adding an explicit modeling component for the missingness (or coarsening) mechanism is the *adaptive imputation and maximization (*AIM*)* procedure proposed by Jaeger (2006b). This procedure is structurally very similar to the EM algorithm, but instead of constructing in an E-step the expected completion given a current model, one computes in an AI-step the fractional completion that minimizes the Kullback-Leibler divergence relative to the current model. As shown by Jaeger (2006b), the resulting procedure maximizes the likelihood function that is free of any assumptions on the missing data mechanism.

Figure 1 gives an illustration of the analogies and difference between the AIM and EM procedures. Underlying complete data is generated by a two-dimensional Gaussian distribution (a). Data sampled from this distribution is subject to a missingness mechanism such that values in the first component (x-axis in the plots) are likely to be missing if they fall into one of two narrow bands centered at -0.8 and 0.8, respectively, and second component values (y-axis) are likely to be missing at the lower end of the value range (an exact specification is given in Section 7.1.6). This leads to coarsened data (b) consisting of a distribution of the fully observed cases (shown as a heatmap), and the two marginal distributions for each of the two components from those cases where the other component is not observed (shown as two density curves on top of the corresponding axes). Both EM and AIM impute fractional completions for these incomplete observations, leading to distributions of imputed completions (c). Combining these imputed completions of the partially observed cases with the fully observed data cases, leads to the imputed complete data sets (d) from which then parameters will be estimated (i.e., (d) is simply the sum of (c) and the fully observed cases of (b)). The E step of the EM algorithm is constrained to fractional completions that follow the underlying parametric model, so that here the fractional completion for every missing value will follow a Gaussian distribution. This leads to an overall rather smooth, "near Gaussian" distribution of the imputed completions ((c) for EM). The AI imputations of AIM, on the other hand, are not constrained in any form. The underlying objective in their construction is that the resulting imputed complete data (d) is consistent with the underlying parametric model of a Gaussian distribution. The imputed completions ((c) for AIM) are therefore filling in the "gaps" in the fully observed data cases, such that the combination of the two becomes approximately Gaussian ((d) for AIM).

In this paper we present an in-depth study of the AIM algorithm, and a detailed comparison with EM. In Section 2 we first establish the foundations of likelihood-based inference from coarse data, and derive a representation of the likelihood function that incorporates no assumption on the data coarsening mechanism. This representation leads to the AIM algorithm. Section 3 then establishes theoretical consistency results for maximum likelihood inference both under the no-assumptions likelihood function, and the likelihood that incorporates the *CAR* assumption. The latter result implies a fundamental consistency guarantee for the EM algorithm, which, to the best of our knowledge, has not been documented in the literature before. In Section 4 the AIM procedure is formally introduced, and its relationship with the EM algorithm clarified. This analysis will resolve an apparent contradiction between our results and earlier work, where the AIM algorithm had already

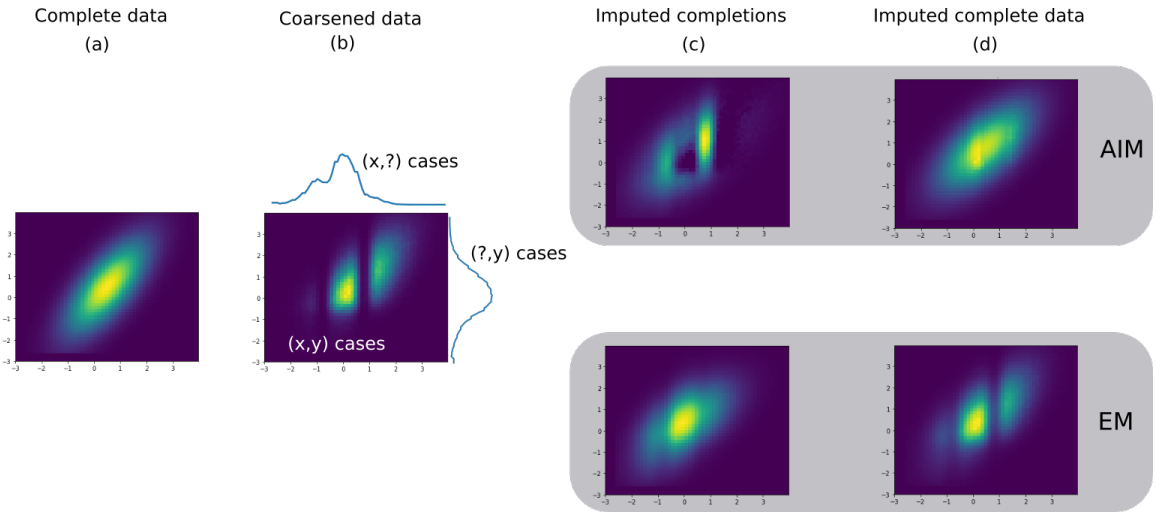

Figure 1: AIM and EM imputations

been considered under the name of *alternating KL minimization procedure*, and where it has been regarded as just an alternative, equivalent representation of the EM algorithm (Csiszár and Tusnády, 1984; Gunawardana and Byrne, 2005). Section 5 derives convergence results for the AIM procedure that closely mirror available convergence guarantees for EM.

The main theoretical results in this paper are for the case of discrete data from a finite sample space. Extending these results in full generality to the case of numeric data appears infeasible, since the underlying coarse data model would then involve probability distributions over the powerset of the reals, for which we even lack the necessary measure theoretic foundations. Therefore, for dealing with continuous data, we use a suitable restricted notion of coarse data and develop a principled approach for discretizing numeric data, such that the theoretical guarantees we obtain for the discrete case become applicable also for continuous data.

In the second part of this paper we conduct an empirical investigation into the practical performance of the AIM and EM algorithms when learning from *CAR* and non-*CAR* data. We consider the two scenarios of learning the parameters of a Gaussian distribution (as illustrated in Figure 1), and of learning the parameters of a Bayesian network. For both scenarios we first develop a suitable implementation of the generic AIM approach. The evaluation shows that for non-*CAR* data AIM generally provides more accurate parameter estimates than EM, whereas the converse holds when data is actually *CAR*. We will also see that a simple combination of EM and AIM achieves results that are generally at least as good as those obtained by EM, and better than EM in non-*CAR* cases.

This paper is an extended version of (Jaeger, 2006b) which originally introduced the AIM procedure[1]. Theorem 3 of the current paper, the general AIM algorithm, and an initial implementation for parameter learning in Bayesian networks was already presented

1. The author wished he had a good excuse/explanation for the slight delay that occurred in preparing this extended version, but none could be found.

in (Jaeger, 2006b). The theoretical results of Sections 3-5 in this paper are new, as are the extension to numeric data, and the experimental design and analysis.

2. Likelihood Inference for Incomplete Data

We assume that the underlying complete data distribution is represented by a random variable X with values in a finite state space $W = \{w_1, \ldots, w_n\}$. X may be a multivariate random variable, i.e. $\boldsymbol{X} = (X_1, \ldots, X_m)$, and $W = W(X_1) \times \cdots \times W(X_m)$, where $W(X_i)$ is the state space of X_i. We use bold font $\boldsymbol{X}$ to make the multi-variate nature of variables explicit.

We denote with $\Delta(W) = \{P = (p_1, \ldots, p_n) \in [0,1]^n : \sum_i p_i = 1\}$ the set of all probability distributions on W. A parametric model for the distribution of X consists of a parameter space $\Theta \subseteq \mathbb{R}^k$ and a mapping $\theta \mapsto P_\theta \in \Delta(W)$. We write $P_\Theta := \{P_\theta \mid \theta \in \Theta\} \subseteq \Delta(W)$. We are concerned with the problem of learning the true parameter θ^* of the distribution of X from incomplete observations of X.

In the general coarse data model incomplete observations of X can be given by any subset of the state space W. Formally, these observations are the values of a random variable Y with state space 2^W. We denote 2^W with $\mathcal{Y}$ when we want to emphasize its nature as the sample space of Y. It is assumed that the observations Y always contain the true value of X (i.e. the data is incomplete, not incorrect). Therefore, the joint sample space of X and Y is

$$\Omega(W) := \{(w, U) \mid w \in W, U \subseteq W : w \in U\}. \tag{1}$$

The joint distribution of X and Y then can be parameterized by P_θ and parameters

$$\lambda_{w,U} = P(Y = U \mid X = w) \quad ((w, U) \in \Omega(W)).$$

Thus, the parameter space of all possible coarsening mechanisms (for the given state space W) is

$$\Lambda_{sat} := \{(\lambda_{w,U})_{(w,U)\in\Omega(W)} \mid \forall w \in W : \sum_{U : w \in U} \lambda_{w,U} = 1\}. \tag{2}$$

The joint distribution for (X, Y) given by $\theta \in \Theta$ and $\lambda \in \Lambda_{sat}$ is denoted $P_{\theta,\lambda}$, and the marginal distribution of Y (the observed data distribution) by $P_{\theta,\lambda}^{\downarrow\mathcal{Y}}$. The parameter space Λ_{sat} represents the *saturated* (*SAT*) coarsening model, i.e. the one that does not encode any assumptions on how the data is coarsened.

Specific assumptions on the coarsening mechanism can be made by restricting admissible λ-parameters to some subset of Λ_{sat}. The most commonly made assumption is the *missing at random* (*MAR*) assumption for unobserved variables. In the coarse data framework, this becomes the *coarsened at random* (*CAR*) assumption (Heitjan and Rubin, 1991). As pointed out by Jaeger (2005b), one actually has to distinguish a *weak* and a *strong* version of the *CAR* assumption. For the context of this paper, the strong version is the more relevant one, since this is the assumption that justifies the *EM* algorithm. Formulated as a restriction on the parameters Λ, (strong) *CAR* is

$$\Lambda_{car} := \{\lambda \in \Lambda_{sat} \mid \forall U \forall w, w' \in U : \lambda_{w,U} = \lambda_{w',U}\}. \tag{3}$$

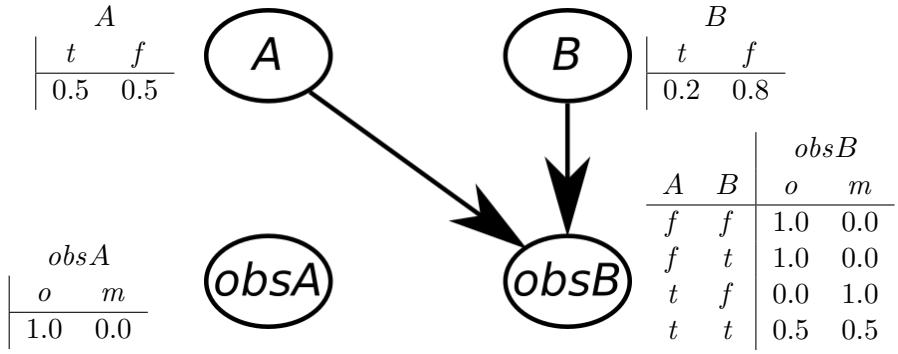

Figure 2: Basic Example

Thus, for any possible incomplete observation U: all complete data points w, w' compatible with U are equally likely to be coarsened to U.

Example 1 *The simplest class of coarsening mechanisms are those that are captured by the* grouped data model. *These models are given by a partitioning $U_1, \ldots, U_k$ of W, and a coarse data variable Y with $P(Y = U_i \mid X = w) = 1$ for the (unique) U_i containing w. In particular, grouped data models are always* CAR *because, by definition then for $w, w' \in U_i : 1 = P(Y = U_i \mid X = w) = P(Y = U_i \mid X = w')$.*

An important special type of grouped data models are latent variable models*: in these models $\boldsymbol{X} = (O_1, \ldots, O_l, L_1, \ldots, L_m)$, where variables O_i are always observed, and the (latent) variables L_j are never observed. The partitioning of W then is defined by the possible joint observations $\boldsymbol{o}$ of the O_i, i.e. consists of the sets of the form*

$$U_{\boldsymbol{o}} = \{(\boldsymbol{o}, \boldsymbol{l}) \mid \boldsymbol{l} \in \times_j W(L_j)\}.$$

The following example will be used for illustration throughout the paper.

Example 2 *Figure 2 shows a Bayesian network with two binary nodes A, B, and two* observation nodes obsA,obsB*. The distribution of interest here is the joint distribution of A and B, i.e. in our general terminology: $X = (A, B)$ and $W = \{t, f\} \times \{t, f\}$. The distribution of X is parameterized by $\Theta = \{(\theta_A, \theta_B) \mid \theta_A, \theta_B \in [0, 1]\}$, where $\theta_A := P(A = t), \theta_B := P(B = t)$ (in Figure 2: $\theta_A = 0.5$, $\theta_B = 0.2$). The observation nodes have the two possible values o (observed) and m (missing) and thereby represent a coarsening mechanism: when* obsA$= m$*, or* obsB$= m$*, then the value of A, respectively B will be recorded as missing. According to the model, A always is observed, and B can only be missing when $A = t$. The model only allows for four distinct observations. The observations, their representation as subsets $U \subseteq W$, and their probabilities are shown in Table 1.*

Table 2 shows the coarsening model in terms of the $\lambda_{w,U}$ parameters. Entries "nd" mean that the parameter is undefined because the given (w, U) pair does not belong to $\Omega(W)$. This data is not CAR*, because $\lambda_{w_3,U_1} \neq \lambda_{w_4,U_1}$.*

We are interested in learning the parameters θ of the complete data distribution from the observed values $\boldsymbol{U} = (U^{(1)}, \ldots, U^{(N)})$ of iid random variables $Y^{(1)}, \ldots, Y^{(N)}$. As a notational convention we use superscripts in parentheses to denote sample indices in order to

Observation	U	$P(Y = U)$
$A = t, B =?$	$U_1 = \{(t,f),(t,t)\}$	0.45
$A = t, B = t$	$U_2 = \{(t,t)\}$	0.05
$A = f, B = t$	$U_3 = \{(f,t)\}$	0.1
$A = f, B = f$	$U_4 = \{(f,f)\}$	0.4

Table 1: Example 2: Probabilities of observations

w	U_1	U_2	U_3	U_4
$w_1 = \{f,f\}$	nd	nd	nd	1
$w_2 = \{f,t\}$	nd	nd	1	nd
$w_3 = \{t,f\}$	1	nd	nd	nd
$w_4 = \{t,t\}$	0.5	0.5	nd	nd

Table 2: Example 2: λ parameters

avoid confusion with subscripts used for other purposes, notably the indexing of components in a multivariate variable.

The sample $\boldsymbol{U}$ together in conjunction with assumptions on the coarsening process expressed as a subset $\Lambda \subseteq \Lambda_{sat}$ induces a *profile likelihood* function on the parameter space Θ by maximizing over λ-values:

$$L_\Lambda(\theta \mid \boldsymbol{U}) := \max_{\lambda \in \Lambda} \prod_{i=1}^{N} P_{\theta,\lambda}(Y = U^{(i)}) = \max_{\lambda \in \Lambda} \prod_{i=1}^{N} \sum_{w \in U^{(i)}} P_\theta(X = w) P_\lambda(Y = U^{(i)} | X = w). \quad (4)$$

We refer to (4) also as the *profile(Λ)-likelihood* (if Λ is not assumed to be closed, the supremum should be used instead of the maximum; however, we will only be concerned with closed Λ). We write LL_Λ for the corresponding log-likelihood $logL_\Lambda$. Also, for $\Lambda = \Lambda_{sat}$ and $\Lambda = \Lambda_{car}$ we write $(L)L_{sat}$ and $(L)L_{car}$ rather than $(L)L_{\Lambda_{sat}}$ and $(L)L_{\Lambda_{car}}$.

The result that under the CAR assumption the coarsening mechanism can be ignored derives from the fact that the profile(CAR)-likelihood factors as

$$L_{car}(\theta \mid \boldsymbol{U}) = f(\boldsymbol{U}) L_{FV}(\theta \mid \boldsymbol{U}), \quad (5)$$

where

$$f(\boldsymbol{U}) := \max_{\lambda \in \Lambda_{car}} \prod_{i=1}^{N} \lambda_{U^{(i)}},$$

and L_{FV} is the *face-value likelihood* (Dawid and Dickey, 1977)

$$L_{FV}(\theta \mid \boldsymbol{U}) := \prod_{i=1}^{N} P_\theta(X \in U^{(i)}). \quad (6)$$

We now proceed to give a useful characterization of L_{sat} that similarly as (5) for L_{car}, provides a more manageable basis for optimization than the generic definition (4).

Definition 1 *Let $\boldsymbol{U} = U^{(1)}, \ldots, U^{(N)}$ be a data set.*

a. *A* fractional completion *of* $\boldsymbol{U}$ *is a mapping* c *that assigns to every* $U^{(i)} \in \boldsymbol{U}$ *a probability distribution* $c(U^{(i)})$ *over* $U^{(i)}$*. We also write* $c(U^{(i)}, w)$ *for* $c(U^{(i)})(w)$ *(*$w \in U^{(i)}$*). The fractional completion* c *defines a probability distribution* $P_c := 1/N \sum_{i=1}^{N} c(U^{(i)})$ *on* W*. We denote with* $\mathcal{C}(\boldsymbol{U})$ *the set of all fractional completions of* $\boldsymbol{U}$*. By a slight abuse of notation, we also use* $\mathcal{C}(\boldsymbol{U})$ *for the set of induced distributions* $\{P_c \mid c \in \mathcal{C}(\boldsymbol{U})\}$*.*

b. *Let* $m(\boldsymbol{U})$ *denote the empirical distribution of* $\boldsymbol{U}$ *on* $\mathcal{Y}$*. If* $m(\boldsymbol{U}) = m(\boldsymbol{U}')$*, then* $\mathcal{C}(\boldsymbol{U}) = \mathcal{C}(\boldsymbol{U}')$*. We therefore also write* $\mathcal{C}(m)$ *for the* $\mathcal{C}(\boldsymbol{U})$ *of any* $\boldsymbol{U}$ *with empirical distribution* m*.*

In the following we will be exclusively concerned with fractional completions. To simplify language, we therefore from now on simply use the term 'completion', which is always to be understood as referring to fractional completions.

We use m to denote the empirical distribution of Y in order to point out its possible interpretation as a *basic probability assignment* in the sense of Dempster-Shafer theory. Moreover, the sets $\mathcal{C}(\boldsymbol{U})$ (respectively $\mathcal{C}(m)$) are the sets of *compatible* probability measures in the sense of Dempster (1967). A useful characterization of the sets $\mathcal{C}(m)$ was given by Dempster as follows.

Theorem 2 ((Dempster, 1967)) *Let* m *be a probability distribution on* $\mathcal{Y}$*. Let* ΠW *be the set of permutations of* W*. For* $\pi \in \Pi W$ *and* $U \in \mathcal{Y}$ *let* $\min^{\pi}(U)$ *be the minimal element in* U *according to the ordering* π*. For* $\pi \in \Pi W$ *define* $\pi(m) \in \Delta W$ *by*

$$\pi(m)(w) := \sum_{U : w = \min^{\pi}(U)} m(U) \quad (w \in W).$$

Then

$$\mathcal{C}(m) = \mathrm{conv}\{\pi(m) \mid \pi \in \Pi W\},$$

where conv *denotes the convex hull.*

For probability distributions P, Q, we use $H(P)$ to denote the entropy of P, and $KL(P, Q)$ to denote the Kullback Leibler divergence $\sum_w P(w) \log(P(w)/Q(w))$. When $\mathcal{P} \subseteq \Delta(W)$ is a set of probability distributions, we also write $KL(\mathcal{P}, Q)$ for $min_{P \in \mathcal{P}} KL(P, Q)$.

We can now characterize the profile likelihood under the saturated coarsening model as follows.

Theorem 3 *Let* $\boldsymbol{U} = U^{(1)}, \ldots, U^{(N)}$ *be a data set, and* m *the empirical distribution defined by* $\boldsymbol{U}$ *on* $\mathcal{Y}$*. Then*

$$\frac{1}{N} LL_{\mathrm{sat}}(\theta \mid \boldsymbol{U}) = -H(m) - \min_{c \in \mathcal{C}(\boldsymbol{U})} KL(P_c, P_\theta). \tag{7}$$

The proof of this and all following theorems can be found in Appendix A.

