

Nonlinear Dimensionality Reduction with Missing Data using Parametric Multiple Imputations

Cyril de Bodt, Dounia Mulders, Michel Verleysen, *Fellow, IEEE*, and John A. Lee

Abstract—Dimensionality reduction (DR) aims at faithfully and meaningfully representing high-dimensional data into a low-dimensional (LD) space. Recently developed neighbor embedding DR methods lead to outstanding performances, thanks to their ability to foil the curse of dimensionality. Unfortunately, they can not be directly employed on incomplete data sets, which become ubiquitous in machine learning. Discarding samples with missing features prevents their LD coordinates computation and deteriorates the complete samples treatment. Common missing data imputation schemes are not appropriate in the nonlinear DR context either. Indeed, even if they model the data distribution in the feature space, they can at best enable the application of a DR scheme on the expected data set. In practice, one would instead like to obtain the LD embedding with the closest cost function value on average with respect to the complete data case. As state-of-the-art DR techniques are nonlinear, the latter embedding results from minimizing the expected cost function on the incomplete database, not from considering the expected data set. This paper addresses these limitations by developing a general methodology for nonlinear DR with missing data, being directly applicable with any DR scheme optimizing some criterion. In order to model the feature dependencies, a high-dimensional extension of Gaussian mixture models is first fitted on the incomplete data set. It is afterward employed under the multiple imputation paradigm to obtain a single relevant LD embedding, minimizing the cost function expectation. Extensive experiments demonstrate the superiority of the suggested framework over alternative approaches.

Index Terms—Dimensionality reduction, missing data, neighbor embedding, Gaussian mixtures, multiple imputations.

I. INTRODUCTION

DIMENSIONALITY reduction (DR) aims at computing relevant low-dimensional (LD) representations of high-dimensional (HD) data sets, mainly for exploratory visualization or to foil the curse of dimensionality [1], [2]. The relevance of an embedding typically relies on the HD neighborhood preservation. Different paradigms have emerged in order to formalize this neighborhood preservation principle in terms of mapping from HD to LD coordinates [1], like the reproduction of distances [3] or neighborhoods [4], [5]. Early

Computational resources have been provided by the supercomputing facilities of the Université catholique de Louvain (CISM/UCL) and the Consortium des Équipements de Calcul Intensif en Fédération Wallonie Bruxelles (CÉCI) funded by the Fond de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under convention 2.5020.11. C. de Bodt and D. Mulders are Research Fellows of the FNRS. J. A. Lee is a Senior Research Associate with the FNRS.

C. de Bodt, D. Mulders, M. Verleysen and J. A. Lee are with the Institute for Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM), Electrical Engineering department (ELEN), Université catholique de Louvain, 1348 Louvain-la-Neuve, Belgium (e-mail: {cyril.debodt; dounia.mulders; michel.verleysen; john.lee}@uclouvain.be).

Please cite as: C. de Bodt, D. Mulders, M. Verleysen, and J. A. Lee, "Nonlinear Dimensionality Reduction with Missing Data using Parametric Multiple Imputations," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 4, pp. 1166-1179, Apr. 2019, DOI: 10.1109/TNNLS.2018.2861891.

Article abstract in IEEE Xplore: <https://ieeexplore.ieee.org/abstract/document/8447227>.

2162-237X © 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.

See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

linear models such as principal component analysis (PCA) [6] and classical metric multidimensional scaling (MDS) [3] project the HD vectors by maximizing the variance or dot product preservation. These approaches being limited to linear mappings, nonlinear metric MDS variants [7] refine them by defining (weighted) distance preservation schemes, based on either Euclidean or approximated geodesic measures [8]–[10]. Other studies define affinity matrices driving the computation of the LD embedding [11]–[15]. In the visualization context however, these methods hardly outperform the older techniques [1], [16], [17], possibly because they reduce to applying classical MDS in an unknown feature space [18], [19]. The norm concentration phenomenon indeed particularly affects distance preserving schemes [20]. In contrast, genuine neighbor embedding methods such as Stochastic Neighbor Embedding (SNE) [5] and variants [17], [21] mitigate this phenomenon by defining probabilities [22], leading to outstanding DR performances. These achievements motivated the development of SNE-based techniques with alternative divergences as cost functions [21], [23]–[25], heavy-tailed LD similarities such as in *t*-SNE [17], [26], [27], fast optimization methods [28]–[34], lower-complexity approximations [35]–[38] and multi-scale, meta-parameter-free approaches [39], [40].

On the other hand, incomplete databases become nowadays more and more frequently encountered in machine learning [41]–[44]. While missing values occur for many different reasons depending on the data type, any inference task on such incomplete data set should take the missing entries into account in order to avoid biased conclusions. Discarding the samples with missing features is most often not acceptable as it results in considerable information losses and drastically reduces the data set size as the proportion of missing values increases. Even more importantly in our context, removing these examples would prevent the computation of their LD representations, which greatly alters the SNE-based probabilities of the complete samples and hence their LD positioning.

Quite surprisingly, despite all the recent advances in DR, the field has not yet designed methods directly able to deal with incomplete data sets. In this context, the cost functions optimized by nonlinear DR schemes can not be directly evaluated as they depend on the incomplete HD samples coordinates. Although it has strong limitations for reasons detailed below, a workaround is to consider a complete HD data set instead, resulting from some imputation. The missing data management is hence considered as a preprocessing step. The imputation may attempt to be consistent with the data distribution in

the feature space [45]. As the features of most data sets are not independent, the conditional distribution of the missing features of the samples given their observed features should indeed be considered in order to derive meaningful insights on the data relationships. This can be accounted by replacing the missing data with their conditional mean given the observed values, provided that the data distribution can be modeled [46]. Nearest neighbor-based imputation is another possibility [47]–[49]. In all these cases however, the DR method would at best yield the LD embedding minimizing the cost function evaluated using the expected HD data set. Denoting by Ξ a set of incomplete HD data points and by $\mathbf{X}$ the set of their LD representations, the LD embedding would be

$$\arg \min_{\mathbf{X}} C(\mathbf{X}, \mathbb{E}[\Xi]), \quad (1)$$

where C is the cost function of the DR algorithm. In practice however, one is rather interested in obtaining the LD embedding minimizing the cost function expectation, yielding the closest cost function value on average to the one that would be obtained in the complete data case:

$$\arg \min_{\mathbf{X}} \mathbb{E}[C(\mathbf{X}, \Xi)]. \quad (2)$$

This strongly differs from (1) when C is nonlinear in Ξ as in most recent, state-of-the-art, DR methods which are based on nonlinear transforms of the HD samples, such as SNE similarities, etc. Rather than (2), one could suggest to seek for

$$\mathbb{E}[\arg \min_{\mathbf{X}} C(\mathbf{X}, \Xi)]. \quad (3)$$

This would however merge LD embeddings specifically designed for different complete HD data sets. As $\mathbf{X}$ is subject to rotations, translations and scaling factors, the obtained LD embedding would not reflect any kind of expected relationship in Ξ . More interestingly, one could create an imputed HD data set for each mode of the conditional distribution of the missing values given the observed data and compute their corresponding LD embeddings. Their number would nevertheless rapidly increase with the proportion of missing values. This would result in a huge amount of DR results to analyze and compare for the user, which greatly complicates its exploratory task. This paper hence focuses on recovering a single relevant LD embedding.

Solving (2) requires reliably approximating the cost function expectation on the incomplete database. Whereas considering a single imputed data set would yield a biased estimate for nonlinear cost functions, a consistent approximation can be obtained under the multiple imputation paradigm [50], [51], as long as the conditional distribution of the missing features given the observed ones can be modeled [52]. Gaussian mixture models are particularly attractive for this purpose, as they can approximate arbitrarily well almost any continuous density [53], [54]. Their parameters are usually derived through maximum likelihood estimation [55], which can be easily adapted when facing missing values [56], [57]. The fitting of Gaussian mixtures being notoriously difficult in very HD spaces, efficient HD extensions exist [58], considerably reducing their number of parameters. Further improvements [59], [60] also allow for the application of these models in

several fields [61]–[63]. In particular, they are successfully used for distance estimation in numerical data sets with missing values [64], [65] and adapted extreme learning machines on incomplete data sets [46], [66].

This work proposes to process missing data in DR under the multiple imputation paradigm. In particular, a high-dimensional extension of Gaussian mixture models is first fitted on the incomplete database. Several imputed versions of the data set are then created by sampling the conditional distribution of the missing features of the samples given their observed features. Averaging the nonlinear DR cost function evaluations on each complete data set finally approximates its expectation on the incomplete database, enabling to optimize (2). This provides a meaningful LD representation of each data point, no matter whether it contains missing values or not. Although this paper considers multi-scale SNE [39], the proposed methodology is directly applicable with any other neighbor embedding DR method, and more generally with any DR scheme optimizing some cost function.

This paper is structured as follows: Section II reviews Gaussian mixture models, describes their missing data extension as well as a high-dimensional variant. This section is strongly inspired from [46], [64]–[66]. Section III briefly introduces neighbor embedding DR by summarizing SNE and presenting the multi-scale SNE method. The complete multiple imputation methodology is derived in Section IV. Its application using multi-scale SNE and other neighbor embedding DR schemes is detailed. Section V is dedicated to experimental results comparing different methods dealing with missing data on several data sets in which missing values are introduced. Some conclusions are finally drawn in Section VI.

II. GAUSSIAN MIXTURE MODELS FOR HIGH-DIMENSIONAL AND INCOMPLETE DATA

A G -component Gaussian mixture model of a complete data set $\Xi_c := \{\xi_i\}_{i=1}^N$ with N independent examples $\xi_i \in \mathbb{R}^H$ consists in G multivariate Gaussians with mean vectors $\mu_1, \dots, \mu_G \in \mathbb{R}^H$, covariance matrices $\Sigma_1, \dots, \Sigma_G \in \mathbb{R}^{H \times H}$ and positive, unit-sum weight coefficients $w_1, \dots, w_G \in \mathbb{R}$, which maximize the model log-likelihood

$$\log p(\Xi_c | \theta) = \sum_{i \in \mathcal{I}} \log \left(\sum_{k \in \mathcal{G}} w_k \mathcal{N}(\xi_i | \mu_k, \Sigma_k) \right), \quad (4)$$

where $\mathcal{I} := \{1, \dots, N\}$, $\mathcal{G} := \{1, \dots, G\}$, $\theta := \{w_k, \mu_k, \Sigma_k; k \in \mathcal{G}\}$ and where $\mathcal{N}(\xi_i | \mu_k, \Sigma_k)$ denotes the evaluation of the probability density function of the k^{th} Gaussian over ξ_i [53], [55]. A traditional Expectation-Maximization (EM) algorithm allows maximizing (4). It is detailed for the sake of introducing relevant notations. After initializing the mean vectors, covariance matrices and weight coefficients, the E-step estimates the probabilities for each example to originate from each Gaussian given the current value of θ : for each $i \in \mathcal{I}$ and $k \in \mathcal{G}$,

$$t_{ik} = \frac{w_k \mathcal{N}(\xi_i | \mu_k, \Sigma_k)}{\sum_{j \in \mathcal{G}} w_j \mathcal{N}(\xi_i | \mu_j, \Sigma_j)}, \quad N_k = \sum_{i \in \mathcal{I}} t_{ik}. \quad (5)$$

The M-step then updates θ given these probabilities:

$$\mu_k = \frac{1}{N_k} \sum_{i \in \mathcal{I}} t_{ik} \xi_i, \quad w_k = \frac{N_k}{N}, \quad (6)$$

$$\Sigma_k = \frac{1}{N_k} \sum_{i \in \mathcal{I}} t_{ik} (\xi_i - \mu_k) (\xi_i - \mu_k)^T. \quad (7)$$

These two steps are iterated until convergence of (4) [65], [66].

