A variational approach to density estimation from incomplete data

Vaidotas Šimkus

Master of Science
Artificial Intelligence
School of Informatics
University of Edinburgh
2019

Abstract

Many real-world datasets contain missing values. Meanwhile, state-of-the-art machine learning models are generally incapable of handling incomplete data. In this thesis, we introduce a general method that combines variational inference (VI) and Gibbs sampling literature for estimating probability density models from incomplete data.

Our algorithm, which is called Cumulative Data Imputation (CDI), addresses a combinatorial explosion of inference networks required by standard VI for learning in the context of missing data. Compared to potentially $2^D - 1$ inference networks required by standard VI, CDI only requires approximating D univariate variational distributions, which can all be approximated with a single inference network [1]. Then, a procedure that is based on Gibbs sampler is used to efficiently draw samples of missing values from the joint posterior distribution of the missing variables, using the univariate variational distributions.

Finally, the effectiveness of CDI for learning from incomplete data is empirically evaluated on a factor analysis (FA) density model. The evaluations show that CDI outperforms default missing data handling methods and for low-dimensional or low-to-medium missingness problems can match the performance of a popular expectation maximisation (EM) algorithm. Importantly, CDI is applicable to general density models, for which there is no analytic solution for evaluating the posterior distribution of the missing data variables, which is the main disadvantage of EM.

[1] We used D to denote the dimensionality of the data.

Acknowledgements

I would like to express sincere gratitude to my supervisors Dr. Michael Gutmann and Benjamin Rhodes. Our discussions were always enjoyable and constructive, and I am grateful for their insightful advice and encouragement to always go the extra mile. The time they have invested in this project is very much appreciated.

I would also like to acknowledge that Ben's preliminary work in his Master's thesis is the seed of my project, and has served as a valuable starting point. The work that is related to Ben's has been appropriately cited in the main text.

I also want to extend my thanks to family and friends who have been very patient and supportive throughout the year.

Table of Contents

Chapter 1

Introduction

Missing data is an inevitable part of scientific research. In medical sciences collecting data is often expensive, invasive, and sometimes dangerous to the patients. In social studies, respondents can refuse to partake in parts of an experiment. As a consequence these studies often produce numerous amounts of incomplete data. In a review Burton and Altman (2004) has identified that 81% of cancer prognostic studies contained missing data. Meanwhile, most of the recent advances in machine learning are in methods that require fully-observed data (Goodfellow et al., 2016). Hence, learning from incomplete data is an important challenge that will not only advance the state of machine learning but also drive adoption of machine learning in other scientific fields, where complete data is scarce.

One of the core learning tasks in machine learning is probability density function estimation. The key appeal of density estimation is that the learnt density model[1] can be used on a multitude of downstream tasks, such as, classification, prediction, anomaly detection, data augmentation, as well as, missing data imputation (Goodfellow et al., 2016, Chapter 5). Density estimation is typically solved via Maximum-Likelihood Estimation (MLE), but unfortunately, in the context of missing data, MLE is generally intractable (Section 2.3).

A probabilistically principled approach for handling missing data is to model the missing variables as latent (hidden) variables. Variational inference (VI) (Jordan et al., 1999) has recently become a key method in machine learning for learning latent variable models due to a significant progress that has been made in scaling VI to large-scale datasets (Zhang et al., 2018). One of the pivotal achievements is amortised inference (Gershman and Goodman, 2014) which enabled efficient parameterisation and learn-

[1] We will refer to the models that capture density functions as density models.

ing of the (approximate) posterior distribution of the latent variables via the use of a neural network, also known as inference network. In most standard latent variable models a single posterior distribution is required and hence one inference network is needed. However, in missing data problems there may be $2^D - 1$ non-trivial patterns of missing (latent) variables and hence $2^D - 1$ posterior distributions are required [2]. It is not obvious how this many posterior distributions can be parameterised with a tractable number of inference networks, and hence is an important research question.

To the best of our knowledge there currently exists only one class of models based on Variational Autoencoder (VAE) (Goodfellow et al., 2014) that has been adapted to handle missing data using the variational framework (Nazabal et al., 2018; Wu and Goodman, 2018; Ivanov et al., 2019). However, in the VAE methods conditional independence assumptions are made about the distribution of the missing variables, which may not be valid for some density models. Hence, we require an additional framework for handling missing data for general density models.

We base our work on the unpublished Cumulative Data Imputation (CDI) algorithm by (Rhodes, 2018), which does not make strong assumptions about the distribution of missing data and requires only D inference networks. The main contributions are:

- A modified CDI algorithm that we justify by theoretical guarantees of Gibbs sampler (Geman and Geman, 1984).
- A shared variational model that can approximate all D variational posteriors with a single inference network.
- A comprehensive set of experiments validating the effectiveness of CDI on synthetic and real-world data.

In the next chapter we provide the necessary background on the missing data problem, default methods for handling missing data, maximum-likelihood estimation methods, including variational inference, and Gibbs sampler. Then, in Chapter 3 we present the recent work on missing data handling, the issue with standard variational inference in the context of missing data in more detail, and research questions. In Chapter 4 we present our method and justify its properties. In Chapter 5 we present the empirical evaluation methodology and results. Finally, in Chapter 6 we draw conclusions and suggest several directions for future work.

[2] We use D to denote the dimensionality of the data.

Chapter 2

Background

In this chapter we cover the necessary background to understand the missing data problem and handling methods, as well as, other ideas required to understand CDI. In the next section we start with the definition of missing data mechanism and why it is so important for learning accurate models. We then discuss some of the default missing data handling methods that are often used in practice and outline their deficiencies. Next, we explain why maximum-likelihood estimation is intractable for incomplete data, and present a theoretical framework – the expectation maximisation (EM) that can be used to fit density models in the context of incomplete data. We then discuss variational inference as a relaxation of the EM method, which is able to fit more general density models than EM. Furthermore, we discuss Gibbs sampling as a method for sampling (approximate) joint posterior distributions using simpler conditional distributions. In our work, we will combine variational inference and Gibbs sampler to mitigate the $2^D - 1$ combinatorial explosion of variational distributions. Finally, we present the factor analysis (FA) density model, which is a fundamental model in machine learning and has several properties that make it useful for evaluation of a new method for learning from incomplete data.

2.1 Missing data mechanisms

Concentrating on the observed parts of the data and ignoring the process that caused the missing data can lead to unreliable conclusions. One example of this is a study on the high-rise syndrome in cats (Whitney and Mehlhaff, 1987), which studied injury to the cats that had fallen from second or higher floors. The study reported that the severity of injuries increased up to the 6-th floor but then declined for floors above the

7-th floor. The study speculated that cats reach terminal velocity at five floors, when they stop accelerating and can no longer sense that they are falling, and hence they relax and take a softer landing. An alternative explanation of this behaviour is that the cats that fell from higher floors were less likely to survive and hence were not brought to the veterinarian and were not included in the data. Thus, different assumptions about the data may lead to very different conclusions. Making incorrect missing data assumption can often lead to survivorship bias.

Rubin (1976) proposed a probabilistic approach for modelling missing data. In his work he considers that each variable has an associated probability of being missing and the process that governs the probability of missing data is the missing data mechanism. Following Little and Rubin (2002), we have two types of variables – the data variable $\boldsymbol{x}$, which has the observed $\boldsymbol{x}^o$ and missing $\boldsymbol{x}^m$ parts, and the binary missing-data indicator $\boldsymbol{m}$, which takes value 0 if the variable is missing, and 1 if it is observed. The missing data mechanism is then described via a conditional distribution $p_{\boldsymbol{\xi}}(\boldsymbol{m} \mid \boldsymbol{x})$. Notice that the distribution of the missing data mechanism depends on both observed and missing data values. For an example of missing data that depends on the value of the missing variable consider a sensor that measures the heat of a device – if the device gets hotter than the sensitivity of a sensor, then the sensor might malfunction and not record the data at all. Such data is known to be Missing Not at Random (MNAR).

Rubin proposed two other classes of missing data mechanisms. Missing at Random (MAR), which assumes that the missing data mechanism depends only on the observed part of the data $\boldsymbol{x}^o$. And Missing Completely at Random (MCAR), which assumes that the missing data mechanism is independent of the data.

When estimating the density model parameters $\boldsymbol{\theta}$ via MLE and the data is MAR or MCAR, then the missing data mechanism is ignorable since

$$p_{\boldsymbol{\theta},\boldsymbol{\xi}}(\boldsymbol{x}^o, \boldsymbol{m}) = \int p_{\boldsymbol{\xi}}(\boldsymbol{m} \mid \boldsymbol{x}^o) p_{\boldsymbol{\theta}}(\boldsymbol{x}^m, \boldsymbol{x}^o) \, \mathrm{d}\boldsymbol{x}^m \tag{2.1}$$

$$= p_{\boldsymbol{\xi}}(\boldsymbol{m} \mid \boldsymbol{x}^o) p_{\boldsymbol{\theta}}(\boldsymbol{x}^o) \propto p_{\boldsymbol{\theta}}(\boldsymbol{x}^o). \tag{2.2}$$

On the other hand, if the data is MNAR, then the missing data mechanism cannot be ignored, since that would result in an incorrect likelihood (Little and Rubin, 2002). In the rest of this work we only consider the case of MAR and MCAR, which simplifies the problem by not requiring us to specify a missing data model.

2.2 Default missing data handling methods

2.2.0.1 Complete-case analysis

Complete-case analysis is the default method for handling low fractions of missing data. The rows that are not fully-observed are removed from the dataset. The main advantage of the complete-case analysis is its simplicity and negligible computational overhead. Moreover, in the case of MCAR it yields a subset of the dataset with unbiased estimates of means and variances (van Buuren, 2018, Section 1.3.1).

However, the method can result in significant loss of precision if the fraction of missingness is large. And in the case of data that is not MCAR the estimates of the means are biased proportionally to the fraction of missing data and the difference between the the observed and missing values (Little and Rubin, 2002, Section 3.2).

2.2.0.2 Unconditional mean imputation

Unconditional mean imputation replaces each missing value with an estimate of the observed unconditional mean of the variable. The method is attractive due to its simplicity and is often used as a quick fix to the missing data problem. However, as Little and Rubin (2002, Section 4.2) point out almost every statistic is biased under mean imputation, and hence it is recommended to avoid this method.

2.2.0.3 Regression imputation

Regression imputation is often called conditional mean imputation. The fully-observed part of the data is used to build regression models for each set of missing values in the incomplete part of the data. The fitted regression models are then used to produce deterministic predictions of the missing values. In some cases the regression model can almost perfectly predict the missing values, however it can then be argued that if the value of a missing variable can be predicted exactly via a deterministic transformation of the observed values, then the missing variable does not carry additional information (van Buuren, 2018, Section 1.3.4).

In general cases where the predictions are not perfect, regression imputations often artificially strengthen the correlations and reduce the variability of the filled-in data (van Buuren, 2018, Section 1.3.4). Moreover, in the case of large fraction of missingness there may not be a fully-observed subset of the dataset to fit the regression model for all sets of missing values.

2.2.0.4 Hot-deck imputation

For a missing value in the j-th dimension, hot-deck imputation replaces the value with a random draw from the empirical distribution of values of similar data points that are observed at the j-th dimension. There are many ad-hoc methods for grouping similar data points, however in our experiments we only consider the hot-deck by simple random sampling with replacement, and thus for more advanced hot-deck methods we refer the interested reader to Little and Rubin (2002, Section 4.3.). The advantage of hot-deck imputation is that the imputed dataset is not distorted as much as mean and regression imputation would Little and Rubin (2002, Section 4.3.2.).

2.2.0.5 Multiple imputation

The imputation methods we have considered so far belong to the class of single imputation. The main advantage of single imputation is that it allows us to apply standard learning methods for complete data. The key disadvantage of single imputation is that it does not account for the uncertainty of the imputed values. Therefore, using standard methods on the filled-in data yields results with underestimated standard errors, and confidence intervals that are too narrow (Little and Rubin, 2002, Chapter 4).

Multiple imputation takes the advantage of single imputation and mitigates the disadvantages. In particular, multiple imputation creates several copies of the dataset with the missing values imputed with random draws from a posterior predictive distribution given the observed values. The expanded dataset is then used with standard methods for learning on complete data. Hence, the bias introduced via multiple imputation is negligible and standard errors are less biased (Little and Rubin, 2002, Chapters 4, 5). Multiple imputation has the added advantage that it can take into account the missing data mechanism, and hence does not rely on the usually unrealistic assumption of MCAR data. Multiple imputation is generally recommended when the posterior predictive distribution is available (Little and Rubin, 2002). However, the posterior distribution is generally not readily available, and hence in the next section we review two iterative methods, which exhibit the advantages of multiple imputation.

2.3 Maximum-likelihood estimation

A common method for fitting a probability density model is maximum-likelihood estimation (MLE), which attempts to optimise the model parameters $\boldsymbol{\theta}$ in order to max-

imise the average log-likelihood of the model under data

$$\ell(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_i).^{1} \tag{2.3}$$

However, when parts of the data are missing at random the log-likelihood is implicitly defined via an integral

$$\ell(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \log \int p_{\boldsymbol{\theta}}(\boldsymbol{x}_i^m, \boldsymbol{x}_i^o)\, \mathrm{d}\boldsymbol{x}_i^m. \tag{2.4}$$

In some density models, such as Gaussian distributions, the integral can be solved analytically, thus MLE can be applied directly. However, for general models the integral is intractable and hence an iterative approach called expectation maximisation is used.

2.3.1 Expectation maximisation

Expectation maximisation (EM) is an incremental MLE algorithm for log-likelihood functions that are intractable due to an intractable integral as in (2.4). The origin of an iterative algorithm dates back to Hartley (1958) and was formalised as the EM algorithm by Dempster et al. (1977). The algorithm has since become one of the most popular methods for optimising latent variable models with an observed part $\boldsymbol{x}$ and unobserved part $\boldsymbol{z}$. For a fixed iteration k and fixed parameters $\boldsymbol{\theta}_k$, the objective is to maximise the following evidence lower bound (ELBO) (Barber, 2017, Section 11.2)

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq \mathbb{E}_{p_{\boldsymbol{\theta}_k}(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z})\right] - \mathbb{E}_{p_{\boldsymbol{\theta}_k}(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}_k}(\boldsymbol{z} \mid \boldsymbol{x})\right], \tag{2.5}$$

When optimising the parameters $\boldsymbol{\theta}$, the second term (entropy) is constant with respect to $\boldsymbol{\theta}$, thus only the first term (energy) needs to be computed

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq \mathbb{E}_{p_{\boldsymbol{\theta}_k}(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z})\right] = Q(\boldsymbol{\theta},\boldsymbol{\theta}_k). \tag{2.6}$$

Each iteration of EM is divided into two steps - the E-step and M-step. In the E-step, for fixed $\boldsymbol{\theta}_k$ the expected complete-data log-likelihood in (2.6) is computed analytically. And the M-step finds the new parameters $\boldsymbol{\theta}_{k+1}$ by maximising this log-likelihood

$$\boldsymbol{\theta}_{k+1} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta},\boldsymbol{\theta}_k). \tag{2.7}$$

Alternatively, in the E-step we can create a multiply-imputed dataset with random draws from the posterior distribution $p_{\boldsymbol{\theta}_k}(\boldsymbol{z} \mid \boldsymbol{x})$. Then, the imputed dataset is used

[1] We use N to denote the number of samples in the dataset and $\boldsymbol{x}_i$ to denote individual samples.

to approximate the expectation using the Monte Carlo method. This type of EM is also known as Monte Carlo EM (Wei and Tanner, 1990). The EM algorithm creates a non-decreasing sequence of likelihoods $\ell(\boldsymbol{\theta}_{k+1}) \geq \ell(\boldsymbol{\theta}_k)$. Therefore, the algorithm continues until convergence to a (local) maximum of the observed data log-likelihood.

