MIWAE: Deep Generative Modelling and Imputation of Incomplete Data Sets

Pierre-Alexandre Mattei [1]   **Jes Frellsen** [1]

Abstract

We consider the problem of handling missing data with deep latent variable models (DLVMs). First, we present a simple technique to train DLVMs when the training set contains missing-at-random data. Our approach, called MIWAE, is based on the importance-weighted autoencoder (IWAE), and maximises a potentially tight lower bound of the log-likelihood of the observed data. Compared to the original IWAE, our algorithm does not induce any additional computational overhead due to the missing data. We also develop Monte Carlo techniques for single and multiple imputation using a DLVM trained on an incomplete data set. We illustrate our approach by training a convolutional DLVM on incomplete static binarisations of MNIST. Moreover, on various continuous data sets, we show that MIWAE provides extremely accurate single imputations, and is highly competitive with state-of-the-art methods.

1. Introduction

In data analysis, we often encounter incomplete data sets. For instance, we may choose to observe a limited set of values due to some constraints; sometimes a data acquisition machine will break; more and more often, respondents do not respond to parts of surveys (National Research Council, 2013). This ubiquity, illustrated for example by Josse & Reiter (2018) or by the monograph of Little & Rubin (2002), calls for a sound statistical response.

Over the last decades, numerous statistical approaches to handling incomplete data sets have been proposed—reviewed, e.g., by Schafer (1997), Little & Rubin (2002), and Van Buuren (2018). One way of handling the missing values is to replace them by a single set of imputations. Such a procedure, called *single imputation*, can be performed effi-

ciently using nonparametric methods, in particular k-nearest neighbours (kNN, see e.g. Troyanskaya et al., 2001) or random forests (Stekhoven & Bühlmann, 2011). However, when dealing with high dimensional and complex data distributions *multiple imputation* methods (recently reviewed by Murray, 2018) are often preferred. Such techniques create several sets of imputed values, and consequently allow to quantify the uncertainty of the imputed values. Often, they rely on fitting a parametric model by integrating the missing values out of the likelihood (in a maximum-likelihood setting) or the posterior distribution (in a Bayesian context). Quite often, these approaches are valid under the assumption that the data are *missing-at-random* (MAR, see e.g. Little & Rubin, 2002), which essentially means that the missingness mechanism is independent of the missing values.

Earlier attempts mainly relied on using multivariate Gaussian or Student's t distributions (see e.g. Little & Rubin, 2002, Chapters 11 & 12), with potentially low-rank constraints (e.g. Audigier et al., 2016). More complex models have also been explored, including mixtures (e.g. Di Zio et al., 2007) and nonparametric Bayes (e.g. Si & Reiter, 2013).

1.1. Deep generative models and incomplete data

The main purpose of this work is to present a simple way to fit a deep latent variable model (DLVM, Kingma & Welling, 2014; Rezende et al., 2014) to an incomplete data set, and to show how to use it for single or multiple imputation.

DLVMs are latent variable models that draw their high flexibility from deep neural architectures. They can be seen as nonlinear extensions of factor analysis and have links with robust principal component analysis (Dai et al., 2018) and nonparametric mixture models (Mattei & Frellsen, 2018a). In recent years, DLVMs have had remarkable successes in modelling images (Siddharth et al., 2017), molecules (Kusner et al., 2017; Gómez-Bombarelli et al., 2018; Riesselman et al., 2018), arithmetic expressions (Kusner et al., 2017), and single-cell data (Grønbech et al., 2018; Lopez et al., 2018).

Since DLVMs are generative models, data imputation is a natural application and was actually already considered in the seminal paper of Rezende et al. (2014). However, most of the inferential machinery used within DLVM cannot read-

[1]Department of Computer Science, IT University of Copenhagen, Denmark. Correspondence to: Pierre-Alexandre Mattei <pima@itu.dk>, Jes Frellsen <jefr@itu.dk>.

ily handle incomplete data sets. Indeed, DLVMs are usually fit by maximising lower bounds of the likelihood of the training data. Such bounds include the popular variational autoencoder bound (VAE, Kingma & Welling, 2014; Rezende et al., 2014), and the tighter importance weighted autoencoder (IWAE, Burda et al., 2016) bound. All these bounds assume that the training data are fully observed. Moreover, the conditionals of a DLVM are generally intractable. Some adaptations are therefore required for DLVMs to handle incomplete data sets efficiently. Specifically, the requirements that we would like to meet are:

(a) Having a sound statistical procedure for training DLVMs on incomplete data sets, under some assumptions on the nature of the missingness mechanism.

(b) Not losing the scalability of variational inference algorithms for DLVMs because of the presence of missing data. In particular, we wish to be able to do online learning on incomplete data sets.

(c) Having sound statistical procedures for both single imputation and multiple imputation.

1.2. Previous and concurrent work

Recently, several attempts at training DLVMs in the presence of missing data have been proposed. Yoon et al. (2018) considered a way to do so by building on the generative adversarial networks (GANs) of Goodfellow et al. (2014). Their training procedure involves a min-max game which, in the idealised nonparametric limit, allows to recover the data-generating process under the missing-completely-at-random (MCAR, see e.g. Little & Rubin, 2002) assumption. The MCAR assumption is stronger than the MAR assumption, and asserts that the missing mechanism is independent of the data. Another GAN-based approach using the MCAR assumption is the one of Li et al. (2019). They optimise jointly three objectives, with respect to six neural networks, which appears quite computationally complex, especially given the optimisation difficulties often noticed when training mix-max objectives like the GANs ones.

An interesting aspect of maximum likelihood inference is that it requires only the weaker MAR assumption. Approximate maximum likelihood using variational inference appears therefore like a natural solution to Requirement (a). Nazabal et al. (2018) proposed to slightly change the VAE objective in order to obtain a lower bound of the likelihood of the observed data. Their approach, like ours, uses zero imputation, and their bound is a particular case of ours. They also propose efficient ways of dealing with heterogeneous data sets. An important limitation of their bound is that it cannot be tight, even under the conditions that guarantee tightness of the original VAE bound (these conditions are discussed for example by Cremer et al., 2018). On the other hand, the likelihood bound that we introduce can be tight, in the limit of very large computational power, similarly to the original IWAE bound that we build on. Ma et al. (2018a;b), like Nazabal et al. (2018), revisit the VAE bound in a missing data context. Rather than relying on zero imputation, they elegantly use a more general framework based on permutation invariant neural networks. It is however unclear whether or not their bound can be tight. Regarding (c), these VAE-based approaches use heuristic schemes that may lead to poor results when the posterior approximation used for variational inference is far from the actual posterior.

