Visual Assessment of Clustering Tendency for Incomplete Data

Laurence A. F. Park, James C. Bezdek, *Life Fellow, IEEE*, Christopher Leckie, Ramamohanarao Kotagiri, *Member, IEEE*, James Bailey, and Marimuthu Palaniswami, *Fellow, IEEE*

Abstract—The iVAT (asiVAT) algorithms reorder symmetric (asymmetric) dissimilarity data so that an image of the data may reveal cluster substructure. Images formed from incomplete data don't offer a very rich interpretation of cluster structure. In this paper, we examine four methods for completing the input data with imputed values before imaging. We choose a best method using contaminated versions of the complete Iris data, for which the desired results are known. Then, we analyze two real world data sets from social networks that are incomplete using the best imputation method chosen in the juried trials with Iris: (i) Sampson's monastery data, an incomplete, asymmetric relation matrix; and (ii) the karate club data, comprising a symmetric similarity matrix that is about 86 percent incomplete.

Index Terms—reordered dissimilarity images, VAT, iVAT, visualisation, incomplete data, cluster heat maps, imputation, Karate club data

1 Introduction

LET $O = \{o_1, \ldots, o_N\}$ denote a set of N objects (red wines, babies, fish, etc.). Let $R = [r_{ij}]$ be a matrix of relational values on $O \times O$, r_{ij} being the relation between o_i and o_j. The most common form of R arises as dissimilarity data, say $D = [d_{ij}]$, where d_{ij} is the pair wise dissimilarity between feature vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in $\mathcal{R}^p$, $d_{ij} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|$. If all the vectors are distinct, no $d_{ij} = 0$. But when there are duplicate vectors, say $\boldsymbol{x}_s = \boldsymbol{x}_t$, then $d(\boldsymbol{x}_s, \boldsymbol{x}_t) = 0$; we call this a *legitimate zero*. In either case D is always a complete, symmetric matrix of distances with no missing values. But for other types of (dis)similarity data, $d_{ij} = d(o_i, o_j)$ may not be complete, and may not be symmetric, $d_{ij} \neq d_{ji}$. For example, Sampson's monastery data [1] is asymmetric and incomplete. Breiger et al. [2] give the relationship from Bonhaven to Ambrose the value 2 in Sampson's data, but the value from Ambrose to Bonhaven in the opposite direction is 1. Some relationships are "missing" in this real data. According to Wasserman and Faust [3], this is the most common form of social network data. In what follows we call such data "incomplete." The Karate club data is another example of social network data with missing values [4]. This famous data set is a symmetric social network that links 34 members of a university Karate club. There are 156 links, and 1,000 missing values.

A relation matrix may have legitimate zeroes in it if derived from non-distinct feature vectors, or because the relationship has actual measured values of zero. But when data are missing from a relation, it is a mistake to simply replace the missing values with zeros in order to make an algorithm run. This is the case we are interested in: when entries of R are missing, is there a reasonable way to impute values that improve the analysis of cluster substructure in the data?

In this article, we will examine the how to minimise the effect of missing relational values on determining the clustering tendency. Particularly, we will examine its effect on the iVAT [5] method of visual assessment and examine how we can determine the quality of the visualisation after taking into account the missing values.

Previous work has shown that we have the choice of removing the objects with missing values, imputing the missing values, or adjusting the algorithm to suit missing values. Removing objects with associated missing values is generally a bad idea, since we are also removing information that will assist us. For our case, we have the relational matrix D, therefore removing missing values would lead to removing all distances to any object with a missing relation.

There is not a clear distinction between imputation and algorithm adjustment, since each method aims at using the known data to minimise the effect of missing values. A comparison of k-means and Fuzzy c-means clustering, with various adjustments for missing feature values, are made in [6], [7]. The distance metrics and centroid computation are altered to allow for missing feature values. These methods are not suitable for our case, as we begin with the set of relations, and have no knowledge of the object features.

A comparison of imputation methods for clustering are presented in [8]. The imputation consists of regression and dimension reduction methods, where the known object feature values are used to estimate the missing feature values. The imputation is performed on missing feature values

- L.A.F. Park is with the School of Computing, Engineering and Mathematics, Western Sydney University, Rydalmere, NSW 2116, Austraila.
E-mail: lapark@scem.westernsydney.edu.au.
- J.C. Bezdek, C. Leckie, R. Kotagiri, and J. Bailey are with the Department of Computing and Information Systems, University of Melbourne, Parkville, VIC 3010, Australia. E-mail: jcbezdek@gmail.com, {caleckie, baileyj}@unimelb.edu.au, rao@csse.unimelb.edu.au.
- M. Palaniswami is with the Department of Electrical and Electronic Engineering, University of Melbourne, Parkville, VIC 3010, Australia.
E-mail: palani@unimelb.edu.au.

Manuscript received 13 Apr. 2015; revised 16 June 2016; accepted 31 Aug. 2016. Date of publication 0 . 0000; date of current version 0 . 0000.
Recommended for acceptance by L. Khan.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2016.2608821

where the feature vectors exist in a multidimensional real space. Unfortunately, iVAT requires imputation of relational values, not feature values. Therefore different assumptions must be made about the missing values and we have the added constraint that each imputed value is positive. We show later in the article that regression is not appropriate for relational data.

The representation of a graph is similar to the relational matrix D used by iVAT (where the graph contains similarities rather than dissimilarities), therefore imputation of edge weights for graph clustering may be useful for iVAT. Graph clustering with missing edge weights is considered in [9]. The clustering is performed using a convex program that takes into account the uncertainty of the missing values. Therefore, this work cannot be applied to iVAT (which uses a dedicated reordering algorithm). Imputation of edge weights is examined in [10], where missing values are imputed with either a sample from a uniform random variable, or the expected value from the known data.

This survey of the literature shows that our problem is unique in that we do not want an imputation method to determine missing values or to provide accurate clustering, but we want an imputation method that will assist iVAT in visually presenting the correct number of clusters in the data. Our task is to identify an imputation method for relational data, such that we maximise the iVAT image dependency on the known values, and minimise its dependency on the imputed values. Our research question is "Can we exploit the known values in D to provide imputation for iVAT?" To answer this question, we must explore methods of imputation of dissimilarities, and also examine the effect of the proportion of missing values on the quality of iVAT images.

We provide the following contributions:

- An investigation of the effect of various imputation strategies on the quality of asiVAT images (Section 4).
- The introduction of the use of bias and variance in visual assessment of cluster tendency (Section 4), and the concept of the iVAT summary image to visually assess the variance associated to an imputation technique (Sections 4, 5 and 6).
- An analysis of the cluster tendency of the Sampson and Karate data (Sections 5 and 6).

The article will proceed as follows: Section 2 provides a description of the iVAT and asiVAT algorithms for visualising cluster tendency, and discusses how we will begin investigating imputation strategies. Section 3 provides details of the set of imputation methods that we will investigate. Section 4 describes the how the concepts of bias and variance will assist us in determining the utility of an imputation method for iVAT, and examines the bias and variance associated with each of the imputation methods. Section 5 examines the cluster tendency of the Sampson data using asiVAT with imputation. Finally, Section 6 provides details of the cluster tendency for the Karate data using imputed asiVAT.

2 Background on Visual Clustering Tendency Analysis

```
Require: $D \in \mathbb{R}^{n\times n}$, $d_{ij} \geq 0$, $d_{ii} = 0$
procedure ASIVAT($D$)
    $D \leftarrow (D + D^T)/2$
    ## Compute: VAT matrix $D^\star$
    $K \leftarrow \{1, \ldots, n\}$
    Initialise: $I \leftarrow J \leftarrow \varnothing$; $P \leftarrow \{0, \ldots, 0\}$
    Select $(i, j) \leftarrow \arg\min_{p\in K, q\in K} (d_{pq})$
    $P(1) \leftarrow i$; $I \leftarrow i$; $J \leftarrow K\backslash\{i\}$
    for $r \leftarrow 2$ to $n$ do
        Select $(i, j) \leftarrow \arg\min_{p\in I, q\in J} (d_{pq})$
        $P(r) \leftarrow j$; $I \leftarrow I \cup j$; $J \leftarrow J\backslash\{j\}$
    end for
    $D^\star \leftarrow D_{PP}$ ## Permute rows and columns
    ## Compute: iVAT matrix $D'^\star$
    Initialise: $D'^\star \leftarrow [0]^{n\times n}$
    for $r \leftarrow 2$ to $n$ do
        Select $j \leftarrow \arg\min_{k\in\{1,2,\ldots,r-1\}} (D^\star_{rk})$
        $D'^\star_{rc} \leftarrow D'^\star_{cr} \leftarrow D^\star_{rc}$ where $c = j$
        $D'^\star_{rc} \leftarrow D'^\star_{cr} \leftarrow \max(D'^\star_{rj}, D'^\star_{jc})$,
                where $c = \{1, \ldots, r-1\}\backslash\{j\}$
    end for
    return $D'^\star$
end procedure
```

Fig. 1. The asiVAT algorithm.

The idea of a visual representation of the rows and/or columns of D to reveal structural relationships between pairs of objects began with Loua [11]. Visualisation of relationships in D by examining *reordered dissimilarity images* (RDIs) was first discussed by Czekanoski [12]. Wilkinson and Friendly [13] call the RDI a "cluster heat map". These authors give a good account of the state of the art in 2009, which includes an estimate that this method of data visualisation has appeared in more than 4,000 papers in the last decade.

