Dimension Reduction for Data with Heterogeneous Missingness

Yurong Ling[1] **Zijing Liu**[2] **Jing-Hao Xue**[1]

[1]Department of Statistical Science, University College London, London, UK
[2]Department of Brain Sciences, Imperial College London, London, UK

Abstract

Dimension reduction plays a pivotal role in analysing high-dimensional data. However, observations with missing values present serious difficulties in directly applying standard dimension reduction techniques. As a large number of dimension reduction approaches are based on the Gram matrix, we first investigate the effects of missingness on dimension reduction by studying the statistical properties of the Gram matrix with or without missingness, and then we present a bias-corrected Gram matrix with nice statistical properties under heterogeneous missingness. Extensive empirical results, on both simulated and publicly available real datasets, show that the proposed unbiased Gram matrix can significantly improve a broad spectrum of representative dimension reduction approaches.

1 INTRODUCTION

Dimension reduction (DR) is important for analysing high-dimensional data as it helps reveal underlying structures of the data. A large number of DR methods have been successfully applied to real data, such as principal components analysis (PCA) [Pearson, 1901], dual probabilistic PCA and its non-linear variant Gaussian process latent variable model (GPLVM) [Lawrence, 2005]. However, missing data arise in many applications [Marlin et al., 2007, Shen et al., 2015, Hicks et al., 2017], making it infeasible standard DR methods, which are usually designed for complete data. To address the problems posed by missing data, a broad spectrum of methods have been proposed, including the expectation-maximisation approaches [Little and Rubin, 2019], the direct imputation of observations via either matrix completion [Candès and Recht, 2009, Candes and Plan, 2010, Hastie et al., 2015] or by chained equations [Van Buuren, 2018], and the implicit imputation of covariance matrix [Cho et al., 2017, Zhu et al., 2019].

In this work, we focus on the implicit imputation of the Gram matrix and show that an unbiased estimator of it, in the presence of missing data, offers a significant prospect for enhancing the reliability of many DR procedures. Specifically, a large number of widely used DR methods obtain the low-dimensional projections via the distance matrix or the Gram matrix rather than the data matrix. For example, multidimensional scaling (MDS) seeks to find an embedded low-dimensional structure, of which the distance matrix is as close to the high-dimensional distance matrix as possible [Torgerson, 1952]. In addition, the objective of preserving the distance relationship between data points is shared by many dimension reduction algorithms. Algorithms such as the t-distributed stochastic neighbor embedding (tSNE) [Maaten and Hinton, 2008] and the uniform manifold approximation and projection (UMAP) [McInnes et al., 2018], two favoured visualisation tools in data analysis, build the stochastic relationship between data points in the low-dimensional space based on their original Euclidean distances. Similarly, dual probabilistic PCA and its non-linear variant GPLVM also seek to find the low-dimensional embedding using the Gram matrix [Lawrence, 2005]. It is adequate for performing the aforementioned methods through precise calculation of either the Gram matrix or the distance matrix, due to the linear transformations between these two matrices: the Gram matrix can be obtained by doubly centering the squared Euclidean distance matrix [Van Der Maaten et al., 2009], while there also exists a linear transformation for converting the Gram matrix to the distance matrix, as shown by (14)). Consequently, for the relevant DR approaches, we do not need to impute the missing values as long as we can estimate the distance or Gram matrix reliably in such cases.

Although Cho et al. [2017] and Zhu et al. [2019] studied the eigenvectors and eigenvalues for homogeneous and heterogeneous missing data, respectively, the effect of missing data on the techniques beyond PCA remains unclear. Moreover, the consequences of neglecting missing observations

Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021), PMLR 161:1310–1320.

are not thoroughly studied from a statistical perspective. Therefore, this paper aims to fill in this critical gap by making the following contributions. First, we elucidate how a reliable Gram matrix can ensure the powerful representation using GPLVM, a generalised framework for DR where the Gram matrix can be seen as an estimator of the covariance matrix [Lawrence, 2005] (sec.2.1). Secondly, we show that, owing to missing data, the original computation of the Gram matrix by an inner product matrix is a biased estimator of the covariance matrix with a larger variance under the framework of GPLVM, and we propose unbiased estimators in the cases of homogeneous missingness (sec.2.2) and heterogeneous mechanism (sec.2.3). In addition, we clarify the role of input dimension in the relevant dimension reduction methods, based on its relationship with the variances of the estimators (sec.2.4).

The data illustrated in this paper include image data and single-cell RNA sequencing (scRNA-seq) data, which measure gene expression at a single-cell level and offer a way to investigate the stochastic heterogeneity of complex issues on a near-genome-wide scale [Saliba et al., 2014, Shapiro et al., 2013, Kolodziejczyk et al., 2015]. The comparison is conducted in two aspects: visualisation and clustering results, on both simulated and real datasets (sec.4.4 and sec.4.5). Moreover, we empirically verify that the impact of input dimension is consistent to the results from our theoretical analysis (sec.4.3).

2 PROPOSED UNBIASED ESTIMATORS

In this section, we first show that the Gram matrix of high-dimensional data is an unbiased estimator of a covariance matrix when there is no missing observation and we clarify the importance of accurately computing the Gram matrix under the framework of GPLVM. The missing data model is then introduced and leads to bias in the estimator, and an unbiased estimator is derived in the presence of missing observations. Finally, we elucidate the role of input dimension on DR.

2.1 FOR COMPLETE DATA

Consider a dataset of N observations and D features represented as an $N \times D$ matrix $Y = [\mathbf{y}_1, \ldots, \mathbf{y}_N]^T$, where $\mathbf{y}_i \in \mathbb{R}^D$ is a D-dimensional observation. Under the assumption of GPLVM, every dimension is a realisation of a Gaussian process (GP) indexed by the latent variables $X = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T$, where $\mathbf{x_i} \in \mathbb{R}^d$ and d is the dimension of the latent space (normally $d \ll D$). Let the GP have a mean function $m(\mathbf{x})$ and a covariance function $k(\mathbf{x}, \mathbf{x}')$. For simplicity, $m(\mathbf{x})$ is taken to be the zero function ($m(\mathbf{x}) = 0$). Let $\mathbf{y}_{:,i}$ denote the i-th column of the data matrix Y, GPLVM assumes that $\mathbf{y}_{:,i} \sim \mathcal{N}(\mathbf{0}, K)$, where K is the covariance matrix with $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. GPLVM then aims to find the latent variables by maximising the marginal likelihood of the data:

$$\log p(Y|X,\theta) = \sum_{s=1}^{D} \log\ p(\mathbf{y}_{:,s}|X,\theta), \tag{1}$$

where

$$\log p(\mathbf{y}_{:,s}|X,\theta) = -\frac{1}{2}\mathbf{y}_{:,s}^T K^{-1}\mathbf{y}_{:,s} - \frac{N}{2}\log 2\pi - \frac{1}{2}\log|K|\,.$$

The above formulation provides a probabilistic interpretation of dual PCA in the case of the linear covariance function $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T\mathbf{x}_j$. Notably, GPLVM would be a non-linear model as long as K is obtained by a non-linear covariance function. Here we denote the latent points found by GPLVM by $\hat{X}$ and denote the corresponding covariance matrix by $\hat{K}$ to differentiate them from the true latent points X and true covariance matrix K, respectively.