A. Missing Data Management

Different studies extended the Gaussian mixture model to manage missing data [56], [57]. Missing values refer to well-defined data entries, unavailable for an unknown reason [51]. Let $\Xi = \{\xi_i, O_i\}_{i=1}^N$ denote an incomplete data set with N independent samples $\xi_i \in \mathbb{R}^H$, $O_i \subset \{1, \dots, H\}$ indexing the observed features of ξ_i . The complement of O_i is noted M_i , and $\xi_i^{O_i}$ (resp. $\xi_i^{M_i}$) denotes the values of the observed (resp. missing) features of ξ_i . The set Ξ^O gathers $\xi_i^{O_i}$ for all $i \in \mathcal{I}$.

As to the missing-data mechanism, this work considers “missing at random” data [51]: given Ξ^O , the probability of a data entry to be missing is independent from $\xi_i^{M_i}$ for all $i \in \mathcal{I}$. This missing-data mechanism can be shown to be ignorable [56], meaning that maximum likelihood estimators remain consistent when evaluated on the incomplete database [51]. The stronger “missing completely at random” assumption, requiring the missingness probability of a data entry to also be independent of Ξ^O , is thus not necessary. A Gaussian mixture model for Ξ then maximizes the observed log-likelihood

$$\log p(\Xi^O | \theta) = \sum_{i \in \mathcal{I}} \log \left(\sum_{k \in \mathcal{G}} w_k \mathcal{N}(\xi_i^{O_i} | \mu_k, \Sigma_k) \right) \quad (8)$$

over its parameters θ , where $\mathcal{N}(\xi_i^{O_i} | \mu_k, \Sigma_k)$ refers to the marginal of $\mathcal{N}(\xi_i | \mu_k, \Sigma_k)$ over the observed features of ξ_i . The EM-algorithm described above can be adapted to maximize (8), by further computing the conditional expectations $\tilde{\mu}_{k, M_i | O_i} \in \mathbb{R}^{|M_i|}$ and covariances $\tilde{\Sigma}_{k, M_i | O_i} \in \mathbb{R}^{|M_i| \times |M_i|}$ of the missing features of each sample ξ_i given $\xi_i^{O_i}$ under each Gaussian component $k \in \mathcal{G}$ during the E-step. For all $k \in \mathcal{G}$, $\tilde{\mu}_{k, M_i | O_i}$ and $\tilde{\Sigma}_{k, M_i | O_i}$ are easily derived as the conditional distribution of $\xi_i^{M_i}$ given $\xi_i^{O_i}$ remains a multivariate Gaussian when ξ_i is considered as originating from a Gaussian component [53]. The E-step then develops, for $k \in \mathcal{G}$ and $i \in \mathcal{I}$, as

$$t_{ik} = \frac{w_k \mathcal{N}(\xi_i^{O_i} | \mu_k, \Sigma_k)}{\sum_{j \in \mathcal{G}} w_j \mathcal{N}(\xi_i^{O_i} | \mu_j, \Sigma_j)}, \quad N_k = \sum_{i \in \mathcal{I}} t_{ik}, \quad (9)$$

$$\tilde{\mu}_{k, M_i | O_i} = \mu_k^{M_i} + \Sigma_k^{M O_i} \left(\Sigma_k^{O O_i} \right)^{-1} (\xi_i^{O_i} - \mu_k^{O_i}), \quad (10)$$

$$\tilde{\Sigma}_{k, M_i | O_i} = \Sigma_k^{M M_i} - \Sigma_k^{M O_i} \left(\Sigma_k^{O O_i} \right)^{-1} \Sigma_k^{O M_i}, \quad (11)$$

where $\mu_k^{M_i}$ and $\mu_k^{O_i}$ denote the components of μ_k indexed by M_i and O_i and where $\Sigma_k^{AB_i}$ refers to the sub-matrix of Σ_k with rows indexed by A_i and columns by B_i . Defining

$\tilde{\xi}_{ik} \in \mathbb{R}^H$ with its components indexed by O_i and M_i equal to $\xi_i^{O_i}$ and $\tilde{\mu}_{k, M_i | O_i}$, the M-step then computes, for $k \in \mathcal{G}$,

$$\mu_k = \frac{1}{N_k} \sum_{i \in \mathcal{I}} t_{ik} \tilde{\xi}_{ik}, \quad w_k = \frac{N_k}{N}, \quad (12)$$

$$\Sigma_k = \frac{1}{N_k} \sum_{i \in \mathcal{I}} t_{ik} (\tilde{\xi}_{ik} - \mu_k) (\tilde{\xi}_{ik} - \mu_k)^T + \tilde{\Sigma}_{ik}, \quad (13)$$

where $\tilde{\Sigma}_{ik} \in \mathbb{R}^{H \times H}$ is defined as $\tilde{\Sigma}_{ik}^{M M_i} = t_{ik} \tilde{\Sigma}_{k, M_i | O_i}$ and zero elsewhere, $\tilde{\Sigma}_{ik}^{M M_i}$ denoting the sub-matrix of $\tilde{\Sigma}_{ik}$ with rows and columns indexed by M_i .

One can note that the E-step requires inverting a different matrix for each component and set of observed features, which may scale as $\mathcal{O}(GN)$ in HD data sets with many missing entries. Such data sets are indeed more likely to present a different pattern of observed features for each sample. As performing $\mathcal{O}(GN)$ matrix inversions at each E-step has a significant cost, some studies suggest to employ the sweep inversion operator [51], [65], which is used for the experiments in this paper. The sweep operator definition along as its application to compute the E-step is described in the appendix.

B. High-dimensional Extension

As the covariance matrices dimension is $H \times H$, the number of parameters of Gaussian mixture models scales in $\mathcal{O}(GH^2)$. This makes the fitting of such models difficult in HD spaces, requiring much more samples than available. However adopting LD representations of each Gaussian function circumvents this problem, leading to the family of high-dimensional data clustering (HDDC) methods [58]. The latter constrain the structure of the covariance matrices computed in the M-step to reduce the number of model parameters. This work considers the most general HDDC model. For each component $k \in \mathcal{G}$, the eigenvalue decomposition (EVD) of Σ_k is first computed after each M-step: $\Sigma_k = \mathbf{Q}_k \Lambda_k \mathbf{Q}_k^T$, where the eigenvalues are sorted in decreasing order on the diagonal of Λ_k . A new $H \times H$ diagonal matrix $\bar{\Lambda}_k$ is then defined as, for $j \in \{1, \dots, H\}$,

$$(\bar{\Lambda}_k)_{jj} = \begin{cases} (\Lambda_k)_{jj} & \text{if } j \leq d_k, \\ \frac{1}{H-d_k} \sum_{l=d_k+1}^H (\Lambda_k)_{ll} & \text{otherwise,} \end{cases} \quad (14)$$

for some $d_k \in \{1, \dots, H\}$. Covariance matrix Σ_k defined in (13) is finally replaced with $\tilde{\Sigma}_k = \mathbf{Q}_k \bar{\Lambda}_k \mathbf{Q}_k^T$. Since the variance along an eigenvector is equal to its associated eigenvalue, the variance in the subspace spanned by the eigenvectors of the $H - d_k$ last eigenvalues of Σ_k is modeled using a single parameter [58]. Hyper-parameter d_k is determined with the scree test [67]: d_k is set to the smallest integer between 1 and H such that the subsequent successive eigenvalues differences are smaller than a threshold $\eta_k \geq 0$, for which a heuristic derivation is detailed in Section II-D. The eigenvalues are hence replaced with their mean as soon as they become similar to each other. This ensures that $\bar{\Lambda}_k$ remains close to Λ_k , suggesting that the HDDC approximation is reasonable. This motivates the use of the scree test for HDDC models [58].

C. EM Algorithm Time Complexity

A single EM iteration includes the computation of the E- and M-steps, as well as the application of the HDDC method. The latter entails the derivation of the EVD of the covariance matrices, from which their inverses are deduced. This scales in $\mathcal{O}(GH^3)$ time. About the M-step, updating the covariance matrices costs $\mathcal{O}(GNH^2)$. As to the E-step, evaluating t_{ik} for a complete sample scales in $\mathcal{O}(H^2)$. When ξ_i contains missing values, the sweep inversion operator allows obtaining t_{ik} , $\tilde{\mu}_{k,M_i|O_i}$ and $\tilde{\Sigma}_{k,M_i|O_i}$ in $\mathcal{O}(|O_i|^2|M_i| + |O_i||M_i|^2 + |M_i|^2)$ time. As $|O_i| = H - |M_i|$, performing the E-step for all samples and Gaussians hence scales in $\mathcal{O}\left(GH \sum_{i \in \mathcal{I}} H \max(1, |M_i|) + |M_i|^2\right)$. The time complexity of an EM iteration then expresses as $T_{MG} := \mathcal{O}\left(GH \left(H^2 + NH + H \left(\sum_{i \in \mathcal{I}} |M_i|\right) + \sum_{i \in \mathcal{I}} |M_i|^2\right)\right)$.

D. Metaparameter Tuning

Thresholds η_k as well as the number G of components are hyper-parameters which can be determined using the Bayesian Information Criterion (BIC) [66]. If the number P_θ of parameters of the model characterized by θ is smaller than N , the BIC can be expressed as $-2\log \mathcal{L}(\theta) + P_\theta \log N$ [68]. Otherwise, it is not defined. The numbers P_θ of parameters of the HDDC models are expressed in [58] as a function of H , G and d_k for all $k \in \mathcal{G}$. The BIC is preferred to Akaike's Information Criterion (AIC) [69] as the latter typically selects too complex models [70]. As mentioned in previous studies [65], [66], it is crucial to avoid considering overly complex models, as P_θ grows quickly with G and may rapidly become larger than N when dealing with HD data [65]. This would yield Gaussian mixture models particularly subject to overfitting of the observed data, affecting the quality of the multiple imputations. This motivates employing the BIC criterion [66], which has moreover been shown to be efficient in the context of HDDC models [58]. A simple heuristic then allows determining G [66]: first fit two models with one and two components, then with three if the BIC decreases from $G = 1$ to $G = 2$, and so on as long as the BIC decreases. However, for each considered value of G , the thresholds η_k should still be determined. Adopting a similar strategy as for G , η_k is defined as a decreasing fraction of the covariance trace in this paper: $\eta_k = \zeta \cdot \text{tr}(\Sigma_k)$, where ζ is first set to 0.5 and then repeatedly divided by two until the BIC stops decreasing. Note that ζ is common to all Gaussian components and that when it becomes small, d_k will tend towards H . HDDC applied to low- or average-dimensional data is hence able to consider full covariances, when N is large enough.