The main disadvantage of the EM algorithm is that the E-step may not always be tractable. Thus, in the next section we look into variational inference as a method to relax the objective.

2.3.2 Variational inference

Variational inference (VI) (Jordan et al., 1999) is a successor method of the EM algorithm, which relaxes the EM lower bound by uncoupling the parameters of the model and posterior via a variational posterior approximation. In VI, we must specify an auxiliary variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x})$ family with $\boldsymbol{\phi}$ as parameters that will be optimised to approximate the true posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})$. Hence the problem of intractable posterior inference in EM is cast as a tractable optimisation problem. By replacing $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})$ with $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x})$ in (2.5) we get the variational ELBO or the (negative) variational free energy

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x})}\right] = \mathcal{J}(\boldsymbol{\theta},\boldsymbol{\phi}). \tag{2.8}$$

In theory, the bound is tight whenever $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}) = p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})$. This can be shown more formally by rewriting the free energy as (Barber, 2017, Section 11.2.)

$$\mathcal{J}(\boldsymbol{\theta},\boldsymbol{\phi}) = -D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})) + \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) \tag{2.9}$$

Rearranging the equation we get

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) = D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})) + \mathcal{J}(\boldsymbol{\theta},\boldsymbol{\phi}), \tag{2.10}$$

where D_{KL} is the Kullback-Leibler divergence (Kullback and Leibler, 1951). D_{KL} is a non-negative quantity and is 0 whenever $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}) = p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})$. Since the KL-divergence is non-negative and for fixed $\boldsymbol{x}$ the sum in (2.10) is constant, maximising the free energy is equivalent to minimising the KL-divergence. Therefore, if the family of variational distributions is sufficiently flexible, then $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x})$ will closely approximate $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x})$.

Mean-field variational inference

A classical approach to parameterising the variational distribution q is mean-field variational inference (MFVI) (Zhang et al., 2018). For a dataset $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ and

corresponding unobserved variables $\mathcal{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n\}$ the mean-field assumption factorises the variational distribution

$$q(\mathcal{Z} \mid \mathcal{D}) = \prod_{i=1}^{N} q(\boldsymbol{z}_i \mid \boldsymbol{x}_i; \boldsymbol{\phi}_i). \tag{2.11}$$

Hence, for each data point we have to specify and optimise local parameters $\boldsymbol{\phi}_i$. There are several issues with this approach. First, the number of parameters scales linearly with the size of the dataset. Second, to perform inference on new observations, the optimisation procedure for the new parameters would have to be run. Finally, it is computationally inefficient to optimise a separate parameter set for each data point.

Amortised inference

Amortised inference (Gershman and Goodman, 2014) is an alternative to MFVI that is typically more efficient and scalable to large datasets. The main idea in amortised inference is that instead of optimising the local parameters directly, we can replace them by a deterministic function from the observed data to the parameters of the posterior, and optimise the globally shared parameters $\boldsymbol{\phi}$ of the function. In practice, the function is often a neural network, also called an inference network (Kingma and Welling, 2013). The advantages of amortised inference are two-fold. First, the optimisation of globally shared parameters is typically more efficient. And second, we can immediately perform inference on new observations using the shared inference network.

Optimising the variational parameters

In classical VI, the ELBO in (2.8) would be first derived analytically and then optimised. However, the analytic approach is typically restricted to conditionally conjugate distributions (see Hoffman et al., 2013), and does not work in general for amortised inference. The problem is that the ELBO contains an expectation with respect to $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x})$, hence differentiating with respect to $\boldsymbol{\phi}$ is difficult. The two approaches that allow us to use amortised inference are the reparameterisation trick (Kingma and Welling, 2013) and black-box variational inference (Ranganath et al., 2013).

The reparameterisation trick simplifies the computation of the gradient via a deterministic mapping of samples from an independent auxiliary noise distribution. For example, samples from a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance Σ can be obtained as follows

$$\boldsymbol{z} = \boldsymbol{\mu} + L\boldsymbol{\varepsilon} \text{ , where } \boldsymbol{\varepsilon} \sim \mathcal{N}(0, I) \text{ and } \Sigma = LL^{\top}. \tag{2.12}$$

We can rewrite the expectation over $\boldsymbol{z}$ of any function f as an expectation over $\boldsymbol{\varepsilon}$

$$\mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\phi}}(\boldsymbol{z})}\left[f(\boldsymbol{z})\right] = \mathbb{E}_{\boldsymbol{\varepsilon}\sim\mathcal{N}(0,I)}\left[f(g(\boldsymbol{\varepsilon};\boldsymbol{\phi}))\right], \tag{2.13}$$

where g is a deterministic mapping from the noise distribution to the sampled distribution, and thus shares parameters $\boldsymbol{\phi}$. Provided that f and g are differentiable, we can easily differentiate this expectation by taking the derivative inside of the expectation. The reparameterisation trick can be applied to any distribution that can be expressed as a deterministic mapping of a simpler independent distribution.

The black-box variational inference (BBVI) approach is more general than the reparameterisation trick as it can be applied to any variational distribution family. However, the cost of using BBVI is higher gradient variance, and hence requires variance reduction methods. In our work, we employ the reparameterisation trick, and hence refer the interested reader to Ranganath et al. (2013); Zhang et al. (2018) for more details on BBVI.

2.4 Gibbs sampler

The common theme between multiple imputation, expectation maximisation, and variational inference is that each method attempts to fill in the unobserved values with random draws from a posterior predictive distribution. In particular, in EM and VI this corresponds to evaluating the expected complete data log-likelihood. A popular alternative for sampling from the (approximate) posterior distribution is a class of Markov chain Monte Carlo (MCMC) methods (Robert and Casella, 2004).

The main idea in MCMC methods is to generate a sequence of dependent samples via a transition operation, such that the generated Markov chain ultimately samples the true distribution (Murray, 2007). Gibbs sampling (Geman and Geman, 1984) is an MCMC method, which sequentially resamples each dimension j of a joint density $p(\boldsymbol{x})$ via the conditional distribution $p(x^j \mid \boldsymbol{x}^{\setminus j})$. The chain created by sequential Gibbs sampling, under certain conditions, is provably ergodic (Geman and Geman, 1984, Theorem A) and converges to the distribution from which the conditionals were derived. Sampling from a posterior distribution $p(\boldsymbol{x}^m \mid \boldsymbol{x}^o)$ can be achieved by fixing the observed variables $\boldsymbol{x}^o$ to the observed values and sampling only the values in $\boldsymbol{x}^m$.

In the literature, MCMC methods and variational inference are often treated as alternative methods for similar estimation problems. MCMC is often recommended when approximately exact samples are required and compute power is not an issue.

VI generally scales better to large datasets and is recommended when fast posterior inference is required. In this work, we use them as complementary methods. We use VI to learn the univariate conditional distributions, and use Gibbs sampler to estimate samples from a more complex approximate joint posterior distribution.

Parallel sampler

Due to strong dependencies between consecutive samples, standard Gibbs sampling can be slow to converge for high-dimensional distributions. When the probability graph is known, the Gibbs updates can be parallelised by exploiting the conditional independencies to identify groups of variables that can be sampled in parallel (Gonzalez et al., 2011). Unfortunately, in general the conditional independencies are not known and there is no other Gibbs sampler that can guarantee ergodicity of the Markov chain (Angelino et al., 2016). However, a class of Hogwild Gibbs methods have been empirically shown to achieve good performance (Angelino et al., 2016) even without the theoretical guarantees. Hence, to speed up the convergence of our method on high-dimensional data we will use the earliest type of Hogwild Gibbs method, known as Synchronous Gibbs sampler (Geman and Geman, 1984). The method essentially ignores the dependencies of the data and samples all variables in parallel anyway.

For more details on parallel sampling schemes for Gibbs sampler refer to Angelino et al. (2016).

2.5 Factor analysis

Factor analysis (FA) is a fundamental probability density model in statistical analysis, which originated in psychology for human intelligence research in the seminal paper (Spearman, 1904). Since its introduction the method has been widely adopted in psychology, biology, operations research, finance, and many other fields.

Factor analysis attempts to find unobserved factors $\boldsymbol{z}$ that explain the correlations in usually higher-dimensional observable data $\boldsymbol{x}$. The prior distribution of the factors is assumed uncorrelated standard Gaussian $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; 0, I)$ (I is the identity matrix), and the conditional distribution of $\boldsymbol{x}$ is

$$p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; F\boldsymbol{z} + \boldsymbol{\mu}, \Psi), \tag{2.14}$$

where the parameters $\boldsymbol{\theta}$ are the observable mean vector $\boldsymbol{\mu}$, the factor matrix F, and a diagonal observation noise matrix Ψ.

The generative process of $\boldsymbol{x}$ can be written as

$$\boldsymbol{x} = F\boldsymbol{z} + \boldsymbol{\mu} + \boldsymbol{\varepsilon} \text{ , where } \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Psi). \tag{2.15}$$

Hence, the covariance of $p(\boldsymbol{x})$ is

$$\begin{aligned} \Sigma_{\boldsymbol{x}} &= \text{Cov}(F\boldsymbol{z} + \boldsymbol{\mu} + \boldsymbol{\varepsilon}) && (2.16)\\ &= \text{Cov}(F\boldsymbol{z}) + \text{Cov}(\boldsymbol{\mu}) + \text{Cov}(\boldsymbol{\varepsilon}) && (2.17)\\ &= FIF^{\top} + \Psi && (2.18) \end{aligned}$$

Note that the choice of the prior probability $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; 0, I)$ does not affect the flexibility of the FA model – any correlation in the prior can be absorbed into F in (2.18).

In recent machine learning practice often more advanced non-linear models are used, such as Variational Autoencoders (VAEs) (Kingma and Welling, 2013). However, factor analysis model possesses some useful qualities for the evaluation of a new training algorithm in the presence of missing data. In particular, Williams et al. (2018) have derived the posterior distribution $p(\boldsymbol{z} \mid \boldsymbol{x}^o) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_{\boldsymbol{z}|\boldsymbol{x}^o}, \Sigma_{\boldsymbol{z}|\boldsymbol{x}^o})$ conditioned on incomplete data with parameters

$$\Sigma_{\boldsymbol{z}|\boldsymbol{x}^o} = \left(I + F^{\top} M \Psi^{-1} M F\right)^{-1} \tag{2.19}$$

$$\boldsymbol{\mu}_{\boldsymbol{z}|\boldsymbol{x}^o} = \Sigma_{\boldsymbol{z}|\boldsymbol{x}^o} F^{\top} M \Psi^{-1} (\boldsymbol{x} - \boldsymbol{\mu}), \tag{2.20}$$

where M is a diagonal matrix with the diagonal element m^{jj} being 0 if the value of x^j is missing, and 1 if it is observed [2]. Hence, in a FA model we can evaluate the posterior of the missing values $p(\boldsymbol{x}^m \mid \boldsymbol{x}^o)$ in closed-form.

Moreover, the authors proposed several closed-form approximations to the exact posterior, which may serve as a useful baseline. The full covariance approximation (FCA) sets the missing values $\boldsymbol{x}^m$ to the corresponding values in the observable mean vector $\boldsymbol{\mu}$ to give $\tilde{\boldsymbol{x}}$, so the elements corresponding to the missing dimensions in $\tilde{\boldsymbol{x}} - \boldsymbol{\mu}$ are zero in (2.20) and hence do not affect the $\boldsymbol{\mu}_{\boldsymbol{z}|\boldsymbol{x}^o}$. Then, it computes the posterior as if the data was complete by setting the diagonal of M to 1 in (2.19) and (2.20).

The scaled covariance approximation (SCA) similarly sets missing values to the observable mean. And then linearly interpolates the posterior covariance of $\boldsymbol{z}$ between the prior and the posterior for fully observed data depending on the fraction of missing data. Refer to Williams et al. (2018) for more details.

[2] We will use the superscript j to refer to the j-th dimension. So x^j corresponds to the j-th element of $\boldsymbol{x}$.

Chapter 3

Related work and research questions

3.1 Related work

Recently, the machine learning community has expressed a renewed interest in incomplete data handling. The approaches can be classified into two groups: deterministic and probabilistic methods. Deterministic methods attempt to predict a single best imputation value, whereas probabilistic methods attempt to learn the posterior predictive distribution over the missing values. In the manner of multiple imputation theory we focus on the probabilistic methods.

Deep generative models such as Variational Autoencoders (VAEs) (Kingma and Welling, 2013) and Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) have been at the forefront of research and proven to be powerful tools in many application domains. Early work on VAEs has shown great capability to inpaint incomplete images (Vincent et al., 2008), but required complete data at training. Recently, progress has been made in handling incomplete data during training by making certain simplifying assumptions either about the inference network $p(\boldsymbol{z} \mid \boldsymbol{x})$ or the generative network $p(\boldsymbol{x} \mid \boldsymbol{z})$ (Vedantam et al., 2017; Nazabal et al., 2018; Wu and Goodman, 2018; Ivanov et al., 2019). In the context of GANs, Generative Adversarial Imputation Network (GAIN) has recently been proposed (Yoon et al., 2018). GAIN is composed of two networks, a generative network, which attempts to learn $p(\boldsymbol{x}^m \mid \boldsymbol{x}^o)$, and a discriminator network, which attempts to distinguish which values are observed and which are imputed, which corresponds to predicting the missingness mask $\boldsymbol{m}$. VAE and GAN models are the current state-of-the-art, however, neither is a silver bullet to solve all problems. For example, VAEs assume a conditional independence of the observable data $\boldsymbol{x}$ given $\boldsymbol{z}$, hence in the context of missing data $\boldsymbol{x}^m$ is conditionally independent

of $\boldsymbol{x}^o$ given $\boldsymbol{z}$, which may not be a reasonable assumption for some problems. Moreover, GAN training often suffers from density mode collapse (Salimans et al., 2016), which means that the posterior density $p(\boldsymbol{x}^m \mid \boldsymbol{x}^o)$ may be too narrow. Hence, it is desirable to diversify the machine learning toolbox with approaches based on additional frameworks.

Expectation maximisation (EM) is another tool for handling incomplete data directly by treating the missing variables as latents. The application to missing data has been mentioned in the seminal paper by Dempster et al. (1977) and later demonstrated for learning a Gaussian mixture model by Ghahramani and Jordan (1994) from incomplete data. However, the inference of posterior density $p(\boldsymbol{x}^m \mid \boldsymbol{x}^o)$ required by the E-step is generally intractable.

Kim (2011) proposed an EM-type Fractional Imputation (FI) algorithm. In FI the dataset is first imputed with multiple draws from an arbitrary proposal distribution. Then, the algorithm iterates between computing fractional weights of the imputed samples in the spirit of importance weights, and updating the density parameters via pseudo maximum likelihood estimation weighting the likelihood of each sample by its fractional weight. Unlike EM, it is not necessary to impute the dataset with new samples at each iteration, only the fractional weights are recomputed. Uehara et al. (2019) demonstrated the algorithm for learning unnormalised density models and suggested a heuristic for choosing appropriate proposal distribution based on the observed data. However, choosing an appropriate proposal distribution can still be difficult in practice.