The following Kullback-Leibler (KL) divergence, between the two Gaussians [Kullback and Leibler, 1951] equivalent to (1) up to a constant independent of X, clarifies the objective of GPLVM:

$$\begin{aligned}\mathbf{KL}(\mathcal{N}(z \mid \mathbf{0}, \frac{1}{D}G) \,||\, \mathcal{N}(z \mid \mathbf{0}, K) = \frac{1}{2}\log|K| - \\ \frac{1}{2}\log\left|\frac{1}{D}G\right| + \frac{1}{2}\mathrm{tr}(\frac{1}{D}GK^{-1}) - \frac{N}{2},\end{aligned} \tag{2}$$

where $G = YY^T$ is the Gram matrix. Thus, GPLVM seeks a matrix of latent points $\hat{X}$, which generate the covariance matrix $\hat{K}$, with $\hat{K}_{ij} = k(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j)$, as close to $\frac{1}{D}G$ as possible in terms of KL divergence. Moreover, $\frac{1}{D}G$ can be regarded as an estimator of the covariance matrix in $\mathcal{N}(\mathbf{0}, K)$. Since $G_{ij} = \sum_{s=1}^{D} y_{is}y_{js}$, under the assumption that each column of Y follows $\mathcal{N}(\mathbf{0}, K)$, the asymptotic properties of $\frac{1}{D}G$ can be summarised by the Lindeberg–Lévy central limit theorem (CLT) as [Lehmann, 2004]

$$\begin{aligned}\sqrt{D}(\frac{G_{ij}}{D} - K_{ij}) \overset{dist.}{\rightarrow} \mathcal{N}(\mathbf{0}, K_{ii}K_{jj} + K_{ij}^2),\ i \neq j, \\ \sqrt{D}(\frac{G_{ii}}{D} - K_{ii}) \overset{dist.}{\rightarrow} \mathcal{N}(\mathbf{0}, 2K_{ii}^2), i = j,\ \text{as } D \to \infty.\end{aligned} \tag{3}$$

That is, when there is no event of missingness, not only $\frac{1}{D}G$ is an unbiased estimator of the covariance matrix K, but also, with a higher dimension D, $\frac{1}{D}G$ is an estimator of higher accuracy because the variance shrinks with D.

2.2 FOR DATA WITH HOMOGENEOUS MISSINGNESS

The nice asymptotic properties in (3) may not hold as missing data exist. It is hence necessary to investigate the consequences of directly using the Gram matrix obtained from data with missing values. Let $\tilde{Y}$ denote the complete data matrix without missing entries and Y the partially observed

data matrix. We define a $N \times D$ binary revelation matrix Ω with 1 representing the corresponding entry in $\tilde{Y}$ being observed and 0 for the missing entry. Here, we assign a value of 0 to missing entries, as Cho et al. [2017], Zhu et al. [2019] did in their work, to calculate the Gram matrix with missing observations by YY^T. Therefore, we have $Y = \tilde{Y} \circ \Omega$, where $\circ$ denotes the element-wise product.

We first study a simple case where values in a dataset are missing independently and completely at random (MCAR) with the homogeneous probability. Under the framework of GPLVM, we assume that the partially observed data matrix Y is generated column-wise from $\mathcal{N}(\mathbf{0}, K)$, followed by the event of homogeneous missingness. The occurrences of missing observations are assumed to follow independent Bernoulli distributions with the homogeneous probability $1-p$ $(0 < p \leq 1)$. We also assume that the missing observations are independent of the Gaussian processes given p. The detailed statistical model is

$$\begin{aligned} \tilde{\mathbf{y}}_{:,s} &\sim \mathcal{N}(\mathbf{0}, K), \\ h_{is} \mid \tilde{y}_{is} &\sim \text{Bernoulli}(p), \\ y_{is} &= \begin{cases} \tilde{y}_{is} & \text{if } h_{is} = 1, \\ 0 & \text{if } h_{is} = 0, \end{cases} \end{aligned} \tag{4}$$

where $\tilde{y}_{is}$ is the true value in the i-th observation and s-th feature $(s = 1, \ldots, D;\ i = 1, \ldots, N)$.

From the above model, we get $\mathbb{E}(y_{is}y_{js}) = p^2 K_{ij}$ and $\text{Var}(y_{is}y_{js}) = p^2 K_{ii}K_{jj} + K_{ij}^2(2p^2 - p^4)$ for $i \neq j$; and $\mathbb{E}(y_{is}^2) = pK_{ii}$ and $\text{Var}(y_{is}^2) = K_{ii}^2(3p - p^2)$. With the CLT, the asymptotic properties of $\frac{1}{D}G$, as $D \to \infty$, are

$$\begin{aligned} &\sqrt{D}(\frac{G_{ij}}{D} - p^2 K_{ij}) \overset{dist.}{\to} \mathcal{N}(\mathbf{0}, p^2 K_{ii}K_{jj} + K_{ij}^2(2p^2 - p^4)), \\ &i \neq j; \\ &\sqrt{D}(\frac{G_{ii}}{D} - pK_{ii}) \overset{dist.}{\to} \mathcal{N}(\mathbf{0}, K_{ii}^2(3p - p^2)), i = j. \end{aligned} \tag{5}$$

Therefore, $\frac{1}{D}G$ is no longer an unbiased estimator of K for data with missing values. As we mentioned in sec.2.1, the Gram matrix is of importance to dual PCA and GPLVM. Hence the latent points found by GPLVM with the original Gram matrix G can be misleading. A straightforward solution to the problem is to use an unbiased estimator instead of $\frac{1}{D}G$. Based on the above analysis in (5), it is easy to see that an unbiased estimator for K is a matrix $\tilde{G}$ where

$$\tilde{G}_{ij} = \frac{G_{ij}}{Dp^2}\ (i \neq j) \text{ and } \tilde{G}_{ii} = \frac{G_{ii}}{Dp}.$$

2.3 FOR DATA WITH HETEROGENEOUS MISSINGNESS

In spite of the simplicity of the assumption presented in sec.2.2, the homogeneous missing data model conflicts with the fact that the missingness probability is often linked to some internal or external factors in reality. For instance, there usually exists an inverse relationship between the true values and the corresponding missingness probabilities in scRNA-seq data [Pierson and Yau, 2015]; in the recommendation system, whether a user rates a movie is determined by their preference and the movie's genre. We therefore consider the case where the missingness probabilities are heterogeneous; that is, values are missing not at random (MNAR). Further, we propose an unbiased estimator in such a situation. Now the statistical model is

$$\begin{aligned} \tilde{\mathbf{y}}_{:,s} &\sim \mathcal{N}(\mathbf{0}, K), \\ h_{is} \mid \tilde{y}_{is} &\sim \text{Bernoulli}(p_{is}), \\ y_{is} &= \begin{cases} \tilde{y}_{is} & \text{if } h_{is} = 1, \\ 0 & \text{if } h_{is} = 0, \end{cases} \end{aligned} \tag{6}$$

where $1 - p_{is}$ denotes the probability of missingness for the i-th observation and s-th feature, and $0 < p_{is} \leq 1$ $(s = 1, \ldots, D;\ i = 1, \ldots, N)$.