E. EM Initialization

The application of the EM algorithm entails the initialization of the Gaussian means, covariances and weights. This work considers uniform weights w_k at the beginning of the algorithm, while the means μ_k are randomly selected among the complete samples. If there are less than G such samples, some incomplete ones are drawn and sample mean imputation

is used to fill their missing features [65]. The covariances Σ_k are all initialized to the sample covariance of the complete examples (complete-case analysis) when there are more than $H + 1$ complete samples, and to a diagonal matrix with the sample variance of the features otherwise [65]. This ensures the positive definiteness of Σ_k , which is not the case when computing the covariances separately for all pairs of features using the available examples [64] (available-case analysis).

The EM algorithm being sensitive to the parameter initialization, several runs should be conducted with different initializations to explore several local maxima. In this paper, for each considered values of G and ζ , the algorithm is run 10 times with distinct initializations and the one yielding the smallest BIC is kept. Each run of the algorithm stops as soon as one of the following conditions is met: (i) the log-likelihood difference between two EM iterations is smaller than 10^{-5} , (ii) 400 iterations have been performed, or (iii) one of the covariance matrices has only close-to zero eigenvalues.

F. Conditional Mixture Model Sampling

Multiple imputation schemes require to sample the conditional distribution of the missing features of each example ξ_i given its observed features. As the data set is modeled using a Gaussian mixture model, this conditional distribution also follows a Gaussian mixture with G components and with means $\tilde{\mu}_{k,M_i|O_i}$, covariances $\tilde{\Sigma}_{k,M_i|O_i}$ and weights t_{ik} [66]. The conditional mean imputation of $\xi_i^{M_i}$ then consists in replacing it with $\sum_{k \in \mathcal{G}} t_{ik} \tilde{\mu}_{k,M_i|O_i}$. On the other hand, a random sample from the conditional distribution can be obtained by first drawing a component index k from $\mathcal{G}$ with a t_{ik} probability, then sampling a vector y_{ik} from a $|M_i|$ -dimensional Gaussian distribution, with zero mean and identity covariance, and finally computing $\tilde{\mu}_{k,M_i|O_i} + C_{ik}y_{ik}$, where C_{ik} is the Cholesky factor of $\tilde{\Sigma}_{k,M_i|O_i}$ [66].

III. NEIGHBOR EMBEDDING

DR aims to compute LD representations of HD data sets which best preserve their topologies. This means that the data points LD neighborhoods should reflect the HD ones. Several paradigms tackled this problem, from the rigid distance preservation methods to the successful neighbor embedding approaches [1]. Among these, Stochastic Neighbor Embedding (SNE) first defined probabilities of being neighbors [5], while more recent multi-scale approaches yield impressive results [39]. SNE is first reviewed at the beginning of this section, and its multi-scale version is presented afterward.

A. SNE

Let again $\Xi_c = [\xi_i]_{i=1}^N$ denote a set of N (complete) points in a H -dimensional metric space (HDS). Let $\mathbf{X} = [\mathbf{x}_i]_{i=1}^N$ be its representation in a P -dimensional metric space, called the LD space (LDS), with $P \leq H$. The HD and LD distances between the i^{th} and j^{th} points are noted δ_{ij} and d_{ij} . Starting

from them, SNE defines single-scale HD and LD similarities between pairs of data points, for $i \in \mathcal{I}$ and $j \in \mathcal{I} \setminus \{i\}$:

$$\sigma_{ij} = \frac{\exp(-\pi_i \delta_{ij}^2/2)}{\sum_{k \in \mathcal{I} \setminus \{i\}} \exp(-\pi_i \delta_{ik}^2/2)}, \quad \sigma_{ii} = 0, \quad (15)$$

$$s_{ij} = \frac{\exp(-p_i d_{ij}^2/2)}{\sum_{k \in \mathcal{I} \setminus \{i\}} \exp(-p_i d_{ik}^2/2)}, \quad s_{ii} = 0. \quad (16)$$

Precision π_i is set by binary search to reach a user-defined perplexity K_* for the distribution $\sigma_i = \{\sigma_{ij}, j \in \mathcal{I} \setminus \{i\}\}$:

$$\pi_i \text{ s.t. } \log K_* = - \sum_{j \in \mathcal{I} \setminus \{i\}} \sigma_{ij} \log \sigma_{ij}, \quad (17)$$

whereas the LD precisions p_i are set to 1. The perplexity measures the soft Gaussian neighborhood size. In order to find $\mathbf{X}$, SNE minimizes the sum of the Kullback-Leibler (KL) divergences between the distributions σ_i and $s_i := \{s_{ij}, j \in \mathcal{I} \setminus \{i\}\}$ for each $i \in \mathcal{I}$ using gradient-based optimization algorithms [5]:

$$C_{SNE} = \sum_{i \in \mathcal{I}, j \in \mathcal{I} \setminus \{i\}} \sigma_{ij} \log(\sigma_{ij}/s_{ij}). \quad (18)$$

B. Multi-scale SNE

As K_* is typically small, SNE is based on local transformations of the HD distances. It thus focuses on the preservation of small neighborhoods, most HD similarities being indeed close to zero which annihilates the influence of the associated LD similarities on the SNE cost function. Relying on local distance transformations may hence be questioned as it prevents considering the average- and large-scale interactions while defining the global structure of the embedding, K_* being most of the time arbitrary. This considerably affects the preservation of long-range neighborhoods. On the other hand, employing too large a perplexity leads to nearly-uniform HD similarities, impairing the preservation of small neighborhoods.

However, combining several SNE similarities defined using increasing perplexities enables to capture both local and global relationships and to focus on various data scales at the same time. This reasoning leads to the multi-scale SNE method [39]. By introducing an additional scale index h , the single-scale similarities are now defined, for $i \in \mathcal{I}$ and $j \in \mathcal{I} \setminus \{i\}$, as

$$\sigma_{hij} = \frac{\exp(-\pi_{hi} \delta_{ij}^2/2)}{\sum_{k \in \mathcal{I} \setminus \{i\}} \exp(-\pi_{hi} \delta_{ik}^2/2)}, \quad \sigma_{hii} = 0, \quad (19)$$

$$s_{hij} = \frac{\exp(-p_{hi} d_{ij}^2/2)}{\sum_{k \in \mathcal{I} \setminus \{i\}} \exp(-p_{hi} d_{ik}^2/2)}, \quad s_{hii} = 0. \quad (20)$$

Multi-scale (ms) similarities are then expressed as

$$\sigma_{ij,ms} = \frac{1}{L} \sum_{h=1}^L \sigma_{hij}, \quad s_{ij,ms} = \frac{1}{L} \sum_{h=1}^L s_{hij}, \quad (21)$$

where L is the considered number of scales. In order to determine the precisions π_{hi} and p_{hi} for the neighborhoods around each ξ_i and $\mathbf{x}_i$, perplexity values covering the whole range from 1 to N can be selected, without any user intervention. As suggested in [39], starting from a small perplexity like $K_* = 2$, indexed perplexities can be defined as $K_{h*} = 2^{h-1} K_*$,

where $1 \leq h \leq L = \lfloor \log_2(N/K_*) \rfloor$, with $\lfloor \cdot \rfloor$ representing the rounding operator. Precision π_{hi} is then computed as in SNE, using perplexity K_{h*} . On the other hand, the LD precisions should be determined according to the exponential evolution of the perplexities [39]. One possibility is $p_{hi} = p_h = K_{h*}^{-2/P}$. Further details for the choice of K_{h*} can be found in [39].

The cost function of multi-scale SNE remains as in SNE. Since the entropy of the HD similarities is fixed, it reduces to

$$C_{MsSNE} = - \sum_{i \in \mathcal{I}, j \in \mathcal{I} \setminus \{i\}} \sigma_{ij,ms} \log s_{ij,ms}. \quad (22)$$

As the cost functions of SNE-like methods are difficult to optimize, especially when small perplexities are employed, the use of multi-scale optimization is typically recommended, by first only considering the larger scales during the optimization and then introducing smaller ones one at the time [39].

As there are $\mathcal{O}(\log N)$ scales, evaluating the multi-scale similarities, hence the multi-scale SNE gradient, costs $\mathcal{O}(N^2 \log N)$. Whereas SNE-like methods succeed best in preserving neighborhoods with sizes close to or smaller than K_* but poorly reproduce larger ones, the parameter-free multi-scale approach preserves the neighborhoods in a broader range of sizes. This behavior has been shown to be of great interest to retain local as well as global structures in the data [39].

IV. MULTIPLE IMPUTATIONS FOR DIMENSIONALITY REDUCTION WITH MISSING DATA

Given an incomplete data set Ξ with N independent examples, multi-scale SNE can not be directly applied as the examples $\xi_i \in \mathbb{R}^H$ contain unobserved features. However, following (2), the expectation, over the missing values, of (22) may be minimized. It can indeed be expressed as

$$\mathbb{E}[C_{MsSNE}] = - \sum_{i \in \mathcal{I}, j \in \mathcal{I} \setminus \{i\}} \mathbb{E}[\sigma_{ij,ms}] \log s_{ij,ms}. \quad (23)$$

Only the expected HD similarities must hence be evaluated in order to apply multi-scale SNE. This can be performed by modeling the Ξ distribution in $\mathbb{R}^H$ through a Gaussian mixture model. One can then sample Ψ times the conditional distribution of the missing features of each sample ξ_i given its observed features. This creates Ψ complete data sets $\Xi_c^1, \dots, \Xi_c^\Psi$. Their multi-scale HD similarities $[\sigma_{ij,ms}^1]_{i,j=1}^N, \dots, [\sigma_{ij,ms}^\Psi]_{i,j=1}^N$ may hence be computed, allowing to approximate the expectation of the multi-scale HD similarities of Ξ :

$$\mathbb{E}[\sigma_{ij,ms}] \approx \frac{1}{\Psi} \sum_{l=1}^{\Psi} \sigma_{ij,ms}^l. \quad (24)$$

This approach follows the multiple imputation paradigm for dealing with missing data [50]–[52], [66]. Notice that computing the multi-scale HD similarities of the complete data set resulting from the imputation of the missing values of each example with their conditional means given the observed features does not approximate the expectation of the multi-scale HD similarities of Ξ , as they are non-linear in the HD coordinates. It would yield the LDS of (1) instead of (2).

The procedure for applying multi-scale SNE to a data set Ξ with missing values can be summarized as follows:

- 1) Fit a Gaussian mixture model on Ξ (Section II);

- 2) Use this model to extract Ψ random samples of the conditional distribution of the missing features of each example ξ_i given its observed ones, in order to create Ψ complete data sets $\Xi_c^1, \dots, \Xi_c^\Psi$;
- 3) Compute their multi-scale HD similarities and the expected multi-scale HD similarities of Ξ using (24);
- 4) Use the approximation to minimize the expected multi-scale SNE cost function (23) over the LD coordinates.

The expected multi-scale SNE gradient can be evaluated in the same way as the cost function, by employing (24).

Computing the multi-scale HD similarities of the Ψ imputed data sets scales in $\mathcal{O}(\Psi N^2 \log N)$ time. Their average enables to tune the LD coordinates thanks to gradient-based optimization algorithm, each gradient evaluation costing $\mathcal{O}(N^2 \log N)$. The total time complexity of the procedure then scales in $\mathcal{O}(T_{MG} + \Psi N^2 \log N)$, where T_{MG} is defined in Section II-C. The number Ψ of random samplings is set to 100 during the experiments of this paper, but the results are stable across variations over this metaparameter.