A natural way to meet Requirement (c) would be to use the conditional distribution of the missing data. Given a trained DLVM, some Markov chain Monte Carlo (MCMC) schemes have been developed to sample from these conditionals (Rezende et al., 2014; Mattei & Frellsen, 2018a). However, they require to have access to a complete training set, which is very limiting. A way to design a model with easily approximable conditionals was recently proposed by Ivanov et al. (2019). Although their primary concern was to treat cases that involve complete training sets (like image inpainting), they propose a heuristic modification of their objective to handle missing data during training.

1.3. Contributions

In this paper, we address Requirement (a) in Section 2 by building a simple modification of the IWAE objective of Burda et al. (2016), the *missing data importance-weighted autoencoder (MIWAE) bound*. This objective is a lower-bound of the likelihood of the observed data that can be tight in the limit of very large computational power. Optimising MIWAE, by doing approximate maximum likelihood on the observed data, is sound in a MAR scenario. The computational cost of our algorithm does not exceed the one of the original IWAE, therefore satisfying Requirement (b).

Moreover, once a DLVM is trained using MIWAE, we show how to use it for missing data imputation. Specifically, we show how to approximate the conditional expectation of any function of the missing data using importance sampling—a particular application is single imputation. We also show how to perform multiple imputation using sampling importance resampling. While approximate, these two algorithms, presented in Section 3, will give the right answer in the limit of an infinite number of samples, therefore addressing Requirement (c) in a grounded way.

2. Training DLVMs with missing data

We start with some i.i.d. data stored in a matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T \in \mathcal{X}^n$. We assume that p different features are present in the data, leading to the decomposition $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_p$ (for example, if all features are contin-

uous, $\mathcal{X} = \mathbb{R}^p$). When some data are missing, we split each sample $i \in \{1, \ldots, n\}$ into the observed features $\mathbf{x}_i^o$ and the missing features $\mathbf{x}_i^m$. The indices of the missing features are stored in binary vectors $\mathbf{m}_i \in \{0, 1\}^p$ such that $m_{ij} = 0$ if feature j is observed for sample i, and $m_{ij} = 1$ if feature j is missing.

We wish to explain these potentially high-dimensional data using some latent variables $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^T \in \mathbb{R}^{n \times d}$. While the dimension d of the latent space will often be smaller than the dimension of the input space $\mathcal{X}$, this will not always be the case.

2.1. Deep latent variable models

DLVMs (Kingma & Welling, 2014; Rezende et al., 2014) assume that $(\mathbf{x}_i, \mathbf{z}_i)_{i \leq n}$ are driven by the following generative model:

$$\begin{cases} \mathbf{z} \sim p(\mathbf{z}) \\ p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) = \Phi(\mathbf{x}|f_{\boldsymbol{\theta}}(\mathbf{z})), \end{cases} \quad (1)$$

where

- $p(\mathbf{z})$ is the *prior distribution of the latent variables*, which is often a standard Gaussian distribution (e.g. Kingma & Welling 2014; Rezende et al. 2014), but might be much more complex and learnable (e.g. Tomczak & Welling 2018).

- $(\Phi(\cdot|\boldsymbol{\eta}))_{\boldsymbol{\eta} \in H}$ is a parametric family of distributions over $\mathcal{X}$ called the *observation model*. Often, it is a very simple family such as the Gaussian distribution if $\mathcal{X}$ is continuous, products of multinomial distributions if $\mathcal{X}$ is binary. It can even simply be the family of Dirac distributions—this is for example the case for generative adversarial networks (Goodfellow et al., 2014) or flow-based generative models (e.g. Dinh et al. 2017; Kingma & Dhariwal 2018).

- $f_{\boldsymbol{\theta}} : \mathbb{R}^d \to H$ is a function called the *decoder* (or the *generative network*), and is parametrised by a deep neural network whose weights are stored in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Prior knowledge about the problem at hand may be distilled into the architecture of $f_{\boldsymbol{\theta}}$: for example, by using convolutional layers when dealing with images (e.g. Salimans et al. 2017) or recurrent networks when dealing with sequential data (e.g. Bowman et al. 2016; Gómez-Bombarelli et al. 2018).

Assumptions. We make the assumption that the data are MAR, which is crucial to legitimate inference strategies based on the observed likelihood. Moreover, we make the fairly general assumption that the observation model $(\Phi(\cdot|\boldsymbol{\eta}))_{\boldsymbol{\eta} \in H}$ is such that all its marginal and conditional distributions are easy to sample and evaluate, which is true for commonly used observation models like the Gaussian

or the Bernoulli ones. In particular, these assumptions will conveniently guarantee that the quantity $p_{\boldsymbol{\theta}}(\mathbf{x}^o|\mathbf{z})$ is easy to compute.

2.2. Amortised Variational Inference with missing data

DLVMs are usually trained using approximate maximum likelihood techniques that maximise lower bounds of the log-likelihood function. Under the MAR assumption, the relevant quantity to maximise is the likelihood of the observed data $\mathbf{x}_1^o, \ldots, \mathbf{x}_n^o$, which is equal to

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^n \log p_{\boldsymbol{\theta}}(\mathbf{x}_i^o) = \sum_{i=1}^n \log \int p_{\boldsymbol{\theta}}(\mathbf{x}_i^o|\mathbf{z})p(\mathbf{z})d\mathbf{z}. \quad (2)$$

Unfortunately, the integrals involved in this expression make direct maximum likelihood impractical. However, it is possible to derive tractable tight lower bounds of $\ell(\boldsymbol{\theta})$ whose maximisation is much easier. More precisely, we will base our inference strategy on the IWAE approach of Burda et al. (2016), which combines ideas from *amortised variational inference* (Gershman & Goodman, 2014; Kingma & Welling, 2014; Rezende et al., 2014) and *importance sampling* (see e.g. Gelman et al. 2013, Section 10.4).

To build a lower bound using amortised variational inference, we will define a parametrised conditional distribution $q_{\boldsymbol{\gamma}}(\mathbf{z}|\mathbf{x}^o)$ called the *variational distribution* that will *play the role of a proposal distribution close to the intractable posterior* $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}^o)$.