By using the statistical model in (6), we get the following two propositions regarding the estimator of K in such case.

Proposition 1 *Let the probabilities of missingness for the s-th feature in the i-th and j-th observations to be $1 - p_{is}$ and $1 - p_{js}$ respectively, where $i = 1, \ldots, N$, $s = 1, \ldots, D$. By assuming that the observed data matrix Y is generated according to the model in* (6)*, we have, for $i \neq j$,*

$$\begin{aligned} \mathbb{E}[y_{is}y_{js}] &= p_{is}p_{js}K_{ij}, \\ \text{Var}[y_{is}y_{js}] &= p_{is}p_{js}K_{ii}K_{jj} + K_{ij}^2(2p_{is}p_{js} - p_{is}^2 p_{js}^2); \end{aligned} \tag{7}$$

and for $i = j$,

$$\begin{aligned} \mathbb{E}[y_{is}y_{js}] &= \mathbb{E}[y_{is}^2] = p_{is}K_{ii}, \\ \text{Var}[y_{is}y_{js}] &= \text{Var}[y_{is}^2] = K_{ii}^2(3p_{is} - p_{is}^2). \end{aligned} \tag{8}$$

Based on Proposition 1, it is straightforward to conclude that $\frac{1}{D}G$ is a biased estimator of K. Consequently, it is necessary to correct the bias so as to get reliable $\hat{K}$ and $\hat{X}$.

Proposition 2 *By adopting the same assumption and notation as those in Proposition 1, we obtain an unbiased estimator $\tilde{G}$ of K with bounded variances. Specifically, for $i \neq j$ we have $\tilde{G}_{ij} = \frac{G_{ij}}{\sum_{s=1}^{D} p_{is}p_{js}}$, and $\text{Var}[\tilde{G}_{ij}]$ is given by*

$$\frac{K_{ii}K_{jj}\sum_{s=1}^{D} p_{is}p_{js} + K_{ij}^2 \sum_{s=1}^{D}(2p_{is}p_{js} - p_{is}^2 p_{js}^2)}{\left(\sum_{s=1}^{D} p_{is}p_{js}\right)^2}. \tag{9}$$

The bounds of $\text{Var}[\tilde{G}_{ij}]$ are given by

$$\frac{K_{ii}K_{jj} + K_{ij}^2}{D\bar{p}_{ij}} \leq \text{Var}(\tilde{G}_{ij}) \leq \frac{K_{ii}K_{jj}}{\bar{p}_{ij}D} + \frac{K_{ij}^2}{D}\left(\frac{2}{\bar{p}_{ij}} - 1\right), \tag{10}$$

where $0 < \bar{p}_{ij} = \frac{1}{D}\sum_{s=1}^{D} p_{is}p_{js} \leq 1$. *Note that the equality holds if and only if* $p_{is} = 1$, *for all* i *and* s, *which means no event of missing observations.*

For the diagonal entries, $\tilde{G}_{ii} = \frac{G_{ii}}{\sum_{s=1}^{D} p_{is}}$, *and* $\mathrm{Var}[\tilde{G}_{ii}]$ *is*

$$\frac{K_{ii}^2 \sum_{s=1}^{D} p_{is}(3 - p_{is})}{\left(\sum_{s=1}^{D} p_{is}\right)^2}. \tag{11}$$

The bounds of $\mathrm{Var}[\tilde{G}_{ii}]$ *are given by*

$$\frac{2K_{ii}^2}{D\bar{p}_i} \leq \mathrm{Var}(\tilde{G}_{ii}) \leq \frac{K_{ii}^2}{D}\left(\frac{3}{\bar{p}_i} - 1\right), \tag{12}$$

where $\bar{p}_i = \frac{1}{D}\sum_{s=1}^{D} p_{is}$. *Again, the equality holds if and only if* $p_{is} = 1$, *for all* i *and* s.

Based on Proposition 2, we conclude that $\tilde{G}$ is bias-corrected with the bounds of variance decreasing with D. Furthermore, under a mild condition, $\tilde{G}$ is a consistent estimator.

Proposition 3 *Let* $x_{ij,s} = y_{is}y_{js}$ *and* $\mu_{ij,s} = \mathbb{E}(x_{ij,s}) = p_{is}p_{js}K_{ij}$. *If* $\sum_{s=1}^{D} p_{is}p_{js} \asymp D$, *then*

$$Z_{ij,D} = \frac{\sum_{s=1}^{D}\left(x_{ij,s} - \mu_{ij,s}\right)}{\left[\sum_{s=1}^{D}\mathrm{Var}(x_{ij,s})\right]^{\frac{1}{2}}} \xrightarrow{dist.} \mathcal{N}(0,1), \tag{13}$$

as D *approaches infinity. Here we define* $\sum_{s=1}^{D} p_{is}p_{js} \asymp D$ *if there exist constants* $0 < m < M < \infty$, *and an integer* n_0 *such that* $m < \frac{\sum_{s=1}^{D} p_{is}p_{js}}{D} < M$, *for all* $D > n_0$.

Corollary 1 *If* $\sum_{s=1}^{D} p_{is}p_{js} \asymp D$, *the proposed unbiased estimator* $\tilde{G}$ *converges in probability to the true covariance matrix* K *as* D *approaches infinity.*

Proposition 3 and Corollary 1 suggest that the unbiased estimator $\tilde{G}$ of K could be beneficial for a method using the Gram matrix as input since it converges to the ground-truth covariance matrix in the presence of missing observations if $\sum_{s=1}^{D} p_{is}p_{js} \asymp D$. In reality, $\frac{\sum_{s=1}^{D} p_{is}p_{js}}{D} < M$ for any constant $M > 1$, since p_{ij}'s ≤ 1. Furthermore, there exists a constant $m > 0$ such that $m < \frac{\sum_{s=1}^{D} p_{is}p_{js}}{D}$ as long as the probabilities of non-missingness p_{ij}'s are bounded from below. Thus, the condition $\sum_{s=1}^{D} p_{is}p_{js} \asymp D$ is readily satisfied in practice. When applying the proposed estimator to DR methods, we use $\tilde{G}$ rather than $\frac{1}{D}G$ to improve the performance.