It should be stressed that, even though this work focuses on the use of multi-scale SNE [39], the proposed methodology can be directly applied using any other neighbor embedding method, by approximating appropriate expectations. More generally, any DR scheme optimizing some criterion can be employed, as the cost function expectation, as well as its gradient, can always be estimated under the multiple imputation framework after fitting a Gaussian mixture model on the incomplete data set. The rest of this section details the application of the methodology with the SNE, t -SNE, Jensen-Shannon embedding (JSE) and Neighborhood Retrieval Visualizer (NeRV) methods.

SNE can be used in the same way as multi-scale SNE, by estimating the expectations $\mathbb{E}[\sigma_{ij}]$ for all $i \in \mathcal{I}$ and $j \in \mathcal{I} \setminus \{i\}$ instead of $\mathbb{E}[\sigma_{ij,ms}]$ and by minimizing, over the LD positions, the expected SNE cost function

$$\mathbb{E}[C_{SNE}] \propto - \sum_{i \in \mathcal{I}, j \in \mathcal{I} \setminus \{i\}} \mathbb{E}[\sigma_{ij}] \log s_{ij}, \quad (25)$$

the entropy of the HD similarities being constant with respect to $\mathbf{X}$. Besides symmetrizing SNE HD similarities, t -SNE employs a Student t -distribution with one degree of freedom in the LDS, circumventing the crowding problem [17]:

$$\sigma_{ij,t} = \frac{\sigma_{ij} + \sigma_{ji}}{2N}, \quad s_{ii,t} = 0, \quad (26)$$

$$s_{ij,t} = \frac{1}{(1 + d_{ij}^2) \sum_{k \in \mathcal{I}, l \in \mathcal{I} \setminus \{k\}} (1 + d_{kl}^2)^{-1}}. \quad (27)$$

The t -SNE cost function remains as in SNE. Its expectation

$$\mathbb{E}[C_{t-SNE}] \propto - \sum_{i \in \mathcal{I}, j \in \mathcal{I} \setminus \{i\}} \mathbb{E}[\sigma_{ij,t}] \log s_{ij,t}, \quad (28)$$

can be minimized by estimating the HD similarities expectation as previously. As to JSE [25], it relies on SNE similarities,

$$\sigma_{ij,j} = \sigma_{ij}, \quad s_{ij,j} = s_{ij}, \quad \text{for all } i, j \in \mathcal{I}, \quad (29)$$

but finds the LD coordinates by minimizing the sum of generalized Jensen-Shannon divergences:

$$C_{JSE} = \sum_{i \in \mathcal{I}, j \in \mathcal{I} \setminus \{i\}} \frac{1}{1 - \kappa} \sigma_{ij,j} \log(\sigma_{ij,j} / z_{ij}) + \frac{1}{\kappa} s_{ij,j} \log(s_{ij,j} / z_{ij}), \quad (30)$$

where $z_{ij} = \kappa \sigma_{ij,j} + (1 - \kappa) s_{ij,j}$, with $\kappa \in [0, 1]$ being a metaparameter weighting both terms in the divergence. The expectation of JSE cost function is then proportional to

$$\mathbb{E}[C_{JSE}] \propto \sum_{i \in \mathcal{I}, j \in \mathcal{I} \setminus \{i\}} \frac{-1}{1 - \kappa} \mathbb{E}[\sigma_{ij,j} \log(z_{ij})] + \frac{1}{\kappa} s_{ij,j} \log(s_{ij,j}) - \frac{1}{\kappa} s_{ij,j} \mathbb{E}[\log(z_{ij})]. \quad (31)$$

Given some current LD coordinates, the HD similarities $[\sigma_{ij,j}^l; l = 1, \dots, \Psi]_{i,j=1}^N$ of the Ψ complete data sets resulting from the multiple imputations and defining $z_{ij}^l = \kappa \sigma_{ij,j}^l + (1 - \kappa) s_{ij,j}$, the expected JSE cost function is estimated using

$$\mathbb{E}[\sigma_{ij,j} \log z_{ij}] \approx \frac{1}{\Psi} \sum_{l=1}^{\Psi} \sigma_{ij,j}^l \log z_{ij}^l, \quad (32)$$

$$\mathbb{E}[\log z_{ij}] \approx \frac{1}{\Psi} \sum_{l=1}^{\Psi} \log z_{ij}^l. \quad (33)$$

The expected JSE gradient can be similarly approximated.

NeRV is based on the same HD similarities as SNE, but slightly different LD ones [21]: for all $i \in \mathcal{I}$, $j \in \mathcal{I} \setminus \{i\}$,

$$\sigma_{ij,n} = \sigma_{ij}, \quad \sigma_{ii,n} = s_{ii,n} = 0, \quad (34)$$

$$s_{ij,n} = \frac{\exp(-\pi_i d_{ij}^2 / 2)}{\sum_{k \in \mathcal{I} \setminus \{i\}} \exp(-\pi_i d_{ik}^2 / 2)}. \quad (35)$$

The LD precisions are hence set to the HD ones. Moreover, NeRV also minimizes a mixture of KL divergences:

$$C_{NeRV} = \sum_{i \in \mathcal{I}, j \in \mathcal{I} \setminus \{i\}} (1 - \kappa) \sigma_{ij,n} \log(\sigma_{ij,n} / s_{ij,n}) + \kappa s_{ij,n} \log(s_{ij,n} / \sigma_{ij,n}), \quad (36)$$

where $\kappa \in [0, 1]$ again weights the different terms. The cost function expectation is thus proportional to

$$\mathbb{E}[C_{NeRV}] \propto \sum_{i \in \mathcal{I}, j \in \mathcal{I} \setminus \{i\}} (\kappa - 1) \mathbb{E}[\sigma_{ij,n} \log(s_{ij,n})] + \kappa \mathbb{E}[s_{ij,n} \log(\sigma_{ij,n} / s_{ij,n})], \quad (37)$$

the LD similarities remaining inside the expectations as their precisions depend on the HD ones. Given some LD positions, the expected NeRV cost function can hence be approximated by estimating the expectations in (37) using the HD similarities of the Ψ imputed data sets, as for JSE. The NeRV gradient can be evaluated likewise. The multi-scale versions of t -SNE, NeRV and JSE [39], [40] can also be employed in the same way, by replacing their single-scale similarities with multi-scale ones.

V. EXPERIMENTAL RESULTS

In order to validate the proposed methodology, several methods are experimentally compared on various data sets, mainly from the UCI Machine Learning Repository [71]: Breast Tissue (BrTis.), Concrete Compressive Strength (CCS)

[72], Connectionist Bench on Vowel Recognition and Deterding Data (ConV.), Concrete Slump Test (ConcSl.) [73], a random subset of Optical Recognition of Handwritten Digits (Digits) with 50 samples per digit, Energy efficiency (EE) [74], Ecoli, Forest Fires (FFir.) [75], Fertility (Fert.) [76], Glass Identification (GIId.), Indian Liver Patient (ILP), Ionosphere (Iono.), Iris, Istanbul Stock Exchange (IstSt.) [77], Leaf [78], Parkinsons (Park.) [79], Planning Relax (PIRel.) [80], Seeds, Servo, User Knowledge Modeling (UsKn.) [81], Wholesale customers (WholC.) [82], Wine and Yacht Hydrodynamics (YaHyd.). Fig. 4 indicates the number of samples N and dimension H of each database. An artificial data set (Sphere) with 1500 dots randomly distributed on a unit-radius three-dimensional sphere is also considered. The data sets being all complete, missing values are randomly and independently introduced during the experiments, corresponding to some fixed proportion of the whole data set, while ensuring that each example and feature contain at least one observed value. The completeness of the data sets enables to quantify the quality of the methods after discarding some data according to different criteria described in Section V-A. The compared methods are:

- Multiple imputations using a mixture of multivariate Gaussians (MIMG), which is the suggested methodology;
- Conditional mean imputation using a mixture of multivariate Gaussians (IMG), which fits a Gaussian mixture model on the incomplete data set and then impute its missing values by their conditional mean given the observed features of their examples in order to apply multi-scale SNE on the resulting data set;
- Multiple imputations using a single multivariate Gaussian (MIG) and conditional mean imputation using a single multivariate Gaussian (IG), which are equivalent to MIMG and IMG except that they use only one Gaussian to model the data set;
- Mean imputation (IME), which first computes the mean of each feature across its observed values in the data, then fills the missing data using the estimated means of their corresponding features and finally applies multi-scale SNE on the imputed data set;
- Incomplete-case k -nearest neighbor imputation (INN) [48], which imputes each missing feature $j \in M_i$ of every sample ξ_i by the mean of the j^{th} feature of its k nearest neighbors in Ξ in order to afterward apply multi-scale SNE on the resulting data set. A data sample ξ_i is eligible as a nearest neighbor if and only if $O_i \subset O_l$ and $j \in O_l$. The distances between ξ_i and its eligible nearest neighbors are evaluated by only considering the features indexed in O_i . When the set of eligible nearest neighbors is empty, the imputation relies on the sample mean of the j^{th} feature. INN improves the complete-case k -NN imputation [47] as the latter only considers the complete instances (i.e. $O_i = \{1, \dots, H\}$) as eligible nearest neighbors, which greatly limits their number especially when dealing with large proportions of missing values. The experiments in this paper consider up to 5 nearest neighbors, as suggested by [48], [65].

Comparing MIMG to IMG highlights the difference between

optimizing the expected cost function and reducing the dimension of the expected HD data set. IG and MIG provide insights on the importance of the data set distribution modeling when confronted to IMG and MIMG. IMe ignores both the feature relationships and the evaluation of the multi-scale SNE cost function expectation, and may thus be considered as the baseline method. In the same spirit as IG and IMG, INN aims to model the data distribution but only considers a single imputed data set. It should be noted that, when applied on some incomplete data set, MIMG and IMG (resp. MIG and IG) use the same Gaussian mixture model (resp. Gaussian model).

The time complexity of IMG is similar to MIMG but with $\Psi = 1$, while IG and MIG behave likewise but with $G = 1$. On the other hand, the simplicity of IMe enables to reach a $\mathcal{O}(NH + N^2 \log N)$ time complexity. Finally, INN must first find the eligible nearest neighbors for each sample, which costs $\mathcal{O}(N^2 H)$. As the imputation is then straightforward, the method scales in $\mathcal{O}(N^2 H + N^2 \log N)$ time.

Multi-scale SNE is used by all methods with $K_* = 2$ and $L = \lfloor \log_2(N/K_*) \rfloor$. Two-dimensional embeddings are searched ($P = 2$), and the L-BFGS optimization algorithm [83] is employed along with multi-scale optimization [39]. The LD embedding is initialized using an origin-centered, isotropic Gaussian distribution with a 10^{-4} standard deviation [17]. Euclidean distances are used in the HDS and LDS.

A. Quality Assessment

The studied methods are evaluated both by assessing the resulting DR performances and by quantifying the approximation quality of the “true” (i.e. complete-case) multi-scale HD similarities $[\sigma_{ij,ms}]_{i,j=1}^N$ by their estimate. These two measures can be computed thanks to the completeness of the data sets, in which missing values are simulated.