A closely related method to EM is the data augmentation (DA) algorithm (Tanner and Wong, 1987). DA is a Markov chain Monte Carlo method for Bayesian inference with missing data [1]. The DA algorithm alternates between two steps – an I-step (imputation) in which the missing data are imputed with samples from the posterior distribution of missing values $p(\boldsymbol{x}^m_{t+1} \mid \boldsymbol{x}^o, \boldsymbol{\theta}_t)$, and a P-step (posterior), in which the new parameters $\boldsymbol{\theta}_{t+1}$ are drawn from $p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{x}^o, \boldsymbol{x}^m_{t+1})$. Tanner and Wong (1987) have shown that the sequence of updates converges to the stationary distribution of $p(\boldsymbol{x}^o, \boldsymbol{\theta})$. Note that the I-step is equivalent to the E-step of Monte Carlo EM, and hence it is also generally not tractable for all density models.

[1] In Bayesian inference the whole likelihood function is estimated, in contrast to MLE where we try to only find the maximum likelihood.

3.2 Variational inference for incomplete data

A natural alternative for intractable E-step in EM is variational inference. With variational inference we transform the intractable inference of the posterior distribution problem into a more tractable optimisation problem. The variational ELBO for incomplete data is (for derivation see Appendix A.1)

$$\mathcal{J}_{\text{C-ELBO}}(\boldsymbol{\theta}, \boldsymbol{\phi}_i) = \frac{1}{N} \sum_{i}^{N} \mathbb{E}_{q_{\boldsymbol{\phi}_i}(\boldsymbol{x}_i^m | \boldsymbol{x}_i^o)} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_i^m, \boldsymbol{x}_i^o)}{q_{\boldsymbol{\phi}_i}(\boldsymbol{x}_i^m \mid \boldsymbol{x}_i^o)} \right], \tag{3.1}$$

where $\boldsymbol{\phi}_i$ corresponds to the parameters of the variational distribution of the i-th realisation.

However, the definition of $\mathcal{J}_{\text{C-ELBO}}$ requires $2^D - 1$ variational distributions, one for each pattern of missingness, hence we call it the combinatorial-ELBO. Maintaining a separate variational model for each pattern of missingness is undesirable for two reasons. First, it is computationally inefficient due to the lack of parameter sharing between different data points. Second, it quickly becomes intractable for higher-dimensional data.

A simplifying solution is to specify a joint variational distribution over all variables, including the observed variables, where all conditional distributions can be computed in closed form. However, the choice of distributions where the conditionals can be computed in closed form can be too restrictive.

3.2.1 Difficulties in amortised inference of missing data

In practice, amortised inference networks (see Section 2.3.2) are often used to facilitate efficient parameterisation and learning. In the VI literature usually a single variational distribution is required, which is then parameterised by a single inference network. In contrast, it is not obvious how to parameterise $2^D - 1$ variational distributions with a tractable number of inference networks.

In a special case of a multivariate Gaussian posterior one could specify a single inference network to predict all posterior covariance coefficients and posterior means, and then select the outputs corresponding to the missing value dimensions to get $q_{\boldsymbol{\phi}}(\boldsymbol{x}^m \mid \boldsymbol{x}^o)$. In order to use neural network architectures, such as fully-connected or convolutional neural networks, the dimensionality of the input to the neural network should be fixed. Hence, we could set the input as all the variables and set the missing values to zero following Nazabal et al. (2018); Ivanov et al. (2019). This is easily

tractable in multivariate Gaussian posteriors, but it is not generally tractable for all choices of posterior distributions.

Alternatively, analogous to mean-field VI (Section 2.3.2), one could assume independence of the missing variables given the observed

$$q_{\boldsymbol{\phi}}(\boldsymbol{x}_i^m \mid \boldsymbol{x}_i^o) = \prod_{j \in \boldsymbol{m}_i} q_{\boldsymbol{\phi}}(x_i^j \mid \boldsymbol{x}_i^o), \tag{3.2}$$

where $\boldsymbol{m}_i$ denotes the set of indices of the missing values in the i-th sample. The factorisation requires us to specify D variational posteriors, one for each dimension, and hence the ELBO is now tractable in general. However, since the samples $\boldsymbol{x}^m$ from the variational posterior are used to optimise the density model, the assumptions in the variational distribution transfer to the density model. Hence, imputing potentially correlated variables with samples of independent variables will introduce additional bias to the model by artificially reducing correlations in the data.

Moreover, the introduced bias increases with fraction of missingness in the dataset because imputed values $\boldsymbol{x}^m$ from the variational posteriors dominate the training samples $\boldsymbol{x}$ when training the model. It is therefore evident that the conditional independence assumption in (3.2) is too strong to learn accurate density models from highly incomplete datasets [2].

Hence, specifying an efficient and accurate variational distribution model for missing data is an important open question.

3.3 Research questions and goals

We have so far established two fundamental properties that the model of variational distributions must possess in order to efficiently learn accurate density models:

- It must mitigate the combinatorial explosion of $2^D - 1$ variational posteriors.
- It must not make unrealistic independence assumptions.

It is not immediately obvious how both of these objectives can be accomplished at once.

In this thesis we adapt the unpublished Cumulative Data Imputation (CDI) algorithm (Rhodes, 2018), which mitigates the combinatorial explosion of variational posteriors to D variational distributions that can be parameterised with D inference

[2] We also prove this claim empirically in Section 5.3.1.

networks. In the original work, the author empirically demonstrated the proposed algorithm's ability to recover from incomplete data the probabilistic graph structure characterised by a truncated Gaussian model. Nevertheless, the work did not provide a theoretical justification of the convergence properties of the algorithm, nor did it quantify the impact CDI had on the quality of the fitted model. Hence, providing theoretical justification and a comprehensive evaluation of the (modified) CDI algorithm is an important part of our work. Our work is presented in the following chapters

Cumulative Data Imputation We present the modified CDI algorithm, provide a theoretical justification of convergence based on the convergence of Gibbs sampler, and discuss approaches to scaling CDI to high-dimensional data.

Simulations We evaluate CDI for fitting a factor analysis density model on two synthetic and one real datasets. We are interested in (1) evaluating the impact of CDI on the fitted density model quality and the quality of the variational approximation of the true posterior distribution, (2) comparing CDI to the use of default missing data handling methods, and (3) evaluating the accuracy of point-estimates from the learnt variational models.

Chapter 4

Cumulative data imputation

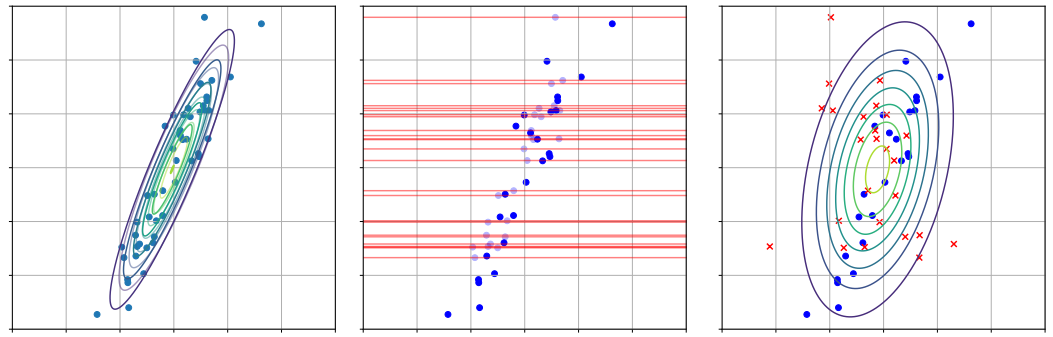
In this chapter we present an approximate probability density estimation method for incomplete data, called the cumulative data imputation (CDI), based on variational inference (VI) (Jordan et al., 1999) and Gibbs sampling (Geman and Geman, 1984). Our work is based on an unpublished algorithm in Rhodes (2018) of the same name. We propose a variational objective for normalised densities and an iterative data imputation scheme that corresponds to Gibbs sampling, as well as, a comprehensive evaluation in the following chapter.

CDI is an iterative algorithm that alternates between training the density model parameters and updating imputations of missing data. We begin by illustrating the motivation for CDI with a toy example, which corresponds to a special case of EM. We then present the modified CDI algorithm and the variational optimisation objective. Finally, we discuss the convergence properties of the algorithm and approaches for scaling CDI to high-dimensional data.

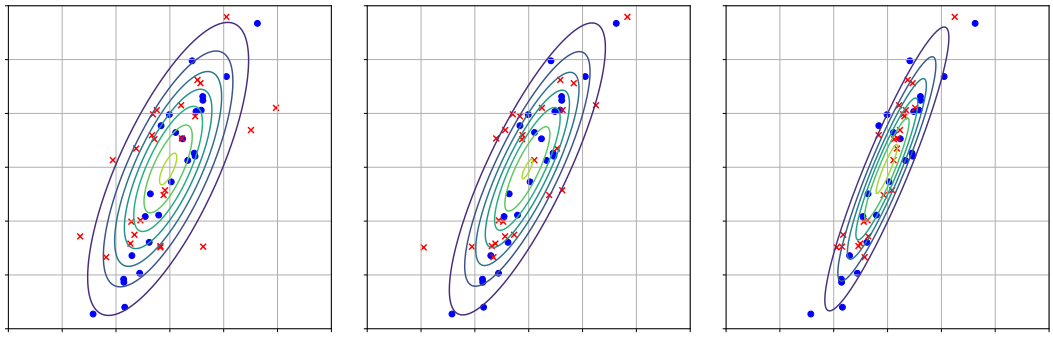
4.1 Illustrative motivation for CDI

Consider a two-dimensional dataset where each datapoint $\boldsymbol{x} = (x^0, x^1)$ comes from a bivariate Gaussian distribution (Figure 4.1a). Assume an idealised problem, where half of the values in the x^0 dimension are missing (Figure 4.1b), and we want to fit a bivariate Gaussian density model $p_{\boldsymbol{\theta}}$ to the observed data.

Since the posterior $p_{\boldsymbol{\theta}}(x^0 \mid x^1)$ can be computed in closed form, we can fit the density model by adopting a Monte Carlo EM approach (Section 2.3.1). At first we impute the missing values from an unfit posterior $p_{\boldsymbol{\theta}}(x^0 \mid x^1)$, denoted by red crosses in Figure 4.1c. Next, we use the completed dataset in maximum likelihood estimation



(a) True density and samples. (b) Incomplete samples. (c) 1st epoch.



(d) 2nd epoch. (e) 3rd epoch. (f) 10th epoch.

Figure 4.1: An illustration of an EM-type algorithm on a simple problem using true model posteriors. Figure 4.1a shows the true distribution and 50 realisations. Figure 4.1b shows the data where half of the values in the x^0 dimension are missing – the red lines refer to the possible imputation values. Figures 4.1c–4.1f visualise the algorithm at epochs 1, 2, 3, and 10. Each of the figures displays the fully observed values as blue dots, the imputed values (at the start of the epoch) as red crosses, and the contours of a fitted Gaussian model on the imputed and fully-observed data.

of the parameters $\boldsymbol{\theta}$ to fit the density (see contours in Figure 4.1c).

At the end of the first iteration the correlation of the two variables in the learnt model is biased towards a diagonal Gaussian due to sub-optimal initialisation. However, if we repeat the iteration again, the imputed values from the updated posterior are closer to the true distribution, and hence the fitted density model is also closer to the true density than after the first iteration (Figure 4.1d). Continuing for more iterations we observe that the density model approaches the true density (Figure 4.1f).

The toy example has one desirable property – we only require univariate posteriors. Hence, in the case of approximate posteriors we would only need to specify D variational distributions for a problem where only one value is missing per data

Algorithm 1 Cumulative data imputation (CDI) algorithm

Require: Density model $p_{\boldsymbol{\theta}}$ and D conditional distributions $q_{\boldsymbol{\phi}^j}(x^j \mid \boldsymbol{x}^{\setminus j})$.

Require: Data set $\mathcal{D}$ with observed and missing subsets X^o and X^m.

Initialisation: Impute each missing value in X^m with an initial guess.

while $p_{\boldsymbol{\theta}}$ not converged **do**

Select mini-batch $\mathcal{B}$ from $\mathcal{D}$.

Optimise $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ using $\mathcal{J}_{\text{CDI-ELBO}}$ on $\mathcal{B}$.

for $\boldsymbol{x} = (\boldsymbol{x}^m, \boldsymbol{x}^o)$ in $\mathcal{B}$ **do**

Select a variable x^j from the missing component $\boldsymbol{x}^m$.

$x^j \leftarrow \tilde{x}^j \sim q_{\boldsymbol{\phi}^j}(x^j \mid \boldsymbol{x}^{\setminus j})$

end for

end while

point. Obviously, this is an unrealistic assumption and we would like to extend the use of univariate posteriors for data points with multiple missing values. Notice that the univariate posterior $p_{\boldsymbol{\theta}}(x^0 \mid x^1)$ used in the above example corresponds to the conditional distribution used in Gibbs sampler (Section 2.4). In a more general case, if the data has multiple missing values per data point and we know all univariate conditional distributions, then Gibbs sampling could be used to impute with joint samples of the missing data. The rest of this chapter presents the CDI algorithm, which follows a similar EM-type approach, approximates the Gibbs sampler conditionals with variational distributions, and uses them to sample joint missing value imputations.

4.2 The algorithm

The CDI algorithm (Algorithm 1) is divided into three stages: initialisation, parameter optimisation, and data imputation. During initialisation we fill in the missing values with an initial guess. The initial guess may depend on the probability distribution being learnt but generally we can use the unconditional empirical mean, a random sample from the unconditional empirical distribution, or any initial imputation scheme which imputes values which have a positive probability under the true distribution. Then the algorithm alternates until convergence between fitting the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, and dataset imputation.