Corollary 2 *If* $\sum_{s=1}^{D} p_{is}p_{js} \asymp D$, *the estimator* G/D *converges in probability to the true covariance matrix* K *if and only if* $\lim_{D\to\infty} \frac{\sum_{s=1}^{D} p_{is}p_{js}}{D} = 1$.

Corollary 2 implies that the estimator G/D would still converge to K if the fraction of missing values is small enough as compared to 1. All proofs in this section are provided in the supplementary material. Note that the mathematical terms in the bias and the variances presented in Proposition 1 and Proposition 2, respectively, would be more involved if we set the missing observations to a non-zero constant, but the statistical properties remain the same.

2.4 IMPACT OF THE INPUT DIMENSION D

As shown in (3), (5) and Proposition 2, the variance of $\frac{1}{D}G$ and the bounds of $\mathrm{Var}(\tilde{G})$ are inversely proportional to the input dimension D. Hence, higher input dimension can lead to more accurate results of dimension reduction, from decreasing the variances of the estimators. Moreover, $\frac{1}{D}G$ would be close enough to K when there exists no missing entries in Y, as long as the dimension is high enough such that the corresponding variances approach zero, so is $\tilde{G}$ for data with missingness. Although $\tilde{G}$ is an unbiased estimator, as shown by Proposition 2, the variance of $\tilde{G}$ are greater than that of $\frac{1}{D}G$ in the presence of missing observations. In other words, in order to reach the same accuracy, more dimensions are required in the presence of missing data, compared with the case of complete data.

3 APPLICATION OF THE PROPOSED UNBIASED ESTIMATOR

In practice, the heterogeneous probabilities of missingness are unknown. Hence, we need to estimate them before applying the proposed estimator to real datasets. The procedure of estimation is proposed as follows: first compute $\mathbf{p_F} \in \mathbb{R}^D$ and $\mathbf{p_S} \in \mathbb{R}^N$, which are the vectors containing the proportion of non-missing observations for each feature and for each sample, respectively; then, the entries in the outer product of two vectors $\mathbf{p_S} \otimes \mathbf{p_F}$ scaled by a constant are treated as the matrix of estimated non-missingness probabilities for the data matrix. The detail of estimators is provided in sec.S.3 of the supplementary material.

Once all the p_{ij} are estimated, we compute $\tilde{G}$ and then substitute $\frac{G}{D}$ in the relevant Gram-matrix-based dimension reduction methods, such as PCA and GPLVM, to correct the bias. In addition, $\tilde{G}$ can benefit the approaches designed taking advantage of the distance structure, owing to the linear transformation between the squared Euclidean distance

matrix E^2 and G:

$$E^2 = \text{diag}(G)\mathbf{1}^T + \mathbf{1}\text{diag}(G)^T - 2G, \qquad (14)$$

where $\text{diag}(G)$ is a column vector of the diagonal elements in G. Considering that the bias corrected $\tilde{G}$ could result in negative values via (14), we propose an alternative way to enhance the distance-matrix-based methods such as tSNE and UMAP: first do PCA with the bias-corrected Gram matrix $\tilde{G}$, and then calculate the distance matrix in the PC space. To ensure a good estimate of the distance, we keep all the PCs with non-negative eigenvalues.

There is an additional challenge when handling real scRNA-seq data: the positions of missing entries remain unknown. There exist highly-frequent zero expression values in scRNA-seq data. Some zeros indicate the true biological non-expression while others are due to the corresponding missing values, which are called dropouts [Hicks et al., 2017, Li and Li, 2018]. To address the mentioned problem, we propose a simple yet reliable ensemble-learning strategy to infer the positions of missing entries (dropouts) in data matrix, as illustrated in Figure 1. Specifically, we first identify similar cells via clustering. A zero count is then regarded as true biological non-expression if most values of the same gene in the corresponding cluster are zero, otherwise it is taken to be a missing value. This identification procedure is performed multiple times using different clustering methods and different numbers of clusters to ensure reliable results. We reach the final decision by the majority voting: a zero count is considered as the true non-expression if more than half results confirm this. The proposal of this procedure is inspired by the principle introduced in scImpute that a zero count may reflect real biological variability if the corresponding gene has constantly low expression in similar cells [Li and Li, 2018]. After the true non-expression or dropout events are identified, we compute the probability of being a dropout across both observations (cells) and features (genes) as we mentioned before.

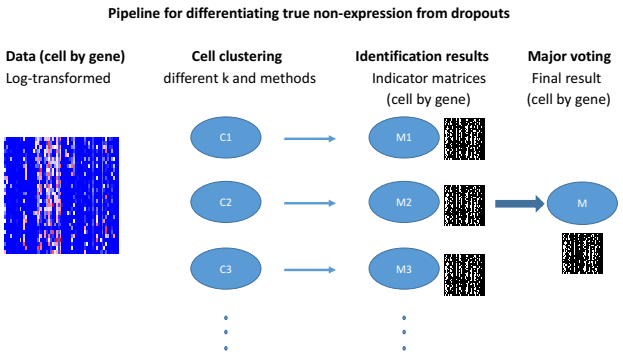


Figure 1: Pipeline for identifying dropout events. From left to right: 1) log-transformed data matrix; 2) clustering with different clustering methods and different numbers of clusters (k); 3) indicator matrices from different clustering results; and 4) combining all indicator matrices (results) by the majority voting.

Table 1: Real datasets used in this paper.

Dataset	# clusters/classes	N	D	Ref
Pollen (scRNA-seq)	11	301	21045	[Pollen et al., 2014]
Deng (scRNA-seq)	10	286	20484	[Deng et al., 2014]
Treutlein (scRNA-seq)	8	405	15893	[Treutlein et al., 2016]
Koh (scRNA-seq)	10	651	41594	[Loh et al., 2016]
Usoskin (scRNA-seq)	4	622	17571	[Usoskin et al., 2015]
Kumar (scRNA-seq)	3	361	16092	[Kumar et al., 2014]
Olivetti faces (image)	40	400	4096	[Samaria and Harter, 1994]
fashion MNIST (image)	10	1000	784	[Xiao et al., 2017]
wine (UCI)	3	178	13	[Dua and Graff, 2017]

4 EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we show the superiority of the proposed unbiased estimator in terms of the clustering accuracy and visualisation quality on both simulated and real datasets. The results in sec.2.4 regarding the role of the input dimension are empirically verified with the simulated data. The code to reproduce these experiments is available at `https://github.com/yurongling/DR-for-Data-with-Missingness.git`

4.1 DATASETS

Nine publicly available real datasets from different domains are selected for comparing different methods: 6 scRNA-seq datasets, 2 image datasets, and 1 dataset from the UCI repository. The characteristics of each dataset are provided in Table 1; the data pre-processing and availability are provided in sec.S.1 and sec.S.2 of the supplementary material, respectively. Clusters in each real scRNA-seq dataset are for different cell types. We sample 1000 images from the test set of the fashion MNIST dataset for comparison by preserving the percentage of samples of each class to reduce the computational complexity of some benchmark imputation methods. Note that the wine dataset possesses only a small number of features (13).