The approximation quality of the HD similarities is quantified by the sum of the KL divergences between the “true” and approximated similarities for each data point,

$$Q_{KL} = \sum_{i \in \mathcal{I}, j \in \mathcal{I} \setminus \{i\}} \sigma_{ij,ms} \log(\sigma_{ij,ms} / \hat{\sigma}_{ij,ms}), \quad (38)$$

where $\hat{\sigma}_{ij,ms}$ is the approximation of $\sigma_{ij,ms}$ obtained by the considered method. In the MIMG case, it is equal to the right-hand side of (24).

As to the DR quality assessment, some studies developed criteria measuring the HD neighborhood preservation in the LDS [16], [84]. This principle is indeed considered as driving the DR quality [21], [85], [86]. The HD and LD ranks of ξ_j and $\mathbf{x}_j$ with respect to ξ_i and $\mathbf{x}_i$ can first be defined as

$$\rho_{ij} = |\{k \in \mathcal{I} | \delta_{ik} < \delta_{ij} \text{ or } (\delta_{ik} = \delta_{ij} \text{ and } k < j)\}|, \quad (39)$$

$$r_{ij} = |\{k \in \mathcal{I} | d_{ik} < d_{ij} \text{ or } (d_{ik} = d_{ij} \text{ and } k < j)\}|. \quad (40)$$

The K -ary neighborhoods of ξ_i and $\mathbf{x}_i$ can then be respectively expressed as $\nu_i^K = \{j \in \mathcal{I} | 1 \leq \rho_{ij} \leq K\}$ and $n_i^K = \{j \in \mathcal{I} | 1 \leq r_{ij} \leq K\}$. They index the K nearest neighbors of ξ_i and $\mathbf{x}_i$ in the HDS and LDS. Their average normalized agreement is computed as

$$Q_{NX}(K) = \sum_{i \in \mathcal{I}} \frac{|\nu_i^K \cap n_i^K|}{KN} \in [0, 1]. \quad (41)$$

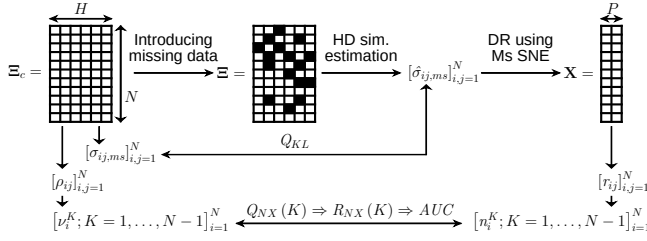


Fig. 1. Quality assessment employed in the experiments of this paper. The depicted data sets have one row per sample and one column per feature. Black entries indicate missing values. The notation “sim.” abbreviates “similarities”.

As $\mathbb{E}[Q_{NX}(K)]$ equals $K/(N-1)$ for random LD points,

$$R_{NX}(K) = \frac{(N-1)Q_{NX}(K) - K}{N-1-K} \quad (42)$$

rescales $Q_{NX}(K)$ to enable the comparison of different neighborhood sizes [25]. The $R_{NX}(K)$ curves are typically displayed with a log-scale for K as local neighborhoods usually prevail. Evaluated in $\mathcal{O}(N^2 \log N)$ [84], the area under the resulting curve

$$AUC = \left(\sum_{K=1}^{N-2} K^{-1} \right)^{-1} \cdot \sum_{K=1}^{N-2} R_{NX}(K)/K \in [-1, 1] \quad (43)$$

increases with the DR quality, quantified at all scales with an emphasis on small ones [39].

Both Q_{KL} and the AUC of the computed LD embeddings are used to evaluate the methods during the experiments. Their computation is depicted in Fig. 1. Note that for Q_{KL} , the smaller the better whereas for the AUC , the larger the better.

B. Experimental Setting

Missing values are introduced in each data set according to 30 different proportions p_{mis} of its number NH of entries, from 1% to 30%. For each considered proportion p_{mis} of missing values, 200 incomplete data sets are created and all methods are applied on each one of them. Before introducing missing values, Z-score standardization is applied on the data sets containing features with different ranges (i.e. all except Sphere and Digits). This ensures that missing data for different features have the same importance, which makes the repetitions for a fixed p_{mis} comparable. The repetitions for each p_{mis} enable computing paired statistical tests assessing whether the average performances of the methods are significantly different according to both Q_{KL} and the AUC . For each one of these two scores, two-tailed paired t -tests are employed if their 200 values can be considered to be normally distributed for each p_{mis} and each method, according to D’Agostino and Pearson’s omnibus normality test [87]. Wilcoxon tests are otherwise used. The level of significance is set to $\alpha = 0.01$.

Comparing the performances of several methods for each p_{mis} however consists of multiple hypothesis tests. The Holm–Bonferroni correction is hence applied for each p_{mis} to bound by α the probability to consider as statistically significant at least one non-significant difference [88].

C. Results

Fig. 2 illustrates the results on the Yacht Hydrodynamics ($N = 308$, $H = 6$) and Leaf ($N = 340$, $H = 14$) data sets. On Yacht Hydrodynamics, MIMG is consistently the best method on average for all proportions p_{mis} of missing values, according to both criteria, except for a very limited number of p_{mis} for most of which it is not statistically different from the corresponding best performing method. MIMG is moreover nearly always statistically significantly superior to the second best method, which is almost systematically IMG for the AUC and switches from IMG to MIG for Q_{KL} . It is noteworthy that even though MIG approximates the multi-scale HD similarities for the largest p_{mis} better than MIMG, MIMG still provides better embeddings according to the AUC . This can be explained as, although MIG has a lower Q_{KL} , MIMG may still better approximate the HD similarity distributions around some data points which can have a larger influence for the DR part. For large p_{mis} , the standard deviations of the performances are not negligible, meaning that the missing values positions in the data set strongly influence their impact: if the observed data sufficiently well describe the distribution of the missing features given the observed ones, which is for instance more likely with incomplete samples fairly spread in the data distribution, the distribution modeling of the missing values is easier, enabling better performances. Otherwise, if the missing entries exhibit specific structures, e.g. if they concentrate in some features for some values of observed data, less information may be available about the missing values distribution, affecting the performances. Despite large standard deviations, most results are statistically significant; the performances of the different methods hence vary consistently with respect to the missing-data pattern. IMe and IG are unsurprisingly always outperformed by the other methods. The superiority of MIG over IG and of MIMG over IMG moreover stresses the importance of optimizing the expected cost function over the missing data, in order to derive the closest result to the complete data case on average. One can finally observe that the number of Gaussian components chosen in the mixture-based models tends to slightly increase with the proportion p_{mis} of missing values. The growing uncertainty may indeed affect the quality of the BIC criterion for model selection. As the number of observed data decreases with p_{mis} , fitting more Gaussian components may result in slight overfitting of the observed data. This justifies the Q_{KL} increase of MIMG and IMG for large proportions of missing values.

As to the Leaf data set, a similar analysis can be conducted. According to both criteria, MIMG is either always the best method or not statistically different from IMG. The impressive performance drop of INN with p_{mis} is a noticeable difference with Yacht Hydrodynamics. The Leaf data set having more features, increasing the proportion of missing values indeed considerably affects the determination of the nearest neighbors. The number of Gaussian components still slightly increases with p_{mis} . However, no overfitting behavior is observed in this case: the performances of the Gaussian mixture-based methods remain blatantly superior to the ones of the other schemes. They are even more robust with respect to the missing values,

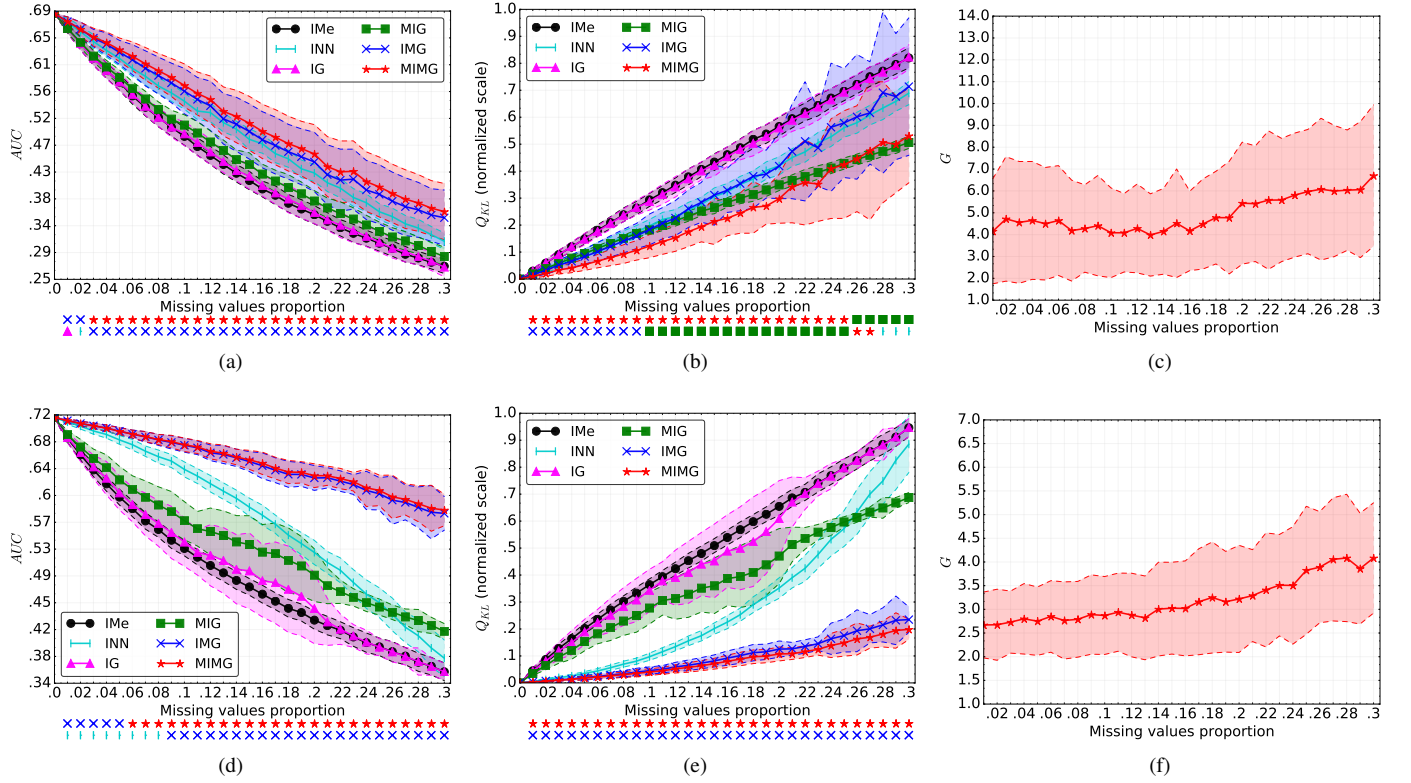


Fig. 2. (a) depicts the AUC of each method on the Yacht Hydrodynamics data set as a function of the proportion p_{mis} of missing values. The curves give the average AUC over the 200 repetitions for each p_{mis} , while the shaded areas delimit one standard deviation around the mean. The first row of markers below the x -axis indicates the best method on average for each p_{mis} . The second row presents, for each p_{mis} , the best method which is statistically significantly outperformed by the best method on average for the corresponding p_{mis} . (b) is similar, but for the sum of the KL divergences between the “true” and approximated multi-scale HD similarities. The y -axis is normalized, as the absolute value of Q_{KL} brings little insights. (c) indicates, for each p_{mis} , the average number of Gaussian components selected in MIMG and IMG when applying the BIC-based heuristic, the shaded area depicting one standard deviation around the mean. (d), (e) and (f) are analogous to (a), (b) and (c) but for the Leaf data set.

since the gap between the methods is increasing with p_{mis} . This confirms that a DR scheme handling missing values should seek for the best possible model of the data distribution.