Specifically, this conditional distribution will be defined as

$$q_{\boldsymbol{\gamma}}(\mathbf{z}|\mathbf{x}^o) = \Psi(\mathbf{z}|g_{\boldsymbol{\gamma}}(\iota(\mathbf{x}^o))), \quad (3)$$

where

- ι is an *imputation function chosen beforehand* that transforms $\mathbf{x}^o$ into a complete input vector $\iota(\mathbf{x}^o) \in \mathcal{X}$ such that $\iota(\mathbf{x}^o)^o = \mathbf{x}^o$.

- The set $(\Psi(\cdot|\boldsymbol{\kappa}))_{\boldsymbol{\kappa} \in \mathcal{K}}$ is a parametric family of simple distributions over $\mathbb{R}^d$, called the *variational family*.

- The function $g_{\boldsymbol{\gamma}} : \mathcal{X} \to \mathcal{K}$, called the *inference network* or the *encoder*, is parametrised by a deep neural network whose weights are stored in $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}$. Its role will be to transform each data point into the parameters of Ψ.

Following the steps of Burda et al. (2016), we can use the distribution $q_{\boldsymbol{\gamma}}$ to build approachable stochastic lower bounds of $\ell(\boldsymbol{\theta})$. More specifically, given $K \in \mathbb{N}^*$, we define the *missing data importance-weighted autoencoder (MIWAE) bound*

$$\mathcal{L}_K(\boldsymbol{\theta}, \boldsymbol{\gamma}) =$$
$$\sum_{i=1}^{n} \mathbb{E}_{\mathbf{z}_{i1},\ldots,\mathbf{z}_{iK} \sim q_{\boldsymbol{\gamma}}(\mathbf{z}|\mathbf{x}_i^o)} \left[\log \frac{1}{K} \sum_{k=1}^{K} \frac{p_{\boldsymbol{\theta}}(\mathbf{x}_i^o|\mathbf{z}_{ik}) p(\mathbf{z}_{ik})}{q_{\boldsymbol{\gamma}}(\mathbf{z}_{ik}|\mathbf{x}_i^o)} \right].$$
(4)

Jensen's (1906) inequality ensures that $\mathcal{L}_K(\boldsymbol{\theta}, \boldsymbol{\gamma}) \leq \ell(\theta)$, which means that $\mathcal{L}_K$ is indeed a lower bound of the observed log-likelihood. This inequality and the analogy with importance sampling motivate the name MIWAE for $\mathcal{L}_K$.

When $K = 1$, the bound resembles the variational autoencoder (VAE, Kingma & Welling, 2014; Rezende et al., 2014) objective, so we call this bound the *MVAE bound*. In very interesting concurrent work, Nazabal et al. (2018) derived independently the MVAE bound. They also discuss efficient strategies to design observation models that handle heterogeneous data sets.

The MIWAE bound is exactly the expected value of what we would get if we were to estimate the log-likelihood by approximating the integrals $p_{\boldsymbol{\theta}}(\mathbf{x}_1), \ldots, p_{\boldsymbol{\theta}}(\mathbf{x}_n)$ present in Eqn. (2) using importance sampling, with proposal distributions $q_{\boldsymbol{\gamma}}(\mathbf{z}|\mathbf{x}_1^o), \ldots, q_{\boldsymbol{\gamma}}(\mathbf{z}|\mathbf{x}_n^o)$. Regarding these importance sampling problems, it follows from a classical result of Monte Carlo integration (see e.g. Robert & Casella 2004, Section 3.3.2) that the optimal (in the sense of minimising the variance of the estimate) proposal distributions would exactly be the posterior distributions $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}_1^o), \ldots, p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}_n^o)$. For this reason, for all $i \in \{1, \ldots, n\}$, we may interpret (and use) $q_{\boldsymbol{\gamma}}(\mathbf{z}|\mathbf{x}_i^o)$ as an approximation of the posterior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}_i^o)$. The nature of this approximation is investigated further in section 2.4.

2.3. Building Blocks of the MIWAE bound

The use of the imputation function ι, which fills in the missing values in each data point in order to feed it to the encoder, is the main originality of the MIWAE bound. Intuitively, it would be desirable to use an accurate imputation function. However, thanks to the properties of importance sampling, we will see in the next subsection that *using a very rough imputation is acceptable*. In particular, in all the experiments of this paper, the function ι *will simply be the zero imputation function*, which replaces missing values by zeroes. Nazabal et al. (2018) also proposed to use zero imputation together with the MVAE bound. More complex imputations can of course be used. In particular it would also be possible to use a parametrised imputation function that could be learned by maximising $\mathcal{L}_K$. An interesting idea in that direction was recently proposed in a VAE context by Ma et al. (2018b), who advocate the use of a permutation invariant neural network. Interestingly, the zero imputation function is a particular case of their proposal.

However, in this paper, we argue that the use of importance sampling considerably alleviate the drawbacks of having a rough imputation function, such as the zero imputation one.

Regarding the variational family, the original IWAE paper of Burda et al. (2016) and used diagonal Gaussians (following Kingma & Welling, 2014 and Rezende et al., 2014). However, it is worth noticing that, similarly to the IWAE, our MIWAE bound could be used for any reparametrisable variational family, like the ones listed by Figurnov et al. (2018). Moreover, an interesting alternative to Gaussians would be to consider elliptical distributions, as recently proposed and advocated by Domke & Sheldon (2018). Indeed, such distributions allow to model different tail behaviours for the proposals $q_{\boldsymbol{\gamma}}(\mathbf{z}|\mathbf{x}_i^o)$, which is desirable in an importance sampling context. Specifically, in our experiments on UCI data sets, we used products of Student's t distributions.

In general, the encoder can simply be a fully connected network, as in Burda et al. (2016). However, when dealing with images or sounds, it will often be more reasonable to use a convolutional network, as in Salimans et al. (2017). When the data are sequential, recurrent networks are also an option (see e.g. Bowman et al. 2016; Gómez-Bombarelli et al. 2018).

2.4. Properties of the MIWAE bound

The main difference between our bound and the original IWAE bound of Burda et al. (2016) is that ours lower bounds the likelihood of an incomplete data set—using a proposal distribution that approximates $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}^o)$—while the original IWAE objective is a lower bound of the complete likelihood—whose proposal approximates $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$. When there is no missing data, our bound reduces to the original IWAE bound. *The MIWAE bound can thus be seen as a simple generalisation of the IWAE bound that can accommodate to missing data.* As a result, the properties of the MIWAE bound described in this subsection are directly obtained by leveraging recent theoretical work on the IWAE bound (Cremer et al., 2017; Nowozin, 2018; Rainforth et al., 2018; Domke & Sheldon, 2018; Tucker et al., 2019).