In addition to real datasets, we also simulate a dataset with 3 clusters for investigating the role of the input dimension. The complete dataset is first simulated with the Probabilistic PCA (PPCA) [Tipping and Bishop, 1999], which can be regarded as a GPLVM with a linear kernel. Then, the data matrix with missing observations is generated with a missingness mechanism mentioned below.

Missing value generation mechanism. Apart from scRNA-seq datasets, all datasets we employ are complete. We therefore adopt a missingness mechanism to generate missing positions. Specifically, $\text{P}(\Omega_{ij} = 0) = b_i q_j$, for $i \in [N]$, $j \in [D]$, where iid $b_1, \ldots, b_N \sim U[0.4, 0.6]$, and iid $q_1, \ldots, q_D \sim U[0.7, 0.9]$. The fraction of missingness is around 0.4.

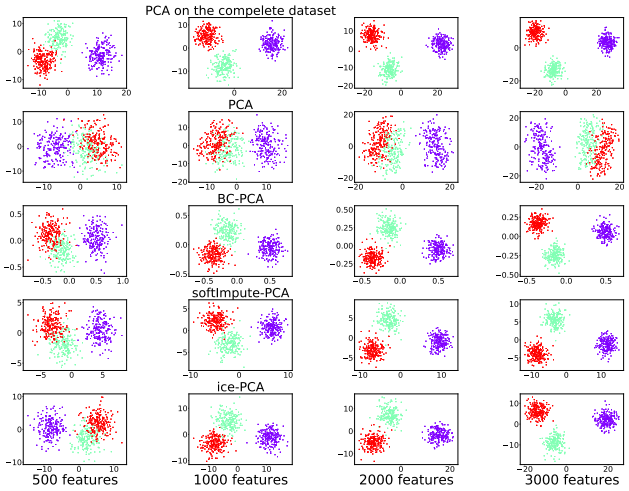


Figure 2: 2D scatter plots of the simulated data with or without missing observations, obtained by four different methods (from top to bottom: PCA on the complete dataset, PCA, BC-PCA, softImpute-PCA, and ice-PCA on the dataset with missing values), and with different numbers of genes (from left to right: 500, 1000, 2000 and all 3000 features). Colours indicate the cluster labels. Comparing columns: generally more input features, better separation between different clusters. Comparing rows: BC-PCA, softImpute-PCA, and ice-PCA lead to much more distinct clusters on the dataset with heterogeneous missingness.

4.2 BENCHMARKS

DR methods. To demonstrate the applicability and the effectiveness of the proposed estimator, we consider four DR methods: PCA, GPLVM, tSNE, and UMAP. PCA and GPLVM are representative Gram-matrix-based methods, while tSNE and UMAP are widely-used approaches depending on the distance matrix. In both the simulated and real experiments, we compare them with their bias-corrected variants proposed in this paper, where the Gram matrix is replaced by $\tilde{G}$ or the distance matrix is calculated in the PCA space obtained from $\tilde{G}$ as discussed in sec.3. The prefix *BC-* (bias-corrected) of each method is to denote its bias-corrected variant.

Imputation methods. The widely-used imputation methods softImpute [Hastie et al., 2015] and imputation by chained equations (ice) [Van Buuren, 2018] are applied to the datasets, followed by performing the DR methods on the imputed data matrices. softImpute is proposed with the low-rank assumption and performs missing values imputation using iterative soft-thresholded SVD's, while ice uses a strategy that models each feature with missing values as a function of other features in a round-robin fashion. We use the prefix *softImpute-*/*soft-* and *ice-* to represent the corresponding DR methods applied to the data matrix imputed by softImpute and ice, respectively. When these two imputation approaches are applied to real scRNA-seq data where the missing positions are unidentified, we input the missing positions inferred by the proposed pipeline.

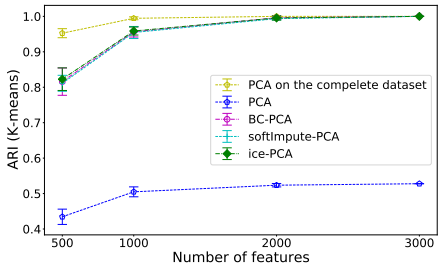


Figure 3: ARI of the k-means clustering results with different DR approaches on the simulated dataset with or without missing values.

Approaches designed specifically for scRNA-seq data For real scRNA-seq data, we also integrate two approaches that deal with the missing data problem of scRNA-seq data: CIDR [Lin et al., 2017] and scImpute [Li and Li, 2018]. CIDR tries to recover the Euclidean distance matrix while scImpute aims at imputing the missing values. The imputed data matrix produced by scImpute is directly fed into the mentioned four benchmarks for extracting low-dimensional components. tSNE and UMAP take the imputed Euclidean distance matrix yielded by CIDR as input. On the other hand, we transform the imputed distance matrix to the Gram matrix by doubly-centering, which is then input into GPLVM and PCA, respectively. We use the prefix *CIDR-* and *sc-*/*scImpute-* to denote the corresponding DR methods integrated with CIDR and scImpute, respectively,

Evaluation. We evaluate the performances from two perspectives: clustering and visualisation. For clustering-based evaluation, we use k-means clustering in the reduced space and the number of clusters is set to the same as the ground truth. The clustering results are evaluated in terms of the adjusted rand index (ARI) between the cluster/class labels obtained from the original publication and the inferred clustering labels. Since the missing positions determined by the proposed pipeline could be variable, we replicate the procedure of first performing DR and then applying k-means 20 times on the real scRNA-seq datasets for a more reliable comparison. Regarding the visualisation-based evaluation, we reduce the input data into two dimensions and visually compare the visualisations. For implementation details, see sec.S.4 of the supplementary material.