First, the algorithm updates the model parameters $\boldsymbol{\theta}$ and the parameters $\boldsymbol{\phi}^j$ of the univariate variational distributions $q_{\boldsymbol{\phi}^j}(x^j \mid \boldsymbol{x}^{\setminus j})$ for all dimensions j that have missing

values, using the $\mathcal{J}_{\text{CDI-ELBO}}$ defined below and a mini-batch $\mathcal{B}$. Using the standard derivation of variational ELBO (equivalent to the derivation in A.1), we derive the CDI training objective – an ELBO for the univariate variational posterior $q_{\boldsymbol{\phi}^j}(x_i^j \mid \tilde{\boldsymbol{x}}_i^{\setminus j})$ assuming that we have data $\tilde{\boldsymbol{x}} = (\tilde{\boldsymbol{x}}^m, \boldsymbol{x}^o)$ where the missing values are filled-in with a previous guess from an approximate joint distribution

$$\mathcal{J}^{j,i}_{\text{U-ELBO}}(\boldsymbol{\theta}, \boldsymbol{\phi}^j) = \mathbb{E}_{q_{\boldsymbol{\phi}^j}(x_i^j \mid \tilde{\boldsymbol{x}}_i^{\setminus j})} \left[\log \frac{p_{\boldsymbol{\theta}}(x_i^j, \tilde{\boldsymbol{x}}_i^{\setminus j})}{q_{\boldsymbol{\phi}^j}(x_i^j \mid \tilde{\boldsymbol{x}}_i^{\setminus j})} \right], \tag{4.1}$$

where the index i corresponds to an individual sample in the dataset, and j is a dimension of the sample with missing value. However, there is a problem with using this ELBO directly – we would only optimise one variational distribution at a time[1], hence in high-dimensional datasets the updates might be scarce and CDI will converge slowly. We propose to take an average over all $\mathcal{J}^{j,i}_{\text{U-ELBO}}(\boldsymbol{\theta}, \boldsymbol{\phi}^j)$ for all missing values $j \in \boldsymbol{m}_i$, such that the updates are more frequent, which is similar to the approach taken in Wu and Goodman (2018)

$$\mathcal{J}^{i}_{\text{CDI-ELBO}}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \frac{1}{|\boldsymbol{m}_i|} \sum_{j \in \boldsymbol{m}_i} \mathcal{J}^{j,i}_{\text{U-ELBO}}(\boldsymbol{\theta}, \boldsymbol{\phi}^j), \tag{4.2}$$

where $\boldsymbol{\phi}$ is the set of parameters $\boldsymbol{\phi}^j$ for all variational distributions. Hence, the full CDI-ELBO is given by

$$\mathcal{J}_{\text{CDI-ELBO}}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{|\boldsymbol{m}_i|} \sum_{j \in \boldsymbol{m}_i} \mathcal{J}^{j,i}_{\text{U-ELBO}}(\boldsymbol{\theta}, \boldsymbol{\phi}^j). \tag{4.3}$$

In fact, $\mathcal{J}_{\text{CDI-ELBO}}$ could be generalised, such that the average is taken over a subset of the missing values $\boldsymbol{k}_i \subseteq \boldsymbol{m}_i$ at a time, allowing us to tune the computational cost of an update at each epoch. The parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ can then be optimised via stochastic gradient ascent on $\mathcal{J}_{\text{CDI-ELBO}}$.

Next, the algorithm must select a missing variable x^j from $\boldsymbol{x}^m$ for each observation in mini-batch $\mathcal{B}$ to be updated with a sample from the variational posterior, which corresponds to a single step of Gibbs sampling. In Gibbs sampler literature the update order is often called the scan order. There are two common types of scan order – random scan and systematic scan. In random scan, the variables are selected independently and uniformly at random at each iteration. In systematic scan, the variables are sequentially selected in a pre-defined order. The scan order can have a significant impact on the convergence of the Markov chain on some problems (He et al., 2016),

[1] The original CDI algorithm also optimises one variational distribution at a time.

however, the asymptotic convergence is guaranteed for both. The implementation of the random scan is generally simpler, and hence we choose random scan in this work.

Once the missing variables are selected, the algorithm completes the iteration by resampling the selected variables in the current mini-batch from the updated variational posterior distributions and updates the values of the selected variables in the dataset. Note that the original algorithm imputed the selected missing values with the expectation $\mathbb{E}_{q_{\boldsymbol{\phi}^j}(x^j|\tilde{\boldsymbol{x}}^{\setminus j})}[x^j]$, however, as noted before in Ghahramani and Jordan (1994), this fails even for the simple case of multivariate Gaussian density model, since the imputations will always lie along a line, and hence bias the estimate of covariance.

4.3 Convergence of CDI

In CDI we perform a single update from the variational posterior at each iteration, which corresponds to a single step of standard Gibbs sampling. Consider an alternative version of the algorithm, where instead of updating the dataset with a single update from the variational distribution at the end of an iteration, we perform multiple steps of Gibbs sampling from a randomly initialised state at the start of the iteration to fill in the dataset with samples from the joint posterior. Also assume that the variational posteriors $q_{\boldsymbol{\phi}^j}(x^j \mid \boldsymbol{x}^{\setminus j})$ approximate the true posteriors $p_{\boldsymbol{\theta}}(x^j \mid \boldsymbol{x}^{\setminus j})$ (we discuss the necessary conditions in the next section). We can then use the convergence proof of Theorem A by Geman and Geman (1984) which states that under mild assumptions for any starting state the Markov chain converges to the stationary distribution from which the conditional distributions were derived. Hence, assuming that the Gibbs sampler is run for long enough, the imputed missing values are draws from the approximate joint posterior $q_{\boldsymbol{\phi}}(\boldsymbol{x}^m \mid \boldsymbol{x}^o)$. Then, the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are updated to maximise the log-likelihood on the completed data, which corresponds to the M-step of an EM-type approach.

Obviously running many steps of Gibbs sampler at each iteration can be prohibitively expensive. The difference between the alternative algorithm above, which performs multiple steps of standard Gibbs sampling in each iteration, and CDI is that in CDI we only perform one step of Gibbs sampling at each iteration and persist the current state of the Markov chain by imputing the missing values in the dataset. Note that the sampling distribution changes between each iteration of CDI. However, the existing state of the Markov chain should be closer to a sample from the true distribution than a randomly initialised state, hence we can use the existing chain to derive

samples from the updated distribution. A similar MCMC sampling approach that persists Markov chain for a changing sampling distribution has been used in persistent contrastive divergence by Tieleman (2008).

We can also show that if the Markov chain on the imputed values is at the stationary distribution $(x_{t-1}^j, \boldsymbol{x}_{t-1}^{\setminus j}) \sim p_{\boldsymbol{\theta}}(\boldsymbol{x})$[2], then resampling $x_t^j \sim p_{\boldsymbol{\theta}}(x_t^j \mid \boldsymbol{x}_{t-1}^{\setminus j})$ will not change the joint distribution and hence CDI converges

$$(x_t^j, \boldsymbol{x}_{t-1}^{\setminus j}) \sim \int p_{\boldsymbol{\theta}}(x_{t-1}^j, \boldsymbol{x}_{t-1}^{\setminus j}) p_{\boldsymbol{\theta}}(x_t^j \mid \boldsymbol{x}_{t-1}^{\setminus j}) \, \mathrm{d}x_{t-1}^j \tag{4.4}$$

$$= p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}^{\setminus j}) p_{\boldsymbol{\theta}}(x_t^j \mid \boldsymbol{x}_{t-1}^{\setminus j}) \tag{4.5}$$

$$= p_{\boldsymbol{\theta}}(x_t^j, \boldsymbol{x}_{t-1}^{\setminus j}). \tag{4.6}$$

4.4 Convergence of the variational posteriors

For arbitrarily chosen conditional distributions $q(x^j \mid \boldsymbol{x}^{\setminus j})$ a unique joint distribution may not even exist, such conditionals are called incompatible (Chen and Ip, 2015). However, if the conditional distributions closely approximate the true posteriors $p_{\boldsymbol{\theta}}(x^j \mid \boldsymbol{x}^{\setminus j})$ of a density model $p_{\boldsymbol{\theta}}$, then the conditional distributions are compatible and Gibbs sampling will have a unique stationary distribution, which approximates the true posterior $p_{\boldsymbol{\theta}}(\boldsymbol{x}^m \mid \boldsymbol{x}^o)$ of the model. Hence, it is important to understand the conditions that are required for the variational distributions $q_{\boldsymbol{\phi}^j}(x^j \mid \boldsymbol{x}^{\setminus j})$ to approximate $p_{\boldsymbol{\theta}}(x^j \mid \boldsymbol{x}^{\setminus j})$ accurately enough.

First, the specified family of univariate variational distributions should include the true univariate posterior distributions or be flexible enough to approximate it with an arbitrary accuracy. The advantage of using univariate variational distributions in CDI is that it is generally easier to specify a good choice of univariate distribution family than to specify a joint variational posterior family. Second, for the posterior distributions $q_{\boldsymbol{\phi}^j}(x^j \mid \boldsymbol{x}^{\setminus j})$ and $p_{\boldsymbol{\theta}}(x^j \mid \boldsymbol{x}^{\setminus j})$ to match exactly, the probability density functions must be equal for all $x^j \in \mathbb{R}$ and $\boldsymbol{x}^{\setminus j} \in \mathbb{R}^D$. Hence, as with all probability density models the match can be exact only in the presence of infinitely many data.

In practice we have finite data that only covers a subspace of the probability space. Therefore, at any time, the variational distribution can only match the model's true posterior distribution up to a limit defined by the number of realisations in the sample space of the true $\boldsymbol{x}^{\setminus j}$. Moreover, when training $q_{\boldsymbol{\phi}^j}(x^j \mid \boldsymbol{x}^{\setminus j})$ only a subset of the dataset is used with samples restricted to those where x^j is missing. This can affect

[2]The subscripts t and $t-1$ correspond to timestep indicators.

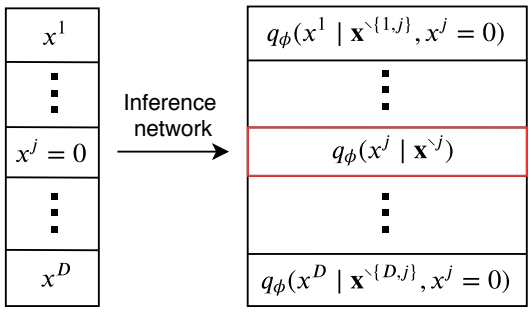


Figure 4.2: Shared variational model for inferring the posterior $q_{\boldsymbol{\phi}}(x^j \mid \boldsymbol{x}^{\setminus j})$.

the accuracy of the variational posterior approximation, particularly at low missingness fractions. In order to mitigate this, one could specify an input-dropout probability p, such that at each iteration with probability p each of the observed values in $\boldsymbol{x}^o$ can be treated as missing and added to the vector $\boldsymbol{x}^m$ in the parameter optimisation step of CDI, however the values of the dropped-out variables in $\boldsymbol{x}^o$ should not be updated during the data imputation step. The hyperparameter p could be tuned to improve the accuracy of the variational distributions.

4.5 Scaling CDI

Training with CDI can potentially be quite expensive for high-dimensional datasets with significant fraction of missingness. First, optimising $\mathcal{J}^i_{\text{CDI-ELBO}}$ in (4.2) requires evaluating $|\boldsymbol{m}_i|$ variational inference networks and probability density models. Second, standard Gibbs updates can be too slow to converge for high-dimensional datasets with large fraction of missingness. In Section 4.2 we mentioned a potential way to tune the cost of $\mathcal{J}^i_{\text{CDI-ELBO}}$ by optimising for a random subset of missing values in $\boldsymbol{x}^m$ and treating the rest as observed. In this section, we consider two additional approaches to scale CDI to high-dimensional data.

Shared variational model

So far we have established that we need D variational posteriors and thus D inference networks. There are several issues with using D inference networks: it can be memory expensive to store the weights of the inference networks and it is not obvious how to efficiently parallelise the computation of $|\boldsymbol{m}_i|$ variational posteriors, such that it leverages efficient matrix operation implementations such as BLAS (Blackford et al., 2002) and implicit parallelism on GPUs.

In fact, we can approximate D variational posteriors with a single neural network with shared parameters $\boldsymbol{\phi}$. We propose a shared variational model for all D variational posteriors in Figure 4.2. The model contains D inputs and $D \times P$ outputs, where P is the number of statistical parameters for one variational distribution. Each input element is associated with a corresponding variable in $\boldsymbol{x}$ and each output is a parameter for one of the D variational posteriors. Hence, to obtain $|\boldsymbol{m}_i|$ posteriors we can make $|\boldsymbol{m}_i|$ copies of $\boldsymbol{x}_i$, one copy for each $j \in \boldsymbol{m}_i$, set the x^j to zero in the corresponding copy of $\boldsymbol{x}_i$, evaluate the variational model to obtain all variational posteriors $q_{\boldsymbol{\phi}}(x^j \mid \boldsymbol{x}^{\setminus j})$ in a single batch, and select the corresponding $q_{\boldsymbol{\phi}}(x^j \mid \boldsymbol{x}^{\setminus j})$ for each j from the output of the shared model. The computation of all variational posteriors now can be parallelised on GPU and leverage BLAS.

Notice that if we do not set $x^j = 0$ then the shared model outputs $q_{\boldsymbol{\phi}}(x_{t+1}^j \mid \boldsymbol{x}_t^{\setminus j}, x_t^j)$ for all $j \in \{1, \ldots, D\}$ in a single pass, where t denotes the current timestep. Hence, by allowing the next value of x^j depend on the previous value we can get all variational posteriors in a single pass of the inference network. The speed-up effect can be significant when $|\boldsymbol{m}_i|$ is large. We refer to this model as single-pass shared model. We conjecture that in high-dimensions the impact of the value of a single variable is low, and hence $q_{\boldsymbol{\phi}}(x_{t+1}^j \mid \boldsymbol{x}_t^{\setminus j}, x_t^j)$ should be close to $q_{\boldsymbol{\phi}}(x^j \mid \boldsymbol{x}^{\setminus j})$.

Parallel Gibbs updates

In Algorithm 1 we consider the case of standard Gibbs updates, where a single missing value for each data point in the mini-batch is updated at each iteration. In high dimensions, standard Gibbs can have slow convergence, hence limiting the convergence speed of CDI. One could run several steps of standard Gibbs sampler at the end of each iteration but this can be too expensive in practice. With the shared model we can efficiently compute all variational posteriors and we would also like to leverage this in the dataset update step.

For high-dimensional data we use the Synchronous Gibbs sampler, in which all missing values are sampled and updated in parallel. With Synchronous Gibbs the missing value imputations are updated more frequently and sampling leverages the parallelism of the shared model, hence it enables faster convergence at little additional expense for sampling. However, Synchronous Gibbs does not guarantee to sample an approximately true joint density. Hence, towards the end of the training, when the training objective stops increasing, we switch from the synchronous sampler to standard single-value per data point updates, which improves the final solution.

Chapter 5

Simulations

In this chapter we evaluate the CDI algorithm for learning a factor analysis (FA) model in the context of missing data. We are primarily interested in mapping out the quality of the learnt model versus the fraction of missing data, as well as the quality of the variational approximation of the true posterior. Then, we also investigate the accuracy of the predictions of the missing values from a joint variational posterior mean.

In the next section we describe the models and methods, as well as, the data used in the evaluations. Then, in the following section we describe the evaluation metrics. Next, we demonstrate empirically that CDI is able to successfully learn probability density models and outperforms default missing data handling methods. Finally, we show that the learnt variational distributions can be used for point-estimate predictions of the missing values with significant accuracy.