Let D be a set of dissimilarity data, $d_{ii} = 0$ for all i, and $D = D^T$. The *visual assessment of tendency* (VAT, [14]) model reorders D to $D^\star$ using the ordering indices of a minimal spanning tree on D, and then displays a grayscale image $I(D^\star)$ of $D^\star$. Each element on the diagonal is zero (black). Off the diagonal, the scaled values range from 0 to 1 (white).

The basic rationale for VAT is that if an object tends to cluster with other objects, then it should also be part of a submatrix of "similarly small" values corresponding to those objects. These submatrices are seen as dark blocks along the diagonal of the VAT image $I(D^\star)$. Contrast can be improved by setting the diagonal to the minimum of the off-diagonal values. For an application of VAT in security administration, see Zhang et al. [15], where VAT is the basis of a product called RoleVAT, an engineering tool for role based access control.

Improved VAT (iVAT, [5]) begins by transforming D to $D' = f(D)$, where f is the feature extraction operation that replaces each d_{ij} in D with the geodesic distance d'_{ij}, followed by VAT reordering of D' to $D'^\star$. The iVAT image $I(D'^\star)$ of many data sets visually represents potential cluster structure in the data much more clearly than VAT images do.

The requirement that $D = D^T$ limits the utility of the VAT/iVAT algorithms. Recently, Havens et al. [16] extended these models to the asymmetric case with asiVAT (asymmetric VAT). The basis of this extension is to transform D to $(D + D^T)/2$ before constructing $D'^\star$. Fig. 1 gives pseudocode for asiVAT, which reduces to iVAT when $D = D^T$

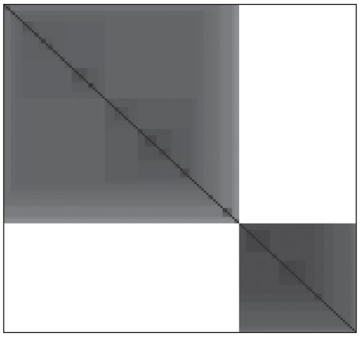


Fig. 2. The iVAT image of the Iris data.

2.1 Selection of an Imputation Method

There are any number of imputation schemes that might be useful. If we test different schemes on a data set for which the results are unknown, there is no way to judge whether a particular method is or is not performing well. In other words, we need some sort of "ground truth" for the imputation methods, so we will process a data set for which the expected results are pretty well known. The test data we will use is the Iris data, collected by Anderson in 1935 [17], and subsequently made famous by Fisher in 1936 [18]. Iris comprises $n = 150$ feature vectors in $p = 4$ dimensions. Each vector in Iris has one of three physical labels corresponding to one of three Iris subspecies: *setosa*, *versicolor*, or *virginica*. The number of clusters in Iris is usually declared as either 2 or 3 depending on the model used to define clusters. Most models identify $c = 2$ clusters in Iris. We begin by calculating the symmetric Euclidean distance matrix D_E. D_E will contain one legitimate pair of zeroes, because $x_{102} = x_{143}$, so $d(\boldsymbol{x}_{102}, \boldsymbol{x}_{143}) = d(\boldsymbol{x}_{143}, \boldsymbol{x}_{102}) = 0$.

Fig. 2 is the iVAT image $I(D_E)$ of D_E. The two dark diagonal blocks of this image clearly suggest that the primary structure in Iris is two clusters. This image comprises the visual ground truth for our imputation experiments. Our plan is to delete randomly selected values in Iris, use the remaining values to impute an approximation $\hat{D}_E$ to D_E, display the iVAT image $I(\hat{D}_E)$ of $\hat{D}_E$, and compare it visually to Fig. 2. (When $\hat{D}_E$ is asymmetric, we will use asiVAT.) Different imputation methods will produce different $\hat{D}_E$s, and thus, different images. We will use both visual and statistical analysis of a number of trials of the basic experiment to select the best imputation method. The Iris experiments can be thought of as a training session for the imputation methods, since we use contaminated versions of this complete data set to select the best imputation method. After we settle the issue of which imputation method to use, we will apply it to the incomplete monastery and Karate club data sets, and compare the resultant images and results to several previous studies.

Note also that all experiments performed on the Iris data were also performed on simulated data containing varying numbers of multidimensional Gaussian clusters (providing iVAT images with varying numbers of blocks along the diagonal). The findings from these experiments were the same as those from the Iris data and so were omitted from this article.

Now we can state the objective of the current study: when D is incomplete, can we impute values for the "missing" values that make the transformed image, say $I(D^{\star\star})$, more useful for assessment of cluster tendency in D?

In the following sections, we use the notation D for the set of known dissimilarity values, $d_{ij} \in D$ for the known dissimilarity between objects o_i and o_j, Δ for the set of unknown dissimilarity values, and $\delta_{ij} \in \Delta$ for the unknown dissimilarity between objects o_i and o_j.

2.2 Related Work

This work concerns with the imputation of missing relations to provide visually correct images when using the many forms of VAT, and therefore only operates on the non-negative domain of features and responses. There has been no prior work on the imputation of relational data for the visualisation that is required by iVAT. Previous work has been proposed with similar but different ideas to those presented in this article, but as stated in the introduction those methods are intended for object data, not relational data, and are used to preserve a given set of statistics rather than provide an appropriate visualisation.

The EM algorithm [19], [20] is used to compute model parameters from incomplete data. The method iteratively computes the expected value of missing information, then uses the known and computed values to obtain maximum likelihood parameters of the model. For our application, it is not clear what form the likelihood function should take, since our goal is the visualisation of the number of clusters. One possibility is to compute the likelihood of an iVAT image based on the clarity of the clusters, but this implies that a parameter exists that controls the clarity. Further work is required to identify the usefulness of the EM algorithm for visualisation.

Imputation using IVEWARE [21] (algorithm in [22]) is the same as our Simple Linear Regression Imputation, but the imputation of the ith variable uses the imputed values of the 1st to $i - 1$th variables, and then re-estimate once all imputed values are computed. IRMI [22] builds upon IVEWARE by first imputing all missing values with approximate values, then iteratively recomputing all missing values until a convergence rule is satisfied. Both of these methods are based on GLM modelling, therefore not appropriate for our imputation of relations (as described in Section 3.2.1).

The MICE algorithm [23] is a framework for imputation that uses Gibbs sampling to compute a distribution over the imputed values for inspection and provides one of the samples as the imputation. The initial inspection process is similar to our computation of the iVAT summary image (Section 4.3), where instead of computing the next imputation from the last, we reinitialise the imputed values, and allow for a burn-in period for each sample. Visualisation of the similarity of the resulting imputations then allow us to decide if there is an appropriate imputation, and where to select it from (using the modes of the summary iVAT image).

Predictive Mean Matching (PMM) [24] is a popular imputation method for the MICE framework. Using this method, regression is performed over the known variables to predict values for missing variables, but instead of imputing using the predicted value, we impute using the known data value with the smallest residual. This ensures that the imputed value remains within the domain of the known values. The concepts from PMM can be adapted to our Kernel regression method, which we will explore in future work.

The code used to compute the iVAT and asiVAT images, as well as the imputation methods can be downloaded from the authors Web site.[1]

3 Imputing Dissimilarities

In this section we will examine three methods of imputing the missing dissimilarity values in D: 1) sampling from a given distribution, 2) prediction from regressing over the known values, and 3) a combination of sampling and regression.

3.1 Imputation by Sampling

There are many ways to impute values for missing data in D. Perhaps the simplest scheme is to interpolate linearly between the missing elements of D. More sophisticated (and better) schemes use random samples drawn from an assumed or estimated probability distribution. In this section, we will examine uniform and bootstrap sampling for imputation.

We do not assume that D is symmetric, but we do assume that it is hollow ($d_{ii} = 0$ for all i) and that $d_{ij} \geq 0$ for pairs (i, j) such that d_{ij} is an observed input value. Since D is a dissimilarity matrix, we will require imputed values δ_{ij} to be greater than or equal to zero. Each of these sampling methods have complexity $O(1)$ with respect to the data.

3.1.1 Uniform Imputation

If we assume that all values within the range of the known dissimilarity values d_{ij} are equally likely candidates for the missing values δ_{ij}, we take samples from the uniform distribution, bounded by the minimum and maximum known dissimilarity values.

The imputed dissimilarity of objects o_i and o_j is computed as:

$$\delta_{ij} \sim \mathrm{Uniform}\left(\min_{d_{ij}\in D}(d_{ij}), \max_{d_{ij}\in D}(d_{ij})\right). \tag{1}$$

For example, consider the dissimilarity matrix:

$$D = \begin{bmatrix} 0 & 1 & 2 & 2 \\ 2 & 0 & 1 & - \\ 1 & 2 & 0 & 1 \\ 2 & 2 & 2 & 0 \end{bmatrix}, \tag{2}$$

where the dash '-' denotes a missing value. The minimum and maximum known dissimilarities are 0 and 2 respectively, therefore, we impute the missing value by taking a random sample from the distribution $\mathrm{Uniform}(0, 2)$, giving us, for example, 1.13.

This method of sampling is simple and does not depend on the objects o_i and o_j. Unfortunately, it uses only the extreme values of the known dissimilarities, and therefore can be quite adversely affected by outliers.

3.1.2 Bootstrapped Imputation

Uniform imputation assumes (somewhat naively) that the distribution of dissimilarities is uniform. Rather than assuming a priori any particular form for the sampling distribution, we can use the known dissimilarities to construct an approximation to the dissimilarity distribution, and sample from this distribution. This process is called bootstrapping, since we are sampling from a sample to obtain further information about its distribution [25]. The imputed dissimilarity of objects o_i and o_j is computed as:

$$\delta_{ij} \sim \mathrm{Bootstrap}(D), \tag{3}$$

where the bootstrap process randomly selects an element of D, where each element has equal probability of being selected, to return as the sample.