Corollary 4 *If* $P_\Theta \cap \mathcal{C}(\boldsymbol{U}) \neq \emptyset$*, then* $\{\theta \in \Theta \mid P_\theta \in \mathcal{C}(\boldsymbol{U})\}$ *is the set of global maxima of* LL_{sat}*.*

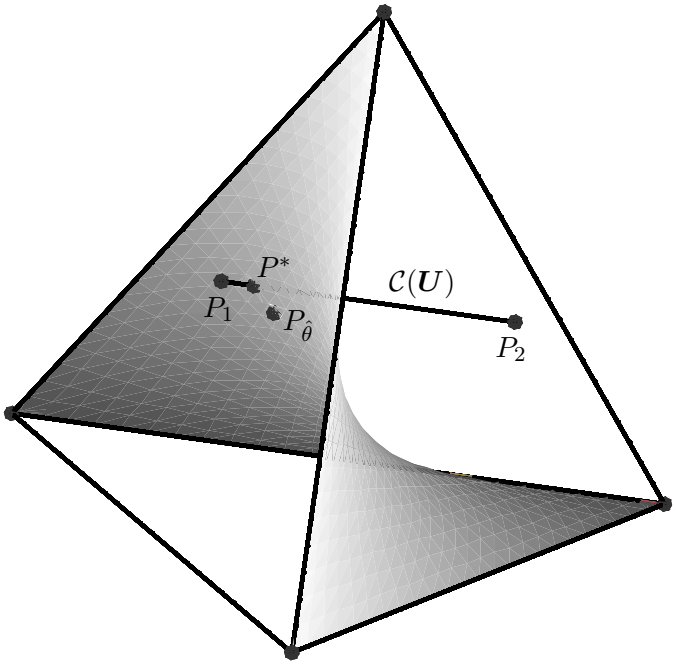

Figure 3: Example 3: Geometric illustration

Example 3 *(Example 2 continued) Suppose we have a large representative sample* $\boldsymbol{U}$ *for* Y*, i.e. the empirical distribution* m *defined by the sample has the expected values given by the third column of Table 1:* $m(U_1) = 0.45, \ldots, m(U_4) = 0.4$*.*

The set $\mathcal{C}(\boldsymbol{U})$ *is the convex hull of the two extreme completions* $P_1 = (0.4, 0.1, 0.45, 0.05)$ *(all unobserved* B *are assumed to be* false*), and* $P_2 = (0.4, 0.1, 0, 0.5)$ *(all unobserved* B *are assumed to be* true*). The probability values in the* P_i*-tuples here are ordered according to the enumeration of the* w_j *in Table 2.*

Figure 2 illustrates the situation: the set $\Delta(W)$ *is a 3-dimensional tetrahedron (embedded in 4-dimensional space). The parametric model* P_Θ *is a 2-dimensional manifold in* $\Delta(W)$*. The set* $\mathcal{C}(\boldsymbol{U})$ *is a line segment with endpoints* P_1, P_2*. As apparent from the figure (and readily verified analytically),* $\mathcal{C}(\boldsymbol{U})$ *intersects* P_Θ *in exactly one point* $P^* = (0.4, 0.1, 0.4, 0.1)$*, corresponding to* P_{θ^*} *for the true parameter* $\theta^* = (0.5, 0.2)$*. Thus, by Corollary 4,* θ^* *is the unique maximum of* LL_{sat}*.*

Turning to LL_{car}*, it is straightforward to maximize* $LL_{\text{FV}} \equiv \sum_{i=1}^{4} m(U_i) \log P_\theta(U_i)$ *analytically, and find that the maximum is attained at* $\hat{\theta} = (0.5, 0.2727)$*, corresponding to* $P_{\hat{\theta}} = (0.3636, 0.1363, 0.3636, 0.1363)$*. Thus, for this non-*CAR *data, maximizing* LL_{sat} *returns the true parameters, while maximizing* LL_{car} *does not.*

3. Consistency

In Example 3 the true parameter θ^* was found by maximization of the profile likelihood function that would be induced by a representative sample. That large samples, with high probability, induce likelihood functions maximized by the true parameter is the *consistency* property of maximum likelihood estimates, and the perhaps strongest justification for the maximum likelihood principle. The classic consistency results ((Wald, 1949), see also (Lehmann and Casella, 1998, Theorems 3.2,3.7)) provide conditions under which a true parameter γ^* will be found (with probability 1) in the limit of large sample size $N \to \infty$. These results, however, require the assumptions that the data is complete, and that the true parameter γ^* is *identifiable*, i.e. $P_{\gamma^*} \neq P_\gamma$ for all $\gamma \neq \gamma^*$.

Redner (1981) generalizes Wald's theorem to the non-identifiable case. In this case, consistency has to be understood as the property that the true parameter is among the maximizers of the likelihood function (in the large sample limit), but not necessarily the unique maximum. Redner's results can be used to analyze the behavior of the sequence $(\hat{\theta}_N, \hat{\lambda}_N)$ of maximum likelihood estimators for (θ^*, λ^*): when a pair (θ, λ) is seen as a parameterization of the distribution of the observed variable Y, then we obtain a complete data, non-identifiable scenario (because typically distinct parameters (θ, λ), (θ', λ') can induce the same distribution on the observation space $\mathcal{Y}$). Redner's results, therefore, establish consistency properties of the full likelihood function $P_{\theta,\lambda}(\boldsymbol{U})$ for jointly estimating θ and λ. Our interest, on the other hand, is to avoid an explicit optimization over λ, and use profile likelihoods (5) and (7) to directly estimate the parameter of interest θ.

In this section we therefore derive consistency results similar in nature to those of Redner (1981), but directly addressing the maximization of the profile(sat) and profile(car) likelihood functions. It is clear that the situation of Example 3, where the true parameter was the unique maximum of LL_{sat}, cannot be expected to be encountered in general. When data is highly incomplete, then the observed sample $\boldsymbol{U}$ will not contain enough information to identify the true parameter (in the very extreme case, the data is vacuous, i.e. $U^{(i)} = W$ for all i. Then $\mathcal{C}(\boldsymbol{U}) = \Delta(W)$, and by Corollary 4 LL_{sat} is constant on Θ). The best we can hope, therefore, is that the true θ^* is one of the maxima of LL_{sat}. The following theorem states that this will almost surely be the case in the large sample limit.

We first formulate a couple of assumptions on the complete data parametric model.

Assumption 1 Θ is a bounded subset of $\mathbb{R}^k$ for some $k \geq 1$.

Assumption 2 The parameterization is continuous: if $\theta_i \to \theta$ in Θ, then $P_{\theta_i} \to P_\theta$ in $\Delta(W)$.

Since we are only considering categorical data, Assumption 1 is almost tautological, and Assumption 2 is true for any reasonable parameterization. For the statement and proof of the following theorem we use the convention that for an event α and a distribution P we also write "α P-a.s" (α holds P-almost surely) instead of $P(\alpha) = 1$.

Theorem 5 *Let $U^{(1)}, U^{(2)}, \ldots$ be observations of iid random variables $Y^{(i)}$, which are distributed according to $m^* := P^{\downarrow\mathcal{Y}}_{\theta^*,\lambda^*}$. Denote by P the joint distribution of all $Y^{(i)}$, and $\boldsymbol{U}_N = (U^{(1)}, \ldots, U^{(N)})$.*

A. *Let $\hat{\theta}_N$ maximize $LL_{\mathrm{sat}}(\cdot \mid \boldsymbol{U}_N)$ ($N \geq 1$). Then*

$$\lim_{N\to\infty} \frac{1}{N}(LL_{\mathrm{sat}}(\hat{\theta}_N \mid \boldsymbol{U}_N) - LL_{\mathrm{sat}}(\theta^* \mid \boldsymbol{U}_N)) = 0 \quad P\text{-a.s.}.$$

B. *Assume that Assumptions 1 and 2 hold, and that $\lambda^* \in \Lambda_{\mathrm{car}}$.*

(i) *Let $\hat{\theta}_N$ maximize $LL_{\mathrm{car}}(\cdot \mid \boldsymbol{U}_N)$ ($N \geq 1$). Then*

$$\lim_{N\to\infty} \frac{1}{N}(LL_{\mathrm{car}}(\hat{\theta}_N \mid \boldsymbol{U}_N) - LL_{\mathrm{car}}(\theta^* \mid \boldsymbol{U}_N)) = 0 \quad P\text{-a.s.}.$$

(ii) *If, furthermore, the system of equations*

$$P_\theta(U) = P_{\theta^*}(U) \quad (U : m^*(U) > 0)$$

has the unique solution $\theta = \theta^*$, then

$$\lim_{N\to\infty} \hat{\theta}_N = \theta^* \quad P\text{-a.s.}$$

Part **B** of the theorem says that face-value likelihood optimization objective underlying the EM algorithm is consistent when the data is actually *CAR*. This is certainly not surprising, and, indeed, an implicit assumption made when using the EM algorithm. To the best of our knowledge, however, while the convergence behavior of the EM algorithm to a (local) maximum of LL_{car} has received much attention, there is a gap in the literature regarding the consistency of these maxima for identifying the true θ^*. Part **A** provides the same guarantee for maxima of LL_{sat} and non-*CAR* data.

4. The AIM Procedure

According to Theorem 3, an optimal parameter $\hat{\theta}$ for LL_{sat} is equivalently characterized as

$$\hat{\theta} = \underset{\theta\in\Theta}{arg\,min} \underset{c\in\mathcal{C}(\boldsymbol{U})}{min} KL(P_c, P_\theta).$$

Based on this characterization, a general procedure for optimizing LL_{sat} is given by the alternating minimization procedure in Table 1. To simplify notation from now on we also write $KL(c, \theta)$ for $KL(P_c, P_\theta)$.

This procedure was proposed by Jaeger (2006a) under the name of Adaptive Imputation and Maximization (AIM)[2], because the AI step imputes a (fractional) data completion $\mathrm{AI}(\theta, \boldsymbol{U})$ that tries to adapt the imputed empirical distribution as closely as possible to the distribution defined by the current parameter setting θ. The general procedure of minimizing a *KL* distance between two sets of distributions by alternating minimization in the first and second argument has previously been proposed and investigated. An early and thorough study of this approach was given by Csiszár and Tusnády (1984). Csiszár and Tusnády also considered maximum likelihood inference from incomplete data as an

2. Actually, it was originally called "Adjusting Imputation and Maximization". The new name is perhaps a little less clumsy, while preserving the acronym.

AIM($\boldsymbol{U}$);
$t := 0$;
Choose initial $\theta_0 \in \Theta$;
repeat
  $c_t := \mathrm{AI}(\theta_t, \boldsymbol{U}) := arg\,min_{c \in \mathcal{C}(\boldsymbol{U})}\, KL(c, \theta_t)$ `/* AI step */`
  $\theta_{t+1} := \mathrm{M}(c_t) := arg\,min_{\theta \in \Theta}\, KL(c_t, \theta)$ `/* M step */`
  $t := t + 1$
until *termination condition applies*;

Algorithm 1: The AIM procedure

application of this alternating minimization procedure, and, indeed, in this case identified it with the EM algorithm. Essentially the same alternating *KL*-minimization procedure also is considered by Neal and Hinton (1998), Heskes et al. (2004), and Gunawardana and Byrne (2005) as an alternative representation of the EM algorithm. As the EM algorithm maximizes LL_{car}, whereas AIM maximizes LL_{sat}, and (as we saw in Example 3) these two maximizations can lead to different results, this may look like a contradiction. In the remainder of this section we will analyze and resolve this apparent conflict, and explain the close relationship between AIM and EM from two different perspectives. This material deals with some rather specific technicalities, and readers mostly interested in the bigger picture may skip it without loosing relevant background information for the following sections.

4.1 AIM and EM in the Flat Data Model

The resolution of the apparent conflict described above lies in the realization that the mentioned works are based on special incomplete data models in which the AIM and EM procedures become equivalent. We next provide a simple condition under which the AI and E steps are identical, and then analyze why this condition is actually fulfilled for the data models assumed in the aforementioned papers. In the following theorem we denote with $E(\theta, \boldsymbol{U}) \in \Delta(W)$ the expected empirical complete data distribution under P_θ given $\boldsymbol{U}$, i.e. $E(\theta, \boldsymbol{U})(x) = 1/N \sum_i P_\theta(x \mid U^{(i)})$.

Theorem 6 *Let $U^{(1)}, \ldots, U^{(N)}$ be a data set. If for all i, j: $U^{(i)} = U^{(j)}$, or $U^{(i)} \cap U^{(j)} = \emptyset$, then*

$$L_{\mathrm{sat}}(\cdot|\boldsymbol{U}) = L_{\mathrm{car}}(\cdot|\boldsymbol{U}), \tag{8}$$

and for all θ:

$$\mathrm{AI}(\theta, \boldsymbol{U}) = \mathrm{E}(\theta, \boldsymbol{U}). \tag{9}$$

Equality (8) is essentially just another instance of the observation in Example 1 that grouped data always is *CAR*. In Example 1 we considered a coarse data distribution that is concentrated on a partition of W. In Theorem 6 we only have a sample that is consistent with the assumption that the underlying distribution follows a grouped data model. This is sufficient to make the two profile likelihoods equal. Equation (9) says that not only do the objective functions of AIM and EM here coincide, but also algorithmically the two methods are equivalent.

The following example shows how the E and AI steps are not equivalent when the conditions of Theorem 6 do not hold.

Example 4 *(Example 3 continued) Let* $\boldsymbol{U}$ *be as in Example 3. Assume that* $\theta_t = (0.5, 0.2)$ *(the true parameters). Since* $P_{\theta_t} \in \mathcal{C}(\boldsymbol{U})$ *(cf. Figure 2), we have that* $\mathrm{AI}(\theta_t)$ *is the completion* c_t *with* $P_{c_t} = P_{\theta_t} = (0.4, 0.1, 0.4, 0.1)$*. Then also* $m(c_t) = \theta_{t+1} = \theta_t$ *again, i.e.* θ_t *is a fixed point of the* AIM *procedure.*

The expected completion of $\boldsymbol{U}$ *under* θ_t *is* c_t *with* $P_{c_t} = (0.4, 0.1, 0.36, 0.14)$*, leading to* $\theta_{t+1} = (0.5, 0.24)$ *in the subsequent* M*-step.*

Neal and Hinton (1998), Heskes et al. (2004) and Gunawardana and Byrne (2005) are basing their work on an incomplete data model that does not explicitly incorporate repeated iid samples. In this what we shall call the *flat data model*, data is represented by a single complete data variable X (possibly multivariate $\boldsymbol{X} = (X_1, \ldots, X_n)$), and a corresponding incomplete data variable Y (possibly given in the form of a multivariate $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, where Y_i can be an observed value of X_i, or a missingness symbol). This abstract view subsumes the case where $\boldsymbol{X}$ is an $N \times m$ matrix of N samples of multivariate random variables (so that $n = N \cdot m$). An assumption about the iid nature of the N samples will then just become part of the overall parametric model for $\boldsymbol{X}$.

On the basis of the flat data model, inference is based on a single observation U of Y. Thus, in our notation, $N = 1$, and the profile likelihood becomes

$$L_\Lambda(\theta \mid U) = \max_{\lambda \in \Lambda} \sum_{w \in U} P_\theta(X = w) P_\lambda(Y = U | X = w). \tag{10}$$

The condition of Theorem 6 then is trivially satisfied, and EM and AIM are the same. Concretely, both for $\Lambda = \Lambda_{sat}$ and $\Lambda = \Lambda_{car}$, (10) is maximized by

$$P_\lambda(Y = U | X = w) = 1 \text{ for all } w \in U. \tag{11}$$

If X actually is composed of N independent samples, i.e., $X = (X^{(1)}, \ldots, X^{(N)})$ and $w = (w^{(1)}, \ldots, w^{(N)})$, then (10) becomes

$$L_\Lambda(\theta \mid U) = \max_{\lambda \in \Lambda} \sum_{w \in U} P_\lambda(Y = U | X = w) \prod_{i=1}^{N} P_\theta(X^{(i)} = w^{(i)}), \tag{12}$$

which still is maximized by (11) for both $\Lambda = \Lambda_{sat}$ and $\Lambda = \Lambda_{car}$.

The important observation now is that while (12) incorporates an iid assumption for the complete data X, it does not incorporate an iid assumption also for the coarsening process, i.e., the assumption that $P_\lambda(Y = U | X = w)$ also factors as a product of N iid coarsening operations applied separately to the N complete data samples. Under that assumption the solution (11) is not valid (unless the observation U actually factors into N components satisfying the conditions of Theorem 6). In order to also include an iid coarsening assumption into the flat data model, the space of admissible coarsening parameters has to be restricted to subsets $\Lambda^{iid}_{sat} \subset \Lambda_{sat}$, respectively $\Lambda^{iid}_{car} \subset \Lambda_{car}$. Maximization over these subsets then gives the profile likelihood (4), which no longer coincides for $SAT$and CAR.

We would argue that when complete data is obtained as a sequence of iid samples $w^{(1)}, \ldots, w^{(N)}$, then it is most natural to assume that a coarsening mechanism also applies independently and uniformly to each individual sample, rather than jointly to the whole data set. This view is also implicit in (Mealli and Rubin, 2015), but see (Corani and Zaffalon, 2008) for arguments to the contrary. Under this assumption, EM and AIM are distinct methods.

4.2 AIM and EM in the Extended Data Model

It is well known that non-*CAR* data can be turned into *CAR* data by including an explicit representation of the coarse observations in the data, and integrating a model for the coarsening process into the complete data model. This has mostly been informally observed in the context of (non-)*MAR* data (e.g. (Little and Rubin, 1987, Section 11.2), (Koller and Friedman, 2009, Exercise 19.2)). In this subsection we make this approach precise for the general *CAR* framework, and clarify the relationship between the AIM algorithm and the EM algorithm executed on the *extended data model* (formally defined below).

In a basic missing-value scenario, as illustrated in Figure 2 and Example 2, the extended data model is given by the inclusion of the observation variables *obsA*, *obsB* into the complete data model. Thus, an example of a complete data case would now be $A = t, B = f, obsA = o, obsB = m$, which then gives rise to the incomplete observation $A = t, B =?, obsA = o, obsB = m$. Similar to what we saw in Section 4.1, the mapping from complete data cases to incomplete observations then follows a simple grouped data model, and hence is *MAR*. Since the complete data representation now contains the observation variables, the parametric complete data model must also include a model for the missingness mechanism. As already indicated in Section 1.3, this often involves the design of relatively specific models for the missingness mechanism. For example, in our Example 2 the model for the missingness mechanism may consist of the conditional independence assumptions encoded by the graphical structure of Figure 2, or even be fully specified via fixed parameters in the conditional probability tables for *obsA* and *obsB*.

In the general coarse data setting, we can define the extended data model simply to be the space $\Omega(W)$ whose elements represent both the underlying complete data w, and the observed U. In order to emphasize the new role of $\Omega(W)$ as a complete data space, we now denote it as W^+, and write $\boldsymbol{X}^+ := (X, Y)$ for a random variable with values in W^+. As before, the distribution of $\boldsymbol{X}^+$ then is jointly parameterized by $\theta \in \Theta$ and $\lambda \in \Lambda$, where $\Lambda \subseteq \Lambda_{sat}$ can still encode any kind of assumptions on the coarsening process for the original complete data variable X (especially non-*CAR* assumptions: $\Lambda \not\subseteq \Lambda_{car}$). A coarse observation $Y = U$ in the original sense is now interpreted as a coarse observation $Y^+ = U^+ := \{(w, U) \in W^+ | w \in U\}$ in the extended space. The conditional distribution of Y^+ given X^+ then is

$$P^+(Y^+ = U^+ | X^+ = (w, U)) = 1 \text{ for all } w \in U, \tag{13}$$

which is *CAR* (similar to what we had in (11)). The parameters θ, λ governing the distribution of X^+, thus, can be learned by maximizing the face-value likelihood (6), which now

can be written as

$$L^+_{FV}(\theta,\lambda|\boldsymbol{U}^+) = \prod_{i=1}^{N} \sum_{(w,U)\in U^{(i)+}} P_{\theta,\lambda}(w,U) = \prod_{i=1}^{N} \sum_{w\in U} P_\theta(X=w)P_\lambda(Y=U|X=w). \quad (14)$$

Maximizing (14) to obtain an estimate $(\hat{\theta},\hat{\lambda}) \in \Theta \times \Lambda$, and then discarding the "nuisance" parameter $\hat{\lambda}$ is equivalent to maximizing the profile likelihood (4).

For the maximization of (14) the EM algorithm can be used. We then denote it by EM$^+$, in order to distinguish it from the EM algorithm operating on the original data space. In particular, for $\Lambda = \Lambda_{sat}$, EM$^+$ and AIM then maximize the same objective. In contrast to what we found in Theorem 6, however, AIM and EM$^+$ are not equivalent algorithmically. A main difference between AIM and EM$^+$ lies in the fact that the AIM iterations are directly defined on the parameter space Θ of interest, and only requires the induced distributions $P_c \in \Delta W$. EM$^+$, in contrast, treats parameters θ and λ equally during the execution of the algorithm, operates on distributions in the larger space W^+, and marginalizes to θ only at the very end. However, if AIM actually maintains $\mathcal{C}(\boldsymbol{U})$ as a completion mapping c defined on $\boldsymbol{U}$ (as is the case in our implementations described in Section 7), then a completion $c \in \mathcal{C}(\boldsymbol{U})$ corresponds to a parameter $\lambda \in \Lambda_{sat}$ via $\lambda_{w,U} = c(U,w)m(U)/P_c(w)$, and the domains of optimization for AIM and EM$^+$ become essentially the same. In (Jaeger, 2006b) a comparison between AIM and EM$^+$ for learning Bayesian network parameters (cf. Section 7.2) is briefly reported. There it was found that EM$^+$ did not scale to larger Bayesian networks, because exact computations of the E step became intractable. AIM, on the other hand, was still able to produce consistent estimates using an implementation of an approximate AI step. For complex models P_θ, both AIM and EM$^+$ will require suitable approximation strategies for the AI and E steps, respectively (and maybe the M step as well). AIM has a potential advantage of supporting approximations on smaller parameter and state spaces. EM$^+$, on the other hand, has the advantage of being applicable to any model $\Lambda \subset \Lambda_{sat}$ for the (non-CAR) coarsening process, whereas AIM only applies to the assumption-free case Λ_{sat}.