The one-sample MVAE bound $\mathcal{L}_1$ may be rewritten

$$\mathcal{L}_1(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \ell(\theta) - \text{KL}\left(\prod_{i=1}^{n} q_{\boldsymbol{\gamma}}(\mathbf{z}_i|\mathbf{x}_i^o) \middle|\middle| \prod_{i=1}^{n} p_{\boldsymbol{\theta}}(\mathbf{z}_i|\mathbf{x}_i^o) \right).$$
(5)

Therefore, maximising $\mathcal{L}_1$ will lead to minimising the Kullback-Leibler divergence between the true posterior distribution of the codes, and the variational distribution. In that sense, the variational approximation obtained by maximising $\mathcal{L}_1$ can be seen as a Kullback-Leibler projection of the posterior distribution on a set of distributions restricted by the choice of the variational family and the architecture of the encoder. For perspectives on the balance between

variational family and encoder architecture, see Cremer et al. (2018).

When $K > 1$, however, the variational approximation that maximises $\mathcal{L}_K$ cannot be seen as a Kullback-Leibler projection of the posterior. Indeed, as studied by Cremer et al. (2017), Naesseth et al. (2018), and Domke & Sheldon (2018) the variational distribution q_γ in (5) is replaced by a more complex distribution q_{IW} which depends both on θ and γ. This fact has both a good consequence—the bound will be tighter since q_{IW} is more complex—and a bad one— $q_\gamma(\mathbf{z}|\mathbf{x}^\text{o})$ might become a poor approximation of $p_\theta(\mathbf{z}|\mathbf{x}^\text{o})$ as K grows. Rainforth et al. (2018) give a formal treatment of this ambivalence, and propose modifications of the IWAE objective that tackle it. Domke & Sheldon (2018, Theorem 3) noted that, when $K \to \infty$, the variational approximation that maximises $\mathcal{L}_K$ may rather be seen as a χ projection rather than a Kullback-Leibler one. These properties are justifications for our use of $q_\gamma(\mathbf{z}|\mathbf{x}_i^\text{o})$ as an approximation of the posterior $p_\theta(\mathbf{z}|\mathbf{x}_i^\text{o})$ in the Monte Carlo schemes that we introduce in Section 3.

Importantly, Burda et al. (2016) proved that, the larger K, the tighter the bound. A direct consequence of their Theorem 1 is that when the posterior distributions have lighter tails than their variational counterparts,

$$\mathcal{L}_1(\theta, \gamma) \leq \mathcal{L}_2(\theta, \gamma) \leq \ldots \leq \mathcal{L}_K(\theta, \gamma) \xrightarrow[K \to \infty]{} \ell(\theta).$$
$$(6)$$

Under some mild moments conditions on the importance weights, the Theorem 3 of Domke & Sheldon (2018) will ensure that $\mathcal{L}_K(\theta, \gamma) =_{K \to \infty} \ell(\theta) + O(1/K)$, where the constant in $O(1/K)$ is linked to the χ divergence between $q_\gamma(\mathbf{z}|\mathbf{x}_i^\text{o})$ and the true posterior. These results show that, provided that we use a large K, optimising the MIWAE bound will lead to optimising a tight bound of the likelihood, even if the imputation function ι gives poor results. This motivates our choice to use simple imputation schemes for ι that can scale easily to large data sets. Specifically, in all the experiments of this paper, we chose to use the very naive but scalable zero imputation function. We will confirm empirically that, as long as K exceeds a few dozens, the accuracy of the imputation function has little effect on training.

In practice, the expectations in Eqn. (4) are hard to compute, but we can obtain unbiased estimates of the gradients of the bound $\mathcal{L}_K$ by sampling from the variational distribution. While this was the original approach of Burda et al. (2016), Roeder et al. (2017) and Rainforth et al. (2018) noticed both experimentally and theoretically that such gradients may have an overly large variance. A solution to this problem, based on a double use of the reparametrisation trick, was recently proposed by Tucker et al. (2019). While we did not implement this trick, it is very likely that using it would improve the optimisation of our MIWAE bound.

3. Missing data imputation

We have derived a statistically sound objective for training DLVMs in the presence of missing data. However, it is often useful to be also able to impute these data, through either single or multiple imputation. From now on, we assume that we have already trained our DLVM using the MIWAE bound, resulting in two known distributions p_θ and q_γ. We assume in this section that we are given a data point $\mathbf{x} \in \mathcal{X}$ composed of some observed features $\mathbf{x}^\text{o}$ and some missing data $\mathbf{x}^\text{m}$. This data point can be taken from the training set, or from another incomplete data set sampled from the same distribution.

3.1. Existing Techniques for Imputation with DLVMs

Since we have already learned a generative model p_θ, a good way of imputing $\mathbf{x}^\text{m}$ would be to sample according to its conditional distribution

$$p_\theta(\mathbf{x}^\text{m}|\mathbf{x}^\text{o}) = \int p_\theta(\mathbf{x}^\text{m}|\mathbf{x}^\text{o}, \mathbf{z}) p(\mathbf{z}|\mathbf{x}^\text{o}) d\mathbf{z}. \qquad (7)$$

The nonlinearity of the decoder makes this conditional distribution hard to assess. However, it is possible to sample from a Markov chain whose stationary distribution is exactly $p_\theta(\mathbf{x}^\text{m}|\mathbf{x}^\text{o})$. Rezende et al. (2014) first proposed a simple scheme to generate samples that approximatively follow $p_\theta(\mathbf{x}^\text{m}|\mathbf{x}^\text{o})$ by starting with some random imputation, which will be autoencoded several times. Their approach was then improved by Mattei & Frellsen (2018a), who derived a Metropolis-within-Gibbs sampler that asymptotically produces samples from $p_\theta(\mathbf{x}^\text{m}|\mathbf{x}^\text{o})$.

These two techniques were developped to use DLVMs trained on complete data, and critically require the availability of a good approximation of the posterior distribution of the complete data $p_\theta(\mathbf{z}|\mathbf{x})$. Indeed, if $p_\theta(\mathbf{z}|\mathbf{x})$ is far from $q_\gamma(\mathbf{z}|\mathbf{x})$, the scheme of Rezende et al. (2014) will converge to a distribution far away from $p_\theta(\mathbf{x}^\text{m}|\mathbf{x}^\text{o})$, while the Metropolis-within-Gibbs sampler of Mattei & Frellsen (2018a) will converge to the right distribution but at an impractical convergence rate.