For example, consider the dissimilarity matrix in (2). Using the known dissimilarity values, we can construct the frequency table and proportion estimates of each dissimilarity:

d_{ij}	0	1	2
Frequency	4	4	7
Sample Proportion	0.267	0.267	0.467

Then, for each missing value, we take a random sample from the set $\{0, 1, 2\}$ using the sample proportions as the probability for each random draw. For example, if we need to impute five missing values, a sample of size five might produce the imputed values 2,1,2,0,0.

This method is independent of the objects o_i and o_j, but is preferred to uniform imputation since it takes into account the distribution of the known dissimilarities.

3.2 Imputation by Regression

The two sampling methods proposed in Section 3.1 compute the imputed value δ_{ij} where δ_{ij} is not conditionally dependent on i and j. This implies that if we permute the set of imputed values, we will not change the likelihood of the resulting imputation, since they are not dependent on their position. The dissimilarity δ_{ij} is the dissimilarity between objects o_i and o_j, and so should be dependent on the values of i and j. In this section, we will examine two regression methods that depend on transitive connections of i and j through intermediate paths between them to compute the unknown dissimilarity δ_{ij}.

3.2.1 Simple Linear Regression Imputation

To compute the unknown dissimilarity δ_{ij} between object pair (i, j), we can make use of the known dissimilarities between the object pairs (i, k) and (k, j). In fact, if we take the vector $\boldsymbol{d}_{i\cdot}$ containing the set of known dissimilarities between o_i and the remaining objects, and find another vector $\boldsymbol{d}_{k\cdot}$ that is approximately equal to $\boldsymbol{d}_{i\cdot}$, then we would expect d_{ij} to approximately equal to d_{kj}.

Using this concept, we can build a linear model, fitting $\boldsymbol{d}_{k\cdot}^{(j)}$ to the response d_{kj}, where $\boldsymbol{d}_{k\cdot}^{(j)}$ is the vector $\boldsymbol{d}_{k\cdot}$, with the jth element removed. This allows us to predict the missing value δ_{ij} using the known values $\boldsymbol{d}_{i\cdot}^{(j)}$:

$$d_{ij} = \langle \boldsymbol{d}_{i\cdot}^{(j)}, \boldsymbol{\beta}^{(j)} \rangle + \epsilon, \tag{4}$$

where $\boldsymbol{d}_{i\cdot}^{(j)}$ is the vector of dissimilarities from object i to the other $n-1$ indices, excluding the dissimilarity between i and j, $\boldsymbol{\beta}^{(j)}$ is the vector of regression coefficients (fitted using the existing relationships between $\boldsymbol{d}_{k\cdot}^{(j)}$ and d_{kj}), $\langle *, * \rangle$ is the

[1] http://www.scem.westernsydney.edu.au/~lapark/impVAT

Euclidean inner product of the two vectors, and ϵ is Normally distributed with constant variance.

For example, using the dissimilarity matrix in (2), we treat the column with the missing value as the dependent variable that we will regress over, and the remaining columns as the independent variables, so that each row is a regression observation. This gives us the simple linear regression:

$$\begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 2 & 0 \\ 2 & 2 & 2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}. \tag{5}$$

Using least squares, we compute $\boldsymbol{\beta}^{(j)} = [-1.67, 1.33, 0.33]$. The value of δ_{ij} is then given by $\langle \boldsymbol{d}_{i\cdot}^{(j)}, \boldsymbol{\beta}^{(j)} \rangle$, giving us $\delta_{ij} = -3$.

But all the entries in D must be non-negative, so this result cannot be used. We can use instead a Generalized Linear Model, with a link function that maps the $(-\infty, \infty)$ domain to $(0, \infty)$, but unfortunately, the occurrence of zeros in the distance matrix causes problems with the model fitting. We could also use the Tobit model, forcing all negative values to zero. But obtaining a negative response from a linear model, where all fitted responses were non-negative, shows that there is a problem with the prediction.

We should also note that fitting a regression model implies that some of the variables have a higher correlation to the response than others, which would be the case for most imputation problems. In our case our data are dissimilarities (all having the same scale, and can be thought of as all coming from the same distribution), therefore information about one given dissimilarity should not be more important than information about another given dissimilarity. There may be sample correlation between dissimilarities related to a specific object, but this correlation may not be present in the population. Instead of fitting a regression model, we can take advantage of the geometric properties of dissimilarities, and hence move on to the more appropriate Kernel Regression.

3.2.2 Kernel Regression Imputation (KR)

Simple linear regression imputes the missing value δ_{ij} based on the similarity of other index pairs to (i, j), but it assumes that the relationship between the dissimilarities is linear. Kernel regression (or kernel smoothing) [26, chap. 6] regresses over a value by computing the weighted sum of the other dissimilarities, where the weight is based on the dissimilarity. Kernel regression is computed as

$$\delta_{ij} = \frac{\sum_{k \neq i} K(\boldsymbol{d}_{i\cdot}, \boldsymbol{d}_{k\cdot}) d_{kj}}{\sum_{k \neq i} K(\boldsymbol{d}_{i\cdot}, \boldsymbol{d}_{k\cdot})}, \tag{6}$$

where the kernel $K(\cdot, \cdot)$ is chosen based on the dissimilarity space. If the dissimilarity space is Euclidean, the Gaussian kernel is an appropriate choice:

$$K(\boldsymbol{x}, \boldsymbol{y}) = \exp\big(-\gamma \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \big), \tag{7}$$

where $\gamma \in \mathbb{R}^+$.

For example, for the dissimilarity matrix in (2), we treat the column with the missing value as the dependent variable that we will regress over, and the remaining columns as the independent variables, where each row is a regression observation. If we let the Gaussian kernel parameter $\gamma = 1$:

$$\begin{aligned} \delta_{2,4} &= \frac{K(\boldsymbol{d}_{2\cdot}, \boldsymbol{d}_{1\cdot}) d_{14} + K(\boldsymbol{d}_{2\cdot}, \boldsymbol{d}_{3\cdot}) d_{34} + K(\boldsymbol{d}_{2\cdot}, \boldsymbol{d}_{4\cdot}) d_{44}}{K(\boldsymbol{d}_{2\cdot}, \boldsymbol{d}_{1\cdot}) + K(\boldsymbol{d}_{2\cdot}, \boldsymbol{d}_{3\cdot}) + K(\boldsymbol{d}_{2\cdot}, \boldsymbol{d}_{4\cdot})} \\ &= \frac{0.086 \times 2 + 0.086 \times 1 + 0.107 \times 0}{0.086 + 0.086 + 0.107} = 0.927. \end{aligned}$$

This is a reasonable value for the range $[0, 2]$ of recorded values in D. To see the effect of changes to the estimate as the kernel parameter gamma changes, we impute the required value for $\gamma = 0.1$, 0.5 and 2.0 and 5.0. The estimates for these choices are 0.993, 0.964, 0.849 and 0.611. Any of these values seems reasonable for the problem at hand. Choosing gamma in $(0, 1]$ results in pretty similar values, but the range for gamma should probably depend on the range of observed values in D. Note that Kernel regression is a set of weighted inner products and so has complexity $O(N^2)$, where N is the number of objects.

3.3 Imputation by Bootstrapped Regression

Kernel regression imputation as described in Section 3.2.2 is useful if D contains only a few missing values. If there are quite a few unknown dissimilarities, we may not have enough information to perform a statistically significant regression. For example, if we have the dissimilarity matrix:

$$D = \begin{bmatrix} 0 & 1 & - & 2 \\ 2 & 0 & 1 & - \\ 1 & - & 0 & 1 \\ - & 2 & 2 & 0 \end{bmatrix}, \tag{8}$$

simple kernel regression requires us to impute the missing values one at a time. If we begin with the missing dissimilarity in the second row, we must be able to compare the second row with the other rows. But the other rows also contain missing values, so some of the weights needed for Equation (6) are not available. To avoid this problem, we propose using *bootstrapped kernel regression* (KR Boot):

1) First initialise the missing values using bootstrap imputation (Section 3.1.2), resulting in a first estimate $\hat{D}$ for D; and then
2) Impute each of the missing values in D using $\hat{D}$ as a placebo for D with the imputation method of choice.

3.4 Imputation by Iterated Bootstrapped Regression

To refine the imputed values further, we can repeat the regression multiple times, using the imputed values from the previous regression as the initial imputed values.

1) First initialise the missing values using bootstrap imputation (Section 3.1.2), resulting in a first estimate $\hat{D}_0$ for D; and then
2) for a in 1 to k:
   a) Impute each of the missing values in $\hat{D}_a$ using $\hat{D}_{a-1}$ as a placebo for D with the imputation method of choice.
3) Use $\hat{D}_k$ as the best estimate of D

It is not clear that iterated regression will improve the utility of the imputed values, so we will examine this

TABLE 1
The Mean and Standard Deviation RMS Differences between Imputed and Actual Values from the Contaminated Iris Data Over 100 Random Trials

Method	$m=5$		$m=10$		$m=50$		$m=100$		$m=500$		$m=1{,}000$	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Uniform	2.689	(0.7391)	2.726	(0.5494)	2.872	(0.2614)	2.818	(0.1762)	2.80	(0.0671)	2.799	(0.057)
Bootstrap	2.236	(0.6037)	2.332	(0.4239)	2.315	(0.2018)	2.318	(0.1315)	2.31	(0.0625)	2.311	(0.047)
KR	0.213	(0.0788)	0.209	(0.0541)	0.249	(0.0302)	0.290	(0.0314)	1.34	(0.4102)	2.211	(0.324)
KR Boot	0.704	(0.1939)	0.689	(0.1276)	0.673	(0.0626)	0.615	(0.0429)	0.42	(0.0221)	0.364	(0.018)

The RMS values are computed for 5, 10, 50, 100, 500, and 1,000 randomly chosen missing values in the dissimilarity matrix D_E.

further in the experiments section. We specifically use Kernel Regression, to obtain *Iterated Bootstrapped Kernel Regression* (IBKR).