5. Convergence

In this section we analyze the convergence behavior of the AIM procedure. The convergence behavior of the EM algorithm has been extensively studied (Dempster et al., 1977; Wu, 1983; Tseng, 2004; Gunawardana and Byrne, 2005). The available theoretical convergence results for the EM algorithm are not very strong in several regards, and do not guarantee the convergence of the parameter sequence defined by EM to a local maximum of the face-value likelihood function (even though such a convergence is mostly observed in practice). In particular, the theoretical results are subject to the following two limitations:

- Only convergence to stationary points of the likelihood function can be shown. These stationary points can be local maxima, saddle points, or even local minima. Example 5 below illustrates a case of "convergence" to the global minimum of the likelihood.
- No actual convergence of the parameter sequence $\theta_0, \ldots, \theta_{i+1} := EM(\theta_i), \ldots$ to some limit point θ^* usually can be proven. Convergence of this sequence to a stationary

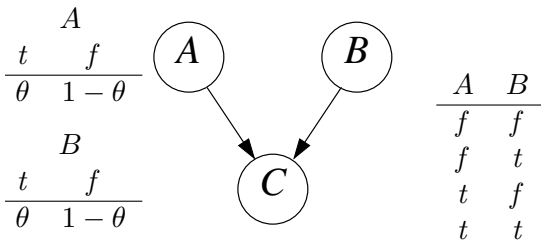

Figure 4: Model for Example 5

point (more correctly: to stationary points) is to be understood in the weaker sense that every accumulation point of the sequence is stationary.

Fortunately, both these limitation of theoretical convergence results are not often observed in practice. Nevertheless, we give the following cautionary example, due to Gunawardana and Byrne (2005).

Example 5 *(Gunawardana and Byrne, 2005, Example 2) Assume a parametric model for the joint distribution of three binary random variables A, B, C as defined by the Bayesian Network shown in Figure 4. The distributions of A and B are defined by a shared parameter θ, and C is deterministic given A and B, so that there is only a single free parameter θ. Assume there is a single incomplete observation $A =?, B =?, C = f$. This induces the face-value likelihood function*

$$L_{\mathrm{FV}}(\theta) = (1-\theta)^2 + \theta^2,$$

which has two global maxima $\theta = 0, \theta = 1$, and one global minimum $\theta = 0.5$. It is easily verified that $\theta = 0.5$ is a fixed point of EM, so that the EM sequence starting with initial point $\theta_0 = 0.5$ remains stuck at the global minimum.

The story for the AIM *procedure here is the same: because $N = 1$, Theorem 6 applies; $\theta = 0.5$ also is a global minimum of the profile(sat) likelihood, and also is a fixed point of the* AIM *operator.*

*While showing the possibility of EM/*AIM *"converging" to the global minimum of the likelihood function, this example does not indicate a serious problem for these procedures in practice, since started at any initial parameter value θ_0 other than $\theta_0 = 0.5$ the procedures will converge to one of the likelihood maxima $\theta = 0$ or $\theta = 1$. In other words, the pathological fixed point $\theta = 0.5$ is not an "attractor" for an EM/*AIM *sequence initialized with some random parameter θ_0.*

We now turn to the convergence of the AIM procedure. Since Gunawardana and Byrne (2005) actually used an alternating *KL*-minimization interpretation of EM to derive their convergence results, one might ask whether AIM convergence has not already been proven in (Gunawardana and Byrne, 2005). This is not the case, because Gunawardana and Byrne (2005) assume their alternating *KL*-minimization procedure to operate in the flat sample space (where EM and AIM are equivalent), and their results are not directly applicable to the case of our main interest, which is executing AIM over the multi-sample space. Nevertheless

an appropriate convergence result for AIM could probably be obtained following very closely Gunawardan and Byrne's line of argument, which (similarly to (Wu, 1983)) is mainly based on a very general convergence theorem by Zangwill (1969). However, since the application of Zangwill's theorem requires to first establish several technical conditions, we prefer to give a direct, stand-alone derivation, which may be more transparent without being significantly more involved. The convergence properties we obtain for AIM are very similar in nature to those known for EM, and subject to the same caveats as mentioned above for the convergence results for EM.

We first introduce notation for the effective domain of optimization:

$$D := \{(c, \theta) \in \mathcal{C}(\boldsymbol{U}) \times \Theta \mid KL(c, \theta) < \infty\} \subseteq \mathbb{R}^{n+k}$$

We note that D depends on the data $\boldsymbol{U}$, and in more complete notation should be written as $D(\boldsymbol{U})$. However, here, and for the remainder of this section, we take $\boldsymbol{U}$ as fixed, and suppress explicit references to $\boldsymbol{U}$ in the notation.

We make the following additional assumption, which essentially is the assumption that the parametric model P_Θ is identifiable.

Assumption 3 For every $c \in \mathcal{C}(\boldsymbol{U})$: if there exists $\theta \in \Theta$ with $KL(c, \theta) < \infty$, then there exists a unique θ^* minimizing $KL(c, \cdot)$ (i.e. M(c) is uniquely defined).

Given $(c, \theta) \in D$ we denote with $\mathrm{AIM}(c, \theta) = (\mathrm{AI}(\theta), \mathrm{M}(\mathrm{AI}(\theta)))$ the result of one round of AI and M updating (thus, $\mathrm{AIM}(c, \theta)$ does not actually depend on c, but it is convenient to treat this as an operator on the space D).

Theorem 7 *Let $(c_0, \theta_0) \in D$, and $(c_{i+1}, \theta_{i+1}) = \mathrm{AIM}(c_i, \theta_i)$ for $i \geq 0$. Every accumulation point $(\hat{c}, \hat{\theta})$ of the sequence $(c_i, \theta_i)_i$ then is a fixed point of the* AIM *operator.*

Theorem 7 does not directly answer the question of interest, which is the nature of the parameter $\hat{\theta}$ in an AIM limit (accumulation) point $(\hat{c}, \hat{\theta})$ in relation to the profile likelihood function. To answer that question one has to clarify the relationship between fixed points of the AIM algorithm and local minima of $KL(\cdot, \cdot)$ on D, and the relationship between local KL-minima and maxima of L_{sat}.

As illustrated by Example 5, there is no guarantee that an AIM fixed point is a local minimum of KL. However, what one can say is that if $(\hat{c}, \hat{\theta})$ is a fixed point in the interior of D, and KL is differentiable on D, then the gradient of KL at $(\hat{c}, \hat{\theta})$ is zero, i.e. $(\hat{c}, \hat{\theta})$ is a minimum, maximum, or saddle-point (cf. also (Gunawardana and Byrne, 2005, Corollary 4)). Moreover, since AIM produces parameter sequences (c_i, θ_i) with non-increasing KL-values, one can expect that a sequence starting from a random initial point (c_0, θ_0) is more likely to be "attracted" by a local KL-minimum, than by a maximum or saddle-point.

Regarding the relationship between KL-minima and L_{sat} maxima, we obtain the following answer.

Theorem 8 *If $(\hat{c}, \hat{\theta}) \in D$ is a local minimum of $KL(c, \theta)$, then $\hat{\theta}$ is a local maximum of L_{sat} on Θ.*

It is known that theoretical limitations of the convergence properties of the EM algorithm are not often encountered in practice. The same appears to be true for AIM: non-convergence (i.e., the existence of multiple accumulation points) is not observed in reality, and the limit points are in realistic scenarios turn out to be local maxima of L_{sat}.

6. Dealing with Continuous Spaces

The theoretical results and algorithmic principles developed in the preceding sections are framed for data in finite sample spaces W only. An extension of the conceptual framework and theoretical results in their full generality to continuous sample spaces is unrealistic, since for a real-valued random variable X the state space for the coarsened variable Y would then be $2^{\mathbb{R}}$. As already observed in (Gill et al., 1997, p.273), we lack the most fundamental foundations to deal with such entities, as "there is no natural topology on this very large space [i.e., $2^{\mathbb{R}}$], no natural Borel σ-algebra". We therefore develop in this section a general framework that allows us to reduce continuous coarse data problems into discrete ones. For this we first introduce a suitable restricted yet important class of continuous coarse data models. We then show that the methods and results we have developed are applicable to suitable discretizations of the continuous coarse data.

6.1 Continuous Coarsening Models

In the continuous case the complete data model consists of a random variable X taking values in $\mathbb{R}^n$ according to a distribution P_θ ($\theta \in \Theta$). We consider a restricted class of coarsening models for continuous variables that are induced by finite coarsening variables (Heitjan, 1994) as follows: let F be a random variable with values in a finite space $\mathcal{F}$ such that there exists a mapping

$$\zeta : \mathbb{R}^n \times \mathcal{F} \to 2^{\mathbb{R}^n}.$$

with $x \in \zeta(x, f)$ for all $x \in \mathbb{R}^n$ and $f \in \mathcal{F}$. Let $P_\gamma(F|X)$ be a conditional probability distribution of F given X defined by a parameter γ from a parameter space Γ.

The following examples show how several canonical cases of coarse continuous data can be represented by discrete coarsening variables.

Example 6 *(Missing values). In the standard missing values case, coarse observations of a multivariate $x \in \mathbb{R}^n$ are missing the values for some of the components of x, whereas the other components are fully observed. Such missingness patterns can be represented by a* missing-data indicator *(Rubin, 1976) $\boldsymbol{f} = (f_1, \ldots, f_n) \in \{0, 1\}^n$ where $f_i = 1$ means that the value of x_i is not observed. This corresponds to defining*

$$\zeta(x, \boldsymbol{f}) = \{\tilde{x} \in R^n | f_i = 0 \Rightarrow \tilde{x}_i = x_i\}.$$

A parametric model $P_\gamma(F \mid X)$ can, for example, be defined component-wise, by specifying for each $i = 1, \ldots, n$ a measurable function for the missingness probability $P(F_i = 1|X_i)$.

Example 7 *(Binned values). In the case of binned measurements, $\mathbb{R}^n$ is partitioned into a set $\mathcal{B}$ of n-dimensional intervals $B_1, \ldots, B_k$. If for all observations, only the bin containing the actual measurement is reported, then $\mathcal{F}$ can be taken to just consist of a single value f,*

and $f(x, f) = B_j$ is the bin containing x. This is a continuous version of a grouped data model (cf. Example 1). If some observations are recorded precisely, and some only in terms of bin membership, then we can choose $\mathcal{F} = \{f_0, f_1\}$ with $f(x, f_0) = \{x\}$, and $f(x, f_1)$ equal to the bin containing x. Examples of where such coarse data patterns can arise include the use of two different versions of a questionnaire, where one version asks the user to fill in an 'Age' statement by entering a number, and another version where the user has to check one of several pre-defined age ranges. Generally, whenever data is obtained by integration from different sources that record corresponding variables using different conventions and levels of precision, one may encounter such (and more complex) coarsening patterns.

6.2 Discretization

We aim to analyze numeric coarse data that can be modeled as described in the previous section by using a discretization approach. We consider discretizations of the real line $\mathbb{R}$ defined by parameters $a, b \in \mathbb{R}$ with $a < b$, and $g \in \mathbb{N}$. A discretization defined by these parameters, denoted $\mathbb{D}_{a,b,g}$, partitions $\mathbb{R}$ into g equal width intervals between the bounds a and b, plus the two unbounded intervals $]-\infty, a]$ and $[b, \infty[$. We also allow the case of the vacuous partition into the single interval $\mathbb{R}$, which we associate with setting $g = 0$. We refer to the parameter g as the *granularity* of the discretization.

Given such partitions defined by parameters $a_i, b_i, g_i (i = 1, \ldots, n)$ for each of the n dimensions, we obtain a finite partition of $\mathbb{R}^n$ into n-dimensional cells [3]. In order to be consistent with our notation for finite state spaces, and suppressing in the notation the parameters defining the discretizations, we denote with W the set of n-dimensional discretization cells. The discretization space W induces a discretization mapping

$$d : \mathbb{R}^n \to W$$

mapping a point $x \in \mathbb{R}^n$ to the element $w \in W$ for which $x \in w$. We extend the discretization mapping to subsets $A \subseteq \mathbb{R}^n$ as

$$\begin{array}{rccl} d: & 2^{\mathbb{R}^n} & \to & 2^W \\ & A & \mapsto & \{w \in W | w \cap A \neq \emptyset\} \end{array} \tag{15}$$

The random variable X with distribution P_θ and the coarsening variable F with conditional distribution $P_\gamma(F|X)$ now define the joint distribution of the W-valued variable $d(X)$, and a 2^W-valued variable Y as

$$P_{\theta,\gamma}(d(X) = w, Y = U) = \int_U \sum_{f \in \mathcal{F}: d(\zeta(x,f)) = U} dP_\theta(x). \tag{16}$$

This, now, is a distribution on the finite state space $\Omega(W)$ as in (1). From observations $\boldsymbol{U} = U^{(1)}, U^{(2)}, \ldots$ of Y we now want to estimate the true parameter θ^* of the continuous X. To justify the application of the techniques and consistency results developed in the previous sections in order to learn θ^* from the discretized data $\boldsymbol{U}$, we need to ascertain a couple of consistency properties between the original and the discretized models.

3. In order to prevent confusion with "bins" as encountered in Example 7, we refer to (multi-dimensional) intervals constructed for discretization purposes as "cells".

First, we note that we can factor $P_{\theta,\gamma}$ as

$$P_{\theta,\gamma}(w,U) = P_{\theta,\gamma}(w)P_{\theta,\gamma}(U|w) = P_\theta(w)P_{\theta,\gamma}(U|w), \tag{17}$$

where the important observation is that the marginal distribution $P_{\theta,\gamma}(w)$ only depends on the distribution of the original X, and thus only on the parameter θ (by a slight abuse of notation, we here use P_θ both for the distribution of X in $\mathbb{R}^n$, and of $d(X)$ in W). The conditional distribution $P_{\theta,\gamma}(U|w)$, on the other hand, depends not only on the parameter γ of the coarsening process, but also on θ. The following proposition establishes that the unconstrained coarsening model Λ_{sat} defined on the basis of the finite space $\Omega(W)$, corresponds exactly to the unconstrained class of finite coarsening variable models.

Theorem 9 *Let Λ_{sat} be as defined by (2) for the discretized space $\Omega(W)$. Then: $\lambda \in \Lambda_{\text{sat}}$ iff there exists a finite set $\mathcal{F}$, and a $\mathcal{F}$-valued random variable F with conditional distribution $P_\gamma(F|X)$, such that for all $\theta \in \Theta$: $P_{\theta,\gamma}(U|w) = \lambda_{w,U}$.*

Our first observation that $P_{\theta,\gamma}(w)$ only depends on the parameter of interest θ, in conjunction with the (trivial) right-to-left direction of Proposition 9 implies the validity of the assumption of Theorem 5 that the observed data $\boldsymbol{U}$ is sampled according to a distribution $m^* = P^{\downarrow\mathcal{Y}}_{\theta^*,\lambda^*}$ with θ^* our parameter of interest, and $\lambda^* \in \Lambda_{sat}$. The guarantee of Theorem 5 that in the large sample limit the true parameter θ^* will be among the optimizers of the profile likelihood LL_{sat}, thus still holds. The converse direction of Proposition 9, furthermore, shows that the restricted class of finite variable coarsening models in the continuous space does not imply any limitations on the coarsening model after discretization, and therefore Λ_{sat} is the appropriate coarsening model for the discretized data if for the original continuous data we make no assumption other than that it can be represented by a finite coarsening variable.

These results apply to all discretizations $\mathbb{D}_{a_i,b_i,g_i}$. We can even consider the vacuous discretization with $g_i = 0$ for all i, leading to the single discretization cell $\mathbb{R}$, and hence completely un-informative discretized observations $U = \mathbb{R}^n$. In this case we then have $\mathcal{C}(\boldsymbol{U}) = \Delta(W)$, and according to Corollary 4 every $\theta \in \Theta$ is a global maximum of $LL_{sat}(\cdot|\boldsymbol{U})$. Clearly, a certain granularity of the discretization is needed in order to still capture in the discretized data information about the distribution P_θ. On the other hand, for a given continuous coarse data set $A^{(1)}, \ldots, A^{(N)}$, too fine-grained discretizations lead to empirical distributions m of the discretized data $\boldsymbol{U} = d(A^{(1)}), \ldots, d(A^{(N)})$ that cannot be fit with $\{P_\theta|\theta \in \Theta\}$ as the underlying complete (continuous) data model. Our optimization objective $min_{c\in\mathcal{C}(\boldsymbol{U})}\, KL(P_c, P_\theta)$ does not provide a stand-alone tool for the selection of the discretization granularity: as a consequence of the decomposition property of $KL()$ stated as Lemma 10 in the appendix, this objective is non-decreasing when moving from discretized data $\boldsymbol{U}$ to data $\boldsymbol{U}'$ that is discretized at a finer level of granularity[4]. We will introduce in Section 7.1.1 a practical solution for the choice of discretization granularity by combining the objective $KL(P_c, P_\theta)$ with a penalty for "underfitting" discretizations.

4. In principle, when considering different discretized data sets $\boldsymbol{U}, \boldsymbol{U}'$ the term $H(m)$ in (3) must also be considered, as it is no longer constant. However, this term, too, is non-decreasing under refinements of the discretization.

7. Experiments

Like the EM algorithm, AIM is a general algorithmic paradigm whose application to concrete parametric models may still require a substantial algorithmic development and implementation effort. In the following we describe and experiment with AIM implementations for Gaussian data, and for parameter learning in Bayesian networks.

7.1 AIM for Gaussian Distributions

For our experiments we consider two types of continuous coarse data:

- *1-dimensional binned Gaussian data*: the complete data follows a 1-dimensional Gaussian distribution $N(\mu, \Sigma)$ with mean μ and variance Σ, which is coarsened by a random binning process as in Example 7. In order to obtain a uniform notation with the following 2-dimensional case, we use Σ rather than σ^2 to denote the variance. We refer to this setting as the *1d-b* scenario.

- *2-dimensional Gaussian data with missing values*: the complete data follows a 2-dimensional Gaussian distribution $N(\mu, \Sigma)$ with mean μ and covariance matrix Σ, which is coarsened by randomly missing values as in Example 6. We refer to this setting as the *2d-m* scenario.

7.1.1 Implementation

The implementation of AIM learning for continuous data consists of two distinct parts: the implementation of AIM for fixed discretizations $\mathbb{D}_{a,b,g}$, and the selection of a suitable discretization.

The core of the algorithm is the implementation of the AI step. As in generalized versions of the EM algorithm (Neal and Hinton, 1998), we do not attempt a full minimization of $KL(c, \theta_t)$ in the AI step (cf. Algorithm 1), but only an update of c_t that leads to a reduction of KL. These updates, in turn, are performed by a sequence of local optimizations.

Let $U^{(1)}, \ldots, U^{(N)}$ be a data set of discretized observations $d(f(x^{(i)}, g^{(i)}))$. Let $U_1, \ldots, U_M$ be the set of distinct observations in the data set, with empirical probabilities $m_1, \ldots, m_M$. In iteration t of the algorithm, we iterate once over the $U_i (1 \leq i \leq M)$, and update the fractional completion of the current U_i as follows. For $w \in U_i$ let

$$p_w := \sum_{j=1}^{i-1} c_t(U_j, w) + \sum_{j=i+1}^{M} c_{t-1}(U_j, w)$$

be the probability assigned to w according to the current completions of the observations other than U_i. Let

$$\boldsymbol{p} = (p_w)_{w \in U_i} \tag{18}$$

be the corresponding vector of probabilities, and

$$\pi := \sum_{w \in U_i} p_w \tag{19}$$

be the total probability mass currently committed to U_i by completions of observations other than U_i. We now calculate $c_t(U_i)$ as the completion that minimizes

$$KL(\frac{\pi \cdot \boldsymbol{p} + m_i c_t(U_i)}{\pi + m_i}, P_\theta | U_i). \tag{20}$$

Thus, we aim to re-distribute the empirical probability mass m_i of U_i over U_i, such that the probability distribution induced by the completions of all $U_j (1 \leq j \leq M)$ minimizes KL distance to the conditional distribution defined by the current P_θ over U_i. The minimization of (20) can be solved exactly (see Appendix B for the details). After termination of the iteration over all U_i, it is ensured that $KL(c_t, \theta_t) \leq KL(c_{t-1}, \theta_t)$.