Fig. 3 further details some of the results obtained on Yacht Hydrodynamics. This data set contains 308 yachts with 22 different hull forms [71]. Each form is represented by 14 boats, which only differ according to their Froude numbers. The 14 Froude numbers are linearly spaced between two values. The embedding obtained by multi-scale SNE on the complete data set hence reveals that the samples are distributed on 22 one-dimensional manifolds. The 14 samples corresponding to some hull form are depicted using the same color in Fig. 3a. When introducing missing values in the data set, the $R_{NX}(K)$ curves in Fig. 3b show that missing data mainly affect the preservation of small neighborhoods, as the curves corresponding to the different methods are lower for small K values than the one obtained when using the complete data set. Fig. 3c illustrates the embedding obtained by MIMG on one of the 200 Yacht Hydrodynamics data set with 5% missing values. We can observe that it tends to recover the structure of the one-dimensional manifolds.

Since an analysis such as provided in Fig. 2 can not be presented for all data sets due to space limitations, Fig. 4 summarizes the remaining results. We can first observe that MIMG is almost systematically top or highly ranked among the different methods according to both the AUC and Q_{KL} .

Moreover, when it is (resp. is not) the best performing method, it is often (resp. often not) statistically different from the second best (resp. best) performing one. The few exceptions to these observations are analyzed in more details below.

ConV. is a data set containing numerous small clusters [71]. Unless the number of available samples becomes very large, this kind of data distribution is not easily described using Gaussian mixtures. Fitting many Gaussian components indeed requires to adjust increasingly many model parameters, even when employing HDDC models. On the other hand, as long as the proportion p_{mis} of missing values remains small, nearest neighbor methods impute the missing data using cluster-specific values. The imputation then follows very well the data distribution and, although a single one is considered, it is sufficiently close to the original data set to preserve its clusters. This explains why INN outperforms MIMG for the first few p_{mis} values. Nevertheless, as p_{mis} increases, the computed nearest neighbors are less and less cluster-specific. This drastically reduces the quality of the single imputation, causing the performances of INN to drop. Even though MIMG can probably not model exactly each cluster, it can still model higher-order relationships in the data when facing an increasing number of missing values. This makes MIMG superior to INN for larger p_{mis} . The same kind of analysis holds for Digits and Park., which also contain many clusters [71]. The performances of INN however decrease

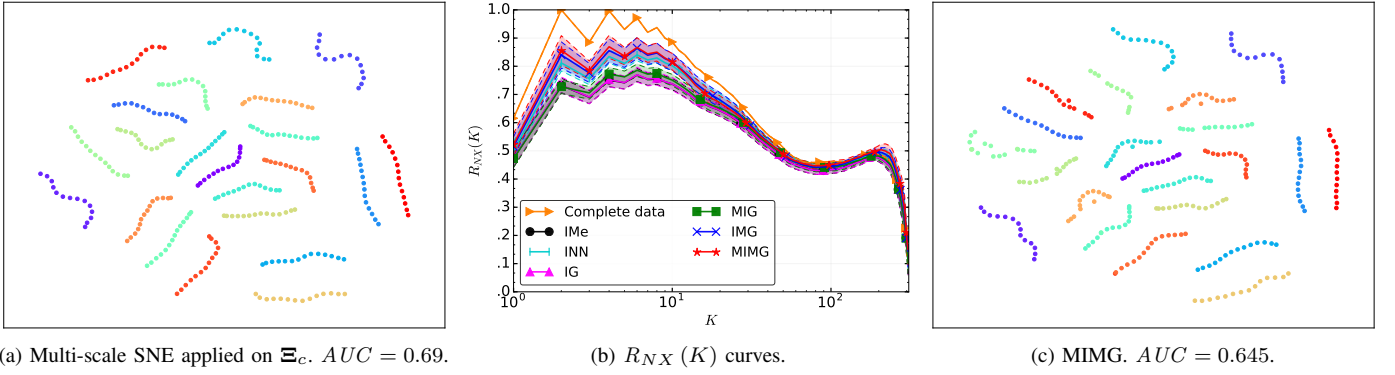


Fig. 3. (a) depicts the two-dimensional embedding obtained when applying multi-scale SNE on the complete Yacht Hydrodynamics data set. (b) provides the associated $R_{NX}(K)$ curve, as well as the average curves obtained by the different methods on 200 incomplete Yacht Hydrodynamics data sets with 5% missing values. As in Fig. 2, the shaded areas delimit one standard deviation around the mean. (c) presents the embedding obtained when applying MIMG on one of the 200 data sets with 5% missing values. The AUC of the embedding is provided below its plot.

even more rapidly as the dimensions of these data sets are larger, which makes the determination of the relevant nearest neighbors much more difficult.

On Ecoli, the opposite trend is observed: INN seems superior to MIMG for larger p_{mis} according to the AUC . Nevertheless, the performances of the two methods are not statistically significantly different for most p_{mis} . The adaptations of Fig. 2a and 2b for Ecoli indeed reveal that the difference between both methods is tiny. Concerning the evolution of Q_{KL} with p_{mis} , MIG better estimates the HD similarities than MIMG for larger values of p_{mis} , similarly to the results for Yacht Hydrodynamics. This is also observed on ILP and Servo.

As to Fert. and UsKn., both multiple imputation schemes are efficient according to Q_{KL} . The superiority of MIG over MIMG may result from potential slight overfitting of the mixture model to the observed data. The AUC difference between MIG and MIMG remain however small when looking at the adaptation of Fig. 2a for these data sets. In particular for Fert., most methods are statistically equivalent for small p_{mis} according to the AUC , possibly because this data set is hardly structured; the missing values management hence does not much impact the HD neighborhood preservation.

For Iono. and the first few p_{mis} of EE, MIMG better approximates the HD similarities than IMG from the Q_{KL} viewpoint. However, IMG is superior according to the AUC , but only with a tight margin when observing the adaptation of Fig. 2a for these data sets. The tiny AUC improvement of IMG over MIMG is probably due to a better estimation of the HD similarity distributions around some data points which have a larger influence on the HD neighborhood preservation.

Generally, multiple imputation schemes tend to outperform their single imputation versions, MIMG and MIG being respectively globally superior to IMG and IG. IMe and IG appear almost systematically as the weakest approaches. The performances of INN quickly drop as p_{mis} increases, especially on data sets with larger dimensions such as Iono., Digits or Park.

VI. CONCLUSION

This paper introduces the missing data processing in DR. After modeling the incomplete data set distribution using a Gaussian mixture extension for HD data, several imputed

versions of the data set are created by sampling the conditional distribution of the missing features of the examples given their observed features. These complete data sets are employed to reliably estimate the DR algorithm cost function expectation over the incomplete database. Its minimization yields the LD embedding with the closest cost function value on average with respect to the complete data case. This multiple imputation approach hence provides a relevant LD representation of each, possibly incomplete, data point. Experimental results show that this procedure outperforms alternative methods while being generally applicable with any DR scheme optimizing some criterion.

The Gaussian mixture model as well as the multiple imputation paradigm inevitably come up with an increase of the computation time complexity compared to single imputation schemes. Their superior performances however justify the use of additional computing resources whenever possible. A situation in which the proposed approach may not be most efficient is when facing extremely HD data: Gaussian mixtures would then be difficult to fit due to their increasing numbers of parameters, even with HDDC methods [58]. This motivates the development of alternative models of the data distribution, such as non-parametric estimations, to study whether they could better deal with the inherent sparsity of very HD spaces and reduce the time complexity of the approach.

APPENDIX

THE SWEEP OPERATOR FOR MATRIX INVERSION

Sweeping a symmetric matrix $\mathbf{E}$ on its row and column l results in a symmetric matrix $\mathbf{E}_s = S_{\{l\}}(\mathbf{E})$ where

$$(\mathbf{E}_s)_{ll} = -(\mathbf{E})_{ll}^{-1}, \quad (\mathbf{E}_s)_{lj} = (\mathbf{E})_{lj} / (\mathbf{E})_{ll} \quad \text{if } j \neq l, \quad (44)$$

$$(\mathbf{E}_s)_{ij} = (\mathbf{E})_{ij} - (\mathbf{E})_{il} (\mathbf{E})_{lj} / (\mathbf{E})_{ll} \quad \text{if } i \neq l \text{ \& } j \neq l. \quad (45)$$

Sweeping over a set of indexes can be performed in any order [51], [65]. Reverse sweeping $\mathbf{E}$ on l yields $\mathbf{E}_r = S_{\{l\}}^{-1}(\mathbf{E})$ equal to $\mathbf{E}_s$ except that $(\mathbf{E}_r)_{lj} = (\mathbf{E}_r)_{jl} = -(\mathbf{E}_s)_{lj}$ if $j \neq l$.

Let us denote by $\tilde{\mu}_{ik} \in \mathbb{R}^H$ a vector whose components indexed by O_i (resp. M_i) are equal to $\mu_k^{O_i} - \xi_i^{O_i}$ (resp. $\mu_k^{M_i}$).