4.3 INPUT DIMENSION INFLUENCES THE PERFORMANCE OF DIMENSION REDUCTION

In order to examine the impact of input dimension on the performance of DR, we randomly select a subset of dimen-

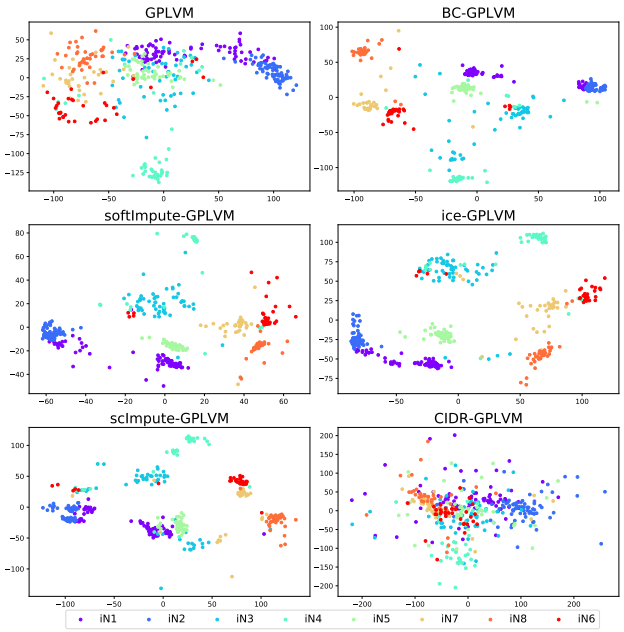


Figure 4: Visualisation of the Treutlein dataset obtained by GPLVM and its variants integrated with the bias correction or imputations.

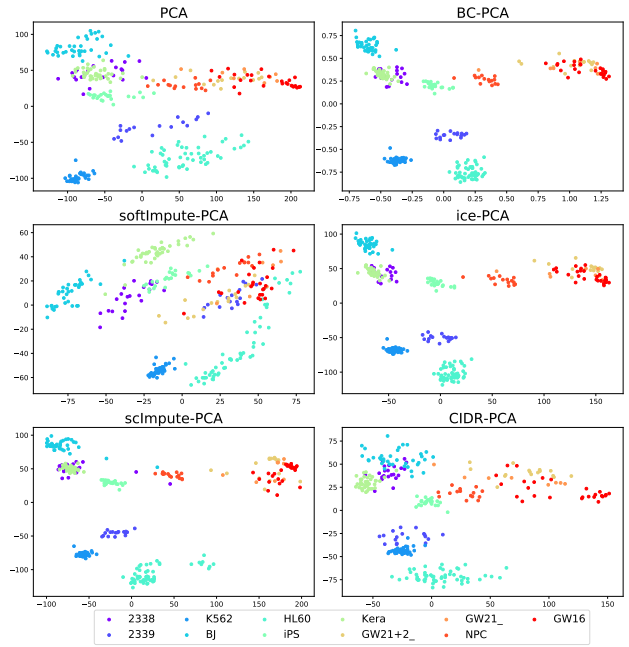


Figure 5: Visualisation of the Pollen dataset obtained by PCA and its variants integrated with the bias correction or imputations.

sions (features) from the simulated data. We then reduce the selected subset of data into two dimensions (2D). The qualities of the produced 2D projections are assessed according to the visualisation and the clustering accuracy. We repeat the aforementioned procedure 10 times and average the ARI. The size of the subset features varies across 500, 1000, and 2000. The performance using all features (3000) is also provided for comparison. Since the simulated dataset is generated in the context of PPCA, we compare only the DR methods based on PCA.

First, we find that, on the simulated datasets with and without missing observations, the visualisation (Figure 2) is of a higher quality with more input features, based on the separation between different clusters. The upward trends of the clustering performances on the simulated datasets shown in Figure 3 are consistent with the visualisation. In addition, more input features lead to a smaller deviation of the ARI.

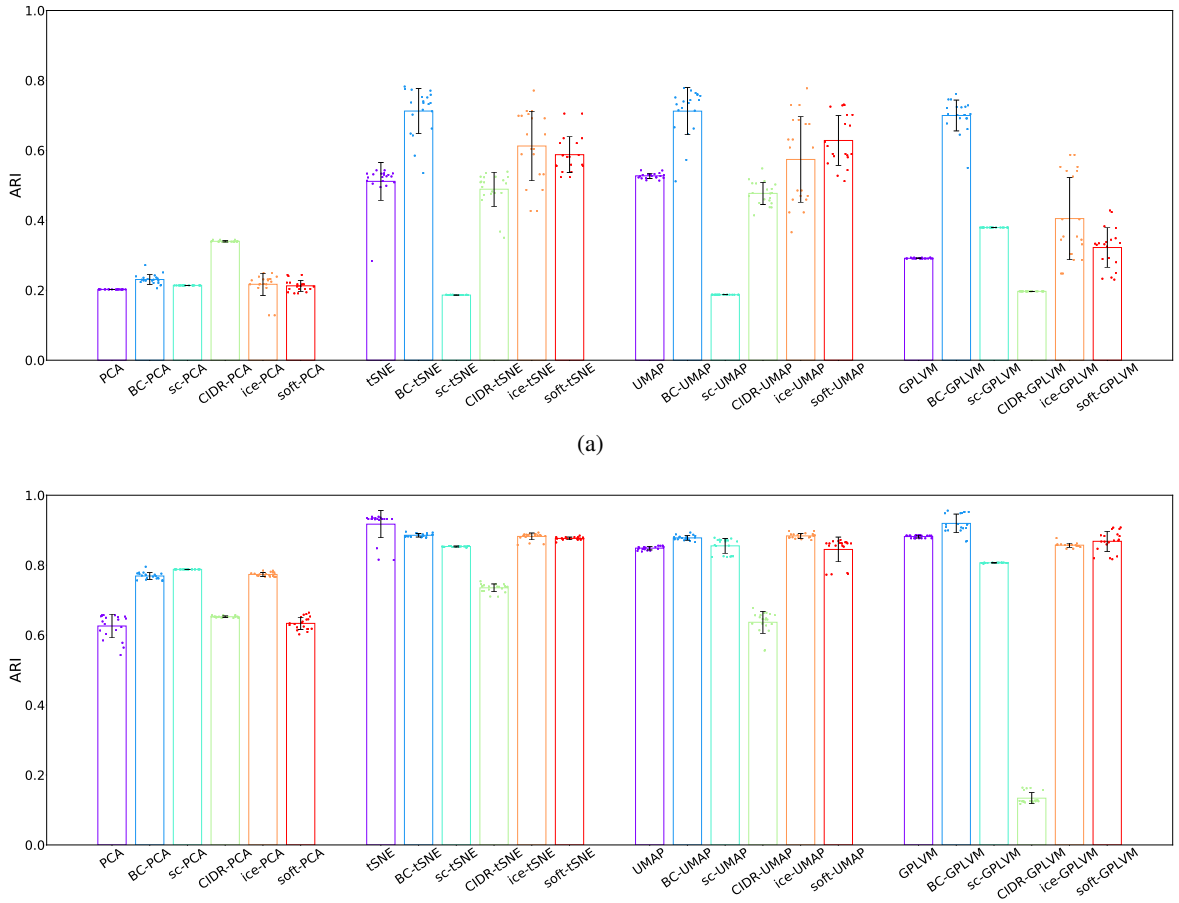
Second, by comparing the performances between the dataset with missing observations and that without missing observations, we find that, with missing data, a higher input dimension is required to reach a performance comparable to that of the complete data. For instance, PCA leads to distinct clusters when the number of input features is 1000 while BC-PCA renders clusters that are overlapping to some degree in such a case.