The maximization step requires to find the maximum likelihood parameters μ, Σ given the probabilities assigned to the discretization cells w by the current completion c_t. This problem has no closed-form solution, and we use gradient descent optimization in the implementation of the M step.

For the selection of the discretization we proceed as follows: we only consider candidate discretizations with the same granularity g in all dimensions (only relevant for the 2d-m setting). The parameters a_i, b_i are always set to $\bar{\mu}_i - 3\bar{\sigma}_i$ and $\bar{\mu}_i + 3\bar{\sigma}_i$, respectively, where $\bar{\mu}_i, \bar{\sigma}_i$ are the empirical mean and standard deviation obtained from the observed values in dimension i. We run AIM for a number of candidate granularities $g_1, \ldots, g_K$. For each granularity, the AIM procedure is restarted several times with different initial parameter settings. The number of restarts is denoted $\#RS$. We then evaluate the outcomes for each granularity based on two criteria: first, we consider the minimal KL-score obtained across the different restarts. Second, we consider the variance in the parameter estimates $\theta_1, \ldots, \theta_{\#RS}$ obtained in the restarts. A high variance is an indication that the discretization is too coarse, allowing for many different optima of the objective. In order to balance these two criteria, we perform a 0-1 normalization of the obtained KL and variance values over the K different granularities. The final score for a given granularity then simply is the sum of the normalized KL and variance values. The parameter values from the best restart (according to KL score) at the granularity with minimal score is returned as the estimate.

7.1.2 Data

In all our experiments we generate data from Gaussian distributions with parameters $\mu = 0.5, \Sigma = 1.0$ in the 1-dimensional case, and $\mu = (0.5, 0.5), \Sigma = ((1,1)(1,2))$ in the 2-dimensional case. All our coarsening models are defined in terms of *coarsening probability functions* $cpf : \mathbb{R} \to [0,1]$ that in the 2d-m setting define the component-wise missingness probabilities $cpf(x) = P_\gamma(F_i = 1 | X = x)$ (cf. Example 6), and in the 1d-b define the binning probabilities $cpf(x) = P_\gamma(F = f_1 | X = x)$ (cf. Example 7). We use two different parametric families of coarsening probability functions. The first is defined by the parametric model

$$cpf(x) = \frac{\lambda_2}{1 + e^{(-\lambda_0(x - \lambda_1))}}, \tag{21}$$

where the denominator with parameters $\lambda_0, \lambda_1 \in \mathbb{R}$ defines a sigmoid-shaped function, and the numerator $\lambda_2 \in [0,1]$ controls the maximum missingness probability. For $\lambda_0 = 0$ this becomes a constant missingness probability of $\lambda_2/2$, leading to CAR models. With coarsening probabilities of the form (21) one obtains high probabilities for missing values for

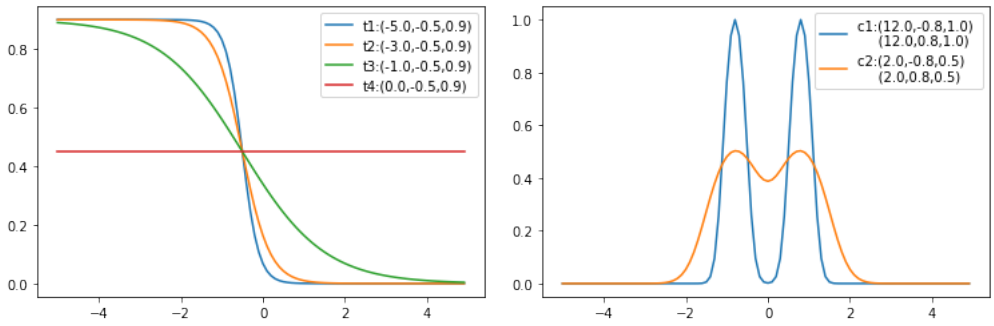

Figure 5: Tail (left) and bi-central (right) coarsening probabilities; λ parameter vectors shown in the legend

either large values of x, or for small values (when $\lambda_0 < 0$). We therefore refer to this model also as *tail coarsening*. Concrete instances $t_1, \ldots, t_4$ from this family used in our experiment are shown in the left plot of Figure 5. They are mostly distinguished by a decreasing slope of the function, which leads to increasingly near-CAR coarsening mechanisms (perfectly CAR for t_4).

The second form we consider is defined by

$$cpf(x) = 2 \cdot \frac{\lambda_2}{1 + e^{(\lambda_0 \cdot (x - \lambda_1)^2)}} \tag{22}$$

defined by $\lambda_0 \in \mathbb{R}^+, \lambda_1 \in \mathbb{R}$ and $\lambda_3 \in [0, 1]$. This gives a bell-shaped coarsening probability centered on λ_1, with a width parameter λ_0, and an upper bound λ_2. We refer to this model as *central coarsening*. In our experiments we use combinations of two central coarsening functions, which we call "bi-central", defined by two sets of $\lambda_1, \lambda_2, \lambda_3$ parameters each. The two bi-central functions c_1, c_2 we use are shown in the right plot of Figure 5. In c_2 the coarsening probabilities are more uniform, leading to a near-CAR nature of the resulting coarse data.

For the 1d-b scenario we also need to specify the bin partitioning of the real line. We conduct experiments with the very simple partitionings

$$\begin{aligned} \mathcal{B}_1 &: \quad]-\infty, \infty[, \\ \mathcal{B}_2 &: \quad]-\infty, -1],]-1, 1],]1, \infty[. \end{aligned}$$

The partitioning $\mathcal{B}_1$ means that when a value is reported only in binned form, then no information is provided other than that a measurement has been made (i.e., in this case the binned value is the same as a missing value). As our experiments will show, the AIM method can actually exploit such information on the mere existence of completely unobserved measurements.

For the 2d-m scenario, we first sample missingness indicator variables F_1, F_2 for the two components using the tail or bi-central coarsening probabilities. For this data we use a version of the EM algorithm for incomplete multi-variate Gaussian data that is restricted to data with at most one missing value in each observation (Little and Rubin, 1987, Section

8.2.1). In cases where we first obtain $F_1 = F_2 = 1$, we therefore randomly pick $i \in \{1, 2\}$ and reset $F_i = 0$ (not missing).

The specification of a 1d-b or 2d-m coarse data generation process in conjunction with a samplesize N constitutes an *experimental setting*.

7.1.3 Evaluation Metric

We measure the accuracy of an estimate $\hat{\mu}, \hat{\Sigma}$ for true parameters μ^*, Σ^* simply by the root mean squared error over all parameter components, where in the 2-d case we sum only over one of the two identical off-diagonal elements of Σ.

7.1.4 Experiment Design

We compare the accuracy of parameter estimates computed by EM, AIM, and a combination of the two. Both EM and AIM are somewhat sensitive to the initial parameter setting θ_0. A simple random initialization will often lead to near-zero probabilities of the data under P_{θ_0}, and hence to numerical underflow problems. We therefore pick initial parameters using *available case analysis (ACA)* on small sub-samples of the data: a random subset $A^{(i_1)}, \ldots, A^{(i_l)}$ of the coarse data set $A^{(1)}, \ldots, A^{(N)}$ is sampled. Then initial parameter values μ_0, Σ_0 are obtained as the empirical values from those $A^{(i_j)}$ where the relevant statistics are fully observed. Concretely, in the 1d-b case, empirical mean and variances are obtained from the exact (non-binned) observations among the $A^{(i_j)}$. In the 2d-m case, the components of μ_0 are obtained from all observations where the respective component is observed, whereas Σ_0 is estimated from the fully observed cases. In order to get a sufficient diversity in the initial parameter settings, we use small samples of size $l = 20$ in this process. As a point of reference, we also report estimation accuracies obtained from ACA on the full data set.

For a given experimental setting as described in Section 7.1.2, we conduct an *experimental run* as follows:

- generate data according to the experimental setting.

Then, for a number of restarts (denoted #RS):

- obtain initial μ_0, Σ_0 by ACA on a subsample of size $l = 20$
- run EM with initial parameters μ_0, Σ_0 resulting in estimates $\hat{\mu}_{EM}, \hat{\Sigma}_{EM}$.
- run AIM with initial parameters μ_0, Σ_0. For the 1d-b experiments, learning is performed for candidate granularities $(g_1, \ldots, g_K) = (3, 5, 10, 20, 50, 100)$. For the 2-d-m experiments, the discretization cell count is quadratic in the granularities, and therefore we only use $(g_1, \ldots, g_K) = (3, 5, 8, 12, 20)$.
- run AIM with initial parameters $\hat{\mu}_{EM}, \hat{\Sigma}_{EM}$ and the same discretization granularities as above.

One such run gives us four estimates, denoted ACA, EM, AIM, and EM-AIM, respectively:

ACA: the ACA estimate from the whole data set (identical in all restarts);

EM: the $\hat{\mu}_{EM}, \hat{\Sigma}_{EM}$ from the restart that obtained the highest face-value likelihood value;

AIM: from the $\#RS \cdot K$ different estimates computed in the run: select the optimal discretization according to the criterion described in Section 7.1.1, and return for this discretization the result from the restart obtaining the lowest KL score;

EM-AIM: as for AIM, but with the EM defined parameter initialization.

An *experiment* for a given experimental setting consists of 10 experimental runs. For each experiment we report the average errors of the four estimates and their standard deviations over the 10 runs.

7.1.5 RESULTS FOR 1D-B

We conduct experiments given by the settings defined by $\mathcal{B}_1$ and $\mathcal{B}_2$ binning, coarsening probability functions $t_1, \ldots, t_4, c_1, c_2$ as shown in Figure 5, sample sizes $N = 100, 1000, 10.000, 100.000, 1.000.000$, and #RS=5.

Figures 6 and 7 show the results for 1d-b experiments with $\mathcal{B}_1$ and $\mathcal{B}_2$ binning, respectively. Each individual plot corresponds to one coarsening probability functions, while different sample sizes are shown on the x-axes in the plots. The (expected) percentage of coarsened data items for the respective coarsening probabilities are shown in the headers of the plots. All plots show the root mean squared errors for the four different estimates, as well as the discretization granularities selected for the AIM and EM-AIM estimates according to our scoring function (all plots show averages and standard deviations over the 10 experimental runs). Tables with the exact numbers underlying these plots are given in Appendix D.

In the case of $\mathcal{B}_1$ binning the ACA estimate also is the unique fixpoint of the EM algorithm. The ACA and EM estimates coincide, and their error curves are on top of each other in Figure 6. For highly non-CAR data (coarsening probability functions t_1, t_2, c_1), AIM is significantly more accurate than EM, albeit with a relatively high standard deviation. For t_1, t_2 the EM-AIM combination achieves some improvement over the initial EM estimates, but the bias induced by the initial EM driven parameter setting is not fully overcome by the subsequent AIM iterations, and the results are less accurate than with pure AIM. When data becomes CAR (or more nearly so) with coarsening probability functions t_3, t_4, c_2, the ACA/EM estimates become very accurate and outperform the AIM estimates. Except for the very small sample size $N = 100$, the EM-AIM combination then is nearly indistinguishable from ACA/EM.

In the case of $\mathcal{B}_2$ binning (Figure 7) the EM algorithm can exploit the limited information contained in the coarse data items, and now mostly outperforms ACA by a large margin. For the two highly non-CAR cases t_1, t_2 AIM still has a small but consistent advantage over EM for the larger sample sizes. This is not the case for the non-CAR c_1. The explanation for the very strong performance of EM here probably is that even though the EM's expectation step here will make "incorrect" imputations for the coarse data items, the relevant statistics obtained from these imputations are still quite accurate. In all experiments with sample sizes $N \geq 10^4$ the EM-AIM combination performs at least as good as EM, and in the non-CAR cases often markedly better (though then sometimes much worse than pure AIM).

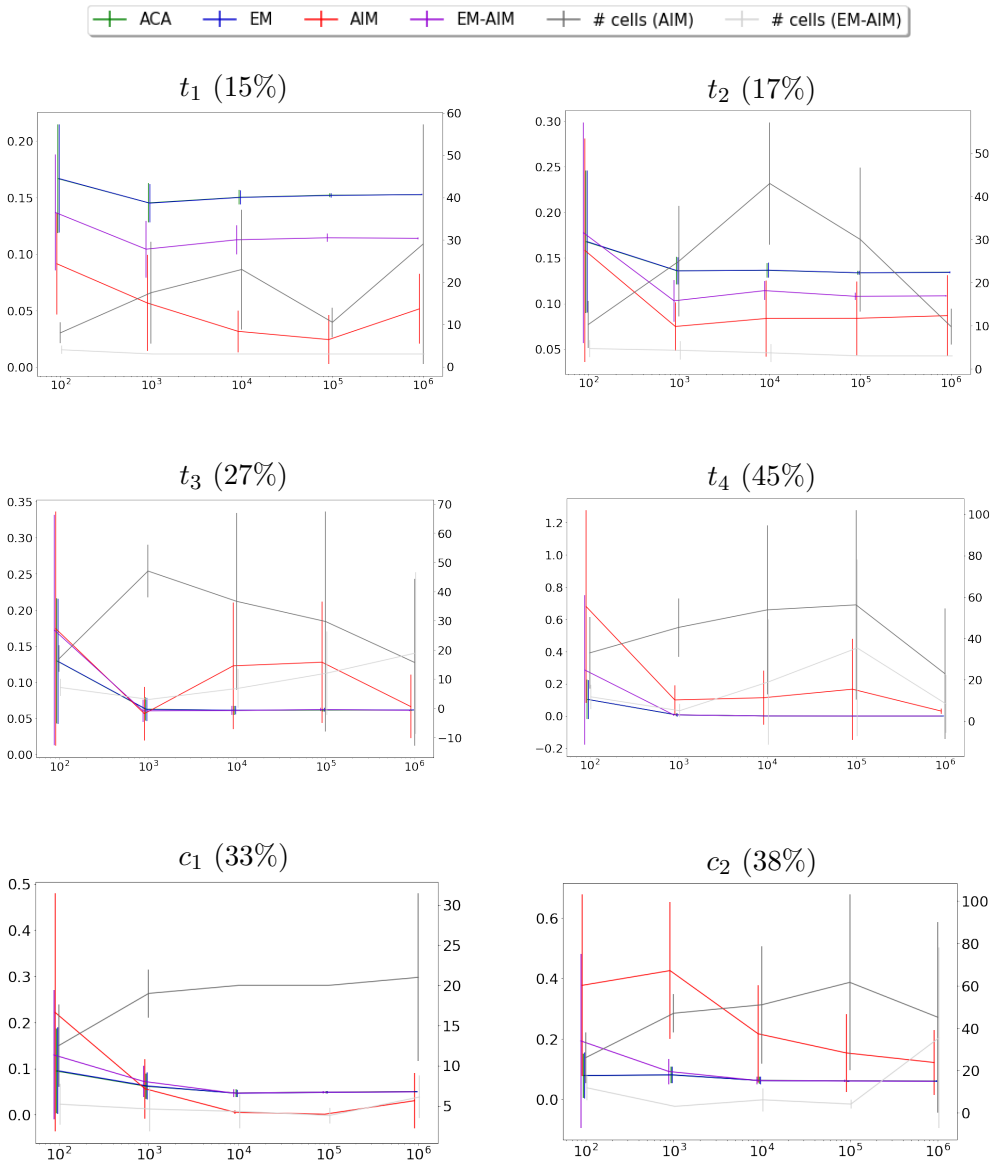

Figure 6: Results for 1d-b with $\mathcal{B}_1$. Left y-axis: error for ACA, EM, AIM and EM-AIM estimates; right y-axis: number of cells; x-axis: sample size in log-scale. Slight horizontal jitter applied to separate the error bars.

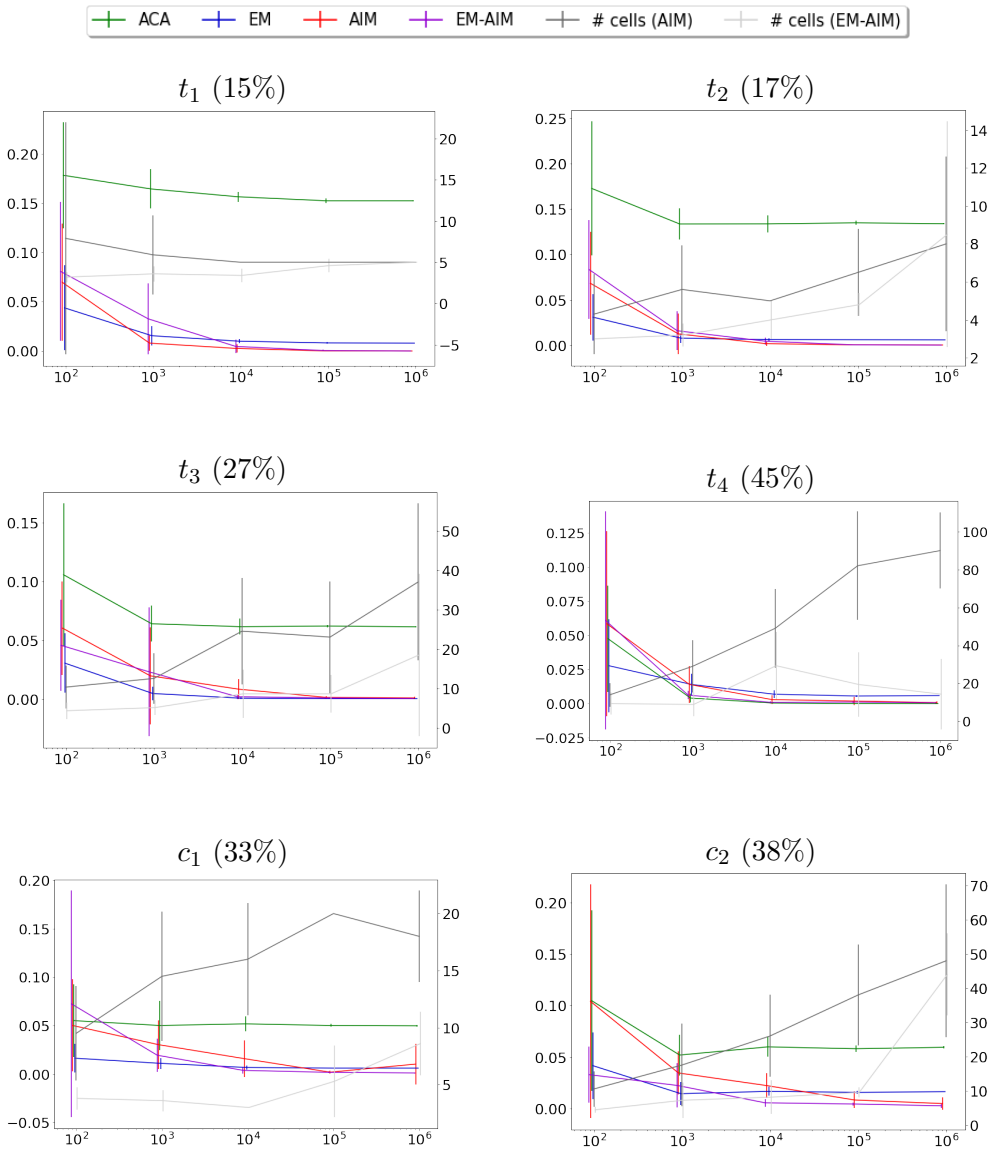

Figure 7: Results for 1d-b with $\mathcal{B}_2$. Left y-axis: error for ACA, EM, AIM and EM-AIM estimates; right y-axis: number of cells; x-axis: sample size in log-scale. Slight horizontal jitter applied to separate the error bars.

Regarding the selected discretization granularity, we observe that in most cases small granularities $g \leq 20$ were sufficient, and our largest candidate value $g = 100$ was rarely selected. The granularities for EM-AIM are smaller than for AIM. This is to be expected: the EM initialization leads to far smaller variance in the initial parameter settings than the ACA initialization, and therefore also to a smaller variance in the final estimates. Thus, the term that penalizes higher variance of estimates in coarser discretization plays a smaller role in EM-AIM than in AIM.

In some experiments we notice an unexpected deterioration of the AIM results at the maximal samplesize $N = 1m$. A closer examination of this phenomenon shows that this can be attributed to sub-optimal choices of the granularities in these cases: the results obtained for any fixed granularity g improve quite consistently with increasing samplesize (these results not pictured in the plots). However, for $N = 1m$ our scoring function for the candidate granularities sometimes led to choices that were too coarse to provide accurate estimates. This problem can be attributed to the fact that the two signals we use in the scoring function become more uniform across granularities as the samplesize increases: the KL score then becomes very small also for large g as the empirical distribution can then be fitted very precisely over these finer discretizations. Also, the variance component in the score becomes rather more uniform across granularities, as the likelihood function shows fewer local optima induced by sampling noise, and thus convergence to one of the global maxima becomes more consistent. As a result, the selection of the granularity becomes more "random" for the largest N. This is also visible in the plots by an increase in the standard deviation of the selected g value at $N = 1m$ compared to $N = 100k$, especially in those cases where there is an increase in the AIM error value.