5.1 Experimental setup

5.1.1 Density model

We choose the factor analysis model as the probability density model for the evaluation of CDI. The FA model allows us to evaluate the true joint posterior $p_{\boldsymbol{\theta}}(\boldsymbol{x}^m \mid \boldsymbol{x}^o)$ in closed-form using (2.18)-(2.20). To derive $p_{\boldsymbol{\theta}}(\boldsymbol{x}^m \mid \boldsymbol{x}^o)$ we use the property of Gaussian marginal distributions (Petersen and Pedersen, 2012) on $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ to get $p_{\boldsymbol{\theta}}(\boldsymbol{x}^m)$, and hence we observe that the derivation of covariance of $p_{\boldsymbol{\theta}}(\boldsymbol{x}^m)$ is the same as in (2.18) but with F and Ψ restricted to submatrices F^m and Ψ^m where only the rows (and columns for Ψ^m) are selected that correspond to the missing dimensions. Then, to get $p_{\boldsymbol{\theta}}(\boldsymbol{x}^m \mid \boldsymbol{x}^o) = \mathcal{N}(\boldsymbol{x}^m; \boldsymbol{\mu}_{\boldsymbol{x}^m|\boldsymbol{x}^o}, \Sigma_{\boldsymbol{x}^m|\boldsymbol{x}^o})$ we use $\boldsymbol{\mu}_{\boldsymbol{z}|\boldsymbol{x}^o}$ and $\Sigma_{\boldsymbol{z}|\boldsymbol{x}^o}$ from (2.19) and (2.20). To

get $\Sigma_{\boldsymbol{x}^m|\boldsymbol{x}^o}$ we simply replace the prior covariance of $\boldsymbol{z}$ in (2.18) with the posterior covariance $\Sigma_{\boldsymbol{z}|\boldsymbol{x}^o}$. And to get $\boldsymbol{\mu}_{\boldsymbol{x}^m|\boldsymbol{x}^o}$ we take the the expectation of (2.15) with respect to the posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}^o)$. Hence, the posterior distribution of missing variables is [1]

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}^m \mid \boldsymbol{x}^o) = \mathcal{N}(\boldsymbol{x}^m; F^m \boldsymbol{\mu}_{\boldsymbol{z}|\boldsymbol{x}^o} + \boldsymbol{\mu}^m, F^m \Sigma_{\boldsymbol{z}|\boldsymbol{x}^o} F^{m\top} + \Psi^m). \tag{5.1}$$

We will refer to this as true joint posterior and use it in the EM algorithm to train an optimal comparison model. Similarly, the univariate posterior $p_{\boldsymbol{\theta}}(x^j \mid \boldsymbol{x}^{\setminus j})$ can also be computed exactly if only one variable is treated as missing at a time, hence we can evaluate CDI where the variational posterior is replaced with a true posterior. Therefore, we can investigate the loss in fitted density model quality due to the variational approximation.

To optimise the density model parameters we use stochastic gradient ascent (SGA) with adaptive moment estimation (Adam) (Kingma and Ba, 2014) on $\mathcal{J}_{\text{CDI-ELBO}}$ in (4.3). Since the marginal observation covariance matrix, defined as $\Sigma_{\boldsymbol{x}} = FF^\top + \Psi$, must be positive semi-definite we must unconstrain the parameters in order to use SGA. The first term is always positive semi-definite, hence only the diagonal noise variance $\text{diag}(\Psi)$ needs to be constrained to non-negative values in order to make the matrix positive semi-definite. We therefore unconstrain $\text{diag}(\Psi)$ during optimisation by reparameterising $\text{diag}(\Psi) = \exp(\xi)$ and optimise with respect to $\boldsymbol{\theta} = (\xi, \boldsymbol{c}, F)$.

5.1.2 Methods

Variational posterior

We specify the variational distributions in a univariate Gaussian family, whose parameters σ^j and μ^j are given the outputs of fully-connected neural networks[2]

$$\sigma^j = \sigma^j_{\boldsymbol{\phi}}(\boldsymbol{x}^{\setminus j}) \quad \text{and} \quad \mu^j = \mu^j_{\boldsymbol{\phi}}(\boldsymbol{x}^{\setminus j}). \tag{5.2}$$

The variational distribution family includes the true posterior $p_{\boldsymbol{\theta}}(x^j \mid \boldsymbol{x}^{\setminus j})$, hence we can expect that the variational posterior should be able to closely approximate the true posterior. And the performance of CDI using the variational posterior should be comparable to using the true posterior.

To approximate the expectations in (4.1) of $\mathcal{J}_{\text{CDI-ELBO}}$ objective, we use Monte Carlo integration (Robert and Casella, 2004) and the reparameterisation trick (Kingma

[1] We further check the result and provide a derivation in Appendix A.2.

[2] For σ^j the outputs of the network return the log-standard deviation, which we then transform with an exponent function to to get σ^j.

and Welling, 2013)

$$x^j = \sigma^j \varepsilon + \mu^j, \quad \varepsilon \sim \mathcal{N}(0,1), \tag{5.3}$$

which enables the backpropagation of gradients via deterministic transformation of an independent noise random variable ε. We optimise the variational model parameters $\boldsymbol{\phi}$ using Adam.

Furthermore, using a univariate Gaussian variational approximation allows us to analytically compute the entropy term in (4.1) (Norwich, 1993)

$$-\mathbb{E}_{q^j(x^j|\tilde{\boldsymbol{x}}^{\setminus j})}\left[\log q^j(x^j \mid \tilde{\boldsymbol{x}}^{\setminus j})\right] = -\mathbb{E}_{\mathcal{N}(x^j;\mu^j,(\sigma^j)^2)}\left[\mathcal{N}(x^j;\mu^j,(\sigma^j)^2)\right] \tag{5.4}$$

$$= \frac{1}{2}\ln(2\pi(\sigma^j)^2) + \frac{1}{2}, \tag{5.5}$$

which can reduce the gradient variance and hence enables faster convergence when training with SGA.

In the experiments we consider two types of inference networks to parameterise the variational distributions. The first type uses an individual inference network for each dimension j and each parameter σ^j and μ^j. Each neural network is a single hidden layer fully-connected network. The second type uses (single-pass) shared model approach from Section 4.5 so that a single inference network is used for all conditional posterior distributions. The shared model also partially shares the parameters between the predictions of σ^j and μ^j – the model consists of two hidden layers, the first layer is shared for inferences of σ^j and μ^j, and the second layer is separated for σ^j and μ^j. All neural networks use leaky ReLU activation functions with negative slope of 0.01.

Baseline methods

In order to evaluate the performance of CDI we also train three baseline models using default missing data handling methods. To train the baseline models we first impute the data once at the start and then fit the FA models using standard MLE for complete data. Hence the baseline models only differ in the imputation method.

Mean baseline. Impute the data with the unconditional univariate mean of the observed values.

Hot-deck baseline. Impute the data with a random draw from an unconditional univariate empirical distribution. This is the most simple case of *hot-deck by simple random sampling with replacement*.

Regression baseline. First, to fit the regression functions on data with multiple missing values the dataset is filled-in with empirical unconditional means. Then, D

non-linear regression functions are trained, one for each dimension given the rest. A regression function for variable x^j is learned given the rest of the variables $\boldsymbol{x}^{\setminus j}$. The function is fitted using stochastic gradient descent with Adam on a subset of training data where the values of x^j are observed. Once a function for the j-th dimension is fitted, the missing values at this dimension are predicted and imputed in the dataset. The regression training then proceeds to train the next regression function in order from the first dimension to last and then terminates. Each regression function is parameterised by a single layer neural network with leaky ReLU activations. Moreover, we train two types of regression functions – one where the regression is performed only on variables $\boldsymbol{x}^{\setminus j}$ and another where a binary missingness mask (1 if the variable is present, 0 otherwise) is passed along with the input.

Summary of methods

In total, we consider eleven approaches to fit the density models in our comparison. The hyperparameters used in training are detailed in Appendix B.

Complete data. The FA model is fitted on complete data.

CDI[3](true). The model is fitted with CDI algorithm, where the true univariate posterior $p_{\boldsymbol{\theta}}(x^j \mid \tilde{\boldsymbol{x}}^{\setminus j})$ is used.

CDI (variational). Uses CDI algorithm with variational approximations. The variational model uses individual networks for all univariate posteriors.

CDI (shared). Same as CDI (variational) but a shared variational model is used.

CDI (single-pass). Same as CDI (variational) but a single-pass shared variational model is used. For high-dimensional datasets we will also use the Synchronous Gibbs sampler at the start of the training to improve CDI convergence on high-dimensional data and then switch to standard Gibbs at the end (Section 4.5).

EM (joint). Model trained using EM, where in the E step the true joint posterior is used from (5.1) and in the M step SGA is used to update the model parameters.

EM (independent). Model trained using EM, where in the E step an exact independent posterior $p_{\boldsymbol{\theta}}(x^j \mid \boldsymbol{x}^o)$ is used similar to (3.2) and in the M step SGA is used. The method is used to demonstrate the loss of model quality due to an independence assumption in the posterior.

[3]For all CDI methods, the data is filled-in with mean imputation at the start of the training.

Mean baseline. The data is filled-in with mean imputation and fitted using standard MLE for complete data.

Hot-deck baseline. The data is filled-in with a random draw from the empirical distribution of observed values, and then fitted using standard MLE.

Regression baseline. The data is filled-in with regression predictions, and then fitted using standard MLE.

Regression with missingness mask baseline.[4] The data is filled-in with regression predictions given $\boldsymbol{x}^{\setminus j}$ and a binary vector indicating the missing values, and then fitted using standard MLE.

5.1.3 Datasets

We evaluate the methods on two synthetic datasets for which the parameters of the ground truth model are known. Access to the ground truth model enable us to perform a comprehensive analysis of the learnt model parameters. Both datasets are split randomly in 80:20 ratio into training and validation subsets. To simulate incomplete data, for each dataset we generate a boolean missingness mask, where 1 corresponds to observed values and 0 indicates missing. The missigness mechanism is MCAR and we consider five fractions of missingness: 16.6%, 33.3%, 50%, 66.6%, and 83.3%.

Toy data. The dataset consists of 10000 independent samples from a 6-dimensional FA model with a 2-dimensional latent space. The parameters of the source FA model are known and are detailed in Appendix B.1.

Frey data[5] Consists of 1965 greyscale pictures of faces from a video with a resolution of 20×28. The images are flattened into a vector and the integer pixel intensities were scaled to lie between 0.15 and 0.85, and then transformed with a logit transformation such that the tails of a fitted Gaussian model correspond to realistic values.

FA-Frey data. The data was generated from a FA model with 43 latent variables, which was pre-trained on fully-observed Frey data. The dataset consists of 3000 independent samples from the FA model.

[4] In the results we only report the regression baseline, which performed the best on the given data.
[5] The dataset is available at `https://cs.nyu.edu/~roweis/data/frey_rawface.mat`.

5.2 Evaluation metrics

In this section we describe the evaluation metrics used in our experiments. The main objective of the evaluation is to investigate the quality of the learnt density model and variational approximations of the true posterior distributions. We are also interested whether the variational distributions learnt via CDI can be used to predict point-estimates of the missing values. Such point estimates might be useful when evaluating new test data point on the fitted density model.

To evaluate the quality of the learnt FA model we consider two metrics. First, we evaluate average log-likelihood of the learnt parameters on fully-observed held-out validation data. The motivation for this metric is that a higher quality of the fitted model is reflected in a higher likelihood when evaluated on data from the true data source distribution. In the case of known ground-truth model, we generate several independent datasets and use them to get mean log-likelihood with standard error on the result.

Second, when the ground truth parameters are known a standard model comparison is to evaluate the mean-squared error (MSE) of the learnt parameters $\boldsymbol{\mu}$, F, and Ψ. However, the factor matrix is not uniquely indentifiable due to factor rotation problem (Barber, 2017, Chapter 21). Instead we compare the eigenvectors of the marginal observation covariance matrix $\Sigma_{\boldsymbol{x}} = FF^\top + \Psi$, which corresponds to the directions of highest variance of the probability density. Since eigenvectors are generally not unique we first standardise to unit length, sort by their eigenvalues, and flip the direction of the eigenvectors of the learnt model such that the sign of the largest element in absolute value matches sign of the corresponding eigenvector element of the true model. We then compute MSE of the transformed eigenmatrix as well as MSE of $\boldsymbol{\mu}$.

Next, we also want to investigate the quality of the variational posterior approximations. Since the variational distribution family is univariate Gaussian and includes the ground truth posterior, it is enough to compare the sufficient statistics μ^j and σ^j. Hence, we compute the posterior parameters μ^j and σ^j on a held-out validation set for all missing values, where the validation set is masked to simulate an incomplete set according to the missingness fraction. Then, the MSE of the posterior parameters is computed for each dimension and averaged, which enables us to examine the quality of the variational distributions.

Finally, we compute the MSE of the predictions from the (approximate) posteriors and regression functions compared to the true values in the original fully-observed

dataset. Note that the imputations during CDI training are not predictions of the missing values – they are random samples from the posterior distribution of the missing values given the observed. In theory, to produce a point estimate we would like to find the mode of the posterior distribution, also known as maximum a posteriori (MAP) estimation. In the case of FA model, we know that the posterior distribution is a Gaussian, hence the mode of the distribution can be estimated via an expectation. One way to estimate this expectation is to sample missing values via Gibbs sampler and take an average of the values in the Markov chain. However, the sampling-based approach might require many samples and hence can be slow in high dimensions. Therefore, in this evaluation we take a different coordinate-wise ascent approach. The approach is similar to Gibbs sampling, but instead of updating each value with a random draw from the conditional distribution, we update each value using the posterior mean. Hence, the final imputation should estimate the mean of the posterior distribution.

5.3 Results

5.3.1 Quality of the learnt density model

Figure 5.1 presents the evaluation of the learnt density model quality on Toy data (left column) and FA-Frey data (right column) [6]. The top row presents the log-likelihood evaluated on a held-out validation set with the standard error bars computed over 20 distinct sets for Toy data and 10 distinct sets for FA-Frey data. The validation sets were generated independently from the ground truth models. The second row demonstrates the mean-squared error of the eigenmatrix of marginal observation covariance $\Sigma_{\boldsymbol{x}}$ for evaluating the quality of factor matrix F and diagonal noise matrix Ψ, and the third row shows the mean-squared error of the mean vector $\boldsymbol{\mu}$.

First, in Toy data log-likelihood plot we observe that EM (independent) performs significantly worse than EM (joint) and CDI methods. This result confirms that the independence assumption in (3.2) is too strong, and a method that produces imputation values from an approximate joint posterior is preferred, such as CDI or EM.

The benefits of both EM and CDI methods are the most apparent in the log-likelihood evaluation – both methods consistently outperform the baseline methods on all fractions of missingness and on both datasets. On Toy data, the CDI methods

[6] The log-likelihood evaluation on the original Frey data yields similar results to the FA-Frey in this section and the figure can be found in Appendix C.1.