However, while training a VAE or an IWAE will provide an approximation of $p_\theta(\mathbf{z}|\mathbf{x})$ via the inference network, using the MIWAE bound rather provides an approximation of $p_\theta(\mathbf{z}|\mathbf{x}^\text{o})$—see Section 2.4. Therefore, the schemes of both Rezende et al. (2014) and Mattei & Frellsen (2018a) seem unfit for our purposes. We will therefore derive a new imputation technique compatible with a DLVM trained using the MIWAE bound.

3.2. Single Imputation with MIWAE

The main idea of our imputation scheme is to leverage the fact that $q_\gamma(\mathbf{z}|\mathbf{x}^\text{o})$ is a good approximation of $p_\theta(\mathbf{z}|\mathbf{x}^\text{o})$.

Let us first focus on the *single imputation problem*: finding a single imputation $\hat{\mathbf{x}}^{\mathrm{m}}$ that is close to the true $\mathbf{x}^{\mathrm{m}}$. If the data are continuous and the ℓ_2 norm is a relevant error metric, then the optimal decision-theoretic choice would be $\hat{\mathbf{x}}^{\mathrm{m}} = \mathbb{E}[\mathbf{x}^{\mathrm{m}}|\mathbf{x}^{\mathrm{o}}]$, which is likely to be intractable for the same reasons $p_{\boldsymbol{\theta}}(\mathbf{x}^{\mathrm{m}}|\mathbf{x}^{\mathrm{o}})$ is. We can actually give a recipe to estimate the more general quantity

$$\mathbb{E}[h(\mathbf{x}^{\mathrm{m}})|\mathbf{x}^{\mathrm{o}}] = \int h(\mathbf{x}^{\mathrm{m}})p_{\boldsymbol{\theta}}(\mathbf{x}^{\mathrm{m}}|\mathbf{x}^{\mathrm{o}})d\mathbf{x}^{\mathrm{m}}$$
$$= \iint h(\mathbf{x}^{\mathrm{m}})p_{\boldsymbol{\theta}}(\mathbf{x}^{\mathrm{m}}|\mathbf{x}^{\mathrm{o}}, \mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}^{\mathrm{o}})d\mathbf{z}d\mathbf{x}^{\mathrm{m}}, \quad (8)$$

where $h(\mathbf{x}^{\mathrm{m}})$ is any absolutely integrable function of $\mathbf{x}^{\mathrm{m}}$. Indeed, this integral can be estimated using self-normalised importance sampling with the proposal distribution $p_{\boldsymbol{\theta}}(\mathbf{x}^{\mathrm{m}}|\mathbf{x}^{\mathrm{o}}, \mathbf{z})q_{\boldsymbol{\gamma}}(\mathbf{z}|\mathbf{x}^{\mathrm{o}})$, leading to the estimate

$$\mathbb{E}[h(\mathbf{x}^{\mathrm{m}})|\mathbf{x}^{\mathrm{o}}] \approx \sum_{l=1}^{L} w_l h\left(\mathbf{x}^{\mathrm{m}}_{(l)}\right), \quad (9)$$

where $(\mathbf{x}^{\mathrm{m}}_{(1)}, \mathbf{z}_{(1)}), \ldots, (\mathbf{x}^{\mathrm{m}}_{(L)}, \mathbf{z}_{(L)})$ are i.i.d. samples from $p_{\boldsymbol{\theta}}(\mathbf{x}^{\mathrm{m}}|\mathbf{x}^{\mathrm{o}}, \mathbf{z})q_{\boldsymbol{\gamma}}(\mathbf{z}|\mathbf{x}^{\mathrm{o}})$ that can be sampled via simple ancestral sampling, and, for all $l \in \{1, \ldots, L\}$

$$w_l = \frac{r_l}{r_1 + \ldots + r_L}, \text{ with } r_l = \frac{p_{\boldsymbol{\theta}}(\mathbf{x}^{\mathrm{o}}|\mathbf{z}_{(l)})p(\mathbf{z}_{(l)})}{q_{\boldsymbol{\gamma}}(\mathbf{z}_{(l)}|\mathbf{x}^{\mathrm{o}})}. \quad (10)$$

This estimate can be computed by performing a single encoding of $\iota(\mathbf{x}^{\mathrm{m}})$ and L decodings of the samples from the variational distribution. This is an additional advantage compared to the schemes of Rezende et al. (2014) and Mattei & Frellsen (2018a), which require one encoding and one decoding at each iteration of the Markov chain.

Often the quantity $\mathbb{E}[h(\mathbf{x}^{\mathrm{m}})|\mathbf{x}^{\mathrm{o}}, \mathbf{z}]$ is available in closed-form. This is for example the case when h is the identity function and the observation model has closed-form conditional means (like multivariate Gaussians or products of univariate distributions). In that situation, a slightly improved estimate is

$$\mathbb{E}[h(\mathbf{x}^{\mathrm{m}})|\mathbf{x}^{\mathrm{o}}] \approx \sum_{l=1}^{L} w_l \mathbb{E}\left[h(\mathbf{x}^{\mathrm{m}})|\mathbf{x}^{\mathrm{o}}, \mathbf{z}_{(l)}\right], \quad (11)$$

where the weights are defined as in Eqn. (10). By virtue of the law of total variance, the variance of this new estimate will always be smaller than the one defined in Eqn. (9).

3.3. Multiple Imputation with MIWAE

Often, it is more interesting to have access to a large number of i.i.d. samples from the conditional distribution than to estimate its moments. Such *multiple imputation* is also approachable using the importance sampling scheme of the previous subsection.

Indeed, using *sampling importance resampling* (see e.g. Gelman et al. 2013, Section 10.4) with the weights defined in Eqn. (10) allows to draw some samples that will be approximatively i.i.d. samples from $p_{\boldsymbol{\theta}}(\mathbf{x}^{\mathrm{m}}|\mathbf{x}^{\mathrm{o}})$ when L is large.

The simple process to sample M samples from the conditional distribution goes at follows. First, for some $L \gg M$, we sample $(\mathbf{x}^{\mathrm{m}}_{(1)}, \mathbf{z}_{(1)}), \ldots, (\mathbf{x}^{\mathrm{m}}_{(L)}, \mathbf{z}_{(L)})$ and compute the weights similarly to the previous subsection. Then, we sample M imputations with replacement from the set $\{\mathbf{x}^{\mathrm{m}}_{(1)}, \ldots, \mathbf{x}^{\mathrm{m}}_{(L)}\}$, the probability of each sample being exactly its importance weight.