Another noteworthy observation is that the AIM estimates actually become less accurate as the coarsening probabilities become more CAR: compare, in particular, t_1 vs. t_4 and c_1 vs. c_2 in the $\mathcal{B}_1$ experiments. To explain this phenomenon and obtain further insights into the working of AIM, we show in Figure 8 detailed snapshots of the initial (top row) and final (bottom row) configurations of AIM for one restart with t_1 coarsening, and two restarts with t_4 coarsening. Here the discretization granularity is $g = 50$, and $N = 100k$. The plots show the current estimates P_θ and completions P_c. In all cases there is a near perfect match ($KL(P_\theta, P_c) \approx 0$) between P_θ and P_c in the final configuration, rendering the two distribution curves indistinguishable. The completion P_c is the sum of the discretized exact observations $d(\{x^{(i)}\})$, and the completions of the discretized vacuous observations $d(]\infty, \infty[)$. The former define a fixed distribution over the discretization cells that the AI operations cannot alter. This component of P_c is plotted by the yellow curves in Figure 8, which are identical in the top and bottom plots. The total mass of the binned (vacuous) observations can be freely distributed by the AI operations over the discretizations cells. The resulting component of P_c is plotted by the red curves in Figure 8. Roughly speaking, the AIM algorithm tries to distribute the available mass for the red curve such that the sum of the red and yellow curves form a Gaussian distribution (cf. Figure 1; the red curves in Figure 8 correspond to the heatmaps in Figure 1 (c)) . In the t_1 case, the yellow curve has a steeper slope on the left side than the right, because the values $x^{(i)}$ that were coarsened to $]-\infty, \infty[$ almost all are in the negative range. The only way to complete the yellow curve to a Gaussian distribution is to distribute a significant amount of probability mass over the interval $[-2, 0]$. This is what consistently happens during the AIM iterations, even

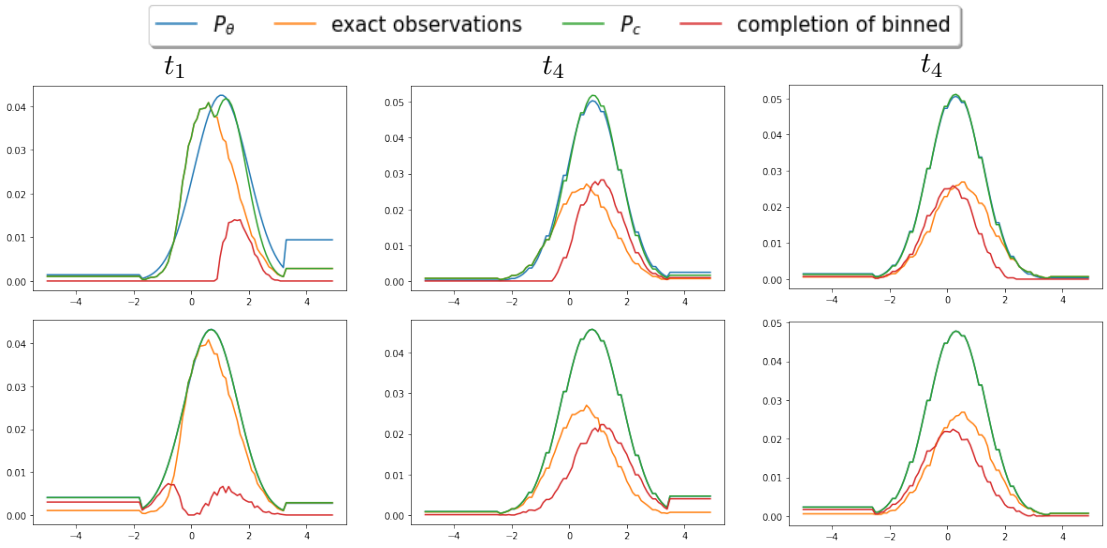

Figure 8: Illustration of AIM iterations, $n = 100000$, $m = 50$, incompl: 0.339

when the initial completion is concentrated on a very different range, as in the run shown in Figure 8. For t_4 coarsening, the distribution defined by the exact observations already forms a Gaussian (and, in fact, approximately the true generating Gaussian, making the ACA/EM estimates very precise in this case). This, in conjunction with the fact that for t_4 the percentage of coarse data items is high, leads to many degrees of freedom to complete the yellow curve to a Gaussian distribution. In the two restarts shown in Figure 8 one restart distributed the disposable probability mass as another near-Gaussian distribution shifted to the right of the yellow curve, while the other restart ended with a near-Gaussian to the left. In both cases, the sum of the yellow and red curves are again a Gaussian, thus both being optimal solutions for the AIM objective.

A different perspective on essentially the same phenomenon is given in Figure 9. It shows the outcomes of three experimental runs with settings defined by $\mathcal{B}_1$ binning, coarsening probability functions c_1, c_2, and sample size $N = 1m$. Different from our the experiments pictured in Figure 6, a large number of restarts $\#RS = 50$ was used. The plots show for each restart in each run (distinguished by red, blue, green colors) the error of the obtained solution on the x-axis vs. the value of the KL score for that solution on the y-axis. The results for the restarts with minimal KL score in each run are marked by larger squares. For the highly non-CAR c_1 coarsening there is a clear monotone relationship between the KL score and the error, and KL score minimization is very successful in identifying low error solutions. In contrast, for c_2 coarsening there is little correlation between KL score and error. The KL scores are overall much smaller than for c_1: the scale on the y-axis is one order of magnitude smaller in the c_2 plot than in the c_1 plot, and the values for the points clustered near the x-axis are several orders of magnitude smaller still. Similar to what we observed for t_4 in Figure 8, this indicates many different near-perfect solutions with widely varying actual errors (note that for c_2 the x-axis is in log-scale, extending to error values > 1).

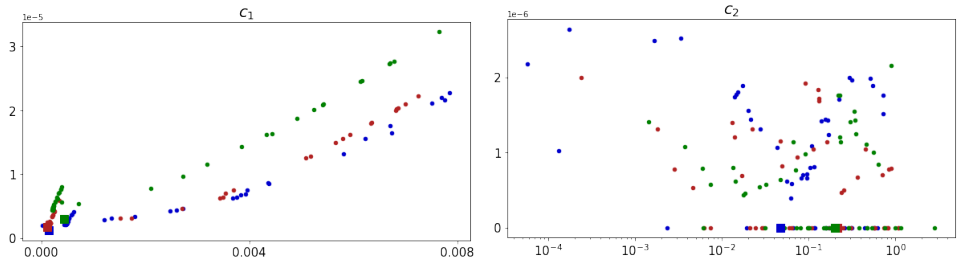

Figure 9: Details for 1d-b with central coarsening and $\mathcal{B}_0$, $N = 1m$. Results from 3 experiments with 50 restarts each. x-axis: error values (log-scale in the right plot), y-axis: KL score

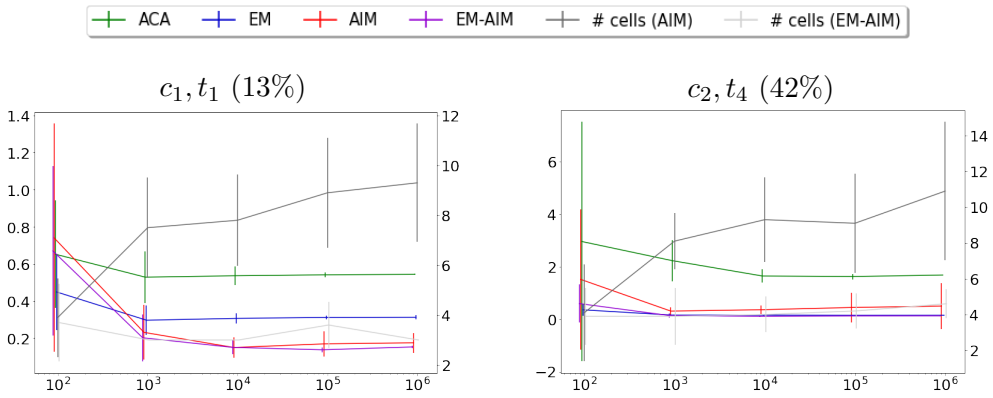

Figure 10: Results for 2d-m experiment

7.1.6 Results for 2d-m

We conduct experiments with two different coarsening models as described in Section 7.1.2. We use again the coarsening probability functions shown in Figure 5 to define the probabilities that values are missing. In the first model c_1 defines the missingness probabilities in the first component, and t_1 the missingness probabilities in the second component. In the second model we use c_2 and t_4, respectively. The intention is that the (c_1, t_1) model is highly non-CAR, whereas the (c_2, t_4) model is near-CAR. Results from one experimental run with the (c_1, t_1) model are pictured in Figure 1. For the purpose of a more detailed visualization, that figure is based on a finer granularity ($g = 50$) than used in the systematic experiments.

Figure 10 shows the results for the two coarsening models and the same range of data set sizes as in the 1d-b experiments. As in the previous experiments, we observe a substantial advantage of AIM over EM for the highly non-CAR data (c_1, t_1), and the opposite for the near-CAR (c_2, t_4). The range of observed errors (y-axis) is significantly larger in the (c_2, t_4) experiments than for (c_1, t_1), which makes the visual comparison between the plots a bit difficult. The precise numbers (cf. Table 6 in Appendix D) show that the gap between EM

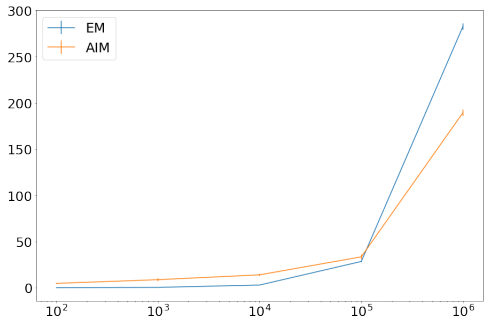

Figure 11: Runtimes for 1d-b experiment with $\mathcal{B}_2$ bins and t_1 coarsening. y-axis: time in ms, x-axis: data set size.

and AIM is actually a little larger in (c_2, t_4) than in (c_1, t_1), and that as already observed in the 1d-b experiments, the AIM results are actually less accurate for the near-CAR data (c_2, t_4) than for (c_1, t_1), with the converse being true for EM. In these experiments the EM-AIM combination was particularly successful, almost always being at least as good, or better than both EM and AIM.

7.1.7 Computation Time

Figure 11 shows exemplary computation times for the experimental setting 1d-b with $\mathcal{B}_2$ bins and t_1 coarsening and our usual range of data set sizes. The time shows averages and standard deviations for a single experiment, which includes the 5 restarts, and, for the AIM case, the iteration over the 6 different candidate granularities. Not least due to this additional iteration over granularities, AIM generally requires more computation time. However, for very large data sets, EM suffers from the disadvantage that it iterates over all data cases, whereas AIM only iterates over the data aggregates associated with the discretization cells. A closer examination of the operations that dominate the computation time reveals that these results are of limited significance, however: the AIM procedure spends about 95-99% of its time in the M step, whereas the KL optimizations of the AI step are quite fast. The EM algorithm spends most of its time in calculating the final face-value likelihoods of solutions obtained in the restarts. In both cases, the computational bottlenecks are operations that require repeated calls to a library function that computes values of the cumulative distribution function of the Gaussian distribution (in our implementation: the `norm` and `multivariate_normal` modules of the `scipy.stats` library). Thus, the use of a different, perhaps less accurate, implementation of the Gaussian cumulative distribution function could lead to a very different picture of the computation times.

7.2 AIM for Bayesian Networks

In a second suite of experiments, we consider Bayesian network models for discrete data, and coarse data that takes the form of data with missing values.

7.2.1 Implementation

As in the implementation for the Gaussian case, we implement the AI step only in an incremental fashion that leads to a reduction of $KL(\cdot, \theta_t)$, rather than a full minimization. A Bayesian network for discrete variables $\boldsymbol{X} = X_1, \ldots, X_n$ and associated state spaces $W(X_i)$ leads to a state space W with cardinality $\prod_i |W(X_i)|$. Compared to the Gaussian case, where we could control the cardinality of W via the discretization granularity, this poses the additional challenge that it becomes infeasible to construct and maintain fractional completions over the full space W. We therefore bias the AI step towards sparse fractional completions $c(U_i)$ with a small support allowing for an explicit representation. Similar to our strategy in the Gaussian case, the reduction of $KL(\cdot, \theta_t)$ is broken down into a sequence of local KL minimization operations for which simple, exact solutions exist. Details of this implementation are described in Appendix C.

In our implementation we use the HUGIN system[5] for basic datastructures and algorithms for Bayesian networks. In particular, we use the HUGIN implementation of the EM algorithm to compute EM parameter estimates.

7.2.2 Data

A given Bayesian network complete data model is augmented with a coarsening model as already seen in Example 2: for each variable $X_i \in \boldsymbol{X}$ a binary variable obs_X_i with states o (observed) and m (missing) is introduced (cf. Example 2). The set of Bayesian network parents of obs_X_i is randomly created such that parents of obs_X_i can be both original variables X_h, and other obs_X_h variables. X_i always is made a parent of obs_X_i, thus encouraging non-MAR missing data patterns.

For all configurations of the parent nodes of obs_X_i a conditional probability value $p = P(obs\_X_i = m \mid Pa(obs\_X_i))$ is sampled from a beta-distribution $\text{Beta}(\alpha, \beta)(p) \sim \Gamma(\alpha + \beta)/(\Gamma(\alpha)\Gamma(\beta))p^{\alpha-1}(1-p)^{\beta-1}$. In order to better control properties of interest, we specify Beta-distributions directly in terms of their mean μ and variance σ^2, rather than their usual α, β parameters. The mean value of the Beta-distribution determines the expected proportion of missing values in the data. The variance controls how far the coarsening process is from being missing at random: $\sigma^2 = 0$ (which does not correspond to a proper Beta-distribution) means that $P(obs\_X_i \mid Pa(obs\_X_i))$ is constant equal to μ, independent of $Pa(obs\_X_i)$, and the resulting coarsening mechanism is MAR. Larger values of σ^2 lead to a greater diversity in the $P(obs\_X_i \mid Pa(obs\_X_i))$ values for different configurations of $Pa(obs\_X_i)$, and thereby to potentially more complex missingness patterns (though there is no guarantee that each coarsening model generated with $\sigma^2 > 0$ is distinctly non-MAR). At $\sigma^2 = 0.15$ (the maximal σ^2 value we consider in the experiments) all conditional probabilities $P(obs\_X_i = m \mid Pa(obs\_X_i))$ are highly concentrated at the extremes 0 and 1.

An *experimental setting* now consists of an underlying Bayesian network for the complete data, a coarsening model defined by the μ, σ parameters of the Beta distribution, and a data set size.

5. `https://www.hugin.com/`

7.2.3 Evaluation Metric

We use a standard indexing scheme for the parameters θ in a Bayesian network and let $\theta = (\theta_{ijk})$, where θ_{ijk} is $P_\theta(X_i = j \mid Pa_i = k)$, i.e. the conditional probability that the ith variable in the network is in state j, given that its parents are in their (joint) kth state. The root means squared error over all model parameters would be a very poor error measure in this case, since different parameters in the Bayesian network may have a very different impact on the distribution that is defined. A meaningful measure for comparing the learned parameter $\hat{\theta}$ with the true parameter θ^* is given by the Kullback-Leibler distance $KL(P_{\theta_0}, P_{\hat{\theta}})$. However, this poses some difficulties due to the sensitivity of KL distance to small deviations at or near zero parameters. For example, when learning parameters for the Alarm network, a typical outcome that we observed is that for a cpt-row with true parameters $\theta^*_{i\bullet k} = (0.01, 0.01, 0.01, 0.97)$ the parameters learned by EM and AIM are, respectively, $\theta^{EM}_{i\bullet k} = (1.1 \cdot 10^{-16}, 3.0 \cdot 10^{-16}, 2.0 \cdot 10^{-16}, 1)$ and $\theta^{AIM}_{i\bullet k} = (0, 0, 0, 1)$. While AIM and EM here seem to agree in their result (very likely, EM was on its way to converge to $(0, 0, 0, 1)$ when it terminated), an evaluation based on $KL()$ would give very different results: since AIM here estimates some parameters as zero which in the true model are non-zero, one obtains $KL(P_{\theta^*}, P_{\theta^{AIM}}) = \infty$. For EM, on the other hand, the difference between the estimates of $\sim 10^{-16}$ and the true values 0.01 here has little influence on the $KL(P_{\theta^*}, P_{\theta^{EM}})$ distance (which, in particular, remains finite).

In order to avoid these problems with KL distance, we conduct our analysis based on the following *weighted absolute error* function:

$$WAE(\theta, \theta') := \sum_i \sum_k P_\theta(Pa_i = k) \sum_j \theta_{ijk} \, | \, \theta_{ijk} - \theta'_{ijk} \, | \tag{23}$$

WAE has a similar structure as KL: by replacing $| \, \theta_{ijk} - \theta'_{ijk} \, |$ with $log(\theta_{ijk}/\theta'_{ijk})$ in (23) one obtains $KL(P_\theta, P_{\theta'})$. Apart from being more robust with respect to parameter values near zero, WAE values also have a more intuitive interpretation: $WAE(\theta, \theta')$ is the average absolute difference between true and estimated parameter values, where the average is weighted according to the frequency with which the parameters are required to compute the probability of a random sample generated from P_θ. Like KL, WAE also is not symmetric, and makes sense only when its two arguments have asymmetric roles as the true/correct parameter (first argument) and its approximation (second argument).

7.2.4 Experiment Design

For a given experimental setting we run 50 experiments along the same lines as described in Section 7.1.4. In each experiment the number of restarts is 10. We again compare the four methods ACA, EM, AIM, and EM-AIM.

7.2.5 Results

In a first experiment we test the implementation on the network of Example 2. In this case we only use the fixed coarsening model described in the example, not the random construction described in Section 7.2.2. The WAE between the generating model and the EM solution $\hat{\theta} = (0.5, 0.2727)$ obtained from ideal data (cf. Example 3) is $0.0727/2 =$

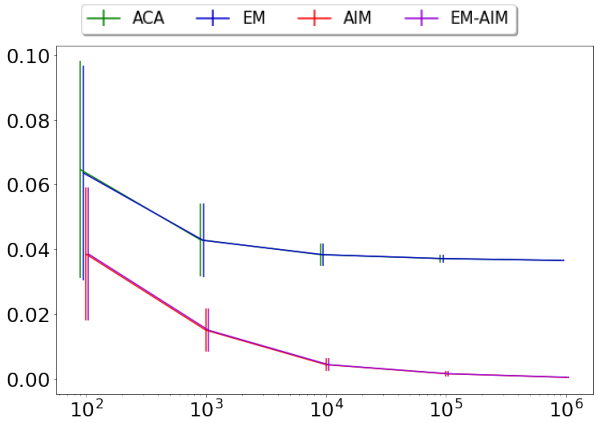

Figure 12: Results for Example 2. y-axis: *WAE* error; x-axis: data set size.

	$N =$				
	100	1000	10k	100k	1m
EM	8	24	128	982	9908
AIM	32	69	103	191	248

Table 3: Computation times (ms) for Asia at $\sigma^2 = 0.1$

0.03636. The results shown in Figure 12 show that both ACA and EM converge to this error, whereas AIM and EM-AIM approach zero error.

We now investigate the performance of EM and AIM depending on the *MAR* characteristics of the data. For this we use as the underlying complete data model the traditional "Asia" network used extensively in the Bayesian network literature. This is a network with 8 binary variables, so that $|W| = 256$. We construct coarsening models defined by Beta parameters $\mu = 0.2$, and σ^2 ranging from 0 to 0.15. Figure 13 shows the *WAE* scores as a function of data set size for different σ^2. As in the experiments with Gaussian data, we observe the expected advantage of EM for *MAR* data ($\sigma^2 = 0$), which turns into an advantage for AIM as the data becomes increasingly non-*MAR*. We do not observe, as in the Gaussian case, that the AIM values actually are more accurate in the non-*MAR* settings (cf. the exact numerical outcomes in Table 7 in the appendix). In spite of the wide and overlapping standard deviation error bars, most of the differences visible in the plots are statistically significant (Wilcoxon test, $\alpha = 0.01$). This is because the errors are highly correlated: some data sets are "easy" for both methods, some are "hard" for both, and the methods with the lower errors on average generally perform consistently better across most data sets.