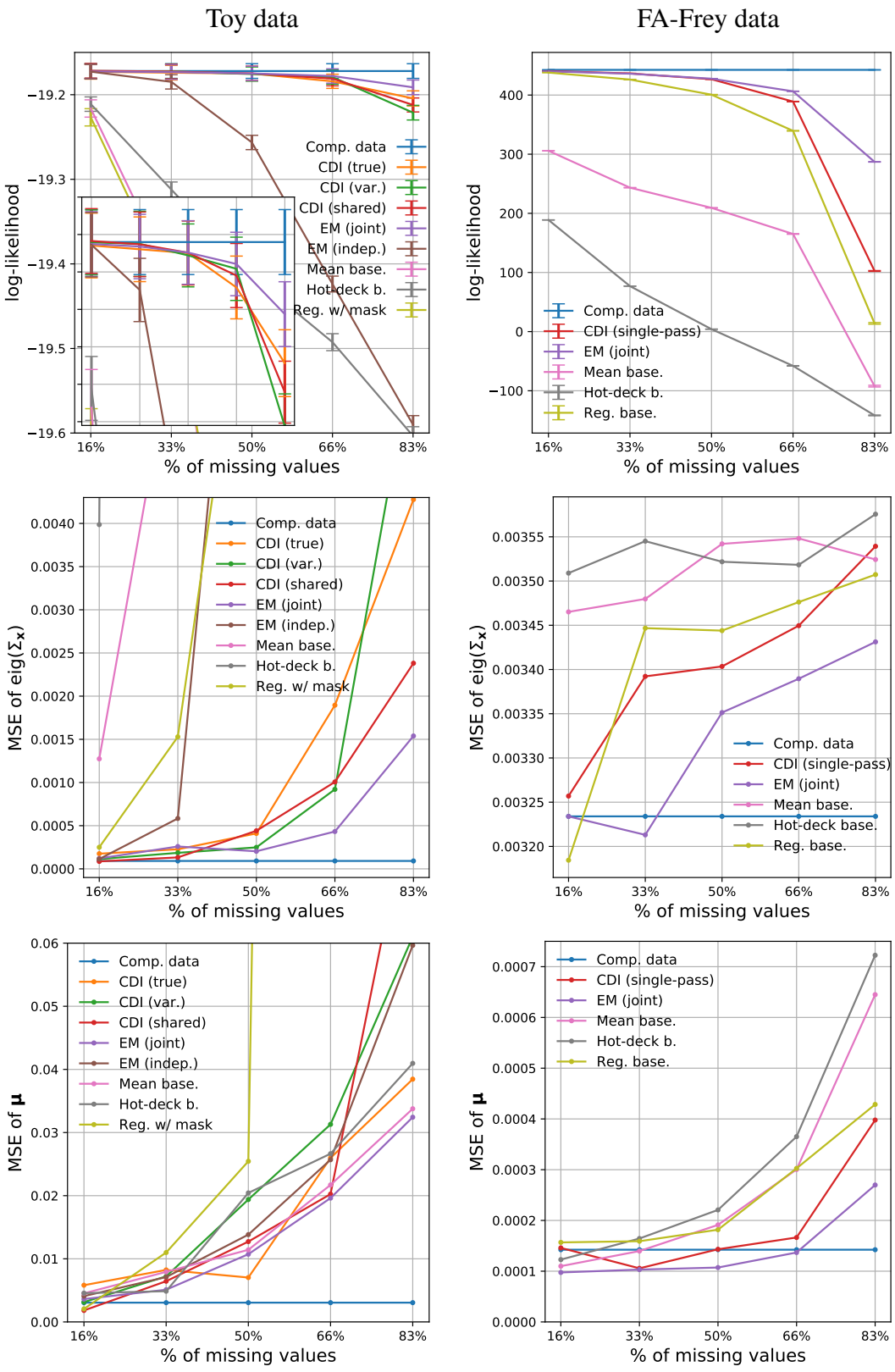


Figure 5.1: Fitted density model quality results of Toy (left) and FA-Frey (right) datasets. Top-down: average log-likelihood on held-out sets with error bars, mean-squared error of the covariance eigenmatrix, and mean-squared error of the mean vector.

perform comparably to EM (joint). A similar trend is observed on FA-Frey data but at missingness fractions of 66% and greater the CDI performance is lower than EM (joint), which is due to slow convergence at high missingness rate and hence early termination of the algorithm.

The performance of baseline methods depend on the dataset, on Toy data the hot-deck baseline is preferable, and on FA-Frey data the regression baseline performs better, whereas the mean baseline is not preferred on any. The significant difference in regression baseline performance on the two datasets can be attributed to the level of variable correlation in each dataset. In particular, the correlations in Toy data are generally weak, whereas in FA-Frey data the nearby pixels are generally strongly correlated. When correlations are strong, regression predictions can be sufficiently accurate and thus the variance of the data is not strongly affected, consequently the regression baseline performs well on FA-Frey data. On the other hand, in the case of Toy data, where the correlations are weak, regression imputations introduce significant bias by strengthening the correlations in the data and hence reducing variance significantly. In comparison, CDI method imputes the values with a random draw from the posterior distribution, hence the introduced bias to the variance of the data is generally smaller.

Moreover, parameter estimation result plots in second and third rows of Figure 5.1 largely confirm that CDI and EM are preferred over the baseline methods. In particular, the MSE for $\text{eig}(\Sigma_{\boldsymbol{x}})$ and $\boldsymbol{\mu}$ of the CDI methods are typically just above the MSE of EM (joint). Although for some fractions of missingness the CDI method has a higher MSE than the baseline methods, the performance of CDI is typically better than the baseline methods. For example, on the Toy data, the CDI (shared) has a higher MSE than the baseline methods on $\boldsymbol{\mu}$ at 83% missingness, however, the CDI (shared) method outperformed all of them on the MSE of $\text{eig}(\Sigma_{\boldsymbol{x}})$. Similarly, on the FA-Frey data, the CDI (single-pass) has a higher MSE on $\text{eig}(\Sigma_{\boldsymbol{x}})$ at missingness rates of 16% and 83% than regression baseline but outperformed the regression baseline on $\boldsymbol{\mu}$ estimation. Overall, the results in Figure 5.1 show that CDI method trained a higher quality density model than the baseline methods in comparison.

Finally, in the Toy data plots we observe that CDI (true), CDI (var.), and CDI (shared) all have comparable performance. Hence, the impact from using a variational posterior was generally small and the shared variational model did not negatively impact the accuracy of the variational posteriors. In addition, CDI (true) and CDI (var.) were too computationally expensive to evaluate on the FA-Frey data, and even CDI(shared) was slow and memory-intensive due to the explosion of the dataset

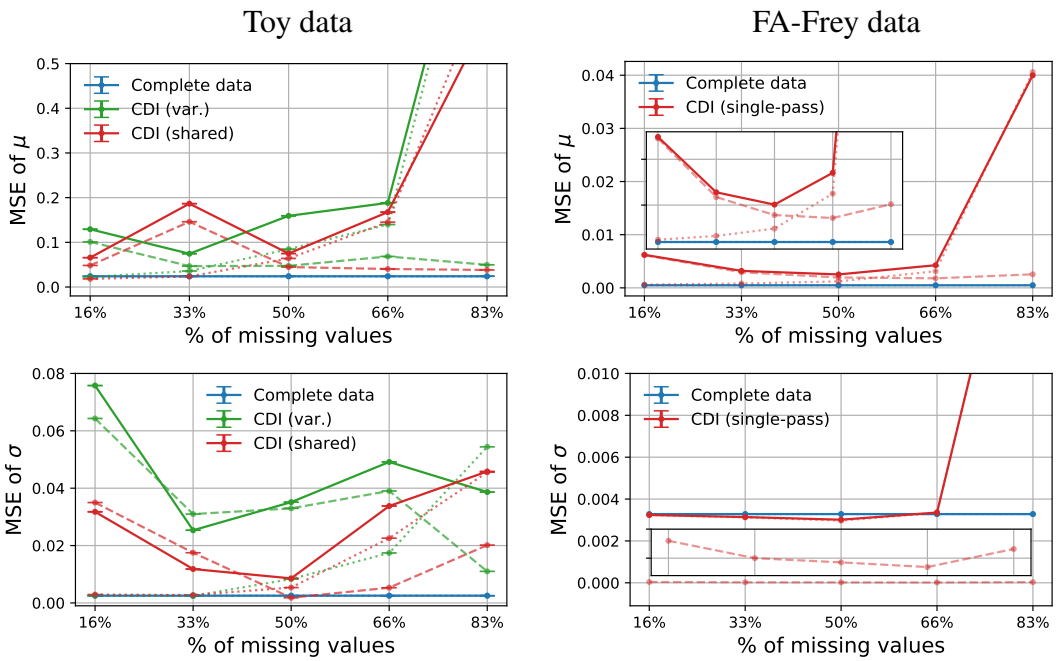


Figure 5.2: Mean-squared error of the variational posterior parameters on Toy (left) and FA-Frey (right) datasets. Top row: mean-squared error of the posterior mean; bottom row: mean-squared error of the posterior standard deviation. The solid curve compares $q_{\boldsymbol{\phi}}(x^j \mid \boldsymbol{x}^{\setminus j})$ and $p_*(x^j \mid \boldsymbol{x}^{\setminus j})$, the dashed curve compares $q_{\boldsymbol{\phi}}(x^j \mid \boldsymbol{x}^{\setminus j})$ and $p_{\boldsymbol{\theta}}(x^j \mid \boldsymbol{x}^{\setminus j})$, and the dotted curve compares $p_*(x^j \mid \boldsymbol{x}^{\setminus j})$ and $p_{\boldsymbol{\theta}}(x^j \mid \boldsymbol{x}^{\setminus j})$.

size, hence the more efficient CDI (single-pass) was used, which performed comparably well when compared to the other methods.

5.3.2 Accuracy of variational approximation

Figure 5.2 presents mean-squared error of the variational posterior mean and standard deviation, where the solid line compares the variational posteriors $q_{\boldsymbol{\phi}}(x^j \mid \boldsymbol{x}^{\setminus j})$ to the true posteriors $p_*(x^j \mid \boldsymbol{x}^{\setminus j})$ computed as in (5.1) using the ground truth model p_*, the dashed line compares $q_{\boldsymbol{\phi}}(x^j \mid \boldsymbol{x}^{\setminus j})$ to the corresponding conditionals of the learnt density model $p_{\boldsymbol{\theta}}(x^j \mid \boldsymbol{x}^{\setminus j})$, and the dotted line compares $p_{\boldsymbol{\theta}}(x^j \mid \boldsymbol{x}^{\setminus j})$ and $p_*(x^j \mid \boldsymbol{x}^{\setminus j})$.

In the Toy data results CDI (var.) and CDI (shared) performed similarly on the estimation of posterior mean but the shared variational model performed better on the estimation of posterior standard deviation. Hence, the shared variational model can be useful not only for implementation efficiency (Section 4.5), but it can also improve learning efficacy by making better use of the available data.

The solid curve in Figure 5.2 shows how well the variational distributions approx-

imate the ground truth posterior. We observe a general trend that the approximation MSE decreases from 16% to 50%, but then increases for larger fractions missingness. The dashed curve presents the quality of the variational approximation with respect to the learnt density model, which is generally close to the solid curve for fractions of missingness from 16% to 50% but then the dashed curve remains low compared to the solid curve for larger fractions of missingness, with a slight increase at 83.3%. The slight increase of the MSE of the dashed curve at large missiness rates could be attributed to the general complexity of the $\mathcal{J}_{\text{CDI-ELBO}}$ objective at large missingness. Finally, the dotted curve compares the posterior distribution on the learnt density model and the ground truth model and hence show how well the learnt model approximates the ground truth. From the dotted curve we observe that the learnt density model approximates the ground truth model worse as the missingness fraction increases.

Notice that the model approximates the ground truth well at low missingness (dotted curve), but the variational approximation of the model's posterior is poor (dashed curve) in comparison. Worse posterior accuracy at low missigness fractions than medium fractions can be attributed to the effective training set size on which a particular $q_{\boldsymbol{\phi}}(x^j \mid \boldsymbol{x}^{\setminus j})$ is trained: lower missingness fraction corresponds to smaller effective training set, and vice versa. As mentioned in Section 4.4 one possible approach to mitigate this is to use an input-dropout in order to artificially increase the effective size of the missing training data when optimising $\mathcal{J}_{\text{CDI-ELBO}}$ on low-missingness data. On the other hand, for large fractions of missingness the variational approximation of the model's posterior is generally good, but the model approximates the ground truth poorly. This suggests that the CDI had not fully converged when it was terminated. We conjecture that the reason for this is because the Markov chain over the missing values has not converged to the stationary distribution, and hence the model could be improved by training longer.

5.3.3 Missing data prediction accuracy

Figure 5.3 presents the mean-squared error of predicted point-estimate values compared to the true value in the original fully-observed dataset. The top row presents the results on Toy data, the middle row – FA-Frey data, and the bottom row – original Frey data. The left column compares the predictions from the variational posterior to regression functions, and the right column compares to closed-form approximations from Williams et al. (2018). The predictions from the variational posterior were pro-

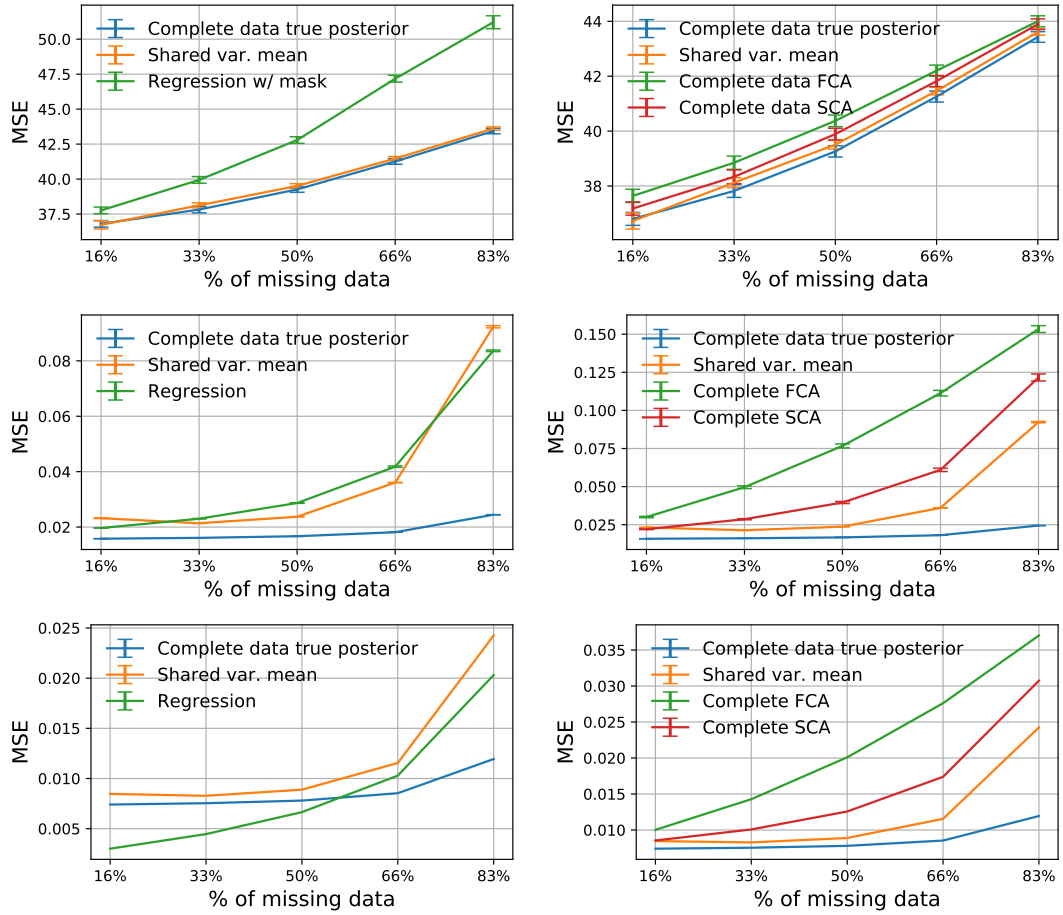


Figure 5.3: Mean-squared error of the predicted missing values evaluated against the true values in Toy data (top), FA-Frey data (middle), and original Frey data (bottom). Compares predictions from variational posterior against regression predictions (left column) and predictions from the closed-form approximations of the joint posterior evaluated on a model that was trained on fully-observed data (right column).

duced by performing 5 cycles of the procedure in Section 5.2 on Toy and FA-Frey data, and 10 cycles on the original Frey data, replacing each missing value with the variational posterior mean in each cycle. On the original Frey data we compared the methods in the same scale as Williams et al. (2018) – on the imputed dataset, we reversed the logit preprocessing transformation with a sigmoid transformation and scaled each pixel intensity between -1 and 1. The results on the original Frey data are also provided in tabular format in Table 5.1.