Overall, the experimental results on the simulated data offer empirical evidence confirming that the input dimension influences the performance of relevant DR approaches, as discussed in sec.2.4.

4.4 BIAS CORRECTION IMPROVES VISUALISATION

Now we examine whether the bias correction exploiting the information of missing data can produce better visualisation. First, we visualise the dimension-reduced data of the simulated datasets without and with bias-correction. Figure 2 shows that PCA is capable of separating different clusters when no missing entry is present in data matrix. However, for data with missing observations, PCA cannot distinguish the subpopulations very clearly (Figure 2). In contrast, BC-PCA shows much more distinct clusters. Furthermore, BC-PCA has comparable performance to softImpute-PCA and ice-PCA in terms of the separation of clusters.

Next, we focus on the comparison between the benchmark DR methods and their bias-corrected versions in terms of the visualisations displayed by them on a wide spectrum of real datasets. Compared with PCA, BC-PCA presents more divergent clusters on the Pollen dataset and the Kumar dataset (Figure 5 and Figure S8 of the supplementary material), and it achieves comparable performance on the other datasets. BC-GPLVM succeeds in separating most clusters on the Treutlein dataset (Figure 4), the Usoskin dataset (Figure S15 of the supplementary material), and the Koh dataset (Figure S14 of the supplementary material), showing a better performance than BC-PCA. It may be due to the nonlinearity of data structure, which is difficult to be captured by a linear dimension reduction method like PCA even after the bias correction. Meanwhile, the degrees of overlapping



(a)

(b)

Figure 6: ARI of k-means with different DR approaches and their variants on the datasets (a) Usoskin, and (b) Pollen.

between clusters are reduced greatly by BC-GPLVM on the fashion MNIST dataset (Figure S9 of the supplementary material) and the Olivettic faces dataset (Figure S10 of the supplementary material). Comparing GPLVM with BC-GPLVM, we find that BC-GPLVM often yields a clearer visualisation. Regarding the distance-matrix-based methods, both BC-tSNE and BC-UMAP clearly show superior visualisation to tSNE and UMAP, respectively.

Last, we compare the proposed bias-corrected estimator with other imputation methods and the approaches specifically handling scRNA-seq data. The visualisations obtained by CIDR are inferior to those produced by the bias-corrected variants in terms of the separation between different groups of cells. Compared with the bias-corrected benchmark methods, scImpute yields more compact clusters on the Koh dataset (Figure S6, Figure S14, Figure S23, and Figure S32 of the supplementary material). However, the within-cluster compactness and between-cluster separation are worse than or comparable to the approaches incorporating the unbiased Gram matrix on the other datasets. When applied to the fashion MNIST dataset and the Olivetti faces dataset, the proposed bias-corrected estimator is comparable to softImpute and ice; see the visualisations shown in Figure S9 and Figure S10 of the supplementary material. However, softImpute presents superior visualisations on the wine dataset (Figure S3, Figure S11, Figure S19, and Figure S28 of the supplementary material). The bias-corrected DR methods fail to separate different clusters in such a case, since our method is proposed to the data that are high-dimensional while the wine dataset has only a small number of features. For the scRNA-seq datasets, the DR methods incorporating the bias correction match or outperform softImpute and ice, respectively. In particular, the visualisations attained by the bias correction on the Usoskin dataset are much more clear than those achieved by softImpute and ice (Figure S15, Figure S24 and Figure S33 of the supplementary material).

To sum up, the superiority shown by the bias-corrected variants suggests that the proposed bias correction is beneficial for displaying better separation of clusters in the presence of missing observations.

4.5 BIAS CORRECTION ENHANCES CLUSTERING

In this subsection, we investigate how the proposed bias correction impacts on the clustering applications. To this end, we first apply different DR methods and their variants to the dataset to extract the low-dimensional points, which are then grouped using the k-means clustering algorithm. The ARI is then calculated as a measure of clustering performance. For the real datasets, the dimension of latent points extracted from PCA and BC-PCA is chosen in terms of the Cattell–Nelson–Gorsuch scree test [Gorsuch and Nelson, 1981], while only two-dimensional projections are produced with the other dimension reduction methods. Note that the dimension determined by the scree test is usually 2 in our experiments.

First, we assess the clustering performance obtained from using all features on the simulated data. On the simulated data without missing observations, the inferred labels obtained from PCA match perfectly with the ground truth labels in terms of ARI (Figure 3). On the simulated data with missing observations, the original PCA is unable to provide distinct clusters in the low-dimensional space (Figure 2), and hence hinders the clustering (Figure 3). On the contrary, their bias-corrected PCA presents nearly perfect ARI values, suggesting that the bias-correction significantly improves the clustering accuracy in such a case.

Next, we compare benchmark DR methods with their variants integrating the bias-correction in terms of the clustering performance on the real datasets, as presented in Figure 6, Figure S35 and Figure S36 of the supplementary material. Consistently with the visualisations, the k-means clustering performance of BC-tSNE and BC-UMAP is better than that of tSNE and UMAP on almost all datasets except the Deng dataset and the wine dataset. Similarly, the cluster labels obtained by BC-GPLVM show a much higher agreement with the ground truth labels than GPLVM on the Treutlein dataset, the Usoskin dataset, the Koh dataset, and the Olivetti faces dataset. For the other datasets, BC-GPLVM accomplishes the ARI values comparable to those of GPLVM. BC-PCA surpasses or is comparable to PCA on all the datasets except the Koh dataset which can be due to the nonlinearity possessed by the datasets and the wine dataset which is not suited for being handled by the proposed estimator, as we discussed in sec.4.4.

Last, the bias-corrected estimators are compared with the imputation methods and the approaches handling scRNA-seq data. It is clear that the k-means results obtained by integrating CIDR with different DR methods are inferior to those attained by the bias correction, according to ARI. Bias-corrected DR methods outperform DR approaches integrated with scImpute on the Usoskin, Treutlein and Pollen datasets. Moreover, all the values of ARI achieved by the proposed bias correction are nearly 1 while the values of ARI of sc-tSNE and sc-UMAP are much lower than 1 on the Kumar dataset. Although BC-tSNE and BC-UMAP perform slightly worse than sc-tSNE and sc-UMAP on the Koh dataset, BC-GPLVM achieves much higher ARI value. On the Deng dataset, scImpute achieves higher ARI values when combined with the benchmark methods in comparison with the bias-corrected versions and the original ones. Generally speaking, the bias-corrected DR methods is better than scImpute on most datasets. When applied to the fashion MNIST dataset and the Olivetti dataset, the methods based on the bias correction often yield higher ARI compared to those integrating ice (Figure S36 of the supplementary material), while their clustering performances are slightly worse than those attained by the methods inputting the data matrix imputed by softImpute. For the scRNA-seq datasets, the bias-corrected approaches outperforms softImpute on the Usoskin dataset, the Pollen dataset, and the Deng dataset. For the other scRNA-seq datasets, the bias-corrected variants accomplishes the ARI values comparable to or slightly worse than those obtained by softImpute.