7.2.6 Computation Time

Table 3 shows the average runtime per restart for the experiment with the Asia network at $\sigma^2 = 0.1$. These numbers are somewhat misleading, however, as for AIM they do not include the transformation of the original raw data set $U^{(1)}, \ldots, U^{(N)}$ into a sequence of distinct observations $U_1, \ldots, U_M$ ($M \leq N$) with associated empirical probabilities $m_1, \ldots, m_M$, over

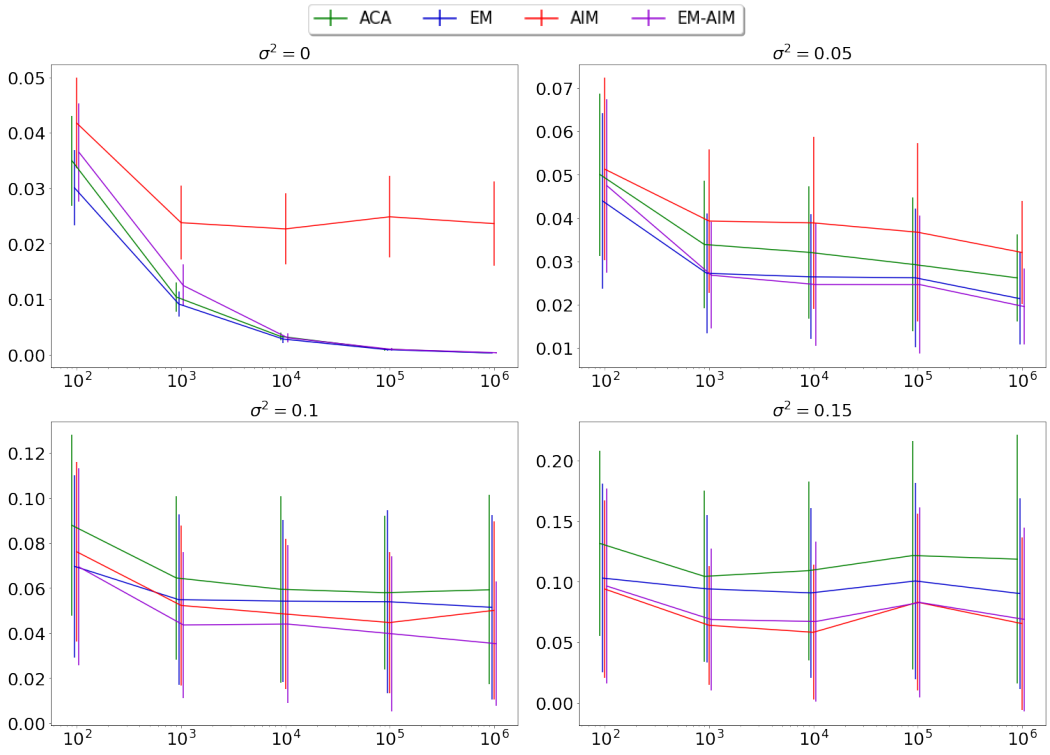

Figure 13: Experiments with Asia network. *WAE* scores (y-axis) vs. samplesize (x-axis) for different coarsening models defined by σ^2 parameter.

which AIM subsequently operates (cf. implementation details described in Appendix C). The EM algorithm, on the other hand, iterates (as in the Gaussian case) over the raw data, and thus scales roughly linearly in N. However, the EM algorithm could also be modified so as to only iterate over the aggregated data U_i, m_i. We therefore should only very cautiously conclude that here the computation times for EM and AIM are somewhat comparable.

This comparable performance does not extend to much larger networks with many more variables than Asia, and thus a much larger size $|W|$ of the state space. For such models the size M of the aggregated data will usually not be much smaller than the original N, because most of the raw observations $U^{(j)}$ will be unique. More importantly, the completions $c(U_i)$ now have to be constructed over the very large W, which in spite of our bias towards sparse completions leads to an optimization problem over a combinatorial space. The fundamental advantage of the EM algorithm over AIM here lies in the fact that when probabilistic inference in the given Bayesian network is tractable, then EM can directly compute the expected sufficient statistics required for the subsequent maximization step, and can bypass an explicit construction of a fractional completion of the data set, thus avoiding that $|W|$ enters as a relevant complexity parameter. Even though AIM, too, in the end only requires the same sufficient statistics as EM, there does not seem to be a way to obtain these statistics of the AIM completions, without actually constructing the completions. This advantage of EM only holds as long as exact inference in the underlying Bayesian network is tractable.

When this is not the case, and the E step has to rely on approximate techniques such as MCMC sampling, then EM, too, has to operate in the space W.

7.3 Summary of Experimental Results

We can summarize the main insights gained from our experiments as follows:

- The theoretical consistency and convergence results for EM and AIM are reflected in the outcomes: AIM is more accurate than EM for learning parameters from non-CAR data, whereas the opposite is true for CAR data.
- In some cases AIM actually benefits from a highly non-CAR nature of the data. Roughly speaking, this is the case when there are few degrees of freedom to complete the coarsened data such that the result complies with the underlying parametric model.
- The discretization approach for continuous data is effective for accurate parameter estimates and computationally quite efficient due to the fact that relatively coarse grained discretizations are sufficient.
- An optimal and efficient choice of discretization granularity is still an open problem.
- Initializing AIM with EM estimates is a quite robust approach to deal with both CAR and non-CAR data.

8. Conclusion

We have studied in detail theoretical properties of the AIM and EM algorithms, and clarified their relationship. We have shown that AIM provides consistent and convergent parameter estimates for non-CAR data. Consistency, here, however has a weaker meaning than in the usual context of maximum likelihood inference for identifiable models: the likelihood function L_{sat} may have multiple maxima, and we are only guaranteed that in the large sample limit the true parameter will be among them. The number and diversity of such maxima depends on the underlying parametric complete data model: under a saturated complete data model, every possible data completion corresponds to a distinct likelihood maximum (Corollary 4). Under more restricted models, on the other hand, the true parameters may be the unique maximum of L_{sat}, and then are reliably learned by AIM (Example 2). Unlike other approaches for dealing with non-CAR data, our approach is completely non-parametric with regard to the unknown coarsening mechanism.

We have developed a concrete instantiation of the AIM paradigm for learning the parameters of Gaussian distributions from binned data and data with missing values, as well as for Bayesian network parameter learning. The source code for the experiments of Section 7.1 is available at `https://github.com/manfred-jaeger-aalborg/aim_for_gauss`. Experimental results demonstrate the differences and respective advantages of AIM and EM learning from CAR and non-CAR data. The combination of EM and AIM learning in the EM-AIM version provides a quite robust approach that for CAR data inherits the accuracy and (to some extent) the computational efficiency of EM, and for non-CAR data improves on the accuracy of EM.

The most important problems for future work are to further optimize the discretization strategy for continuous data, and to answer the question whether it could be possible for AIM to operate directly in the space of sufficient statistics, rather than concrete data completions. Progress on the first problem certainly is possible. An obvious improvement over our current approach would be a more dynamic construction of discretizations instead of the iteration over a fixed set of candidate granularities. The answer to the second problem is conjectured to be negative, however.

Appendix A. Proofs

First, we recall a well-known decomposition property of KL (cf. e.g. (Kullback, 1968, Corollary 3.2)), that we will use repeatedly.

Lemma 10 *Let W be a finite space, and $\mathcal{V} = \{V_1, \ldots, V_k\}$ be a partitioning of W. For a probability distribution P on W denote by $P^{\downarrow\mathcal{V}}$ the marginal distribution of P on $\mathcal{V}$, and by $P \mid V_i$ the conditional distribution on V_i. Then, for any two distribution P, Q:*

$$KL(P, Q) = KL(P^{\downarrow\mathcal{V}}, Q^{\downarrow\mathcal{V}}) + \sum_i P(V_i) KL(P \mid V_i, Q \mid V_i). \tag{24}$$

In particular, if $P \mid V_i = Q \mid V_i$ for all i, then $KL(P, Q) = KL(P^{\downarrow\mathcal{V}}, Q^{\downarrow\mathcal{V}})$.

Next, we note a simple lemma regarding continuity of KL:

Lemma 11 *Let $\boldsymbol{p}, \boldsymbol{q}, \boldsymbol{p}_i, \boldsymbol{q}_i \in \Delta(W)$ $(i \geq 1)$ with $\lim_{i\to\infty} \boldsymbol{p}_i = \boldsymbol{p}$, $\lim_{i\to\infty} \boldsymbol{q}_i = \boldsymbol{q}$, and $KL(\boldsymbol{p}, \boldsymbol{q}) < \infty$, $KL(\boldsymbol{p}_i, \boldsymbol{q}_i) < \infty$ $(i \geq 1)$. Then*

a. $\lim_{i\to\infty} KL(\boldsymbol{p}, \boldsymbol{q}_i) = KL(\boldsymbol{p}, \boldsymbol{q})$

b. $\lim_{i\to\infty} KL(\boldsymbol{p}_i, \boldsymbol{q}) = KL(\boldsymbol{p}, \boldsymbol{q})$

c. *lim inf*$_{i\to\infty} KL(\boldsymbol{p}_i, \boldsymbol{q}_i) \geq KL(\boldsymbol{p}, \boldsymbol{q})$

The proof of **a.** and **b.** is elementary. Part **c.** follows from (Kullback, 1968, Chapter 4, Theorem 2.1). An example where **c.** does not hold with equality is given by $\boldsymbol{p}_i = (1 - 1/i, 1/i), \boldsymbol{q}_i = (1 - e^{-i}, e^{-i})$. Then $\lim_{i\to\infty} \boldsymbol{p}_i = (1, 0), \lim_{i\to\infty} \boldsymbol{q}_i = (1, 0)$, $\lim_{i\to\infty} KL(\boldsymbol{p}_i, \boldsymbol{q}_i) = 1 > KL(\boldsymbol{p}, \boldsymbol{q}) = 0$.

Theorem 3 *Let $\boldsymbol{U} = U^{(1)}, \ldots, U^{(N)}$ be a data set, and m the empirical distribution defined by $\boldsymbol{U}$ on $\mathcal{Y}$. Then*

$$\frac{1}{N} LL_{\text{sat}}(\theta \mid \boldsymbol{U}) = -H(m) - \min_{c \in \mathcal{C}(\boldsymbol{U})} KL(P_c, P_\theta). \tag{7}$$

Proof Throughout this proof, maximizations and minimizations for λ range over Λ_{sat}. We first observe that if $m(U) > 0$ for some U with $P_\theta(U) = 0$, then both sides of (7) are $-\infty$. From now on we assume that no such U exists. Then

$$\begin{aligned}\frac{1}{N} LL_{sat}(\theta \mid \boldsymbol{U}) = \frac{1}{N} max_\lambda \sum_{i=1}^{N} log P^{\downarrow\mathcal{Y}}_{\theta,\lambda}(U^{(i)}) = max_\lambda \sum_{U \subseteq W} m(U) log P^{\downarrow\mathcal{Y}}_{\theta,\lambda}(U) \\ = -H(m) - min_\lambda KL(m, P^{\downarrow\mathcal{Y}}_{\theta,\lambda}).\end{aligned}$$

Let $m^{\uparrow_\theta \Omega}$ be the extension of m to Ω defined by $m^{\uparrow_\theta \Omega}(x, U) = m(U)P_\theta(x \mid U)$. Thus, $m^{\uparrow_\theta \Omega}$ can be understood as the expected joint distribution of X and Y under P_θ given observations m (to be distinguished from the expected complete data, which would be $m^{\uparrow_\theta \Omega}$ marginalized on W). According to Lemma 10 we have for all λ:

$$KL(m, P_{\theta,\lambda}^{\downarrow \mathcal{Y}}) = KL(m^{\uparrow_\theta \Omega}, P_{\theta,\lambda}),$$

and

$$KL(m, P_{\theta,\lambda}^{\downarrow \mathcal{Y}}) \leq KL(m^{\uparrow_* \Omega}, P_{\theta,\lambda}),$$

for any (other) possible extension $m^{\uparrow_* \Omega}$ of m to Ω. Thus, letting $m^{\uparrow_* \Omega}$ range over all possible extensions of m,

$$min_\lambda KL(m, P_{\theta,\lambda}^{\downarrow \mathcal{Y}}) = min_\lambda min_{m^{\uparrow_* \Omega}} KL(m^{\uparrow_* \Omega}, P_{\theta,\lambda}) = min_{m^{\uparrow_* \Omega}} min_\lambda KL(m^{\uparrow_* \Omega}, P_{\theta,\lambda}). \quad (25)$$

Again by Lemma 10, for a given $m^{\uparrow_* \Omega}$, the minimum over λ at the right-hand side of (25) is attained for $\hat{\lambda}$ defined by $\hat{\lambda}_{w,U} := m^{\uparrow_* \Omega}(U \mid x)$, and

$$KL(m^{\uparrow_* \Omega}, P_{\theta,\hat{\lambda}}) = KL(m^{* \downarrow W}, P_{\theta,\hat{\lambda}}^{\downarrow W}) = KL(m^{* \downarrow W}, P_\theta).$$

Thus,

$$min_\lambda KL(m, P_{\theta,\lambda}^{\downarrow \mathcal{Y}}) = min_{m^{\uparrow_* \Omega}} KL(m^{* \downarrow W}, P_\theta) = \min_{c \in \mathcal{C}(\boldsymbol{U})} KL(P_c, P_\theta),$$

which concludes the proof. ■

Theorem 5 *Let $U^{(1)}, U^{(2)}, \ldots$ be observations of iid random variables $Y^{(i)}$, which are distributed according to $m^* := P_{\theta^*,\lambda^*}^{\downarrow \mathcal{Y}}$. Denote by P the joint distribution of all $Y^{(i)}$, and $\boldsymbol{U}_N = (U^{(1)}, \ldots, U^{(N)})$.*

A. *Let $\hat{\theta}_N$ maximize $LL_{\text{sat}}(\cdot \mid \boldsymbol{U}_N)$ ($N \geq 1$). Then*

$$\lim_{N \to \infty} \frac{1}{N}(LL_{\text{sat}}(\hat{\theta}_N \mid \boldsymbol{U}_N) - LL_{\text{sat}}(\theta^* \mid \boldsymbol{U}_N)) = 0 \quad P\text{-a.s.}.$$

B. *Assume that Assumptions 1 and 2 hold, and that $\lambda^* \in \Lambda_{\text{car}}$.*

(i) *Let $\hat{\theta}_N$ maximize $LL_{\text{car}}(\cdot \mid \boldsymbol{U}_N)$ ($N \geq 1$). Then*

$$\lim_{N \to \infty} \frac{1}{N}(LL_{\text{car}}(\hat{\theta}_N \mid \boldsymbol{U}_N) - LL_{\text{car}}(\theta^* \mid \boldsymbol{U}_N)) = 0 \quad P\text{-a.s.}.$$

(ii) *If, furthermore, the system of equations*

$$P_\theta(U) = P_{\theta^*}(U) \quad (U : m^*(U) > 0)$$

has the unique solution $\theta = \theta^*$, then

$$\lim_{N \to \infty} \hat{\theta}_N = \theta^* \quad P\text{-a.s.}$$

Proof The proof consists of two distinct parts for parts **A.** and **B.** of the theorem, with little overlap in the arguments. Both parts, however, rely on the strong law of large numbers applied to the empirical distributions $m(\boldsymbol{U}_N)$, which, in the following, is denoted m_N:

$$\lim_{N\to\infty} m_N = m^* \quad P\text{-a.s.} \tag{26}$$

A. By Theorem 3 we have

$$\frac{1}{N}(LL_{sat}(\hat{\theta}_N \mid \boldsymbol{U}_N) - LL_{sat}(\theta^* \mid \boldsymbol{U}_N)) = KL(\mathcal{C}(\boldsymbol{U}_N), P_{\theta^*}) - KL(\mathcal{C}(\boldsymbol{U}_N), P_{\hat{\theta}_N}). \tag{27}$$

By definition of $\hat{\theta}_N$, (27) is non-negative. Taking into account that $KL() \geq 0$, it is therefore sufficient to show that

$$\lim_{N\to\infty} KL(\mathcal{C}(\boldsymbol{U}_N), P_{\theta^*}) = 0 \quad P\text{-a.s.} \tag{28}$$

We have $P_{\theta^*} \in \mathcal{C}(m^*)$, and by Theorem 2 there exists a representation of P_{θ^*} as a convex combination

$$P_{\theta^*} = \sum_{i=1}^{n!} \kappa_i \pi_i(m^*)$$

($\kappa_i \geq 0; \sum \kappa_i = 1$). From (26) we obtain that for all $\pi \in \Pi W$

$$\lim_{N\to\infty} \pi(m_N) = \pi(m^*) \quad P\text{-a.s,}$$

and

$$\lim_{N\to\infty} \sum_{i=1}^{n!} \kappa_i \pi_i(m_N) = P_{\theta^*} \quad P\text{-a.s.} \tag{29}$$

By definition, $\sum \kappa_i \pi_i(m_N) \in \mathcal{C}(\boldsymbol{U}_N)$. (29) together with Lemma 11 implies that P-a.s.:

$$\lim_{N\to\infty} KL(\mathcal{C}(\boldsymbol{U}_N), P_{\theta^*}) \leq \lim_{N\to\infty} KL(\sum \kappa_i \pi_i(m_N), P_{\theta^*}) = KL(P_{\theta^*}, P_{\theta^*}) = 0.$$

B. To simplify notation, from now on we write $LL_{FV}(\cdot \mid m_N)$ for $1/N\, LL_{FV}(\cdot \mid \boldsymbol{U}_N)$. The proof is based on the following known relationship between L_{FV} and Λ_{car} (cf. (Gill et al., 1997, Theorem 1), (Jaeger, 2005b, Theorem 4.7)): for $P \in \Delta(W)$ and $m \in \Delta(\mathcal{Y})$ the following are equivalent:

(a) P is a global maximum of $L_{FV}(\cdot \mid m)$ in $\Delta(W)$

(b) There exists $\lambda \in \Lambda_{car}$ such that $P_{P,\lambda}^{\downarrow\mathcal{Y}} = m$.

Furthermore, the global maximum of L_{FV} is essentially unique in the sense that

(c) if P, P' are global maxima of $L_{FV}(\cdot \mid m)$ in $\Delta(W)$, then $P(U) = P'(U)$ for all U with $m(U) > 0$.

According to (5) we have for all θ

$$LL_{car}(\theta \mid \boldsymbol{U}_N) - LL_{car}(\theta^* \mid \boldsymbol{U}_N) = LL_{FV}(\theta \mid \boldsymbol{U}_N) - LL_{FV}(\theta^* \mid \boldsymbol{U}_N), \tag{30}$$

and $LL_{car}(\cdot \mid \boldsymbol{U}_N)$ and $LL_{FV}(\cdot \mid \boldsymbol{U}_N)$ are maximized by the same $\hat{\theta}$. Thus, we can prove **B.(i)** by showing that

$$\lim_{N\to\infty} \frac{1}{N}(LL_{FV}(\hat{\theta}_N \mid \boldsymbol{U}_N) - LL_{FV}(\theta^* \mid \boldsymbol{U}_N)) = 0 \quad P\text{-a.s.}, \tag{31}$$

with $\hat{\theta}_N$ a maximum for $LL_{FV}(\cdot \mid \boldsymbol{U}_N)$. Equivalently, in shorter notation:

$$LL_{FV}(\hat{\theta}_N \mid m_N) - LL_{FV}(\theta^* \mid m_N) \to 0 \quad P\text{-a.s.} \tag{32}$$

where from now on a limit $\to$ always is understood to be taken for $N \to \infty$.

By the assumption that $\lambda^* \in \Lambda_{car}$, and the equivalence **(a)**$\Leftrightarrow$ **(b)**, we have that θ^* is a global maximum of $LL_{FV}(\cdot \mid m^*)$. This means that (32) holds at the limit $N = \infty$ in the sense that $LL_{FV}(\hat{\theta}_* \mid m^*) - LL_{FV}(\theta^* \mid m^*) = 0$ where $\hat{\theta}_*$ is any maximum of $LL_{FV}(\cdot \mid m^*)$. What is left for the remainder of the proof is to establish that $LL_{FV}(\cdot \mid m_N)$ is well-behaved on a suitably defined compact neighborhood of θ^*, so that from the equality at the limit we can infer the convergence statement (32).

For $m \in \Delta(\mathcal{Y})$ let $\mathit{support}(m) := \{U \in \mathcal{Y} \mid m(U) > 0\}$, and define

$$\Theta^+ := \{\theta \in \Theta \mid \forall U \in \mathcal{Y} : U \in \mathit{support}(m^*) \to P_\theta(U) > 0\}.$$

Then, for $\theta \in \Theta^+$ and all N: $LL_{FV}(\theta \mid m_N) > -\infty$ P-a.s., and with (26): $LL_{FV}(\theta \mid m_N) \to LL_{FV}(\theta \mid m^*)$ P-a.s. To infer (32) from this we show that there exists a compact subset $\tilde{\Theta} \subseteq \Theta^+$ containing θ^*, such that for all $\theta \notin \tilde{\Theta}$, and all sufficiently large N:

$$LL_{FV}(\theta^* \mid m^*) - LL_{FV}(\theta \mid m_N) > 1 \quad P\text{-a.s.} \tag{33}$$

Then (32) follows, because (33) guarantees that $\hat{\theta}_N \in \tilde{\Theta}$ P-a.s. for all sufficiently large N, and the convergence $LL_{FV}(\theta \mid m_N) \to LL_{FV}(\theta \mid m^*)$ is uniform on $\tilde{\Theta}$.