The left column of Figure 5.3 compares the variational posterior predictions, regression predictions, and predictions from the mean of the true joint posterior $p_{\boldsymbol{\theta}}(\boldsymbol{x}^m \mid \boldsymbol{x}^o)$ evaluated using a model $p_{\boldsymbol{\theta}}$ that was trained on fully-observed (complete) data. On Toy

	16.6%	33.3%	50%	66.6%	83.3%
Shared var. mean	0.8466	0.8275	0.8888	1.1545	2.4252
Complete (true)	0.7404	0.7538	0.7799	0.8533	1.1941
Complete (SCA)	0.8544	1.0076	1.2567	1.7383	3.0738
Complete (FCA)	1.0019	1.4280	2.0095	2.7611	3.7019
Regression	0.3005	0.4459	0.6647	1.0286	2.0307

Table 5.1: Mean-squared error ($\times 10^2$) of the predicted values on the original Frey data.

data, the variational posterior predictions achieved similar accuracy to the predictions from the complete model's joint posterior mean, and significantly outperformed regression predictions. On FA-Frey data, the variational posterior predictions are of comparable accuracy to regression predictions. And on the original Frey data, the variational posterior predictions were worse than regression predictions. The poor performance on the original Frey data can be explained by looking at the performance of the complete model's performance – we observe that even when trained on complete data, the predictions from the true joint posterior mean are poor for low fractions of missingness. This suggests, that a more flexible model may be required to better model the data. Overall, the above results suggest that if the chosen density model is sufficiently flexible and the variational posteriors accurately approximate the true posteriors, then the predictions from the variational posterior can be significantly accurate.

The right column of Figure 5.3 compares the variational posterior predictions to predictions using the closed-form approximations of the true joint posterior – full-covariance approximation (FCA) and scaled-covariance approximation (SCA) from Williams et al. (2018), which were evaluated using a model $p_{\boldsymbol{\theta}}$ that was trained on fully-observed (complete) data. The predictions from the variational posterior were better than from FCA and SCA approximations for all fractions of missingness and on all datasets, even though FCA and SCA used a model that was trained on complete data and the variational distributions were trained on incomplete data via CDI. This suggests, that CDI can be useful even where closed-form approximations, such as FCA and SCA, are available.

Chapter 6

Conclusions

In this thesis we have investigated a variational approach for learning normalised density models from incomplete data, called Cumulative Data Imputation based on Rhodes (2018). The CDI algorithm has several attractive properties. First, it mitigates the combinatorial explosion of variational distributions by requiring only D variational distributions, hence it scales linearly with the dimensionality of the data. And second, the factorisation of the variational distributions does not make false assumptions about the distribution of the missing data.

We proposed several modifications to the original CDI algorithm that improve convergence of the variational distributions and reduce bias (Section 4.2). Then, we provided provided a justification of the convergence of the CDI algorithm based on the convergence proof of the Gibbs sampler (Section 4.3). Next, we discussed the potential issues when training on low-missingness data and proposed a simple drop-out technique to alleviate them (Section 4.4). Finally, we proposed a shared variational model that uses a single inference network for all D variational distributions (Section 4.5), which significantly reduced the computational overhead required to compute multiple variational posteriors and improved data efficiency of the algorithm.

We demonstrated the capability of CDI for learning a factor analysis model from incomplete data. The empirical results show that CDI was able to perform better than many default missing data handling methods, including mean imputation, simple hot-deck imputation, and regression imputation. In fact, CDI was able to train a density model of comparable quality to the EM algorithm for fractions of missingness up to and including 50%. Moreover, we have shown that the quality of the variational posterior approximations depends on the missingness fraction, and provided an approach to improve the approximations at low fractions of missingness. Finally, we have demon-

strated that we can use the variational posteriors to produce point-estimate predictions of the missing values with significant accuracy, and that predictions from the variational posteriors were significantly better than predictions from closed-form posterior approximations for factor analysis by Williams et al. (2018). This shows that the CDI algorithm can be useful even when closed-form approximations of the posterior are available.

There are several limitations: CDI can have slow convergence on high-missingness high-dimensional data, and the optimisation of the CDI objective in (4.3) can be expensive since the computational effort scales linearly with the number of missing values. To mitigate the former, in our evaluations we used the Synchronous Gibbs sampler on high dimensional Frey data to improve the convergence speed of the Markov chain on the missing data. However, the suitability of the Synchronous Gibbs sampler can be problem-specific, hence we see that understanding the convergence criteria and speed on high-dimensional data is an important future question to address. Additionally, it would be interesting to investigate other Hogwild Gibbs methods (Angelino et al., 2016) for the CDI update step, which have empirically shown good results (Newman et al., 2008; Asuncion et al., 2009), to further improve the convergence speed of CDI. To alleviate the computational cost, we have suggested in Section 4.2 that the computational cost can be tuned by optimising for a selected subset of missing values at a time, however it is not clear how that would affect the convergence speed. Hence, mapping-out the trade-off between computational cost of the optimisation step and the convergence speed is another interesting future topic.

Whilst we have evaluated CDI on a factor analysis model, in recent practice more complex and non-linear models, such as, variational autoencoders (VAEs) are more often used. There are several recent versions of VAE that can be trained on incomplete data (Vedantam et al., 2017; Nazabal et al., 2018; Wu and Goodman, 2018; Ivanov et al., 2019). Since VAEs are one of the current state-of-the-art machine learning models, we view the comparison of CDI against VAE-based models for incomplete data to be a promising direction for future research.

Bibliography

Angelino, E., Johnson, M. J., and Adams, R. P. (2016). Patterns of Scalable Bayesian Inference. *Foundations and Trends® in Machine Learning*, 9(2-3):119–247.

Asuncion, A. U., Smyth, P., and Welling, M. (2009). Asynchronous Distributed Learning of Topic Models. In *Advances in Neural Information Processing Systems 21 (NIPS 2008)*, pages 81–88.

Barber, D. (2017). *Bayesian Reasoning and Machine Learning*. Cambridge University Press.

Blackford, L. S., Demmel, J., Dongarra, J., Duff, I., Hammarling, S., Henry, G., Heroux, M., Kaufman, L., Lumsdaine, A., Petitet, A., Pozo, R., Remington, K., and Whaley, R. C. (2002). An Updated Set of Basic Linear Algebra Subprograms (BLAS). *ACM Transactions on Mathematical Software*, 28(2):135–151.

Burton, A. and Altman, D. G. (2004). Missing covariate data within cancer prognostic studies: A review of current reporting and proposed guidelines. *British Journal of Cancer*, 91(1):4–8.

Chen, S. H. and Ip, E. H. (2015). Behaviour of the Gibbs sampler when conditional distributions are potentially incompatible. *Journal of Statistical Computation and Simulation*, 85(16):3266–3275.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741.

Gershman, S. J. and Goodman, N. D. (2014). Amortized Inference in Probabilistic Reasoning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36.

Ghahramani, Z. and Jordan, M. I. (1994). Supervised learning from incomplete data via an EM approach. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 120–127.

Gonzalez, J. E., Low, Y., Gretton, A., and Guestrin, C. (2011). Parallel Gibbs Sampling: From Colored Fields to Thin Junction Trees. *14th International Conference on Artificial Intelligence and Statistics*, 15:324–332.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Networks. In *Advances in Neural Information Processing Systems 27 (NIPS 2014)*.

Hartley, H. O. (1958). Maximum Likelihood Estimation from Incomplete Data. *Biometrics*, 14(2):174.

He, B. D., Sa, C. M. D., Mitliagkas, I., and Ré, C. (2016). Scan Order in Gibbs Sampling: Models in Which it Matters and Bounds on How Much. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, pages 1–9.

Henderson, H. V. and Searle, S. R. (1981). On Deriving the Inverse of a Sum of Matrices. *SIAM Review*, 23(1):53–60.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic Variational Inference. *Journal of Machine Learning Research*, 14:1303–1347.

Ivanov, O., Figurnov, M., and Vetrov, D. (2019). Variational Autoencoder with Arbitrary Conditioning. *arXiv pre-print*, 1806.02382.

Jensen, J. L. W. V. (1906). Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, 30:175–193.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37(2):183–233.

Kim, J. K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika*, 98(1):119–132.

Kingma, D. P. and Ba, J. (2014). Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*.

Kingma, D. P. and Welling, M. (2013). Auto-Encoding Variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*.

Kullback, S. and Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data: Second Edition*. Wiley-Interscience.

Murray, I. (2007). *Advances in Markov chain Monte Carlo methods*. PhD thesis, University College London.

Nazabal, A., Olmos, P. M., Ghahramani, Z., and Valera, I. (2018). Handling Incomplete Heterogeneous Data using VAEs. *arXiv preprint*, 1807.03653.

Newman, D., Smyth, P., Welling, M., and Asuncion, A. U. (2008). Distributed Inference for Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems 20 (NIPS 2007)*, pages 1081–1088.

Norwich, K. H. (1993). *Information, Sensation and Perception*. San Diego: Academic Press.

Petersen, K. B. and Pedersen, M. S. (2012). The Matrix Cookbook. *Technical University of Denmark*, 16(4):1–16.

Ranganath, R., Gerrish, S., and Blei, D. M. (2013). Black Box Variational Inference. In *Proceedings of Artificial Intelligence and Statistics (AISTATS)*.

Rhodes, B. (2018). Variational Noise-Contrastive Estimation. Master's thesis, University of Edinburgh.

Robert, C. P. and Casella, G. (2004). *Monte Carlo statistical methods*. Springer.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, 63(3):581.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved Techniques for Training GANs. In *Advances in neural information processing systems*.

Spearman, C. (1904). "General Intelligence" Objectively Determined and Measured. *The American Journal of Psychology*, 15(2):201.

Tanner, M. and Wong, W. H. (1987). The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, 82(398):528–540.

Tieleman, T. (2008). Training restricted Boltzmann machines using approximations to the likelihood gradient. *Proceedings of the 25th international conference on Machine learning - ICML '08*, pages 1064–1071.

Uehara, M., Matsuda, T., and Kim, J. K. (2019). Imputation estimators for unnormalized models with missing data. *arXiv preprint*, 1903.03630.

van Buuren, S. (2018). *Flexible Imputation of Missing Data*. CRC Press LLC.

Vedantam, R., Fischer, I., Huang, J., and Murphy, K. (2017). Generative Models of Visually Grounded Imagination. *International Conference on Learning Representations (ICLR)*.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning - ICML '08*, pages 1096–1103, New York, New York, USA. ACM Press.

Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *Journal of the American Statistical Association*, 85(411):699–704.

Whitney, W. O. and Mehlhaff, C. J. (1987). High-rise syndrome in cats. *Journal of the American Veterinary Medical Association*, 191(11):1399–403.

Williams, C. K. I., Nash, C., and Nazábal, A. (2018). Autoencoders and Probabilistic Inference with Missing Data: An Exact Solution for The Factor Analysis Case. *arXiv preprint*, 1801.03851.

Wu, M. and Goodman, N. (2018). Multimodal Generative Models for Scalable Weakly-Supervised Learning. In *NeurIPS 2018*.

Yoon, J., Jordon, J., and van der Schaar, M. (2018). GAIN: Missing Data Imputation using Generative Adversarial Nets. In *Proceedings of the 35th International Conference of Machine Learning (ICML)*.

Zhang, C., Butepage, J., Kjellstrom, H., and Mandt, S. (2018). Advances in Variational Inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–23.

Appendix A

Derivations

A.1 Derivation of the ELBO for incomplete data

We derive the variational evidence lower bound (ELBO) for incomplete data using standard variational inference derivation

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_i) = \log \int p_{\boldsymbol{\theta}}(\boldsymbol{x}_i^m, \boldsymbol{x}_i^o)\, \mathrm{d}\boldsymbol{x}_i^m \tag{A.1}$$

$$= \log \int \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_i^m, \boldsymbol{x}_i^o)}{q_{\boldsymbol{\phi}_i}(\boldsymbol{x}_i^m \mid \boldsymbol{x}_i^o)} q_{\boldsymbol{\phi}_i}(\boldsymbol{x}_i^m \mid \boldsymbol{x}_i^o)\, \mathrm{d}\boldsymbol{x}_i^m \tag{A.2}$$

$$= \log \mathbb{E}_{q_{\boldsymbol{\phi}_i}(\boldsymbol{x}_i^m \mid \boldsymbol{x}_i^o)} \left[\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_i^m, \boldsymbol{x}_i^o)}{q_{\boldsymbol{\phi}_i}(\boldsymbol{x}_i^m \mid \boldsymbol{x}_i^o)} \right] \tag{A.3}$$

$$\geq \mathbb{E}_{q_{\boldsymbol{\phi}_i}(\boldsymbol{x}_i^m \mid \boldsymbol{x}_i^o)} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_i^m, \boldsymbol{x}_i^o)}{q_{\boldsymbol{\phi}_i}(\boldsymbol{x}_i^m \mid \boldsymbol{x}_i^o)} \right] = \mathcal{J}^i_{\text{C-ELBO}}(\boldsymbol{\theta}, \boldsymbol{\phi}_i), \tag{A.4}$$

where we use the concavity of logarithm to apply Jensen's inequality (Jensen, 1906). In the equations, $\boldsymbol{\theta}$ corresponds to the parameters of the density model and $\boldsymbol{\phi}_i$ corresponds to the parameters of the posterior distribution of the i-th sample in the dataset

A.2 Derivation of FA posterior

The derivation of the missing variable posterior $p_{\boldsymbol{\theta}}(\boldsymbol{x}^m \mid \boldsymbol{x}^o)$ on a factor analysis model is simple given the posterior of the latents $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}^o) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_{\boldsymbol{z}|\boldsymbol{x}^o}, \Sigma_{\boldsymbol{z}|\boldsymbol{x}^o})$ from (2.19) and (2.20) (Williams et al., 2018).