4 Bias and Variance of Imputed asiVAT Images

For an imputation method to be useful, the asiVAT images generated after imputation must have low bias and variance. The images after imputation must have low variance, in that the location of the missing values should have little effect on the asiVAT image, and they must have low bias, meaning that the mean asiVAT image should provide us with an indication of the (nominally) correct number of clusters.

If an imputation method leads to low bias but high variance, then for a given data set, we would have low confidence in the accuracy of the imputed asiVAT image. If the imputation method leads to low variance but large bias, then we would have confidence that the imputed asiVAT image is *not* correct. Therefore it is important that both of these attributes are minimal.

To examine the bias and variance of the imputed asiVAT images, we require a data set with no missing values and a "known" number of clusters. We can then randomly select and remove dissimilarities, perform the imputation and compute the asiVAT image. By repeating this process, we will obtain a distribution of asiVAT images, that can be examined for bias and variance.

To perform our analysis, we return to the Iris data. The relational matrix is again constructed by computing the euclidean distance between all 150 object pairs, providing us with the 150×150 matrix D_E as above. We then randomly remove off-diagonal elements (i, j) of D_E and their counterparts (j, i) (due to the Iris distance being symmetric) to obtain m MAR missing values. An experiment of this kind allows us to assess the accuracy of our imputation methods because we know the exact values that are "missing," and we think we know the number of clusters (which we presume to be 2) within the data.

For each experiment, we will examine the four imputation methods: uniform, bootstrap, kernel regression, and bootstrapped kernel regression, and examine how they behave as the number of missing values grows.

We choose the kernel parameter $\gamma = (2ns^2)^{-1}$ for kernel regression, where n is the row length and s is the standard deviation of the known values in D before imputation. This choice allows us to normalise the distance between vectors by the length of the vector n and the expected distance between each dissimilarity and the mean dissimilarity (s^2), removing the dependence of the function on the size of D and the magnitude of the elements in D.

We set the kernel parameter differently for bootstrapped kernel regression. After using bootstrap initialisation, we have many values in D that are estimates, therefore we have lower confidence in these values. When using the Gaussian kernel to compute the similarity of row j to row i, if there are only a few bootstrapped values in $\boldsymbol{d}_{j\cdot}$, we can be reasonably confident that the computed similarity is accurate. If row j has many missing points and hence many bootstrapped values, then the similarity computed using the kernel may not be accurate. To account for this, we set the kernel regression parameter to a vector, whose jth element is

$$\gamma_j = \frac{m_j + 1}{2ns^2}, \tag{9}$$

where m_j is the number of missing values for row j. By weighting each element of the vector, we are providing a weighted form of feature selection, that puts more emphasis on the Kernel regression model variables that have been fitted with less imputed values. These weights also act as a form of regularisation that reduces the risk of over-fitting the model to the data.

4.1 Accuracy of Imputed Values

Before we look at the iVAT images generated using each imputation method, we will first examine how well each imputation method is able to predict the missing values. Using the Iris data, we randomly removed 5, 10, 50, 100, 500 and 1,000 values from the matrix D_E and then imputed these values using each of the four methods. The Root Mean Square (RMS) difference between the imputed and true values was then measured and recorded. This process was repeated 100 times to examine the variation in the RMS values. The mean and standard deviation RMS values are presented in Table 1.

The uniform and bootstrap imputation methods provide an approximately constant mean error as the number of missing values increases, with a decreasing standard deviation. The Bootstrap mean error is lower than uniform, most likely due to the bootstrap imputation using the distribution of the data, while uniform does not.

Kernel regression and bootstrapped kernel regression both begin with a lower RMS error. As the number of missing values increases, the mean RMS error of kernel regression approaches the those of bootstrap and uniform. The mean RMS error of bootstrapped kernel regression decreases as the number of missing values increases. The reduction in error is likely to be due to bootstrapped kernel regression making use of all known data in D, while kernel regression can use only a small fraction when there is a large number of missing values.

4.2 Bias in Imputed asiVAT Images

Our next set of experiments evaluate the bias induced by each imputation method when generating the asiVAT image. We do this by examining how well each method of imputation leads to determining the number of clusters, on average, using an asiVAT image. We will examine the accuracy by examining the number of clusters suggested by the asiVAT image and we will examine the bias of each method relative to the number of missing values in D.

Bias in a distribution of asiVAT images can be determined by examining the mean of the set of asiVAT images. If we can determine the (apparently) correct number of clusters from the mean asiVAT image, it implies that the imputation method has induced little bias. But if the number of apparent clusters in the input data differs from the number suggested by the mean asiVAT image, then the imputation method has introduced a large amount of bias in the interpretation of data substructure.

To simulate this situation for detecting bias in the mean asiVAT image, we begin with the Iris data, and:

(1) Set the number of missing values m.
(2) Randomly select m positions in D_E and remove them, resulting in the depleted matrix $D_{E,m}$
(3) Impute the set of m missing values in $D_{E,m}$ to obtain D'_E.
(4) Generate the asiVAT image of D'_E

This process was repeated 100 times to obtain a distribution of asiVAT images for a given m and given imputation method. The mean image was then generated by obtaining the mean intensity at each pixel over the distribution of asiVAT images.

If low bias exists, the mean asiVAT image will show two clusters (which should be similar to the image of D_E in Fig. 2). If there is high bias, the mean image may not show two clusters. When this happens, the likelihood is high that the replacement of m values (on average) in D_E by imputed values alters the actual substructure that asiVAT enables us to visualise.

We split the bias experiments into two sets: the first set contains small numbers of missing values, and the second set contains large numbers of missing values.

4.2.1 Bias for a Small Number of Missing Values

The first set of experiments will compare the bias induced by each of the four imputation/asiVAT methods when a small fraction of the dissimilarities are missing.

We computed the mean asiVAT image over 100 trials using the uniform, bootstrap, kernel regression and bootstrapped kernel regression imputation methods. The images were generated for missing value counts of $m = 10$, 50, 100 and 500. Fig. 3 presents the mean asiVAT image from 100 trials. To put this experiment in perspective, 500 missing values in Fig. 3 corresponds to 2.22 percent of the uncontaminated input values in D_E.

Fig. 3 shows that all of the imputation methods induce low bias for 10 and 50 missing values, but only the kernel regression and bootstrapped kernel regression methods induce low bias for 100 and 500 missing values. Note that both bootstrapped kernel regression and kernel regression behaved similarly, implying that bootstrapping did not greatly affect these asiVAT images.

4.2.2 Bias for a Large Number of Missing Values

In this section, we will perform the same experiments as the previous section, but we increase the number of missing values to $m = 225$, 1,125, 2,250 and 4,500, which is equivalent to 1, 5, 10 and 20 percent of D_E. Due to the large number of missing values, we are unable to perform kernel regression, therefore the experiments in this section use the imputation methods uniform, bootstrap and bootstrapped kernel regression. Fig. 4 presents the mean asiVAT images for the KR boot method: the other methods did not perform well, and are not discussed further for the Iris data.

Only bootstrapped kernel regression provides little bias when there are a large number of symmetric missing values.

Since bootstrapped kernel regression imputation seems to produce fairly unbiased approximations, we test its limits by increasing the number of missing values. Fig. 4 shows the results of our experiments with KR boot ranging from 1 to 60 percent missing values.

We can see in Fig. 4 that bootstrapped kernel regression produces little bias for 30 percent missing values. For 40 percent missing values the cluster structure is still visible, but not clear. For 50 and 60 percent missing values, we find that the bias is too great to suggest the correct cluster block structure.

The results from these experiments show that *bootstrapped kernel regression induces the least bias amongst the set of imputation methods for all numbers of missing values*. We also found that if we have a very small number of missing values (less that, say 5 percent), then any of the proposed imputation methods is sufficient.

4.3 Variance in Imputed asiVAT Images

We found in the previous section that bootstrapped kernel regression induced the least bias in the imputed asiVAT image. In this section, we will examine the variance of the asiVAT image distribution, which identifies how similar a generated asiVAT image will be to the expected image.

The variation in imputed values caused by different missing value positions and random sampling causes variation in the generated asiVAT image. Therefore it is as if we are sampling an asiVAT image from a distribution of asiVAT images, conditioned on the known values in D and the imputation method.

When comparing imputation methods, a robust method will produce an asiVAT image distribution with low variance, so when we sample from the asiVAT distribution, we are more likely to get what we expect. A non-robust method will provide an asiVAT image distribution with high variance, so we will not know what to expect when sampling from the asiVAT image distribution (meaning there may be large changes to the asiVAT image each time we generate it, due to the imputation method not coping with the randomness of uniform or bootstrap sampling).