4. Experiments

In this section, we explore empirically the properties of DLVMs trained with the MIWAE bound, as well as their imputations. General implementation details are provided as supplementary material.

4.1. A MCAR version of MNIST

We illustrate the features of MIWAE by training a DLVM on an incomplete version of the static binarisation of MNIST. We consider a simple setting: with 50% of the pixels missing uniformly at random (in a MCAR fashion). We use the convolutional architecture of Salimans et al. (2015). To assess the validity of our claim that using a naive imputation function is not too harmful when K is moderate, we compare using the zero imputation function, and using an oracle imputation that utilises the true values of the missing pixels. To that end, we focus on two models: one trained with the MIWAE bound with $K = 50$, and one trained using the MVAE bound (i.e. $K = 1$). As baselines, we also train a regular VAE and a regular IWAE using the complete data sets. To compare models, we evaluate estimates of their test log-likelihood obtained using importance sampling with 5000 samples and an inference network refitted on the test set, as suggested by Cremer et al. (2018) and Mattei & Frellsen (2018b).

The results are shown in Fig. 2. In the MVAE case ($K = 1$, which is essentially equivalent to the framework of Nazabal et al., 2018), using the oracle imputation provides a clear improvement. In the MIWAE case (with $K = 50$), both imputations schemes are on par (in accordance with our hypothesis), and outperform MVAE.

Single imputation. We also evaluate the quality of the imputations provided by MIWAE. To this end, we use MVAE and MIWAE with zero imputation, together with the proposed importance sampling scheme for imputation, and compare them to a state-of-the-art single imputation algorithm: missForest (Stekhoven & Bühlmann, 2011, fit using the `missForest` R package with default options). The results are displayed in Fig. 3. When $L \geq 10$, both MVAE

Figure 1. Random incomplete samples from the MNIST training data set, and single imputations obtained by MIWAE (trained with $K = 50$ importance weights, and imputed with $L = 10\,000$ importance weights). The MCAR data set is on the left, and the MAR data set is on the right.
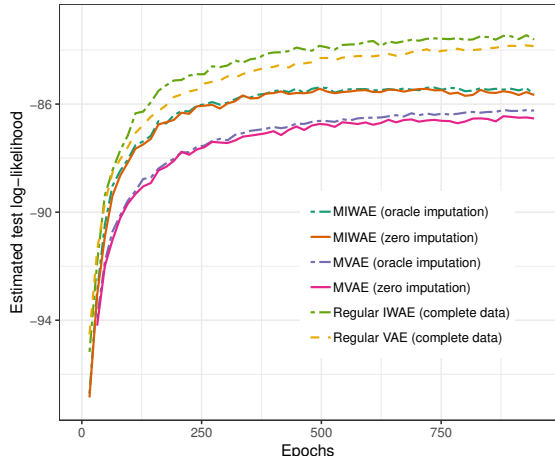

Figure 2. Estimated test log-likelihood of various models trained on binary MNIST as a function of the number of training epochs. The MIWAE model was trained using $K = 50$ importance
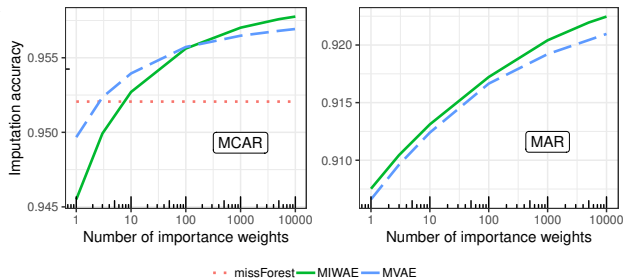

Figure 3. Imputation accuracy as a function of the number of importance weights (L) used in the single imputation scheme for various models trained on the two incomplete versions of MNIST. MissForest is not displayed for the MAR experiment because it provided quite inaccurate imputations (87.61% accuracy).

and MIWAE outperform missForest, and MIWAE provides the most accurate imputations when $L \geq 1\,000$. Some incomplete and imputed digits from the training set are also presented in Fig. 1. On Fig. 3, it appears that importance sampling converge faster for MVAE than for MIWAE. A possible explanation is that MIWAE learns a model where the true posterior is much less Gaussian than the model found by MVAE (this is consistent with the findings of Cremer et al., 2018, Section 5.4). This means that MIWAE will need more samples from a Gaussian proposal to provide accurate results. Another possible explanation is that the optimisation of the encoder is easier when optimising MVAE than MIWAE (see Rainforth et al., 2018), so the encoder found by MIWAE is slighlty worse. Using the the gradient estimates of Tucker et al. (2019) might help in that regard.

Multiple imputation. To evaluate multiple imputation, we consider the task of classifying the incomplete binary MNIST data set. We train a two-layer convolutional network (whose architecture is similar to the dropout one of Wan et al., 2013 and model selection is done using early stopping) using the original data and some imputed versions, and assess the classification performance on the complete test set. The results are shown in Table 1. Regarding MIWAE, we use both the single imputation obtained with $10\,000$ importance weights, and a multiple imputation of 20 complete data sets obtained using sampling importance resampling. For multiple imputation, we average the predictions obtained by training on the 20 imputed data sets. Interestingly, *leveraging multiple imputation is competitive with training on complete data.*

4.2. A MAR (but not MCAR) version of MNIST

We consider a MAR version of MNIST where all bottom halves of the pixels are observed. For each digit, either the top half, top quarter, or second quarter, is missing (depending on the number of white pixels in the bottom half). The detailed simulation scheme is provided as supplementary material. We perform the same single imputation experiments (see Fig. 3), and observe that MIWAE vastly outperforms missForest. Again, importance sampling seems to converge faster for MVAE than MIWAE, but MIWAE is more accurate. Incomplete and imputed digits are presented on the right hand side of Fig. 1. We also inspect visually multiple imputations (Fig. 4) obtained with $50\,000$ importance weights and note that sampling importance resampling allows to efficiently explore several modes of the conditional distribution of the missing pixels.