It remains to construct a suitable $\tilde{\Theta}$. Let

$$m_{min} := \min_{U \in \mathit{support}(m^*)} m^*(U),$$

and

$$U_{min} := \operatorname{argmin}_{U \in \mathit{support}(m^*)} P_{\theta^*}(U).$$

Then we define

$$\tilde{\Theta} := \{\theta \in \Theta^+ \mid \mathit{logP}_\theta(U) \geq \mathit{logP}_{\theta^*}(U_{min}) - q \text{ for all } U \in \mathit{support}(m^*)\},$$

where

$$q := \frac{m_{min} - 2}{m_{min}} \mathit{logP}_{\theta^*}(U_{min}) + \frac{2}{m_{min}}.$$

By assumptions A1 and A2, $\tilde{\Theta}$ is compact, and $\theta^* \in \tilde{\Theta}$ by definition (because $q \geq 0$). It remains to show (33). For this let $\theta \in \Theta \setminus \tilde{\Theta}$, and $U \in \mathit{support}(m^*)$ with $\mathit{logP}_\theta(U) < \mathit{logP}_{\theta^*}(U_{min}) - q$. Directly from the definition of LL_{FV} and U_{min}, we have

$$LL_{FV}(\theta \mid m_N) \leq m_N(U) \mathit{logP}_\theta(U) \tag{34}$$

and

$$LL_{FV}(\theta^* \mid m^*) \geq logP_{\theta^*}(U_{min}), \tag{35}$$

and therefore

$$LL_{FV}(\theta^* \mid m^*) - LL_{FV}(\theta \mid m_N) \geq logP_{\theta^*}(U_{min}) - m_N(U) logP_\theta(U) \tag{36}$$

By the definition of m_{min} and the convergence $m_N(U) \to m^*(U)$, we have that $m_N(U) \geq m_{min}/2$ P-a.s. for all sufficiently large N. With the definition of U, we can therefore bound the right-hand side of (36) P-a.s. for all sufficiently large N:

$$\begin{aligned} logP_{\theta^*}(U_{min}) - m_N(U) logP_\theta(U) &> logP_{\theta^*}(U_{min}) - \frac{m_{min}}{2}(logP_{\theta^*}(U_{min}) - q) \\ &= 1. \end{aligned}$$

Part **B.(ii)** of the Theorem follows directly from **(c)**: under the assumption of **B.(ii)** it follows from **(c)** that θ^* is the unique maximum of $L_{FV}(\cdot \mid m^*)$ (or, equivalently, $LL_{car}(\cdot \mid m^*)$). Together with part **B.(i)** of the Theorem, the convergence $\hat{\theta}_N \to \theta^*$ then follows. ■

Theorem 6 *Let $U^{(1)}, \ldots, U^{(N)}$ be a data set. If for all i, j: $U^{(i)} = U^{(j)}$, or $U^{(i)} \cap U^{(j)} = \emptyset$, then*

$$L_{\text{sat}}(\cdot|\boldsymbol{U}) = L_{\text{car}}(\cdot|\boldsymbol{U}), \tag{8}$$

and for all θ:

$$\text{AI}(\theta, \boldsymbol{U}) = \text{E}(\theta, \boldsymbol{U}). \tag{9}$$

Proof

In order to maximize L_{sat} we can set for all $(w, U) \in \Omega(W)$: $\lambda_{w,U} = 1$ if $U = U^{(i)}$ for some $i = 1, \ldots, N$, $\lambda_{w,\{w\}} = 1$ for all $w \notin \cup_{i=1}^{N} U^{(i)}$ (this specification is just for completeness; it has no impact on the profile likelihood), and $\lambda_{w,U} = 0$ for all other (w, U). Due to the assumptions on the structure of the $U^{(i)}$, this is well-defined, the resulting λ parameters are in Λ_{car}, and thus (8) holds.

The proof of (9) is another direct consequence of Lemma 10 applied to $\mathcal{V} = \{U^{(1)}, \ldots, U^{(N)}\} \cup W \setminus \cup_i U^{(i)}$, which, according to the assumption of the theorem, is a partitioning of W. Since for all $V_i \in \mathcal{V}$, and all $P_c \in \mathcal{C}(\boldsymbol{U})$ we have that $P_c(V_i)$ is just the empirical probability of V_i in the sample $\boldsymbol{U}$, we have that the term $KL(P^{\downarrow\mathcal{V}}, Q^{\downarrow\mathcal{V}})$ as well as the factors $P(V_i)$ in (24) are constant for all $P \in \mathcal{C}(\boldsymbol{U})$. $KL(\mathcal{C}(\boldsymbol{U}), P_\theta)$, therefore, is minimized by choosing P_c with $P_c \mid V_i = P_\theta \mid V_i$, which is obtained by completing each $U^{(i)} \in \boldsymbol{U}$ as $P_\theta(\cdot \mid U^{(i)})$. The resulting P_c then is just $E(\theta, \boldsymbol{U})$. ■

Theorem 7 *Let $(c_0, \theta_0) \in D$, and $(c_{i+1}, \theta_{i+1}) = \text{AIM}(c_i, \theta_i)$ for $i \geq 0$. Every accumulation point $(\hat{c}, \hat{\theta})$ of the sequence $(c_i, \theta_i)_i$ then is a fixed point of the* AIM *operator.*

Proof Let $(\hat{c}, \hat{\theta})$ be an accumulation point, and $(c_{i_j}, \theta_{i_j})_j$ be a subsequence of $(c_i, \theta_i)_i$ that converges to $(\hat{c}, \hat{\theta})$. Assume that $aim(\hat{c}, \hat{\theta}) \neq (\hat{c}, \hat{\theta})$. Then either

$$\mathrm{AI}(\hat{\theta}) \neq \hat{c} \text{ and } KL(\mathrm{AI}(\hat{\theta}), \hat{\theta}) < KL(\hat{c}, \hat{\theta}), \tag{37}$$

or

$$\mathrm{AI}(\hat{\theta}) = \hat{c},\ \mathrm{M}(\hat{c}) \neq \hat{\theta} \text{ and } KL(\hat{c}, \mathrm{M}(\hat{c})) < KL(\hat{c}, \hat{\theta}). \tag{38}$$

Here the strict inequality in (37) is due to the fact that $KL(\cdot, \hat{\theta})$ obtains a unique minimum on the convex set $\mathcal{C}(\boldsymbol{U})$; the strict inequality in (38) is due to Assumption 3.

First consider case (37). Since $\theta_{i_j} \to_j \hat{\theta}$, we have by lemma 11 a. and c. and Assumption 2:

$$KL(\mathrm{AI}(\hat{\theta}), \theta_{i_j}) \to_j KL(\mathrm{AI}(\hat{\theta}), \hat{\theta}) < KL(\hat{c}, \hat{\theta}) \leq lim_j KL(c_{i_j}, \theta_{i_j})$$

For the application of lemma 11 we here require assumption Assumption 2, which allows us to infer from the convergence properties of sequences in $\Delta(W)$ given by lemma 11 these convergence statements for sequences in Θ. Note, too, that since the sequence $KL(c_{i_j}, \theta_{i_j})$ is non-increasing, the *lim inf* of lemma 11 becomes a limit.

Thus, for sufficiently large j:

$$KL(\mathrm{AI}(\hat{\theta}), \theta_{i_j}) < KL(c_{i_{j+1}}, \theta_{i_{j+1}}) \leq KL(c_{i_{j+1}}, \theta_{i_j}),$$

which contradicts the definition of $c_{i_{j+1}}$ as $\mathrm{AI}(\theta_{i_j})$. In the case of (38) a contradiction is obtained in the same manner, using lemma 11 b. instead of a.

■

Theorem 8 *If $(\hat{c}, \hat{\theta}) \in D$ is a local minimum of $KL(c, \theta)$, then $\hat{\theta}$ is a local maximum of L_{sat} on Θ.*

Proof If $(\hat{c}, \hat{\theta})$ is a local minimum of $KL(\cdot, \cdot)$ on D, then $\hat{c}$ is a global minimum of $KL(\cdot, \hat{\theta})$ on $\mathcal{C}(\boldsymbol{U})$ (due to the strict convexity of $KL(\cdot, \hat{\theta})$). Now assume that $\hat{\theta}$ is not a local maximum of L_{sat}. Then there exists a sequence $\theta_i \to \hat{\theta}$ with $L_{sat}(\theta_i) > L_{sat}(\hat{\theta})$ for all i, and hence by Theorem 3, $KL(c_i, \theta_i) < KL(\hat{c}, \hat{\theta})$, where $c_i = \mathrm{AI}(\theta_i)$. We may assume that the sequence (c_i, θ_i) converges to some limit $(\bar{c}, \bar{\theta}) \in D$ (otherwise select a convergent sub-sequence). Then $\bar{\theta} = \hat{\theta}$, and by Lemma 11

$$KL(\hat{c}, \hat{\theta}) \geq \liminf_i KL(c_i, \theta_i) \geq KL(\bar{c}, \bar{\theta}) = KL(\bar{c}, \hat{\theta}).$$

Since $\hat{c}$ uniquely minimizes $KL(\cdot, \hat{\theta})$, we obtain $\hat{c} = \bar{c}$. Thus, each neighborhood of $(\hat{c}, \hat{\theta})$ contains (c_i, θ_i) with $KL(c_{c_i}, c_{\theta_i}) < KL(\hat{c}, \hat{\theta})$, contradicting the assumption that $(\hat{c}, \hat{\theta})$ is a local KL minimum. ■

Theorem 9 *Let Λ_{sat} be as defined by (2) for the discretized space $\Omega(W)$. Then: $\lambda \in \Lambda_{\mathrm{sat}}$ iff there exists a finite set $\mathcal{F}$, and a $\mathcal{F}$-valued random variable F with conditional distribution $P_\gamma(F|X)$, such that for all $\theta \in \Theta$: $P_{\theta,\gamma}(U|w) = \lambda_{w,U}$.*

Proof The right to left direction is trivial: the conditional distribution $P_{\theta,\gamma}(Y|d(X))$ is an element of Λ_{sat} by definition. For the converse direction, let $\lambda \in \Lambda_{sat}$. We obtain a representation of λ in the form $P_{\theta,\gamma}(Y|d(X))$ by a canonical construction: let $\mathcal{F} = 2^W$. We identify elements $U \in \mathcal{G}$ with the subsets $\bigcup U := \{x \in w : w \in U\} \subseteq \mathbb{R}^n$, and define $\zeta(x, U) = \bigcup U$. Define $P_\gamma(U|x) := \lambda_{d(x),U}$ if $x \in \bigcup U$, and $P_\gamma(U|x) := 0$, otherwise. Then, since $P_\gamma(U|x)$ is constant on the cell $d(x)$, we have that for all $x \in w = d(x)$: $P_{\theta,\gamma}(U|w) = P_\gamma(U|x) = \lambda_{w,U}$. ■

Appendix B. Details on the AIM Implementation for Gaussian Data

AIM($\boldsymbol{A}$,c);

1 Discretize $\boldsymbol{A}$ with granularity c to obtain the set $\boldsymbol{U} = (U_1, \ldots, U_M)$ of distinct discretized observations $d(A^{(j)})$ with empirical probabilities $m(U_i)$;
2 Set initial parameters $\theta_0 = (\mu_0, \Sigma_0)$ `/* using ACA or EM */`
3 Initialize $c_0(U_i, w) = 0$ for all i and all $w \in U_i$;
4 $t = 0$;
5 **repeat**
6 $\quad t = t + 1$;
$\quad$ `/* Incremental AI step */`
7 $\quad$ **for** $i = 1, \ldots, M$ **do**
8 $\quad\quad$ Calculate $\boldsymbol{p}$ and π according to (18) and (19);
9 $\quad\quad$ $\boldsymbol{p}^* =$BestCompletion$(\boldsymbol{p}, P_\theta|U_i)$;
10 $\quad\quad$ $c_t(U_i) = \frac{1}{1-\pi}(\boldsymbol{p}^* - \boldsymbol{p})$;
11 $\quad$ **end**
$\quad$ `/* M step */`
12 $\quad$ Set θ_t by minimizing $KL(c_t, \cdot)$ by gradient descent;
13 **until** *termination condition applies*;
14 **return** θ_t

Algorithm 2: Pseudo code for AIM using discretization for coarse continuous data

Algorithm 2 shows the high-level structure of the AIM implementation for Gaussian data (indeed, this is a completely generic algorithm for learning from coarse numeric data via discretization). The main loop operates as described in Section 7.1.1. Lines 9 and 10 solve the optimization problem (20) using the BestCompletion algorithm, shown in Algorithm 3.

BestCompletion takes as input a sub-probability vector $\boldsymbol{p} = (p_1, \ldots, p_k)$, i.e., a vector of probability values $p_i \in [0, 1]$ with $\sum_i p_i \leq 1$, and a probability vector $\boldsymbol{q} = (q_1, \ldots, q_k)$ of the same length with $\sum_i q_i = 1$. It computes the probability vector $\boldsymbol{p}^*$ that completes $\boldsymbol{p}$ into a probability vector (i.e., $p^*_i \geq p_i$ for all i, and $\sum_i p^*_i = 1$), such that $KL(\boldsymbol{p}^*, \boldsymbol{p})$ is minimized for all possible completions. BestCompletion incrementally constructs $\boldsymbol{p}^*$ by distributing a part of the remaining available probability mass (initially $1 - \sum_i p_i$) as

/* Input: $\boldsymbol{p} = (p_1, \ldots, p_k)$: sub-probability vector */
/* $\boldsymbol{q} = (q_1, \ldots, q_k)$ probability vector. */
/* Output: probability vector $\boldsymbol{p}^*$ with $\boldsymbol{p}^* \geq \boldsymbol{p}$ that minimizes $KL(\cdot, \boldsymbol{q})$ */
BestCompletion$(\boldsymbol{p}, \boldsymbol{q})$;
remaining$= 1 - \sum_i p_i$;
$\boldsymbol{p}^* = \boldsymbol{p}$;
while *remaining*> 0 **do**
  let *rmin1* and *rmin2* be the smallest and second smallest of the ratios p_i^*/q_i $(i = 1, \ldots, k)$;
  $R = \{i : p_i^*/q_i = rmin1\}$;
  for $j \in R$: $d_j = rmin2 \cdot q_j - p_j^*$;
  $D = \sum_{j \in R} d_j$;
  if $D < remaining$ **then**
    for $j \in R$: $p_j^* = p_j^* + d_j$;
    $remaining = remaining - D$
  else
    $Q = \sum_{j \in R} q_j$;
    for $j \in R$: $p_j^* = p_j^* + remaining \cdot \dfrac{q_j}{Q}$;
    $remaining = 0$
  end
end
return $\boldsymbol{p}^*$

Algorithm 3: Pseudo code for local optimization routine

follows: the set R of indices i is determined for which the ratio p_i^*/q_i is minimal (lines 4,5). Allocating probability mass to these components is optimal, because the partial derivatives

$$\frac{\partial}{\partial p_i^*} KL(\boldsymbol{p}^*, \boldsymbol{q}) = \ln \frac{p_i^*}{q_i} + 1$$

are monotone in this ratio (the KL function here in a general sense allowing sub-probability vectors in the first argument), and therefore adding probability mass to these components decreases KL the most (or increases it the least). Line 6 then determines how much probability mass d_j is required for each component $j \in R$ in order to equal the smallest ratio p_h^*/q_h for components $h \notin R$. If sufficient remaining probability mass is available to reach that level for all $j \in R$, then the required probability mass is allocated to the components in R, the remaining mass updated, and another iteration will be performed (lines 7, 8-10). Otherwise, the remaining mass is distributed over the components in R such that they maintain a uniform ratio (lines 11-14).

Appendix C. Details on the AIM Implementation for Bayesian Networks

```
   /* Input:  U = (U^(1), ..., U^(N)):  observed incomplete data items with
      empirical probabilities m(U_i)                                    */
   /* bnet:  Bayesian network structure                                 */
   AIM(U, bnet);
   /* c:  completion mapping U × W → ℝ with Σ_{x∈W} c(U_i, x) = m(U_i)  */
   /* for all U_i θ:  Parameter setting for bnet                        */
 1 initialize c ;
 2 θ = M(c);
 3 while !terminate do
 4 |   oldce = KL(P_c, P_θ);
 5 |   c = IAI(U, θ, c);
 6 |   θ = M(c);
 7 |   newce = KL(P_c, P_θ);
 8 |   if (oldce − newce < threshold) then
 9 |   |   terminate = true;
10 |   end
11 end
12 return θ;
```

Algorithm 4: AIM for Bayesian Networks

The overall structure of the algorithm is shown in Algorithm 4. We here assume that the data is already given by the empirical probabilities $m(U_i)$ of M observed distinct incomplete data items U_i (obtained from some original sample of size $N \geq M$). A completion c then is represented as a mapping $\boldsymbol{U} \times W \to \mathbb{R}$ with $\sum_{x \in U_i} c(U_i, x) = m(U_i)$. The implementation maintains a sparse representation of the mapping c consisting of a list of the nonzero values $c(U_i, x)$ for each U_i.

The algorithm starts by constructing an initial completion c (the initialization of the parameters θ then is performed by a first M step, line 2). This first completion can be constructed in two ways, leading to two different versions of the algorithm:

AIM-ACA (available case analysis): An initial value for the cpt row $\theta_{i\bullet k} = P_\theta(X_i \mid Pa_i = k)$ is obtained from the empirical count in the data cases where X_i and Pa_i are fully observed. If there are no such data cases with the parent configuration $Pa_i = k$, then $\theta_{i\bullet k}$ is initialized as the uniform distribution over $W(X_i)$. In order to obtain diverse initializations, different small random subsets of $\boldsymbol{U}$ are used in the random restarts for the computation of the ACA statistics.

EM-AIM (EM guided): First parameters θ^{EM} are learned using the EM algorithm. For each U_i then a unique completion x_i is sampled from $c_{\theta^{EM}}(X \mid U_i)$. These initial one-point completions are refined to a fractional completion by performing IAI steps until convergence to a stable fractional completion.

After initialization, a main loop of IAI and M steps is executed until the decrease in $KL(c, \theta)$ falls below a user-defined threshold.