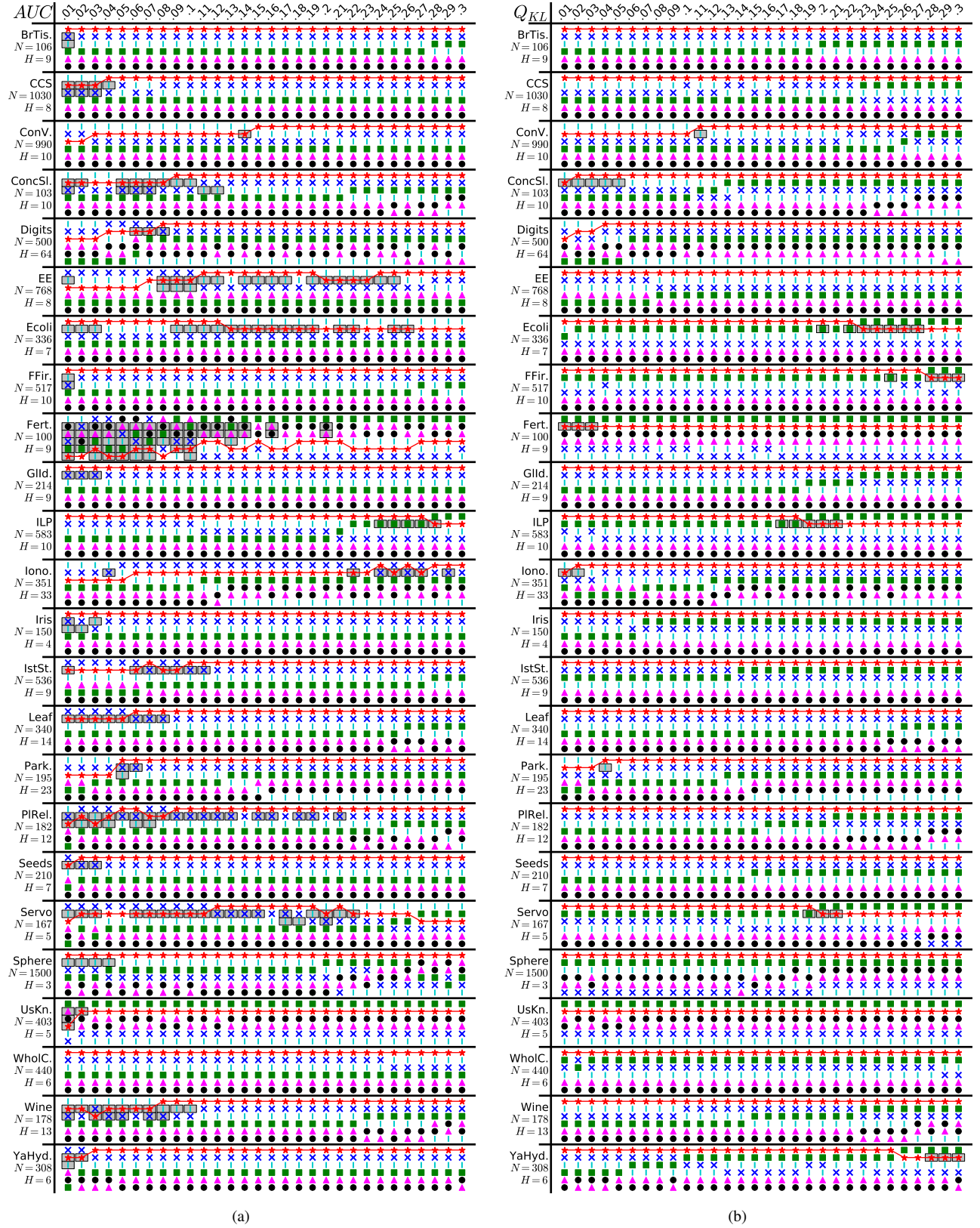


Fig. 4. Summary of the performances in terms of (a) AUC and (b) Q_{KL} for all data sets. The results for each data set are displayed horizontally, using six rows of markers. These refer to the same six methods as in Fig. 2. There is one column per proportion p_{mis} of missing values. For each data set and p_{mis} , the six markers are sorted vertically from the best performing method on average at the top (i.e. largest average AUC or smallest average Q_{KL}) to the worst average one at the bottom (i.e. smallest average AUC or largest average Q_{KL}). The averages are computed across the 200 replications for each p_{mis} . The markers corresponding to MIMG are linked. A grey shaded area is displayed behind a marker when the corresponding method is **not** statistically significantly outperformed by the best method on average for the associated p_{mis} , except for the top marker of the columns, as a method can obviously not be significantly different from itself.

It can be shown [51], [65] that, for $i \in \mathcal{I}$ and $k \in \mathcal{G}$,

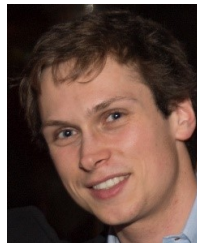
$$S_{O_i} \begin{pmatrix} \Sigma_k & \tilde{\mu}_{ik} \\ \tilde{\mu}_{ik}^T & 0 \end{pmatrix} = S_{M_i}^{-1} \begin{pmatrix} -\Sigma_k^{-1} & \Sigma_k^{-1} \tilde{\mu}_{ik} \\ \tilde{\mu}_{ik}^T \Sigma_k^{-1} & -\tilde{\mu}_{ik}^T \Sigma_k^{-1} \tilde{\mu}_{ik} \end{pmatrix} \quad (46)$$

contains $\tilde{\Sigma}_{k, M_i | O_i}$ in its rows and columns indexed by M_i , $\tilde{\mu}_{k, M_i | O_i}$ in its last column at the rows indexed by M_i and $(\xi_i^{O_i} - \mu_k^{O_i})^T (\Sigma_k^{O_i})^{-1} (\mu_k^{O_i} - \xi_i^{O_i})$ in its bottom-right entry, useful to evaluate t_{ik} . As $\Sigma_k^{-1} = \mathbf{Q}_k \bar{\Lambda}_k^{-1} \mathbf{Q}_k^T$, the right-hand side of (46) is faster to compute when $|O_i| > |M_i|$ [65].

REFERENCES

- [1] J. A. Lee and M. Verleysen, *Nonlinear dimensionality reduction*. Springer Science & Business Media, 2007.
- [2] K. Bunte, M. Biehl, and B. Hammer, "A general framework for dimensionality-reducing data visualization mapping," *Neural Computation*, vol. 24, no. 3, pp. 771–804, 2012.
- [3] I. Borg and P. J. F. Groenen, *Modern Multidimensional Scaling: Theory and applications*. Springer Science & Business Media, 2005.
- [4] T. Kohonen, "The self-organizing map," *Proc. IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.
- [5] G. Hinton and S. Roweis, "Stochastic neighbor embedding," in *NIPS*, vol. 15, 2002, pp. 833–840.
- [6] I. T. Jolliffe, "Principal component analysis and factor analysis," in *Principal component analysis*. Springer, 1986, pp. 115–128.
- [7] J. W. Sammon, "A nonlinear mapping for data structure analysis," *IEEE Trans. Comput.*, vol. 100, no. 5, pp. 401–409, 1969.
- [8] J. B. Tenenbaum, "Mapping a manifold of perceptual observations," in *Advances in neural information processing systems*, 1998, pp. 682–688.
- [9] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.
- [10] J. A. Lee, A. Lendasse, and M. Verleysen, "Nonlinear projection with curvilinear distances: Isomap versus curvilinear distance analysis," *Neurocomputing*, vol. 57, pp. 49–76, 2004.
- [11] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Advances in neural information processing systems*, 2002, pp. 585–591.
- [12] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
- [13] B. Nadler, S. Lafon, I. Kevrekidis, and R. R. Coifman, "Diffusion maps, spectral clustering and eigenfunctions of fokker-planck operators," in *Advances in neural information processing systems*, 2006, pp. 955–962.
- [14] B. Schölkopf, A. Smola, and K.-R. Müller, "Kernel principal component analysis," in *International Conference on Artificial Neural Networks*. Springer, 1997, pp. 583–588.
- [15] J. Suykens, T. Van Gestel, J. Vandewalle, and B. De Moor, "A support vector machine formulation to PCA analysis and its kernel version," *IEEE Trans. on neural networks*, vol. 14, no. 2, pp. 447–450, 2003.
- [16] J. A. Lee and M. Verleysen, "Quality assessment of dimensionality reduction: Rank-based criteria," *Neurocomputing*, vol. 72, no. 7, pp. 1431–1443, 2009.
- [17] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.
- [18] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural computation*, vol. 10, no. 5, pp. 1299–1319, 1998.
- [19] J. Ham, D. D. Lee, S. Mika, and B. Schölkopf, "A kernel view of the dimensionality reduction of manifolds," in *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, p. 47.
- [20] D. Francois, V. Wertz, and M. Verleysen, "The concentration of fractional distances," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 7, pp. 873–886, 2007.
- [21] J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski, "Information retrieval perspective to nonlinear dimensionality reduction for data visualization," *Journal of Machine Learning Research*, vol. 11, no. Feb, pp. 451–490, 2010.
- [22] J. A. Lee and M. Verleysen, "Shift-invariant similarities circumvent distance concentration in stochastic neighbor embedding and variants," *Proc. International Conference on Computational Science (ICCS 2011)*, vol. 4, pp. 538–547, 2011.
- [23] J. Venna and S. Kaski, "Nonlinear dimensionality reduction as information retrieval," in *Artificial Intelligence and Statistics*, 2007, pp. 572–579.
- [24] K. Bunte, S. Haase, M. Biehl, and T. Villmann, "Stochastic neighbor embedding (SNE) for dimension reduction and visualization using arbitrary divergences," *Neurocomputing*, vol. 90, pp. 23–45, 2012.
- [25] J. A. Lee, E. Renard, G. Bernard, P. Dupont, and M. Verleysen, "Type 1 and 2 mixtures of kullback–leibler divergences as cost functions in dimensionality reduction based on similarity preservation," *Neurocomputing*, vol. 112, pp. 92–108, 2013.
- [26] Z. Yang, I. King, Z. Xu, and E. Oja, "Heavy-tailed symmetric stochastic neighbor embedding," in *Advances in neural information processing systems*, 2009, pp. 2169–2177.
- [27] A. Gisbrecht, A. Schulz, and B. Hammer, "Parametric nonlinear dimensionality reduction using kernel t-SNE," *Neurocomputing*, vol. 147, pp. 71–82, 2015.
- [28] M. Carreira-Perpinán, "The elastic embedding algorithm for dimensionality reduction," in *ICML*, vol. 10, 2010, pp. 167–174.
- [29] K. Bunte, B. Hammer, T. Villmann, M. Biehl, and A. Wismüller, "Neighbor embedding XOM for dimension reduction and visualization," *Neurocomputing*, vol. 74, no. 9, pp. 1340–1350, 2011.
- [30] L. v. d. Maaten, "Fast optimization for t-SNE," in *NIPS Workshop on Challenges in Data Visualization*, 2010.
- [31] M. Vladymyrov and M. Carreira-Perpinán, "Partial-hessian strategies for fast learning of nonlinear embeddings," in *ICML*, vol. 29, 2012, pp. 345–352.
- [32] J. Peltonen and Z. Lin, "Multiplicative update for fast optimization of information retrieval based neighbor embedding," in *International Workshop on Machine Learning for Signal Processing*, 2013, pp. 1–6.
- [33] M. Vladymyrov and M. Carreira-Perpinán, "Entropic affinities: Properties and efficient numerical computation," in *ICML*, 2013, pp. 477–485.
- [34] A. Gisbrecht, B. Mokbel, and B. Hammer, "Linear basis-function t-SNE for fast nonlinear dimensionality reduction," in *The International Joint Conference on Neural Networks*. IEEE, 2012, pp. 1–8.
- [35] Z. Yang, J. Peltonen, and S. Kaski, "Scalable optimization of neighbor embedding for visualization," in *ICML (2)*, 2013, pp. 127–135.
- [36] L. v. d. Maaten, "Accelerating t-SNE using tree-based algorithms," *J. of machine learning research*, vol. 15, no. 1, pp. 3221–3245, 2014.
- [37] M. Vladymyrov and M. Carreira-Perpinán, "Linear-time training of nonlinear low-dimensional embeddings," in *Artificial Intelligence and Statistics*, 2014, pp. 968–977.
- [38] C. de Bodt, D. Mulders, M. Verleysen, and J. A. Lee, "Extensive assessment of Barnes-Hut t-SNE," in *ESANN*, 2018, pp. 135–140.
- [39] J. A. Lee, D. H. Peluffo-Ordóñez, and M. Verleysen, "Multi-scale similarities in stochastic neighbour embedding: Reducing dimensionality while preserving both local and global structure," *Neurocomputing*, vol. 169, pp. 246–261, 2015.
- [40] C. de Bodt, D. Mulders, M. Verleysen, and J. A. Lee, "Perplexity-free t-SNE and twice Student t-SNE," in *ESANN*, 2018, pp. 123–128.
- [41] P. J. García-Laencina, J.-L. Sancho-Gómez, A. R. Figueiras-Vidal, and M. Verleysen, "K nearest neighbours with mutual information for simultaneous classification and missing data imputation," *Neurocomputing*, vol. 72, no. 7, pp. 1483–1493, 2009.
- [42] G. Doquire and M. Verleysen, "Feature selection with missing data using mutual information estimators," *Neurocomputing*, vol. 90, pp. 3–11, 2012.
- [43] A. Baraldi and C. Enders, "An introduction to modern missing data analyses," *Journal of school psychology*, vol. 48, no. 1, pp. 5–37, 2010.
- [44] P. D. Allison, "Missing data: Quantitative applications in the social sciences," *British Journal of Mathematical and Statistical Psychology*, vol. 55, no. 1, pp. 193–196, 2002.
- [45] S. Van Buuren, J. Brand, C. Groothuis-Oudshoorn, and D. Rubin, "Fully conditional specification in multivariate imputation," *J. of statistical computation and simulation*, vol. 76, no. 12, pp. 1049–1064, 2006.
- [46] Q. Yu, Y. Miche, E. Eirola, M. Van Heeswijk, E. Séverin, and A. Lendasse, "Regularized extreme learning machine for regression with missing data," *Neurocomputing*, vol. 102, pp. 45–51, 2013.
- [47] E. Hruschka, E. Hruschka, and N. Ebecken, "Evaluating a nearest-neighbor method to substitute continuous missing values," in *Australasian Joint Conference on Artificial Intelligence*, 2003, pp. 723–734.
- [48] J. Van Hulse and T. M. Khoshgoftaar, "Incomplete-case nearest neighbor imputation in software measurement data," *Information Sciences*, vol. 259, pp. 596–610, 2014.
- [49] I. Wasito and B. Mirkin, "Nearest neighbour approach in the least-squares data imputation algorithms," *Information Sciences*, vol. 169, no. 1, pp. 1–25, 2005.
- [50] D. B. Rubin, *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, 2004, vol. 81.
- [51] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*. John Wiley & Sons, 2014.