	Test accuracy	Test cross-entropy
Zero imp.	0.9739 (0.0018)	0.1003 (0.0092)
missForest imp.	0.9805 (0.0018)	0.0645 (0.0066)
MIWAE single imp.	0.9847 (0.0009)	0.0510 (0.0035)
MIWAE multiple imp.	**0.9870 (0.0003)**	**0.0396 (0.0003)**
Complete data	0.9866 (0.0007)	0.0464 (0.0026)

Table 1. Test accuracy and cross-entropy obtained by training a convolutional network using the imputations obtained with the MCAR version of MNIST. The numbers are the mean of 10 repeated trainings with different seeds and standard deviations are shown in parentheses.

	Banknote	*Breast*	*Concrete*	*Red*	*White*	*Yeast*
MIWAE	**0.446 (0.038)**	**0.280 (0.021)**	**0.501 (0.040)**	**0.643 (0.026)**	**0.735 (0.033)**	**0.964(0.057)**
MVAE	0.593 (0.059)	0.318 (0.018)	0.587(0.026)	0.686 (0.120)	0.782 (0.018)	0.997 (0.064)
missForest	0.676 (0.040)	0.291 (0.026)	0.510 (0.11)	0.697 (0.050)	0.798 (0.019)	1.41 (0.02)
PCA	0.682 (0.016)	0.729 (0.068)	0.938 (0.033)	0.890 (0.033)	0.865 (0.024)	1.05(0.061)
kNN	0.744 (0.033)	0.831 (0.029)	0.962(0.034)	0.981 (0.037)	0.929 (0.025)	1.17 (0.048)
Mean	1.02 (0.032)	1.00 (0.04)	1.01 (0.035)	1.00 (0.03)	1.00 (0.02)	1.06 (0.052)

Table 2. Mean-squared error for single imputation for UCI data sets (mean and standard deviations over 5 randomly generated data sets).

4.3. Single imputation of UCI data sets

We now address the single imputation of various continuous data sets from the UCI database (Dua & Efi, 2017). All data sets are corrupted by removing half of the features uniformly at random in a MCAR fashion (we also present a MAR but not MCAR experiment in the supplementary material). As competitors, we chose the mean imputation, which is an often inaccurate baseline, the kNN approach of Troyanskaya et al. (2001) fit using the VIM R package (Kowarik & Templ, 2016), with the number of neighbours chosen in $\{5, ..., 15\}$ using 5-fold cross validation; the miss-Forest approach of Stekhoven & Bühlmann (2011) (fit using the missForest R package with default options); a technique based on principal component analysis (PCA, Josse & Husson, 2012a), fit using the missMDA R package (Josse & Husson, 2016), where the number of principal components is found using cross validation (Josse & Husson, 2012b).

We also trained two DLVMs with the same general features: the intrinsic dimension d is fixed to 10, which may be larger than the actual number of features in the data, but DLVMs are known to automatically ignore some latent dimensions (Dai et al., 2018); both encoder and decoder are multi-layer perceptrons with 3 hidden layers (with 128 hidden units) and tanh activations; we use products of Student's t for the variational family (following Domke & Sheldon, 2018) and the observation model (following Takahashi et al., 2018). We perform $500\,000$ gradient steps for all data sets; no regularisation scheme is used, but the observation model is constrained so that the eigenvalues of its covariances are larger than 0.01 (as suggested by Mattei & Frellsen, 2018a),

Figure 4. Digits from the MAR version of MNIST. The first digit of each row is from the original (complete) data, the second digit is its incomplete counterpart, and the last ten digits are multiple imputations obtained by MIWAE.

when the data are scaled, this means that the code explains at most 99% of the variance of the features. One of the DLVM was trained using the MVAE bound, which is equivalent to the objective of Nazabal et al. (2018), and one using the MIWAE bound with $K = 20$. We used both models to impute the data sets by estimating the conditional means with $L = 10\,000$ importance samples.

The results, averaged over 5 randomly corrupted repetitions, are presented in Table 2. MIWAE provides more accurate imputations than all competitors. Interestingly, going from $K = 1$ (i.e., MVAE) to $K = 20$ significantly improves the accuracy of the imputations. Additional experiments available as supplementary material suggest that the MIWAE imputations are also more accurate when the data are not MCAR. None of the other recent DLVM-based approaches is within the competitors of this experiment. However, Ivanov et al. (2019) reimplemented the GAN-based approach of Yoon et al. (2018) to compare it with their VAE-based model. They concluded that none of them significantly outperformed missForest, which may still be considered a state-of-the-art algorithm. We used the same architecture for all data sets; it is therefore likely that architecture search would significantly improve these results.

5. Conclusion and future work

We have proposed a simple way of performing approximate maximum likelihood training for a DLVM with an incomplete data set, by maximising an objective called the MIWAE bound. We have also shown how to use the obtained model for single and multiple imputation. A first way of improving the MIWAE optimisation would be to use the low variance gradients recently proposed by Tucker et al. (2019) for similar objectives. Another avenue would be to study more refined imputation functions, even though we have shown that zero imputation does a good job together with MIWAE. For instance, we could try to use the permutation invariant networks advocated by Ma et al. (2018a;b). An important limitation is that MIWAE can only deal with MAR data. Trying to extend to the missing-not-at-random case might be possible by explicitly modelling the missing data mechanism.

References

Audigier, V., Husson, F., and Josse, J. Multiple imputation for continuous variables using a Bayesian principal component analysis. *Journal of Statistical Computation and Simulation*, 86(11):2140–2156, 2016.

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Jozefowicz, R., and Bengio, S. Generating sentences from a continuous space. *Proceedings of CoNLL*, 2016.

Burda, Y., Grosse, R., and Salakhutdinov, R. Importance weighted autoencoders. In *International Conference on Learning Representations*, 2016.

Cremer, C., Morris, Q., and Duvenaud, D. Reinterpreting importance-weighted autoencoders. *International Conference on Learning Representations (Workshop track)*, 2017.

Cremer, C., Li, X., and Duvenaud, D. Inference suboptimality in variational autoencoders. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Dai, B., Wang, Y., Aston, J., Hua, G., and Wipf, D. Connections with robust PCA and the role of emergent sparsity in variational autoencoder models. *The Journal of Machine Learning Research*, 19(1):1573–1614, 2018.

Di Zio, M., Guarnera, U., and Luzi, O. Imputation through finite Gaussian mixture models. *Computational Statistics & Data Analysis*, 51(11):5305–5316, 2007.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real NVP. *International Conference on Learning Representations*, 2017.

Domke, J. and Sheldon, D. Importance weighting and varational inference. *Advances in Neural Information Processing Signals*, pp. 4475–4484, 2018.