The most important aspect of the algorithm is the implementation of the IAI step, shown in Algorithm 5. This IAI procedure receives as an argument the completion c computed in the previous iteration. This initial c is modified iteratively to further reduce $KL(c, P_\theta)$ with regard to the current θ values. The iterations are over all current completions of all incomplete data items (lines 1 and 2—*support*$(c(U_i))$ denotes the set of all x with $c(U_i, x) > 0$). For every such pair, U_i, x the completion c is updated by re-allocating some of the weight $c(U_i, x)$ to another completion x^* of U_i. This re-allocation is performed by the *bestshift* subroutine shown in Algorithm 6 Since it would be computationally infeasible to consider all possible alternative completions x^* (which would be exponentially many in the number of missing values in U_i), only x^* are considered that differ from the current x in the value of exactly one variable (reminiscent of Gibbs sampling). Given such a candidate alternative completion, the problem is to determine the part of $c(U_i, x)$ that should be reallocated to x^*. One obtains analytically that for minimizing $KL(\cdot, P_\theta)$ one optimally needs to re-allocate probability mass

$$\mathit{bestshiftamount}(x, x^*, c, \theta) = \frac{P_c(x)P_\theta(x^*) - P_c(x^*)P_\theta(x)}{P_\theta(x) + P_\theta(x^*)}$$

from x to x^*. This value can be greater than $c(U_i, x)$ (when other incomplete data items U_j also assign some weight to x), for which reason in line 4 *shiftamount* is computed by taking the minimum of *bestshiftamount* and $c(U_i, x)$. Furthermore, *bestshiftamount* can be negative (one should rather reallocate weight from x^* to x). This possibility is not considered by the *bestshift* routine (because it will then be considered when executing *bestshift* with x^* as an argument). Therefore, line 4 also excludes negative values by taking a maximum with 0. In this way, lines 3-11 determine the optimal way to re-allocate part of the weight $c(U_i, x)$ to some alternative completion x^* (differing by one value from x). However, always performing such a reallocation would quickly lead to intractably large sets of support of the completion c (the number of pairs U_i, x with $c(U_i, x) > 0$ could double in each call of *AI*). For this reason, the shift of weight from x to x^* only is performed if either x^*

already is in the support of $U_i's$ current completion function $c(U_i)$, or the KL-reduction exceeds a user defined threshold *new_completion_penalty* (line 12-14). This threshold can be used to trade-off time and space requirement against accuracy of the approximate AIM implementation.

```
   /* Input:  U = (U^(1), ..., U^(N)):  observed incomplete data items      */
   /* θ:  current Bayesian network parameter setting                        */
   /* c:  current completion mapping                                        */
   IAI(U, θ, c) ;
1  for i = 1, ..., N do
2  |  for all x ∈ support(c(U_i)) do
3  |  |  c = BESTSHIFT(U_i, x, c, θ);
4  |  end
5  end
6  return c
```

Algorithm 5: Incremental AI step

```
   BESTSHIFT(U, x, c, θ);
1  bestcegain=0;
2  bestc = c;
3  for all completions x* that differ from x in the value of one variable missing in U
   do
4  |  shiftamount = max(0, min(c(U, x), bestshiftamount(x, x*, c, θ));
5  |  cnewcand= the completion that differs from c by allocating:
   |    cnewcand(U, x) = c(U, x) − shiftamount
   |  cnewcand(U, x*) = c(U, x*) + shiftamount;
6  |  cegain = KL(P_c, P_θ) − KL(P_cnewcand, P_θ);
7  |  if (cegain > bestcegain) then
8  |  |  bestcegain = cegain;
9  |  |  bestc = cnewcand;
10 |  end
11 end
12 if bestcegain>new_completion_penalty or support(bestc(U))=support(c(U)) then
13 |  return bestc
14 else
15 |  return c
16 end
```

Algorithm 6: Local improvement step

Appendix D. Detailed Experimental Results

Coarsening	N	ACA	EM	AIM	EM-AIM
t_1	100	0.167 ± 0.048	0.167 ± 0.048	0.092 ± 0.045	0.137 ± 0.051
	1000	0.145 ± 0.017	0.145 ± 0.017	0.057 ± 0.043	0.104 ± 0.025
	10000	0.150 ± 0.006	0.150 ± 0.006	0.032 ± 0.018	0.113 ± 0.013
	100000	0.152 ± 0.002	0.152 ± 0.002	0.025 ± 0.022	0.115 ± 0.004
	1000000	0.153 ± 0.000	0.153 ± 0.001	0.052 ± 0.031	0.114 ± 0.001
t_2	100	0.168 ± 0.078	0.168 ± 0.078	0.158 ± 0.122	0.178 ± 0.121
	1000	0.136 ± 0.015	0.136 ± 0.015	0.075 ± 0.026	0.103 ± 0.023
	10000	0.136 ± 0.008	0.137 ± 0.008	0.084 ± 0.042	0.114 ± 0.010
	100000	0.134 ± 0.002	0.134 ± 0.002	0.084 ± 0.041	0.108 ± 0.004
	1000000	0.135 ± 0.001	0.134 ± 0.001	0.087 ± 0.044	0.108 ± 0.001
t_3	100	0.129 ± 0.087	0.129 ± 0.086	0.174 ± 0.162	0.172 ± 0.159
	1000	0.063 ± 0.016	0.062 ± 0.016	0.056 ± 0.037	0.060 ± 0.015
	10000	0.061 ± 0.006	0.061 ± 0.006	0.123 ± 0.088	0.061 ± 0.006
	100000	0.061 ± 0.002	0.062 ± 0.002	0.128 ± 0.084	0.062 ± 0.002
	1000000	0.061 ± 0.001	0.061 ± 0.001	0.066 ± 0.044	0.061 ± 0.001
t_4	100	0.103 ± 0.123	0.102 ± 0.120	0.680 ± 0.598	0.286 ± 0.463
	1000	0.007 ± 0.010	0.007 ± 0.009	0.100 ± 0.090	0.008 ± 0.006
	10000	0.000 ± 0.000	0.000 ± 0.000	0.113 ± 0.168	0.000 ± 0.000
	100000	0.000 ± 0.000	0.000 ± 0.000	0.166 ± 0.314	0.000 ± 0.000
	1000000	0.000 ± 0.000	0.000 ± 0.000	0.031 ± 0.016	0.000 ± 0.000
c_1	100	0.095 ± 0.092	0.096 ± 0.094	0.222 ± 0.257	0.130 ± 0.140
	1000	0.062 ± 0.028	0.062 ± 0.029	0.056 ± 0.065	0.072 ± 0.034
	10000	0.047 ± 0.009	0.046 ± 0.008	0.005 ± 0.004	0.047 ± 0.008
	100000	0.049 ± 0.003	0.049 ± 0.002	0.001 ± 0.001	0.049 ± 0.002
	1000000	0.050 ± 0.001	0.050 ± 0.001	0.031 ± 0.060	0.050 ± 0.001
c_2	100	0.078 ± 0.074	0.078 ± 0.076	0.377 ± 0.301	0.193 ± 0.289
	1000	0.080 ± 0.027	0.080 ± 0.027	0.426 ± 0.226	0.091 ± 0.042
	10000	0.062 ± 0.012	0.061 ± 0.012	0.216 ± 0.161	0.062 ± 0.012
	100000	0.060 ± 0.003	0.060 ± 0.004	0.153 ± 0.129	0.060 ± 0.004
	1000000	0.060 ± 0.001	0.059 ± 0.002	0.121 ± 0.107	0.059 ± 0.001

Table 4: Results for 1d-b $\mathcal{B}_1$ (Figure 6). Sum of squared errors with standard deviation

Coarsening	N	ACA	EM	AIM	EM-AIM
t_1	100	0.159 ± 0.060	0.043 ± 0.033	0.084 ± 0.109	0.072 ± 0.071
	1000	0.160 ± 0.025	0.018 ± 0.010	0.007 ± 0.005	0.008 ± 0.007
	10000	0.151 ± 0.007	0.012 ± 0.002	0.001 ± 0.001	0.003 ± 0.004
	100000	0.154 ± 0.003	0.013 ± 0.001	0.000 ± 0.000	0.000 ± 0.000
	1000000	0.152 ± 0.001	0.013 ± 0.001	0.000 ± 0.000	0.000 ± 0.000
t_2	100	0.182 ± 0.067	0.039 ± 0.045	0.093 ± 0.158	0.051 ± 0.058
	1000	0.138 ± 0.026	0.011 ± 0.008	0.025 ± 0.027	0.015 ± 0.010
	10000	0.132 ± 0.006	0.007 ± 0.001	0.002 ± 0.002	0.003 ± 0.002
	100000	0.133 ± 0.002	0.007 ± 0.000	0.002 ± 0.002	0.004 ± 0.003
	1000000	0.134 ± 0.001	0.007 ± 0.000	0.001 ± 0.001	0.003 ± 0.003
t_3	100	0.081 ± 0.028	0.021 ± 0.027	0.205 ± 0.335	0.117 ± 0.308
	1000	0.066 ± 0.017	0.016 ± 0.013	0.086 ± 0.050	0.026 ± 0.021
	10000	0.063 ± 0.005	0.008 ± 0.001	0.054 ± 0.057	0.010 ± 0.002
	100000	0.062 ± 0.002	0.008 ± 0.001	0.062 ± 0.063	0.011 ± 0.002
	1000000	0.061 ± 0.000	0.008 ± 0.000	0.099 ± 0.063	0.009 ± 0.002
t_4	100	0.085 ± 0.090	0.117 ± 0.153	0.698 ± 0.963	0.139 ± 0.135
	1000	0.007 ± 0.005	0.017 ± 0.014	0.132 ± 0.086	0.017 ± 0.010
	10000	0.001 ± 0.001	0.012 ± 0.005	0.067 ± 0.048	0.015 ± 0.005
	100000	0.000 ± 0.000	0.012 ± 0.001	0.058 ± 0.065	0.009 ± 0.003
	1000000	0.000 ± 0.000	0.012 ± 0.001	0.121 ± 0.173	0.011 ± 0.002
c_1	100	0.055 ± 0.038	0.016 ± 0.015	0.050 ± 0.048	0.072 ± 0.117
	1000	0.050 ± 0.025	0.011 ± 0.006	0.030 ± 0.025	0.019 ± 0.017
	10000	0.052 ± 0.008	0.007 ± 0.002	0.016 ± 0.019	0.004 ± 0.003
	100000	0.050 ± 0.002	0.006 ± 0.000	0.002 ± 0.001	0.002 ± 0.001
	1000000	0.050 ± 0.001	0.006 ± 0.000	0.010 ± 0.021	0.001 ± 0.000
c_2	100	0.105 ± 0.088	0.042 ± 0.032	0.104 ± 0.114	0.033 ± 0.027
	1000	0.052 ± 0.020	0.014 ± 0.011	0.034 ± 0.021	0.022 ± 0.021
	10000	0.060 ± 0.010	0.016 ± 0.004	0.022 ± 0.012	0.005 ± 0.004
	100000	0.058 ± 0.003	0.016 ± 0.001	0.008 ± 0.008	0.004 ± 0.001
	1000000	0.059 ± 0.001	0.016 ± 0.000	0.005 ± 0.006	0.002 ± 0.002

Table 5: Results for 1d-b $\mathcal{B}_2$ (Figure 7). Sum of squared errors with standard deviation

t_1, c_1	100	0.652 ± 0.290	0.449 ± 0.205	0.741 ± 0.615	0.671 ± 0.456
	1000	0.529 ± 0.140	0.297 ± 0.080	0.231 ± 0.147	0.202 ± 0.127
	10000	0.536 ± 0.050	0.307 ± 0.028	0.150 ± 0.054	0.150 ± 0.038
	100000	0.541 ± 0.012	0.312 ± 0.010	0.169 ± 0.068	0.137 ± 0.012
	1000000	0.544 ± 0.004	0.313 ± 0.012	0.175 ± 0.053	0.153 ± 0.005
t_4, c_2	100	2.959 ± 4.552	0.357 ± 0.233	1.518 ± 2.674	0.596 ± 0.727
	1000	2.231 ± 0.785	0.154 ± 0.021	0.309 ± 0.149	0.139 ± 0.073
	10000	1.646 ± 0.251	0.155 ± 0.019	0.361 ± 0.164	0.111 ± 0.031
	100000	1.626 ± 0.104	0.147 ± 0.007	0.446 ± 0.560	0.124 ± 0.023
	1000000	1.682 ± 0.018	0.149 ± 0.008	0.500 ± 0.874	0.128 ± 0.013

Table 6: Results for 2d-m (Figure 10)

n	(μ, σ^2)	#RN	#RS	ACA	EM	AIM	EM-AIM
100	(0.2,0.0)	50	10	0.035 ± 0.008	0.030 ± 0.007	*0.042 ± 0.008*	*0.036 ± 0.009*
1000	(0.2,0.0)	50	10	0.010 ± 0.003	0.009 ± 0.002	*0.024 ± 0.007*	*0.012 ± 0.004*
10000	(0.2,0.0)	50	10	0.003 ± 0.001	0.003 ± 0.001	*0.023 ± 0.006*	*0.003 ± 0.001*
100000	(0.2,0.0)	50	10	0.001 ± 0.000	0.001 ± 0.000	*0.025 ± 0.007*	*0.001 ± 0.000*
1000000	(0.2,0.0)	50	10	0.000 ± 0.000	0.000 ± 0.000	*0.024 ± 0.008*	*0.000 ± 0.000*
100	(0.2,0.05)	50	10	0.050 ± 0.019	0.044 ± 0.020	*0.051 ± 0.021*	*0.047 ± 0.020*
1000	(0.2,0.05)	50	10	0.034 ± 0.015	0.027 ± 0.014	*0.039 ± 0.017*	0.027 ± 0.012
10000	(0.2,0.05)	50	10	0.032 ± 0.015	0.026 ± 0.014	*0.039 ± 0.020*	*0.025 ± 0.014*
100000	(0.2,0.05)	50	10	0.029 ± 0.015	0.026 ± 0.016	*0.037 ± 0.021*	*0.025 ± 0.016*
1000000	(0.2,0.05)	50	10	0.026 ± 0.010	0.021 ± 0.011	*0.032 ± 0.012*	*0.020 ± 0.009*
100	(0.2,0.1)	50	10	0.088 ± 0.040	0.070 ± 0.041	0.076 ± 0.040	0.069 ± 0.044
1000	(0.2,0.1)	50	10	0.064 ± 0.036	0.055 ± 0.038	0.052 ± 0.036	$\mathbf{0.044 \pm 0.032}$
10000	(0.2,0.1)	50	10	0.059 ± 0.041	0.054 ± 0.036	0.048 ± 0.033	$\mathbf{0.044 \pm 0.035}$
100000	(0.2,0.1)	50	10	0.058 ± 0.034	0.054 ± 0.041	0.045 ± 0.031	$\mathbf{0.040 \pm 0.034}$
1000000	(0.2,0.1)	50	10	0.059 ± 0.042	0.051 ± 0.041	0.050 ± 0.040	$\mathbf{0.035 \pm 0.028}$
100	(0.2,0.15)	50	10	0.132 ± 0.076	0.103 ± 0.078	0.094 ± 0.073	$\mathbf{0.096 \pm 0.080}$
1000	(0.2,0.15)	50	10	0.104 ± 0.070	0.094 ± 0.061	$\mathbf{0.064 \pm 0.049}$	$\mathbf{0.069 \pm 0.058}$
10000	(0.2,0.15)	50	10	0.109 ± 0.074	0.091 ± 0.070	$\mathbf{0.058 \pm 0.056}$	$\mathbf{0.067 \pm 0.066}$
100000	(0.2,0.15)	50	10	0.122 ± 0.094	0.101 ± 0.081	$\mathbf{0.083 \pm 0.073}$	$\mathbf{0.083 \pm 0.078}$
1000000	(0.2,0.15)	50	10	0.119 ± 0.102	0.090 ± 0.079	$\mathbf{0.065 \pm 0.071}$	$\mathbf{0.069 \pm 0.076}$

Table 7: Result details for Figure 13. Values in the AIM and EM-AIM columns are printed in bold when the advantage over EM is statistically significant (Wilcoxon signed rank test at $\alpha = 0.01$). Values are printed in italics, when conversely the advantage of EM is statistically significant (in some cases the statistical significance is computed based on decimals not represented by the number format used in the table).

References

Rebecca R. Andridge and Roderick J. A. Little. A review of hot deck imputation for survey non-response. *International Statistical Review*, 78(1):40–64, 2010.

Jessa Bekker and Jesse Davis. Beyond the selected completely at random assumption for learning from positive and unlabeled data. *arXiv preprint arXiv:1809.03207*, 2018.

Eric Cator. On the testability of the CAR assumption. *The Annals of Statistics*, 32(5): 1957–1980, 2004.

Giorgio Corani and Marco Zaffalon. Learning reliable classifiers from small or incomplete data sets: The naive credal classifier 2. *Journal of Machine Learning Research*, 9:581–621, 2008.

Timothee Cour, Ben Sapp, and Ben Taskar. Learning from partial labels. *Journal of Machine Learning Research*, 12(May):1501–1536, 2011.

Inés Couso, Didier Dubois, and Eyke Hüllermeier. Maximum likelihood estimation and coarse data. In *International Conference on Scalable Uncertainty Management*, pages 3–16. Springer, 2017.

Robert G. Cowell. Parameter estimation from incomplete data for Bayesian networks. In *Artificial Intelligence and Statistics 99: Proceedings of the 7th International Workshop on Artificial Intelligence and Statistics*, pages 193–196, 1999.

I. Csiszár and G. Tusnády. Information geometry and alternating minimization procedures. *Statistics and Decisions*, Supplement Issue No. 1:205–237, 1984.

A. P. Dawid and James M. Dickey. Likelihood and Bayesian inference from selectively reported data. *Journal of the American Statistical Association*, 72(360):845–850, 1977.

Arthur P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Ser. B*, 39:1–38, 1977.

Richard D. Gill, Mark J van der Laan, and James M. Robins. Coarsening at random: Characterizations, conjectures, counter-examples. In D. Y. Lin and T. R. Fleming, editors, *First Seattle Symposium in Biostatistics: Survival Analysis*, Lecture Notes in Statistics, pages 255–294. Springer-Verlag, 1997.

Romain Guillaume and Didier Dubois. Robust parameter estimation of density functions under fuzzy interval observations. In *9th International Symposium on Imprecise Probability: Theories and Applications (ISIPTA'15)*, pages 147–156, 2015.

Asela Gunawardana and William Byrne. Convergence theorems for generalized alternating minimization procedures. *Journal of Machine Learning Research*, 6:2049–2073, 2005.

Daniel F. Heitjan. Ignorability in general incomplete-data models. *Biometrika*, 81(4):701–708, 1994.

Daniel F. Heitjan and Donald B. Rubin. Ignorability and coarse data. *The Annals of Statistics*, 19(4):2244–2253, 1991.

José Miguel Hernández-Lobato, Neil Houlsby, and Zoubin Ghahramani. Probabilistic matrix factorization with non-random missing data. In *International Conference on Machine Learning*, pages 1512–1520, 2014.

Tom Heskes, Onno Zoeter, and Wim Wiegerinck. Approximate expectation maximization. In *Advances in Neural Information Processing Systems 16*. MIT Press, 2004.

Joel L Horowitz and Charles F Manski. Imprecise identification from incomplete data. In *Proc. of the 2nd Int. Symposium on Imprecise Probabilities and Their Applications (ISIPTA)*, pages 213–218, 2001.

Eyke Hüllermeier. Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization. *International Journal of Approximate Reasoning*, 55(7):1519–1534, 2014.

Eyke Hüllermeier and Weiwei Cheng. Superset learning based on generalized loss minimization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 260–275. Springer, 2015.

Manfred Jaeger. Ignorability in statistical and probabilistic inference. *J. of Artificial Intelligence Research*, 24:889–917, 2005a.

Manfred Jaeger. Ignorability for categorical data. *The Annals of Statistics*, 33(4):1964–1981, 2005b.

Manfred Jaeger. On testing the missing at random assumption. In *17th European Conference on Machine Learning (ECML-06)*, 2006a.

Manfred Jaeger. The AI&M procedure for learning from incomplete data. In *22nd Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 225–232, 2006b.

José M Jerez, Ignacio Molina, Pedro J García-Laencina, Emilio Alba, Nuria Ribelles, Miguel Martín, and Leonardo Franco. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial intelligence in medicine*, 50 (2):105–115, 2010.

Rong Jin and Zoubin Ghahramani. Learning with multiple labels. In *Advances in neural information processing systems*, pages 921–928, 2003.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Solomon Kullback. *Information Theory and Statistics*. Dover Publications, Inc., 1968.

E. L. Lehmann and George Casella. *Theory of Point Estimation*. Springer, 1998.

Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, 1987.

Charles F. Manski. Partial identification with missing data: Concepts and findings. *International Journal of Approximate Reasoning*, 39:151–165, 2005.

Fabrizia Mealli and Donald B Rubin. Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika*, 102(4):995–1000, 2015.

Karthika Mohan and Judea Pearl. On the testability of models with missing data. In *Artificial Intelligence and Statistics*, pages 643–650, 2014.

Karthika Mohan, Judea Pearl, and Jin Tian. Graphical models for inference with missing data. In *Advances in neural information processing systems*, pages 1277–1285, 2013.

Karthika Mohan, Felix Thoemmes, and Judea Pearl. Estimation with incomplete data: The linear case. In *IJCAI*, pages 5082–5088, 2018.

Radford M. Neal and Geoffrey E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In M.I. Jordan, editor, *Learning in Graphical Models*. MIT Press, 1998.

Marco Ramoni and Paola Sebastiani. Parameter estimation in Bayesian networks from incomplete databases. *Intelligent Data Analysis*, 2(2):139–160, 1998.

Richard Redner. Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *The Annals of Statistics*, 9(1):225–228, 1981.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Donald B. Rubin. Multiple imputation in sample surveys – a phenomenological Bayesian approach to nonresponse. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, pages 20–34, 1978.

Donald B. Rubin. Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473–489, 1996.

Shaun Seaman, John Galati, Dan Jackson, and John Carlin. What is meant by "missing at random"? *Statistical Science*, 28(2):257–268, 2013.

Harald Steck. Training and testing of recommender systems on data missing not at random. In *16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 713–722. ACM, 2010.

Paul Tseng. An analysis of the EM algorithm and entropy-like proximal point methods. *Mathematics of Operations Research*, 29(1):27–44, 2004.

Guy Van den Broeck, Karthika Mohan, Arthur Choi, Adnan Darwiche, and Judea Pearl. Efficient algorithms for Bayesian network parameter learning from incomplete data. In *Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 161–170. AUAI Press, 2015.

Abraham Wald. Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 20(4):595–601, 1949.

C. F. Jeff Wu. On the convergence properties of the EM algorithm. *Annals of Statistics*, 11:95–103, 1983.

Willard I. Zangwill. Convergence conditions for nonlinear programming algorithms. *Management Science*, 16(1), 1969.