To examine the variance of the asiVAT image distribution, we take a sample of N asiVAT images using a given imputation method. Each of the N images is based on an

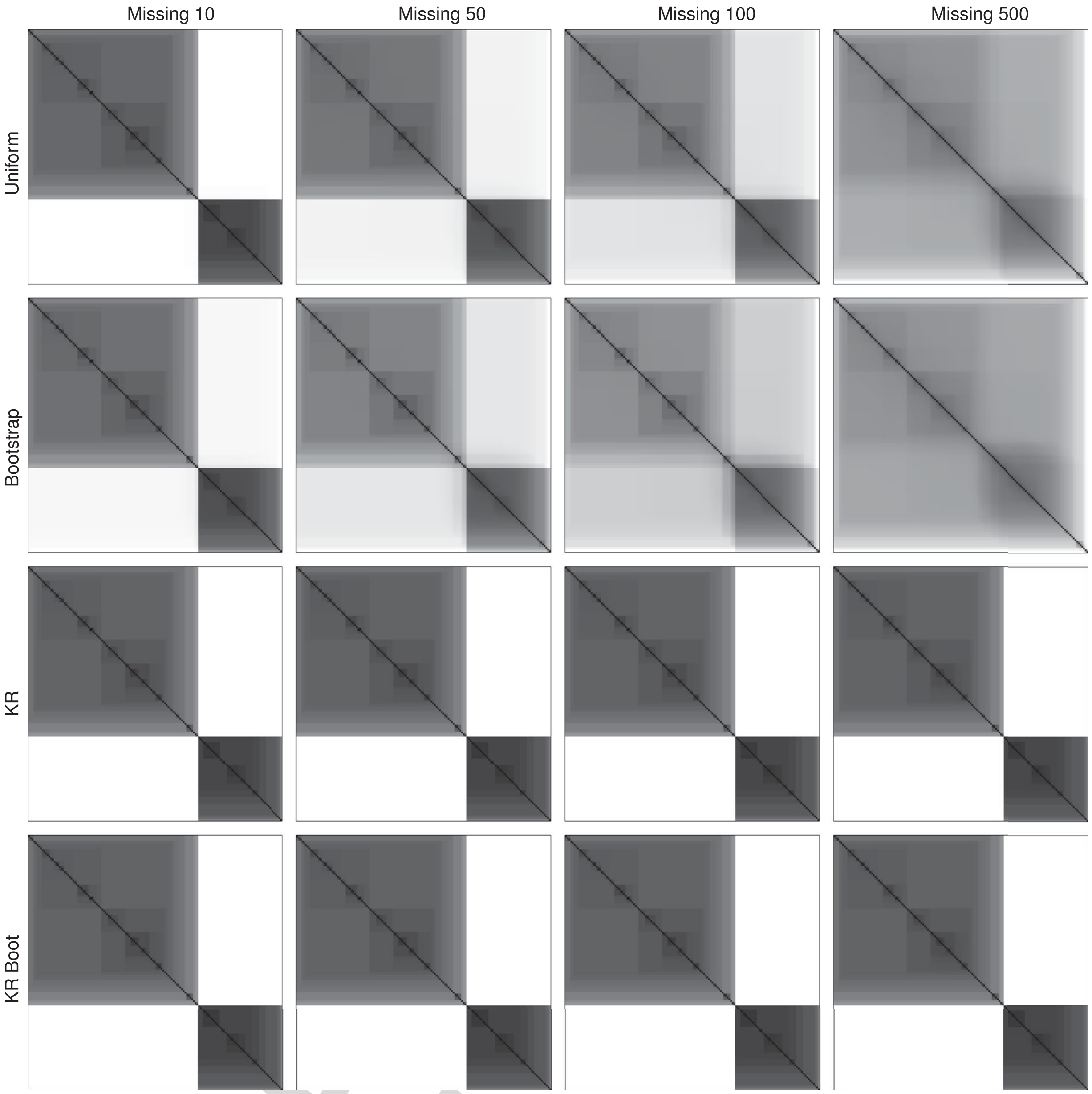


Fig. 3. Mean asiVAT images of 100 matrices on the Iris data for the four imputation methods. The columns show the results of having 10, 50, 100, and 500 missing values from D.

approximation of D_E by a $\hat{D}_E$. An imputation method with low variance will produce N approximations that are similar to each other, while a method with high variance will produce less similar approximations. Next, we describe how to compute an $N \times N$ matrix of correlation coefficients from the N imputation trials.

The asiVAT algorithm builds $\hat{D}_E$ by reordering the objects underlying D_E. In other words, iVAT permutes the n objects in D_E. To compute the difference between two asiVAT images, we compare the difference in the permutations of D_E to $\hat{D}_{Ei}$ and $\hat{D}_{Ej}$. A commonly used method of comparing permutations is Kendall's τ distance, which counts the discordant pairs between two rankings, and is also equal to the minimum number of adjacent transpositions required to transform one of the orderings of objects to the other.

For example, if we impute the missing value in matrix D in equation (2) as 0, we obtain the asiVAT matrix:

$$\mathrm{VAT}\left(\frac{D + D^T}{2}\right) = \begin{bmatrix} 0.0 & 1.0 & 2.0 & 1.5 \\ 1.0 & 0.0 & 1.5 & 1.5 \\ 2.0 & 1.5 & 0.0 & 1.5 \\ 1.5 & 1.5 & 1.5 & 0.0 \end{bmatrix}, \boldsymbol{p}_1 = \begin{bmatrix} 4 \\ 2 \\ 1 \\ 3 \end{bmatrix}, \quad (10)$$

where $\boldsymbol{p}_1$ is the permutation of the rows and columns. Instead, if we impute the missing value in matrix D in Equation (2) as 2, we obtain the asiVAT matrix:

$$\mathrm{VAT}\left(\frac{D + D^T}{2}\right) = \begin{bmatrix} 0.0 & 1.5 & 2.0 & 2.0 \\ 1.5 & 0.0 & 1.5 & 1.5 \\ 2.0 & 1.5 & 0.0 & 1.5 \\ 2.0 & 1.5 & 1.5 & 0.0 \end{bmatrix}, \boldsymbol{p}_2 = \begin{bmatrix} 4 \\ 3 \\ 1 \\ 2 \end{bmatrix}, \quad (11)$$

where $\boldsymbol{p}_2$ is the permutation of the rows and columns. We have left out the path based distance refinement of the asiVAT algorithm to simplify the example. When ignoring the imputed values, the resulting matrices from the asiVAT

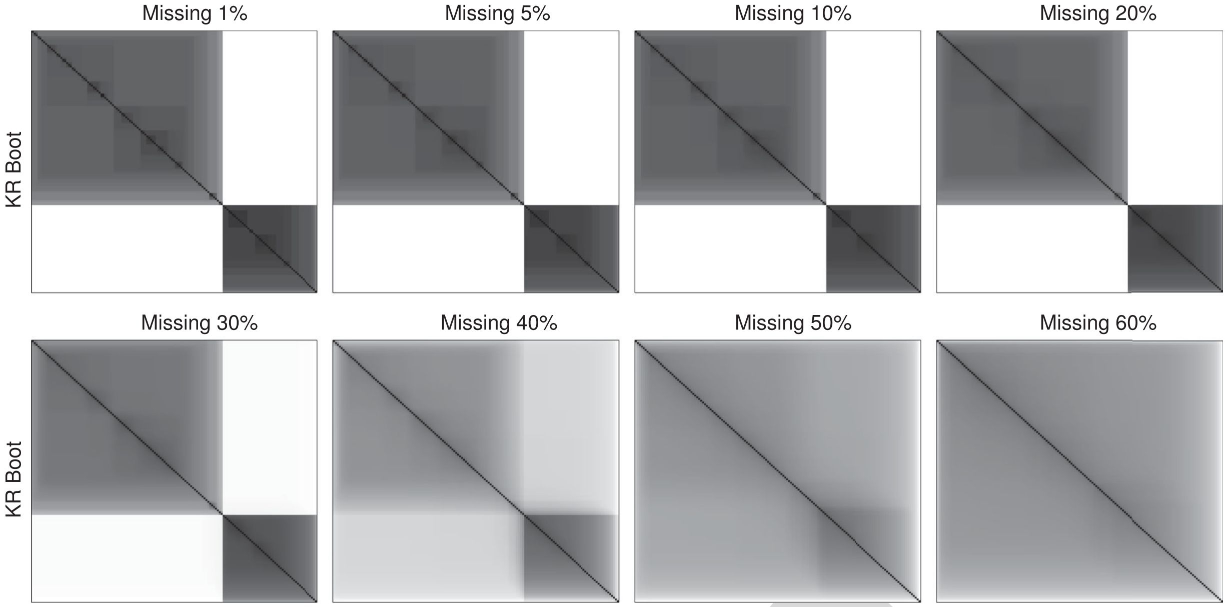


Fig. 4. Mean asiVAT images of 100 matrices on the Iris data for the KR Boot imputation method from 1 to 60 percent missing values.

algorithm are identical up to a permutation, where the permutation is shown by the vector $\boldsymbol{p}$. Therefore, to compare the two asiVAT matrices (images), we compute the Kendall's τ distance between the permutations.

$$\tau(\boldsymbol{p}_1, \boldsymbol{p}_2) = 3, \tag{12}$$

since the pairs (1,2), (2,3), (1,3) are discordant between the two orderings. Note that the domain of τ is $\{0, 1, \ldots, n(n-1)/2\}$, where n is the number of items being ordered.

Given a set of N asiVAT images, we compute Kendall's pairwise dissimilarity between all $N(N-1)/2$ asiVAT distance matrix pairs, to obtain a dissimilarity matrix D_N of distance approximations. *To visualise the similarity between the set of asiVAT images, we can view the iVAT image of the asiVAT dissimilarity matrix* D_N. If this summary iVAT image is all (or mostly all) black, the N asiVAT images are very similar to each other and hence the method is robust. If the image if D_N is grey to white, then there is little similarity between the asiVAT images, meaning that the method is not robust.