Dua, D. and Efi, K. T. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Figurnov, M., Mohamed, S., and Mnih, A. Implicit reparameterization gradients. *Advances in Neural Information Processing Signals*, pp. 439–450, 2018.

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., and Rubin, D. B. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

Gershman, S. and Goodman, N. Amortized inference in probabilistic reasoning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36, 2014.

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., and Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 2018.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

Grønbech, C. H., Vording, M. F., Timshel, P. N., Sønderby, C. K., Pers, T. H., and Winther, O. scVAE: Variational auto-encoders for single-cell gene expression data. *bioRxiv*, 2018. URL https://www.biorxiv.org/content/early/2018/05/16/318295.

Ivanov, O., Figurnov, M., and Vetrov, D. Variational autoencoder with arbitrary conditioning. In *International Conference on Learning Representations*, 2019.

Jensen, J. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta mathematica*, 30(1):175–193, 1906.

Josse, J. and Husson, F. Handling missing values in exploratory multivariate data analysis methods. *Journal de la Société Française de Statistique*, 153(2):79–99, 2012a.

Josse, J. and Husson, F. Selecting the number of components in principal component analysis using cross-validation approximations. *Computational Statistics & Data Analysis*, 56(6):1869–1879, 2012b.

Josse, J. and Husson, F. missMDA: a package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, 70(1):1–31, 2016.

Josse, J. and Reiter, J. P. Introduction to the special section on missing data. *Statistical Science*, 33(2):139–141, 05 2018.

Kingma, D. P. and Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. *arXiv preprint arXiv:1807.03039*, 2018.

Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

Kowarik, A. and Templ, M. Imputation with the R package VIM. *Journal of Statistical Software*, 74(7):1–16, 2016.

Kusner, M. J., Paige, B., and Hernández-Lobato, J. M. Grammar variational autoencoder. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1945–1954, 2017.

Li, S. C.-X., Jiang, B., and Marlin, B. Learning from incomplete data with generative adversarial networks. In *International Conference on Learning Representations*, 2019.

Little, R. J. A. and Rubin, D. B. *Statistical analysis with missing data*. 2nd ed. Wiley-Interscience, Hoboken, NJ, 2002.

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, 2018.

Ma, C., Hernández-Lobato, J. M., Koenigstein, N., Nowozin, S., and Zhang, C. Partial VAE for hybrid recommender system. In *3rd NeurIPS workshop on Bayesian Deep Learning*. 2018a.

Ma, C., Tschiatschek, S., Palla, K., Hernández-Lobato, J. M., Nowozin, S., and Zhang, C. EDDI: Efficient dynamic discovery of high-value information with partial VAE. *arXiv preprint arXiv:1809.11142*, 2018b.

Mattei, P.-A. and Frellsen, J. Leveraging the exact likelihood of deep latent variable models. In *Advances in Neural Information Processing Systems*, pp. 3859–3870, 2018a.

Mattei, P.-A. and Frellsen, Jes. Refit your encoder when new data comes by. In *3rd NeurIPS workshop on Bayesian Deep Learning*. 2018b.

Murray, J. S. Multiple imputation: A review of practical and theoretical findings. *Statistical Science*, 33(2):142–159, 2018.

Naesseth, C., Linderman, S., Ranganath, R., and Blei, D. Variational sequential Monte Carlo. In *International Conference on Artificial Intelligence and Statistics*, pp. 968–977, 2018.

National Research Council. *Nonresponse in social science surveys: A research agenda*. Washington, DC: National Academies Press, 2013.

Nazabal, A., Olmos, P. M., Ghahramani, Z., and Valera, I. Handling incomplete heterogeneous data using VAEs. *arXiv preprint arXiv:1807.03653*, 2018.

Nowozin, S. Debiasing evidence approximations: On importance-weighted autoencoders and jackknife variational inference. In *International Conference on Learning Representations*, 2018.

Rainforth, T., Kosiorek, A., Le, Tuan A., Maddison, C., Igl, M., Wood, F., and Teh, Y. W. Tighter variational bounds are not necessarily better. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 4277–4285, 2018.

Rezende, D., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, pp. 1278–1286, 2014.

Riesselman, A. J., Ingraham, J. B., and Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nature methods*, 15(10):816–822, 2018.

Robert, C. P. and Casella, G. *Monte Carlo statistical methods*. Springer-Verlag New York, 2004.

Roeder, G., Wu, Y., and Duvenaud, D. Sticking the landing: Simple, lower-variance gradient estimators for variational inference. In *Advances in Neural Information Processing Systems*, pp. 6928–6937. 2017.

Salimans, T., Kingma, D. P., and Welling, M. Markov chain Monte Carlo and variational inference: Bridging the gap. In *International Conference on Machine Learning*, pp. 1218–1226, 2015.

Salimans, T., Karpathy, A., Chen, X., and Kingma, D. P. PixelCNN++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *International Conference on Learning Representations*, 2017.

Schafer, J. L. *Analysis of incomplete multivariate data*. Chapman & Hall, London, 1997.

Si, Y. and Reiter, J. P. Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics*, 38(5):499–521, 2013.

Siddharth, N., Paige, B., van de Meent, J.-W., Desmaison, A., Goodman, N., Kohli, P., Wood, F., and Torr, P. Learning disentangled representations with semi-supervised deep generative models. In *Advances in Neural Information Processing Systems*, pp. 5927–5937, 2017.

Stekhoven, D. J. and Bühlmann, P. MissForest: a nonparametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2011.

Takahashi, H., Iwata, T., Yamanaka, Y., Yamada, M., and Yagi, S. Student-t variational autoencoder for robust density estimation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 2696–2702, 2018.

Tomczak, J. and Welling, M. VAE with a VampPrior. In *International Conference on Artificial Intelligence and Statistics*, pp. 1214–1223, 2018.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, 2001.

Tucker, G., Lawson, D., Gu, S., and Maddison, C. J. Doubly reparameterized gradient estimators for Monte Carlo objectives. In *International Conference on Learning Representations*, 2019.

Van Buuren, S. *Flexible imputation of missing data*. 2nd ed. Chapman and Hall/CRC, 2018.

Wan, L., Zeiler, M., Zhang, S., LeCun, Y., and Fergus, R. Regularization of neural networks using DropConnect. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 1058–1066, 2013.

Yoon, J., Jordon, J., and van der Schaar, M. GAIN: Missing data imputation using generative adversarial nets. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 5689–5698, 2018.