Remember that the generative process in FA is

$$\boldsymbol{x} = F\boldsymbol{z} + \boldsymbol{\mu} + \boldsymbol{\varepsilon}\ , \text{ where } \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Psi). \tag{A.5}$$

Notice that the generative process of the missing variables can be written as

$$\boldsymbol{x}^m = F^m \boldsymbol{z} + \boldsymbol{\mu}^m + \boldsymbol{\varepsilon}^m \text{ , where } \boldsymbol{\varepsilon}^m \sim \mathcal{N}(0, \Psi^m), \tag{A.6}$$

where F^m, $\boldsymbol{\mu}^m$, and Ψ^m submatrices of F, $\boldsymbol{\mu}$, and Ψ where only the rows (and columns for Ψ^m) are selected that correspond to the missing variable dimensions. Hence, to get the posterior distribution parameters of $p_{\boldsymbol{\theta}}(\boldsymbol{x}^m \mid \boldsymbol{x}^o) = \mathcal{N}(\boldsymbol{x}^m; \boldsymbol{\mu}_{\boldsymbol{x}^m|\boldsymbol{x}^o}, \Sigma_{\boldsymbol{x}^m|\boldsymbol{x}^o})$ we take the expectation and covariance given the posterior of the latents $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}^o)$

$$\begin{aligned}
\Sigma_{\boldsymbol{x}^m|\boldsymbol{x}^o} &= \mathrm{Cov}_{\boldsymbol{z}|\boldsymbol{x}^o}(F^m \boldsymbol{z} + \boldsymbol{\mu}^m + \boldsymbol{\varepsilon}^m) && \text{(A.7)}\\
&= \mathrm{Cov}_{\boldsymbol{z}|\boldsymbol{x}^o}(F^m \boldsymbol{z}) + \mathrm{Cov}_{\boldsymbol{z}|\boldsymbol{x}^o}(\boldsymbol{\mu}^m) + \mathrm{Cov}_{\boldsymbol{z}|\boldsymbol{x}^o}(\boldsymbol{\varepsilon}^m) && \text{(A.8)}\\
&= F^m \Sigma_{\boldsymbol{z}|\boldsymbol{x}^o} F^{m\top} + \Psi^m && \text{(A.9)}\\
\boldsymbol{\mu}_{\boldsymbol{x}^m|\boldsymbol{x}^o} &= \mathbb{E}_{\boldsymbol{z}|\boldsymbol{x}^o}[F^m \boldsymbol{z} + \boldsymbol{\mu}^m + \boldsymbol{\varepsilon}^m] && \text{(A.10)}\\
&= \mathbb{E}_{\boldsymbol{z}|\boldsymbol{x}^o}[F^m \boldsymbol{z}] + \mathbb{E}_{\boldsymbol{z}|\boldsymbol{x}^o}[\boldsymbol{\mu}^m] + \mathbb{E}_{\boldsymbol{z}|\boldsymbol{x}^o}[\boldsymbol{\varepsilon}^m] && \text{(A.11)}\\
&= F^m \boldsymbol{\mu}_{\boldsymbol{z}|\boldsymbol{x}^o} + \boldsymbol{\mu}^m. && \text{(A.12)}
\end{aligned}$$

Since the marginal distribution of the observable $\boldsymbol{x}$ is a multivariate Gaussian distribution, the result can be checked by comparing it to the conditional of a multivariate Gaussian distribution (Petersen and Pedersen, 2012)

$$p(\boldsymbol{x}_1 \mid \boldsymbol{x}_2) = \mathcal{N}(\boldsymbol{x}_1; \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}), \tag{A.13}$$

where the joint Gaussian is

$$p(\boldsymbol{x}) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix}; \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right). \tag{A.14}$$

To show that our result in (A.9) and (A.12) is equivalent to (A.13) we will use the Woodbury's identity (Petersen and Pedersen, 2012)

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1} \tag{A.15}$$

and related push-through identity (Henderson and Searle, 1981)

$$(A + UCV)^{-1}U = A^{-1}U(C^{-1} + VA^{-1}U)^{-1}C^{-1}. \tag{A.16}$$

First we use the Woodbury's identity on (2.19)

$$\Sigma_{\boldsymbol{z}|\boldsymbol{x}^o} = (I + F^\top M \Psi^{-1} M F)^{-1} \tag{A.17}$$

$$= I^{-1} - I^{-1}F^\top M(\Psi + MFI^{-1}F^\top M)^{-1}MFI^{-1} \tag{A.18}$$

$$= I - F^\top M(\Psi + MFF^\top M)^{-1}MF. \tag{A.19}$$

Substituting into (A.9) we get

$$\Sigma_{\boldsymbol{x}^m|\boldsymbol{x}^o} = F^m(I - F^\top M(\Psi + MFF^\top M)^{-1}MF)F^{m\top} + \Psi^m \tag{A.20}$$

$$= F^m F^{m\top} + \Psi^m - F^m F^\top M(\Psi + MFF^\top M)^{-1}MFF^{m\top}. \tag{A.21}$$

Note that the sum of the first two terms is equivalent to Σ_{11} and the final term is equivalent to $\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ in (A.13). It can also be easily checked that $F^m F^\top M$ is equivalent to Σ_{12} but with extra zero columns, which correspond to the missing dimensions. An equivalent relation applies to $MFF^{m\top}$ and Σ_{21}, since $\Sigma_{12}^\top = \Sigma_{21}$ and $(F^m F^\top M)^\top = MFF^{m\top}$. Similarly $(\Psi + MFF^T M)^{-1}$ is equivalent to Σ_{22}^{-1} but with extra zero columns and rows that corresponds to the missing dimensions. In addition the resulting matrix of $(\Psi + MFF^T M)^{-1}$ has non-zero elements on the diagonals which correspond to the missing dimensions but the elements get multiplied by 0 afterwards when multiplied by $F^m F^\top M$ on the left and $MFF^{m\top}$ on the right, and hence the elements do not contribute to the final product. Hence, we can rewrite (A.21)

$$\Sigma_{\boldsymbol{x}^m|\boldsymbol{x}^o} = F^m F^{m\top} + \Psi^m - F^m F^{o\top}(\Psi^o + F^o F^{o\top})^{-1}F^o F^{m\top}, \tag{A.22}$$

where F^o and Ψ^o submatrices of F and Ψ where only the rows (and columns for Ψ^o) are selected that correspond to the observed variable dimensions. Here $F^m F^{o\top}$ corresponds to Σ_{12}, $F^o F^{m\top}$ corresponds to Σ_{21}, and $\Psi^o + F^o F^{o\top}$ corresponds to Σ_{22}. Hence, (A.22) is equivalent to the conditional covariance expression in (A.13).

And now we show that (A.12) is equivalent to the expression in (A.13). First, we substitute (2.19) to (2.20) and apply the push-through identity

$$\boldsymbol{\mu}_{\boldsymbol{z}|\boldsymbol{x}^o} = \left(I + F^\top M\Psi^{-1}MF\right)^{-1} F^\top M\Psi^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \tag{A.23}$$

$$= I^{-1}F^\top M(\Psi + MFI^{-1}F^\top M)^{-1}\Psi\Psi^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \tag{A.24}$$

$$= F^\top M(\Psi + MFF^\top M)^{-1}\Psi\Psi^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \tag{A.25}$$

$$= F^\top M(\Psi + MFF^\top M)^{-1}(\boldsymbol{x} - \boldsymbol{\mu}). \tag{A.26}$$

Notice again that the columns of $F^\top M$, as well as, columns and rows of $(\Psi + MFF^\top M)^{-1}$ that correspond to missing variable dimensions are zero, hence the corresponding elements of $(\boldsymbol{x} - \boldsymbol{\mu})$ will not contribute to the product. Therefore, we can rewrite (A.26) as

$$\boldsymbol{\mu}_{\boldsymbol{z}|\boldsymbol{x}^o} = F^{o\top}(\Psi^o + F^o F^{o\top})^{-1}(\boldsymbol{x}^o - \boldsymbol{\mu}^o), \tag{A.27}$$

where $\boldsymbol{\mu}^o$ corresponds to the rows of $\boldsymbol{\mu}$ that correspond to the observed dimensions. Substituting (A.27) in (A.12) we get

$$\boldsymbol{\mu}_{\boldsymbol{x}^m|\boldsymbol{x}^o} = F^m F^{o\top}(\Psi^o + F^o F^{o\top})^{-1}(\boldsymbol{x}^o - \boldsymbol{\mu}^o) + \boldsymbol{\mu}^m, \tag{A.28}$$

where $F^m F^{o\top}$ corresponds to Σ_{12} and $\Psi^o + F^o F^{o\top}$ corresponds to Σ_{22}. Hence, (A.28) corresponds to the conditional mean expression in (A.13).

Appendix B

Experimental details

In this appendix we detail the hyperparameters used in all evaluations. Additionally, in Appendix B.1 we provide the details of the ground truth model of Toy data.

B.1 Toy data details

The Toy data consists of 10000 independent draws from a factor analysis model. The data is 6-dimensional and the latent space of the factor analysis model is 2-dimensional. The parameters of the factor analysis model are the following

$$F = \begin{bmatrix} -5 & -2 \\ 4 & 0 \\ -3 & -1 \\ -3 & -3 \\ 1 & 5 \\ -1 & 2 \end{bmatrix} \quad \boldsymbol{\mu} = \begin{bmatrix} 3 \\ -1 \\ 0 \\ 2 \\ -1 \\ 0 \end{bmatrix} \quad \text{diag}(\Psi) = \begin{bmatrix} 50.4794 \\ 30.0988 \\ 6.766 \\ 17.3357 \\ 40.9839 \\ 25.1122 \end{bmatrix}. \tag{B.1}$$

Furthermore, the FA model used in the evaluations also had a 2-dimensional latent space. The regression models were 1-hidden layer neural networks with the hidden layer dimension of 3 and leaky ReLU activations. For each data dimension with missing values, a the regression function was trained for up to 1200 epochs with a learning rate of 10^{-2}, and early stopping that stopped training of the current regression function if validation mean-squared error loss did not improve for more than 100 epochs. For the CDI methods we used mean imputation in the initialisation stage. The evaluations used a batch size of 2048 and Table B.1 describes the rest of the hyperparameter settings.

	16.6%	33.3%	50%	66.6%	83.3%
Complete data					
# of epochs	400	400	400	400	400
learning rate	10^{-1}	10^{-1}	10^{-1}	10^{-1}	10^{-1}
CDI (true)					
# of epochs	400	400	400	2000	2000
learning rate	10^{-1}	10^{-1}	10^{-1}	10^{-1}	10^{-1}
CDI (var.)					
# of epochs	400	400	400	2000	2000
learning rate	10^{-1}	10^{-1}	10^{-1}	10^{-1}	10^{-1}
var. learning rate	10^{-1}	10^{-1}	10^{-1}	7×10^{-2}	7×10^{-2}
hidden layer dim	3	3	3	3	3
CDI (shared)					
# of epochs	400	400	400	2000	2000
learning rate	10^{-1}	10^{-1}	10^{-1}	10^{-1}	10^{-1}
var. learning rate	10^{-2}	5×10^{-2}	10^{-1}	5×10^{-2}	5×10^{-2}
1st hidden layer dim	10	10	10	10	10
2nd hidden layer dim	6	6	6	6	10
EM (joint)					
# of epochs	400	400	400	400	2000
learning rate	10^{-1}	10^{-1}	10^{-1}	10^{-1}	10^{-2}
EM (indep.)					
# of epochs	400	400	400	400	400
learning rate	10^{-1}	10^{-1}	10^{-1}	10^{-1}	10^{-2}
Mean baseline					
# of epochs	400	400	400	400	400
learning rate	10^{-1}	10^{-1}	10^{-1}	10^{-1}	10^{-2}
Hot-deck baseline					
# of epochs	400	400	400	400	400
learning rate	10^{-1}	10^{-1}	10^{-1}	10^{-1}	10^{-1}
Regression baseline					
# of epochs	400	400	400	400	400
learning rate	10^{-1}	10^{-1}	10^{-1}	10^{-1}	10^{-2}

Table B.1: Toy data experiment hyperparameter settings.

	16.6%	33.3%	50%	66.6%	83.3%
Complete data					
# of epochs	1000	1000	1000	1000	1000
learning rate	10^{-2}	10^{-2}	10^{-2}	10^{-2}	10^{-2}
CDI (single-pass)					
# of epochs	1500	1500	1500	2500*	3000*
learning rate	10^{-2}	10^{-2}	10^{-2}	10^{-2}	10^{-2}
var. learning rate	10^{-3}	10^{-3}	10^{-3}	10^{-3}	10^{-3}
1st hidden layer dim	500	500	500	500	500
2nd hidden layer dim	500	500	500	500	500
switch imputation epoch	800	800	1000	2000	2000
EM (joint)					
# of epochs	1000	1000	1000	1500	2000*
learning rate	10^{-2}	10^{-2}	10^{-2}	10^{-2}	10^{-2}
Mean/Hot-deck/Regression baseline					
# of epochs	1000	1000	1000	1000	1000
learning rate	10^{-2}	10^{-2}	10^{-2}	10^{-2}	10^{-2}

Table B.2: FA-Frey and original Frey experiment hyperparameter settings. The asterisks represent experiments that did not fully converge when they were terminated.

B.2 Frey data details

In the experiments on FA-Frey and original Frey data, we fit a FA model with latent dimension of 43. The regression models were 1-hidden layer neural networks with the hidden layer dimension of 100 and leaky ReLU activations. For each data dimension with missing values, a regression function was trained for up to 800 epochs with a learning rate of 10^{-3}, and early stopping that stopped training of the current regression function if validation mean-squared error loss did not improve for more than 50 epochs. For the CDI methods we used mean imputation in the initialisation stage and in the beginning of training we used the Synchronous Gibbs sampler and then switched to standard Gibbs updates. The epoch of the switch from synchronous to standard updates is in 'switch imputation epoch' row. The batch size was 512 and the rest of the hyperparameters are in Table B.2. The experiments which did not fully converge

before termination of training are marked with an asterisk on the '# of epochs' rows.

Appendix C

Additional evaluation results

In this appendix chapter we present additional CDI algorithm evaluation results. In Section C.1 we present additional evaluation on the original Frey faces data.

C.1 Evaluation on Frey data

In this section we present additional evaluation on Frey data. Since the ground truth model is not known, the evaluation is limited to metrics that do not require comparison to the ground truth model. Otherwise the evaluation methodology is the same as the one used for FA-Frey data in sections 5.2 and 5.3.3.

Figure C.1 presents the log-likelihood of the fitted density model on a held-out validation set. The performance of CDI (single-pass) is comparable to EM (joint) and outperformed the baseline models at all fractions of missingness. Which is inline with the results on Toy data and FA-Frey data.

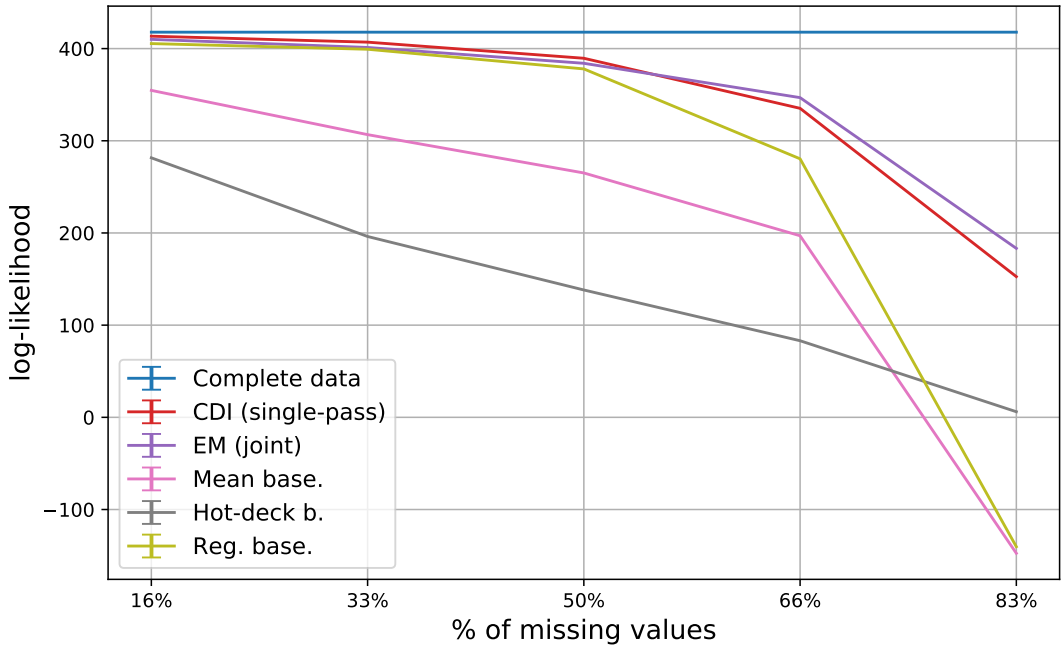


Figure C.1: Average log-likelihood on a held-out set of the original Frey data for missingness fractions from 16% to 83%.