Summary iVAT images $I(D_N)$ using the uniform, bootstrap and bootstrapped kernel regression imputation methods on the Iris data, with 1 and 20 percent missing values are shown in Fig. 5. There is high correlation between the imputed asiVAT images for all methods when 1 percent of the values are missing (shown by the mostly black iVAT summary images). As we increase the percentage of missing values, the correlation decreases. At 20 percent missing values using uniform and bootstrap imputation, the dark blocks are quite a bit lighter, indicating loss of correlation. But when using KR Boot regression, the darkness fades only slightly. *This tells us that the asiVAT images generated using bootstrapped kernel regression are highly correlated and hence the method is robust and has low variance.*

The presence of many small blocks in an iVAT correlation image indicates that there are many small clusters of asiVAT images. This implies that we probably won't obtain similar images when different sets of missing value positions are provided.

If we have no information about the true asiVAT image (the image generated where no missing values exist), we can use the iVAT summary image of the asiVAT dissimilarity matrix to measure confidence in the generated image.

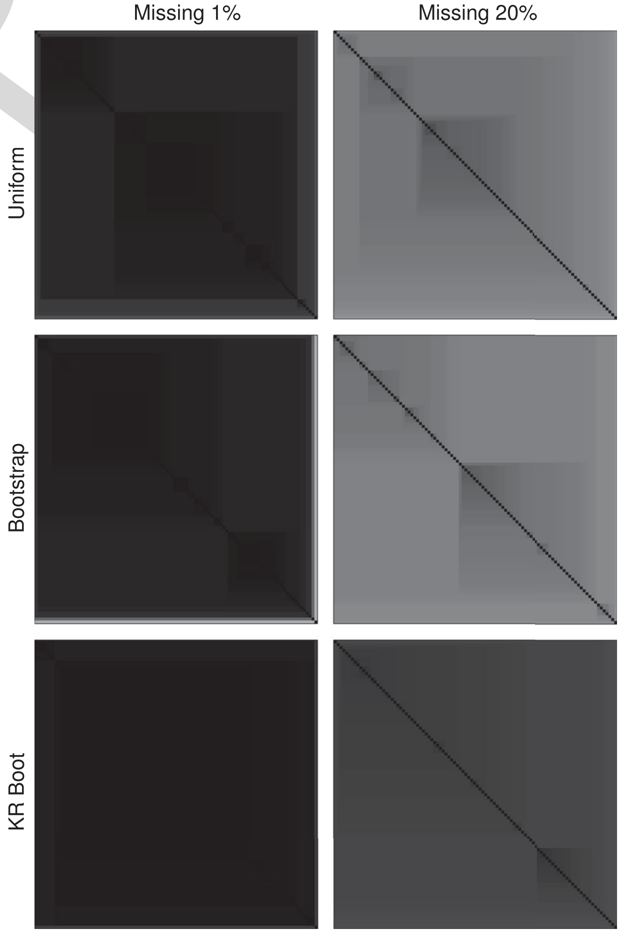


Fig. 5. Summary iVAT images $I(D_N)$ for $N = 100$ imputed asiVAT images on the Iris data.

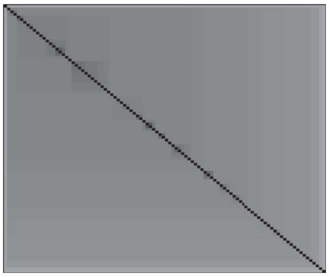


Fig. 6. iVAT image $I(D_N)$ for $N = 100$ imputed asiVAT images using KR Boot on Sampson's data.

High variance implies low confidence, while low variance implies high confidence.

At this point we are done with the Iris data, having used it to identify the best imputation method as KR boot. The examples in the next two sections will consider only this method.

5 Sampson's Monastery Data

Sampson's monastery data contains a set of measured relationships between 18 monks. Each monk rated the top and bottom three brother monks over 4 traits: like, esteem, influence and consistency. A typical question in the poll was "list in order the three brothers who you like the most." Sums over the four traits result in 104 non-zero and 204 missing values off-diagonal. Thus, 68 percent of the data are missing values because they were not collected. These are the values we will impute.

When examining the Iris data, we had two sources of variation, the position of the missing values (that were randomly removed) and the randomness from the imputation method. The Sampson data contains a predefined set of missing value locations, so in this section, any variance in asiVAT images will be due to the randomness in the imputation method only.

The monks were asked to list the top 3 other monks, meaning that only the greater relationships were recorded and not the lesser relationships, so the missing values are associated to lesser relationships. Therefore it would not make sense to use bootstrapped initialisation, since the bootstrapped values will be in the range of the known values. The recorded values in the data are from 5 (the highest), to 1 (the lowest). This implies that the missing relationships should be considered to be between 0 and 1 (less than the known relationships but not negative). Therefore, we initialise the missing values for bootstrapped kernel regression uniformly with values between 0 and 1.

5.1 Imputing Sampson's Data

There is no true asiVAT image for the Sampson data, since it contains missing values, Therefore, we are unable to examine the bias (since it requires comparison to the true asiVAT image). We can examine the variance of KR Boot on Sampson's data, since it only requires comparison to other images made by the same method. We used the process described in Section 4.3 to obtain the summary iVAT image $I(D_N)$ of the dissimilarity between asiVAT images for KR Boot. The resulting image is shown in Fig. 6.

The very light colour and many diagonal blocks in Fig. 6 indicate that KR boot produced many clusters of self-similar approximations to the Sampson Data. This implies that if we repeatedly run imputed asiVAT on the data, we will probably obtain different asiVAT images on each run, providing us with many different visual estimates of the number of clusters in the data, and hence no confidence in the results. This result agrees with our Iris data analysis, where each method had high variance when 60 percent of the dissimilarities were missing.

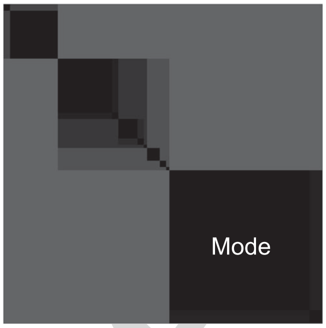


Fig. 7. The summary iVAT image $I(D_N)$ for $N = 100$ imputed asiVAT images obtained using IBKR on the Sampson data. The mode of the asiVAT image distribution corresponds to the largest dark block (lower right) on the diagonal.

5.2 Iterated Imputation

The previous section showed that KR Boot is not effective on Sampson's data, where the likely cause is the high proportion of missing values. In this section, we will examine the effect of iterated bootstrapped kernel regression (IBKR) on Sampson's data. IBKR updates the value of a missing value using the set of known values and the currently assigned values to all other missing values. Therefore imputation at each iteration is dependent on imputation from the previous iteration.[2]

When performing IBKR, it would be ideal if the asiVAT procedure converged. Our experiments indicate that IBKR iterations lead to either convergence or to a limit cycle which alternates between two asiVAT approximations. Faced with this dilemma, we ran IBKR until one of the two conditions emerged for each trial. When cycling occurred, we stopped the iteration at a randomly chosen cycle. We found that 50 iterations was sufficient for these conditions to be met. The resulting iVAT image $I(D_N)$ of $N = 100$ asiVAT dissimilarities is given in Fig. 7.

Fig. 7 has three (hard to see) dark, diagonal blocks. This indicates that IBKR produced three clusterings of very similar approximations over the 100 trials. The largest dark block, in the lower right portion of $I(D_N)$, represents half of the estimated matrices. Comparing Figs. 6 and 7 shows that IBKR estimates produce a much smaller variance than KB Boot estimates when there are a large number of missing values.

5.3 Mode of the asiVAT Image Distribution

The previous section shows that IBKR imputation results in an asiVAT distribution with small variance, but we did not describe how to select an asiVAT image from the distribution.

[2] See Breiger et al. [2] for a similar approach called CONCOR to iterated improvement of cluster analysis in Sampson's data. CONCOR attempts to improve the interpretation of structure in a correlation matrix (but without imputation of missing values) by repeated computation of correlation between the column vectors of the input matrix.

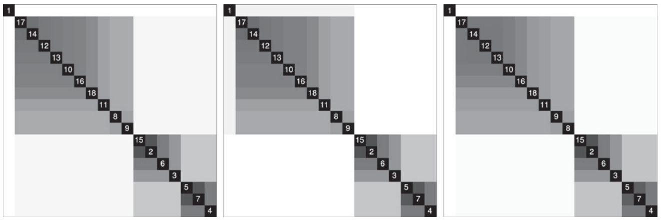


Fig. 8. Three *mode* asiVAT images of Sampson’s data from the largest dark block in Fig. 7.

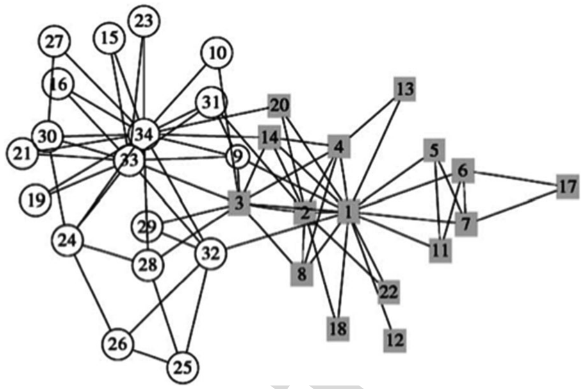


Fig. 10. Newman-Girvan clusters in G [28].