Overall, the clustering results indicate that the bias correction is able to infer the cluster labels that are more consistent with the ground truth and improve the performance of clustering following dimension reduction.

5 CONCLUSION

This paper proposes an unbiased estimator of the covariance matrix in the presence of missing data. The proposed bias-corrected Gram matrix is able to substantially improve the performance of various DR methods. As shown by the theoretical results in this paper, the Gram matrix is a biased estimator in the presence of missing observations and could be adverse for DR, while the proposed unbiased estimator can correct the bias introduced to the Gram matrix by the missingness. Moreover, the bounds of variances ensure the accurate estimation of the ground-truth covariance matrix in the low-dimensional space as long as the input dimension is high enough, and hence guarantees the reliable representation of the high-dimensional data. The experimental results on both simulated and real datasets demonstrate that the proposed unbiased estimator is widely applicable and is able to effectively enhance the performance of both the distance-matrix-based and Gram-matrix-based DR methods.

Acknowledgements

The authors thank the anonymous reviewers for helpful discussions and suggestions.

References

Emmanuel J. Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

Juhee Cho, Donggyu Kim, and Karl Rohe. Asymptotic theory for estimating the singular vectors and values of a partially-observed low rank matrix with noise. *Statistica Sinica*, 27(4):1921–1948, 2017.

Qiaolin Deng, Daniel Ramsköld, Björn Reinius, and Rickard Sandberg. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, 343(6167):193–196, 2014.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

RL Gorsuch and J Nelson. CNG scree test: an objective procedure for determining the number of factors. In *Annual Meeting of the Society for Multivariate Experimental Psychology*, 1981.

Trevor Hastie, Rahul Mazumder, Jason D. Lee, and Reza Zadeh. Matrix completion and low-rank svd via fast alternating least squares. *Journal of Machine Learning Research*, 16(104):3367–3402, 2015.

Stephanie C Hicks, F William Townes, Mingxiang Teng, and Rafael A Irizarry. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics*, 19(4):562–578, 2017.

Aleksandra A. Kolodziejczyk, Jong Kyoung Kim, Valentine Svensson, John C. Marioni, and Sarah A. Teichmann. The technology and biology of single-cell RNA sequencing. *Molecular Cell*, 58(4):610 – 620, 2015. ISSN 1097-2765.

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

Roshan M. Kumar, Patrick Cahan, Alex K. Shalek, Rahul Satija, A. Jay DaleyKeyser, Hu Li, Jin Zhang, Keith Pardee, David Gennert, John J. Trombetta, Thomas C. Ferrante, Aviv Regev, George Q. Daley, and James J. Collins. Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature*, 516(7529):56–61, 2014.

Neil Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of Machine Learning Research*, 6(Nov): 1783–1816, 2005.

Erich Leo Lehmann. *Elements of large-sample theory*. Springer Science & Business Media, 2004.

Wei Vivian Li and Jingyi Jessica Li. An accurate and robust imputation method scimpute for single-cell RNA-seq data. *Nature Communications*, 9(1):997, 2018.

Peijie Lin, Michael Troup, and Joshua W. K. Ho. CIDR: Ultrafast and accurate clustering through imputation for single-cell rna-seq data. *Genome Biology*, 18(1):59, 2017.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2019.

Kyle M. Loh, Angela Chen, Pang Wei Koh, Tianda Z. Deng, Rahul Sinha, Jonathan M. Tsai, Amira A. Barkal, Kimberle Y. Shen, Rajan Jain, Rachel M. Morganti, Ng Shyh-Chang, Nathaniel B. Fernhoff, Benson M. George, Gerlinde Wernig, Rachel E.A. Salomon, Zhenghao Chen, Hannes Vogel, Jonathan A. Epstein, Anshul Kundaje, William S. Talbot, Philip A. Beachy, Lay Teng Ang, and Irving L. Weissman. Mapping the pairwise choices leading from pluripotency to human bone, heart, and other mesoderm cell types. *Cell*, 166(2):451 – 467, 2016.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Benjamin M. Marlin, Richard S. Zemel, Sam Roweis, and Malcolm Slaney. Collaborative filtering and the missing at random assumption. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, page 267–275, Arlington, Virginia, USA, 2007. AUAI Press.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Karl Pearson. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

Emma Pierson and Christopher Yau. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology*, 16(1):241, 2015.

Alex A Pollen, Tomasz J Nowakowski, Joe Shuga, Xiaohui Wang, Anne A Leyrat, Jan H Lui, Nianzhen Li, Lukasz Szpankowski, Brian Fowler, Peilin Chen, et al. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nature Biotechnology*, 32(10): 1053, 2014.

Antoine-Emmanuel Saliba, Alexander J. Westermann, Stanislaw A. Gorski, and Jörg Vogel. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Research*, 42(14):8845–8860, 2014.

F.S. Samaria and A.C. Harter. Parameterisation of a stochastic model for human face identification. In *Proceedings of 1994 IEEE Workshop on Applications of Computer Vision*, pages 138–142, 1994.

Ehud Shapiro, Tamir Biezuner, and Sten Linnarsson. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics*, 14(9): 618, 2013.

Huanfeng Shen, Xinghua Li, Qing Cheng, Chao Zeng, Gang Yang, Huifang Li, and Liangpei Zhang. Missing information reconstruction of remote sensing data: A technical review. *IEEE Geoscience and Remote Sensing Magazine*, 3(3):61–85, 2015.

Michael E. Tipping and Christopher M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61 (3):611–622, 1999. ISSN 13697412, 14679868.

Warren S Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419, 1952.

Barbara Treutlein, Qian Yi Lee, J. Gray Camp, Moritz Mall, Winston Koh, Seyed Ali Mohammad Shariati, Sopheak Sim, Norma F. Neff, Jan M. Skotheim, Marius Wernig, and Stephen R. Quake. Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq. *Nature*, 534(7607):391–395, 2016.

Dmitry Usoskin, Alessandro Furlan, Saiful Islam, Hind Abdo, Peter Lönnerberg, Daohua Lou, Jens Hjerling-Leffler, Jesper Haeggström, Olga Kharchenko, Peter V Kharchenko, et al. Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nature Neuroscience*, 18(1):145, 2015.

Stef Van Buuren. *Flexible imputation of missing data*. CRC press, 2018.

L. Van Der Maaten et al. Dimensionality reduction: a comparative review. *Technical report, Tilburg University, TiCC-TR 2009-005*, 2009.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017.

Ziwei Zhu, Tengyao Wang, and Richard J Samworth. High-dimensional principal component analysis with heterogeneous missingness. *arXiv preprint arXiv:1906.12125*, 2019.