- [52] C. K. Enders, *Applied missing data analysis*. Guilford Press, 2010.
- [53] C. Bishop, *Pattern recognition and machine learning*. Springer, 2006.
- [54] D. M. Titterton, A. F. Smith, and U. E. Makov, *Statistical analysis of finite mixture distributions*. Wiley, 1985.
- [55] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.
- [56] Z. Ghahramani and M. Jordan, "Learning from incomplete data," *Tech. rep., Lab Memo 1509, CBCL Paper 108, MIT AI Lab*, 1995.
- [57] L. Hunt and M. Jorgensen, "Mixture model clustering for mixed data with missing information," *Computational Statistics & Data Analysis*, vol. 41, no. 3, pp. 429–440, 2003.
- [58] C. Bouveyron, S. Girard, and C. Schmid, "High-dimensional data clustering," *Computational Statistics & Data Analysis*, vol. 52, no. 1, pp. 502–519, 2007.
- [59] T. I. Lin, J. C. Lee, and H. J. Ho, "On fast supervised learning for normal mixture models with missing information," *Pattern Recognition*, vol. 39, no. 6, pp. 1177–1187, 2006.
- [60] O. Delalleau, A. Courville, and Y. Bengio, "Efficient em training of gaussian mixtures with missing data," *preprint arXiv:1209.0521*, 2012.
- [61] R. J. Steele, N. Wang, and A. E. Raftery, "Inference from multiple imputation for missing data using mixtures of normals," *Statistical methodology*, vol. 7, no. 3, pp. 351–365, 2010.
- [62] C. Beunckens, G. Molenberghs, and M. G. Kenward, "Direct likelihood analysis versus simple forms of imputation for missing data in randomized clinical trials," *Clinical Trials*, vol. 2, no. 5, pp. 379–386, 2005.
- [63] A. Morris, M. Cooke, and P. Green, "Some solution to the missing feature problem in data classification, with application to noise robust asr," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1998.*, vol. 2. IEEE, 1998, pp. 737–740.
- [64] E. Eirola, G. Doquire, M. Verleysen, and A. Lendasse, "Distance estimation in numerical data sets with missing values," *Information Sciences*, vol. 240, pp. 115–128, 2013.
- [65] E. Eirola, A. Lendasse, V. Vandewalle, and C. Biernacki, "Mixture of Gaussians for distance estimation with missing data," *Neurocomputing*, vol. 131, pp. 32–42, 2014.
- [66] D. Sovilj, E. Eirola, Y. Miche, K.-M. Björk, R. Nian, A. Akusok, and A. Lendasse, "Extreme learning machine for missing data using multiple imputations," *Neurocomputing*, vol. 174, pp. 220–231, 2016.
- [67] R. B. Cattell, "The scree test for the number of factors," *Multivariate behavioral research*, vol. 1, no. 2, pp. 245–276, 1966.
- [68] G. Schwarz *et al.*, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.
- [69] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. 19, no. 6, pp. 716–723, 1974.
- [70] C. Hurvich and C.-L. Tsai, "Regression and time series model selection in small samples," *Biometrika*, vol. 76, no. 2, pp. 297–307, 1989.
- [71] M. Lichman, "UCI Machine Learning repository," 2013. [Online]. Available: <http://archive.ics.uci.edu/ml>
- [72] I.-C. Yeh, "Modeling of strength of high-performance concrete using artificial neural networks," *Cement and Concrete research*, vol. 28, no. 12, pp. 1797–1808, 1998.
- [73] —, "Modeling slump flow of concrete using second-order regressions and artificial neural networks," *Cement and Concrete Composites*, vol. 29, no. 6, pp. 474–480, 2007.
- [74] A. Tsanas and A. Xifara, "Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools," *Energy and Buildings*, vol. 49, pp. 560–567, 2012.
- [75] P. Cortez and A. Morais, "A Data Mining Approach to Predict Forest Fires using Meteorological Data," in *Proc. of the 13th EPIA - Portuguese Conference on Artificial Intelligence*, 2007, pp. 512–523.
- [76] D. Gil, J. L. Girela, J. De Juan, M. J. Gomez-Torres, and M. Johnsson, "Predicting seminal quality with artificial intelligence methods," *Expert Systems with Applications*, vol. 39, no. 16, pp. 12 564–12 573, 2012.
- [77] O. Akbilgic, H. Bozdogan, and M. E. Balaban, "A novel Hybrid RBF Neural Networks model as a forecaster," *Statistics and Computing*, vol. 24, no. 3, pp. 365–375, 2014, DOI 10.1007/s11222-013-9375-7.
- [78] P. F. Silva, A. R. Marcal, and R. M. A. da Silva, "Evaluation of features for leaf discrimination," *Springer Lecture Notes in Computer Science*, vol. 7950, pp. 197–204, 2013.
- [79] M. Little, P. McSharry, S. Roberts, D. Costello, and I. Moroz, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection," *BioMedical Engineering OnLine*, vol. 6, no. 23, 2007.
- [80] R. Bhatt, "Planning-Relax Dataset for Automatic Classification of EEG Signals," *UCI Machine Learning Repository*, 2012.
- [81] H. T. Kahraman, S. Sagioglu, and I. Colak, "Developing intuitive knowledge classifier and modeling of users' domain dependent data in web," *Knowledge-Based Systems*, vol. 37, pp. 283–295, 2013.
- [82] N. Abreu, "Análise do perfil do cliente recheio e desenvolvimento de um sistema promocional," Ph.D. dissertation, ISCTE-IUL, Lisbon, 2011.
- [83] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM Journal on Scientific Computing*, vol. 16, no. 5, pp. 1190–1208, 1995.
- [84] J. A. Lee and M. Verleysen, "Scale-independent quality criteria for dimensionality reduction," *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2248–2257, 2010.
- [85] S. France and D. Carroll, "Development of an agreement metric based upon the rand index for the evaluation of dimensionality reduction techniques, with applications to mapping customer data," in *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. Springer, 2007, pp. 499–517.
- [86] L. Chen and A. Bujá, "Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis," *Journal of the American Statistical Association*, vol. 104, no. 485, pp. 209–219, 2009.
- [87] R. B. d'Agostino, "An omnibus test of normality for moderate and large size samples," *Biometrika*, vol. 58, no. 2, pp. 341–348, 1971.
- [88] J. P. Shaffer, "Multiple hypothesis testing," *Annual review of psychology*, vol. 46, no. 1, pp. 561–584, 1995.



Cyril de Bodt was born in 1993 in Belgium. He received the M.Sc. degree in Applied Sciences (Mathematical Engineering) in 2016 from the Université catholique de Louvain (UCL), Belgium. He is now a Ph.D. student in the UCL Machine Learning Group of the Institute for Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM). His research interests include dimensionality reduction, machine learning, visualization, clustering, optimization and data mining.



Dounia Mulders was born in 1993 in Belgium. She received the M.Sc. degree in Applied Sciences (Mathematical Engineering) in 2016 from the Université catholique de Louvain (UCL), Belgium. As a Ph.D. student in the UCL Machine Learning Group of the Institute for Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM), she studies the central mechanisms underlying pain perception in humans using signal processing tools and machine learning techniques to analyze scalp and intracerebral EEG recordings.



Michel Verleysen is a Professor of Machine Learning at the Université catholique de Louvain, Belgium. He has been an invited professor at the Swiss E.P.F.L. (Ecole Polytechnique Fédérale de Lausanne, Switzerland), the Université d'Evry Val d'Essonne (France), the Université ParisI-Panthéon-Sorbonne and the Université Paris Est. He is an Honorary Research Director of the Belgian F.N.R.S. (National Fund for Scientific Research), and the Dean of the Louvain School of Engineering. He is editor-in-chief of the Neural Processing Letters journal (published

by Springer), chairman of the annual ESANN conference (European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning), past associate editor of the IEEE Trans. on Neural Networks journal, and member of the editorial board and program committee of several journals and conferences on neural networks and learning. He was the chairman of the IEEE Computational Intelligence Society Benelux chapter (2008-2010), and member of the executive board of the European Neural Networks Society (2005-2010). He is author or co-author of more than 250 scientific papers in international journals and books or communications to conferences with reviewing committee. His research interests include machine learning, feature selection, nonlinear dimensionality reduction, visualization, high-dimensional data analysis, self-organization, time-series forecasting and biomedical signal processing.



John Aldo Lee was born in 1976 in Brussels, Belgium. He received the M.Sc. degree in Applied Sciences (Computer Engineering) in 1999 and the Ph.D. degree in Applied Sciences (Machine Learning) in 2003, both from the Université catholique de Louvain, Belgium. His main interests are dimensionality reduction, intrinsic dimensionality estimation, clustering, vector quantization, and various aspects of image processing applied to cancer care. He is a member of the UCL Machine Learning Group and a Senior Research Associate with the Belgian

F.R.S.-FNRS (Fonds National de la Recherche Scientifique). Together with Michel Verleysen, he wrote a monography entitled 'Nonlinear Dimensionality Reduction' published by Springer in 2007. He is currently working in the UCL center of Molecular Imaging, Radiotherapy, and Oncology (MIRO), where he develops specific image processing techniques to delineate organs and targets on medical images, to improve and speed up treatment planning in radiation oncology.