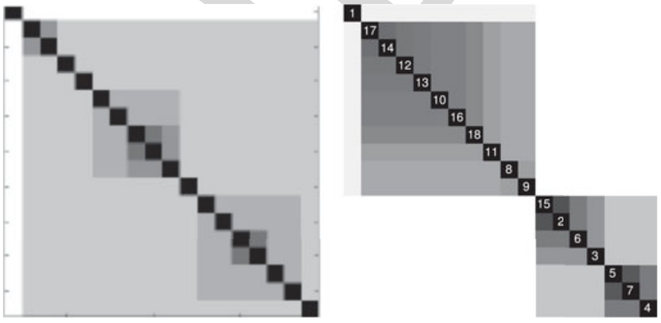
A general asiVAT image is used to estimate the number of clusters in a given data set. The subsets of $I(D_N)$ corresponding to its dark blocks suggest a certain number of clusters (of approximations to the input data matrix). Each of the blocks presumably contains very similar estimates. So, if we can identify a large set of images with low dissimilarity, we can examine them to obtain the suggested number of clusters in the data. We call the largest set of asiVAT approximations with low dissimilarity, the *mode approximation D_M to D, and $I(D_M)$ the mode asiVAT image of the data*.

To identify the number of clusters in Sampson’s data, we examine the set of asiVAT images associated with the mode in Fig. 7. All of the images were very similar, so a sample of three images taken from the mode are presented in Fig. 8. Each of these images show three main clusters, where the first contains one object, the second contains a smaller stronger cluster and the third contains two sub-clusters.

To assess the overall utility of our method, we compare it to several previously published results. First and foremost, Sampson used a combination of several analytical techniques to conclude that the 18 monks were best partitioned into 3 subsets: the “young turks” = {1, 2, 7, 12, 14, 15, 16}: the “loyal opposition” = {4, 5, 6, 8, 9, 10, 11, 13}; and the “outcasts” = {3, 17, 18}.

Brieger et al. [2] and White et al. [27] both get exactly the same conclusions using 3 block models of Sampson’s data. But Brieger et al. also provide several alternate interpretations of Sampson’s data that exhibit subtle changes to the original structure. Their statement in [2] provides some insight that justifies the finer analysis:

> *“the basic Sampson pattern, ..., involves two separate main cliques (the Young Turks and the Loyal Opposition), each possessing the familiar internal organisation of leaders and hangers-on; but there is also a peripheral group, the Outcasts, whose lack of received positive sentiment from the top blocks is analogous to that of the scapegoats”*



(a) Image from [16]. (b) Image using IBKR.

Fig. 9. Comparison of iVAT image in [16] to an IBKR image from Fig. 8 for Sampson’s data. The numbers on the diagonal refer to the monk ID from Sampson’s data.

White et al. [27] posit a slightly different partition, viz., “young turks” = {1, 2, 7, 12, 14, 15, 16}: the “loyal opposition” = {4, 5, 6, 8, 9, 10, 11}; and the “outcasts” = {3, 13, 17, 18}.

Some basic split into three groups appears in nearly every study of Sampson’s data. Fig. 9 compares the leftmost IBKR image from Fig. 8 to the asiVAT image of Sampson’s data that appears as Fig. 7d of [16]. The primary structure in both images is a loner pixel and a large block containing the other 17 monks. View 9b portrays the split of the monks into $c = 3$ clusters, but these three clusters reside within $c = 2$ blocks, suggesting that the primary structure is two clusters, $C_1 = \{1, 8\text{-}14, 16\text{-}18\}$ union $C_2 = \{2\text{-}7, 15\}$. You have to squint to see that monk 1 is inside a much larger block that contains the 10 other monks in C_1. The other larger block C_2 containing the remaining 7 monks supports a split into two clusters $\{2, 3, 6, 15\}$ and $\{4, 5, 7\}$. The point here is not that this is in any sense a “better” interpretation of the monastery data than the many different solutions offered in the literature. Rather, it is to emphasise that IBKR improves the quality of asiVAT images, such as the one in view 9a, that do not use imputation for the missing values.

6 Zachary’s Karate Club Data

The ZACHC Karate club data are a square symmetric set of relational data collected by Zachary [4] that represent the relative strength of the associations (number of situations in and outside the club in which interactions occurred) between the 34 members of a university karate club. The maximum number of interactions is 7, between nodes 26 and 32; the minimum is 1, which occurs several times. These data have appeared in many papers about social networks because the evolution of the relationship between pairs of members in the Karate club—which was known and recorded by Zachary – provides “ground truth” in the sense that the split into two subgroups actually occurred. The question posed by many writers: how many clusters does this network data contain after the split?

Zachary used an information flow model of network conflict resolution to explain the split of this group into $c = 2$ factions. Fig. 10 shows Newman and Girvan’s [28] interpretation of the network after the split as a weighted, undirected graph $G = (V, E, W)$. The principals in the split were

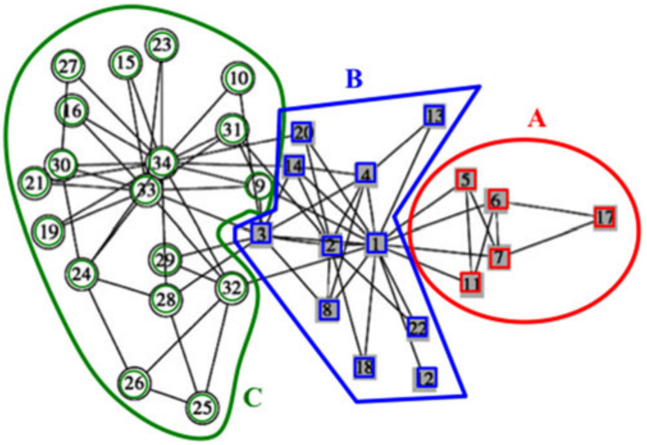


Fig. 11. Zhang et al. clusters in G [29].

the karate instructor (vertex 1) and the president of the club (vertex 34). Square nodes in Fig. 10 represent the instructor's faction and circular nodes depict the president's faction, so these authors also supported the idea that there are $c = 2$ clusters in the objects represented by G.

The belief that this network contains $c = 2$ computationally evident clusters has been challenged many times. For example, Zhang et al. [29], clustered a vector representation of G based on spectral representation of the information in the adjacency matrix of the graph with the standard hard c-means algorithm. Fig. 11 shows their crisp 3-partition of the Karate club data. The clusters $A = \{5, 6, 7, 11, 18\}$, $B = \{1, 2, 3, 4, 8, 12, 13, 14, 18, 20, 22\}$ and $C = \{1, \ldots, 34\} - \{A \cup B\}$ shown there, with minor changes, are accepted by many other writers as being "correct."

Edge weight w_{ij} in G is just the relative strength of interaction between individuals i and j, so the weight matrix W is not a dissimilarity relation. There are 156 edges with weights $w_{ij} > 0$, and 1,000 edges that are not connected, $w_{ij} = 0$, so this data set is about 86.5 percent incomplete. According to our previous determination, IBKR is needed. In order to make this data compatible with iVAT input requirements, it is transformed to the dissimilarity matrix D by setting $d_{ij} = 7 - w_{ij}$. The diagonal of D is set to 0 after this transformation.

Fig. 12 shows the 100×100 summary image $I(D_N)$ made by computing Kendall's Tau over $N = 100$ trials of IBKR on the Karate club data. There are three main dark blocks along the diagonal, indicating three major modes for the corresponding set of 100 asiVAT images, implying uncertainty in the clustering.

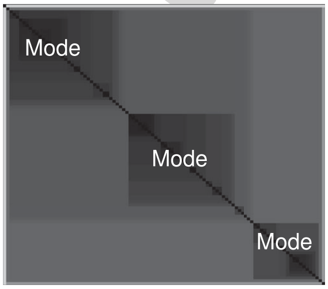


Fig. 12. The summary asiVAT image $I(D_N)$ for 100 trials of IBKR on the Karate Club Data has three fairly equal modes.

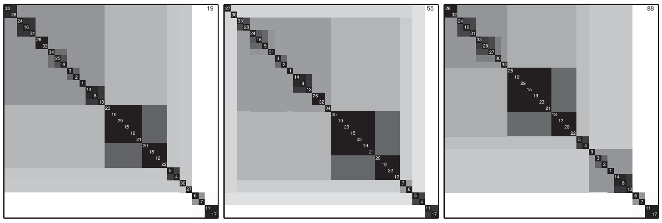


Fig. 13. Central asiVAT images from the three IBKR modes in Fig. 12, showing Image $I(D_{19})$ (left), Image $I(D_{55})$ (center), and Image $I(D_{88})$ (right).

Fig. 13 contains asiVAT images 19, 55 and 88, which are the centre images for each of the three modes seen in Fig. 12. These three images present different interpretations of clusters in the Karate club data.

Fig. 13(left) suggests a primary structure of $c = 3$ clusters, but they are substantially different from the 3 clusters in Fig. 13(centre), The strongest substructure in 13(left) posits $c = 7$ clusters. View 13(centre) has 2 primary clusters. View 13(right) has 3 primary clusters. The two tail clusters at the bottom right in views 13(left) and 13(right) are the same; $\{6, 7\}$ and $\{11, 17\}$, but subclusters in the main block are quite different. We do not assert that any of these views of the Karate club data are "correct." The point of this example is to show how IBKR imputation can be used with asiVAT as a tool for exploratory data analysis.

7 Discussion

It is difficult to evaluate methods of determining the number of clusters in data, therefore it is at least as difficult to evaluate the effectiveness of imputation methods for determining the number of clusters in data. Evaluation could have been performed using simulated data (for example, randomly sampling from multidimensional Gaussian distributions) then randomly removing values to generate the incomplete data. This would allow us to examine how well each imputation method dealt with Gaussian clusters, but we would be unsure as to how well the results generalised to different cluster distributions.

Our approach of investigating the utility of each method on real data, where values were artificially removed, allowed us to examine how each imputation method performed on real data, while the artificial removal of data allowed us to examine the effect of the proportion of missing data. Using real data sets, that have been thoroughly investigated by the clustering research community, meant that we had at least a rudimentary idea of what type of cluster structure might be present.

In any empirical evaluation, we always have the problem of how many data sets to use for evaluation. Our analysis used three data sets: Iris was used to discriminate between the imputation candidates; Sampson's and Zachary's data was used to test the limits of the effective IBKR method. We saw that IBKR provided some uncertainty in its summary image for Sampson's data, but provided a good imputation, and that IBKR showed a higher level of uncertainty for Zachary's data (with three mode blocks). More data could have been used to examine how IKBR copes with different data distributions, but the analysis we have performed provides sufficient evidence to assert that IBKR is a useful tool for cluster analysis and visual assessement methods on incomplete data.

The Summary image gives us an indication of the variation of the iVAT image due to the imputation. While we observed that the variation was associated to uncertainty, further analysis is required to determine the reliability of this association.

8 Conclusion

The iVAT and asiVAT algorithms offer a visual means for organisation of a data set that allows us to estimate the number of clusters within the data. Unfortunately, the iVAT and asiVAT algorithms both require pairwise dissimilarities between all n objects in order to compute the assessment images.

In this article, we investigated four methods of imputation of missing dissimilarity values and their effect on the subsequent asiVAT image. We first investigated the asiVAT image bias and variance induced by each imputation method using the Iris data, and found that uniform and bootstrapped imputation were acceptable for a small number of missing values (up to 5 percent), while bootstrapped kernel regression was acceptable for a larger number of missing values (up to 40 percent). Based on the Iris trials, we recommend KR Boot for data with any number of missing values up to about 40 percent.

We also investigated the effect of imputation on Sampson's monastery data with 68 percent missing values, and Zachary's Karate club data with 86 percent missing values. The large number of missing values led us to propose and investigate iterated KR Boot (IBKR). This modification of KR Boot led to visual interpretations of both data sets that were consistent, but different, from others in the extant literature.

Our IBKR split of Sampson's data in Fig. 9b differs a bit from those of several previous studies, but those studies also disagree with each other in fine detail. This observation holds for the Karate club data too. The important point is that the IBKR method is, to our knowledge, the first method that offers a visual estimate for the number of potential clusters in these data sets, and the IBKR mode images contain fine structure detail that can be extremely useful in sharper estimates of cluster structure. Don't forget, asiVAT does NOT find the clusters – it just suggests how many to look for.

In summary, IBKR promises to be a useful extension of the VAT/iVAT family of cluster heat maps to the case of symmetric or asymmetric input dissimilarity data with missing values. However, much more experimental evidence is needed to validate this variant of the basic method. We will devote a forthcoming paper to a number of additional case studies.

References

[1] S. F. Sampson, "Crisis in a cloister," Univ. Microfilms, Ann Arbor, MI, Tech. Rep. 69–5775, 1969.

[2] R. L. Breiger, S. A. Boorman, and P. Arabie, "An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling," *J. Math. Psychology*, vol. 12, no. 3, pp. 328–383, 1975.

[3] S. Wasserman, *Social Network Analysis: Methods and Applications*, vol. 8. Cambridge, U. K.: Cambridge Univ. Press, 1994.

[4] W. W. Zachary, "An information flow model for conflict and fission in small groups," *J. Anthropological Res.*, vol. 33, pp. 452–473, 1977.

[5] T. C. Havens and J. C. Bezdek, "An efficient formulation of the improved visual assessment of cluster tendency (iVAT) algorithm," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 5, pp. 813–822, May 2012.

[6] L. Himmelspach and S. Conrad, "Clustering approaches for data with missing values: Comparison and evaluation," in *Proc. 5th IEEE Int. Conf. Digit. Inform. Manage.*, 2010, pp. 19–28.

[7] K. Wagstaff, "Clustering with missing values: No imputation required," in *Classification, Clustering and Data Mining Applications*. Berlin, Germany: Springer, 2004.

[8] J. Tuikkala, L. L. Elo, O. S. Nevalainen, and T. Aittokallio, "Missing value imputation improves clustering and interpretation of gene expression microarray data," *BMC Bioinf.*, vol. 9, no. 1, 2008, Art. no. 202.

[9] R. K. Vinayak, S. Oymak, and B. Hassibi, "Graph clustering with missing data: Convex algorithms and analysis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2996–3004.

[10] J. Poland and T. Zeugmann, "Clustering pairwise distances with missing data: Maximum cuts versus normalized cuts," in *Discovery Science*. Berlin, Germany: Springer, 2006, pp. 197–208.

[11] T. Loua, *Atlas Statistique de la Population de Paris*. Paris, France: J. Dejey & cie, 1873.

[12] J. Czekanowski, "Zur differentialdiagnose der Neandertalgruppe," *Korrespondenzblatt der Deutschen Gesellschaft fur Anthropologie, Ethnologie und Urgeschichte*, vol. 40, pp. 44–47, 1909.

[13] L. Wilkinson and M. Friendly, "The history of the cluster heat map," *Amer. Statistician*, vol. 63, no. 2, 2009, Art. no. 179184.

[14] J. C. Bezdek and R. J. Hathaway, "VAT: A tool for visual assessment of (cluster) tendency," in *Proc. Int. Joint Conf. Neural Netw.*, 2002, pp. 2225–2230.

[15] D. Zhang, K. Ramamohanarao, S. Versteeg, and R. Zhang, "Rolevat: Visual assessment of practical need for role based access control," in *Proc. Annu. Comput. Secur. Appl. Conf.*, 2009, pp. 13–22.

[16] T. C. Havens, J. C. Bezdek, C. Leckie, and M. Palaniswami, "Extension of iVAT to asymmetric matrices," in *IEEE Int. Conf. Fuzzy Syst.*, 2013, pp. 1–6.

[17] E. Anderson, "The irises of the GASPE Peninsula," *Bulletin Amer. Iris Soc.*, vol. 59, pp. 2–5, 1935.

[18] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[19] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *J. Royal Statistical Soc. Series B (Methodological)*, vol. 39, pp. 1–38, 1977.

[20] C. J. Wu, "On the convergence properties of the em algorithm," *Ann. Statist.*, vol. 11, pp. 95–103, 1983.

[21] T. E. Raghunathan, J. M. Lepkowski, J. Van Hoewyk, and P. Solenberger, "A multivariate technique for multiply imputing missing values using a sequence of regression models," *Survey Methodology*, vol. 27, no. 1, pp. 85–96, 2001.

[22] M. Templ, A. Kowarik, and P. Filzmoser, "Iterative stepwise regression imputation using standard and robust methods," *Comput. Stat. Data Anal.*, vol. 55, no. 10, pp. 2793–2806, Oct. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.csda.2011.04.012

[23] S. Buuren and K. Groothuis-Oudshoorn, "MICE: Multivariate imputation by chained equations in R," *J. Statistical Softw.*, vol. 45, no. 3, 2011, https://www.jstatsoft.org/issue/view/v045

[24] R. J. Little, "Missing-data adjustments in large surveys," *J. Bus. Econ. Statist.*, vol. 6, no. 3, pp. 287–296, 1988.

[25] B. Efron, "Bootstrap methods: Another look at the jackknife," in *Breakthroughs in Statistics*. Berlin, Germany: Springer, 1992, pp. 569–593.

[26] T. Hastie, R. Tibshirani, and J. J. H. Friedman, *The Elements of Statistical Learning*. New York, NY, USA: Springer, 2001.

[27] H. C. White, S. A. Boorman, and R. L. Breiger, "Social structure from multiple networks. I. Blockmodels of roles and positions," *Amer. J. Sociology*, vol. 81, pp. 730–780, 1976.

[28] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E*, vol. 69, no. 2, 2004, Art. no. 026113.

[29] S. Zhang, R.-S. Wang, and X.-S. Zhang, "Identification of overlapping community structure in complex networks using fuzzy c-means clustering," *Physica A: Statistical Mech. Its Appl.*, vol. 374, no. 1, pp. 483–490, 2007.

Laurence A. F. Park received the BE (Hons.) and BSc degrees from the University of Melbourne, Australia, in 2000, and the PhD degree from the University of Melbourne, in 2004. He joined the Computer Science Department, University of Melbourne, as a research fellow, in 2004, and was promoted to senior research fellow, in 2008. He joined the School of Computing and Mathematics, University of Western Sydney, in 2009, where he is currently as a senior lecturer in computational mathematics and statistics. His research interests include large-scale data mining, machine learning, and information retrieval.

James C. Bezdek (M'80-SM'90-F'92-LF'09) received the PhD degree in applied mathematics from Cornell University, Ithaca, New York, in 1973. His research interests include optimization, pattern recognition, clustering in very large data and social networks, coclustering, and visual clustering. He was the president of the North American Fuzzy Information Processing Society, the International Fuzzy Systems Association, and the IEEE Computational Intelligence Society. He is the founding editor of the *International Journal of Approximate Reasoning* and the *IEEE Transactions on Fuzzy Systems*. He has received the IEEE Third Millennium, IEEE CIS Fuzzy Systems Pioneer, and IEEE TFA Rosenblatt medals and Kampe de Feriet award. He is a life fellow of the IFSA and the IEEE.

Christopher Leckie received the BSc degree, in 1985, the BE degree in electrical and computer systems engineering (with first class honors), in 1987, and the PhD degree in computer science, in 1992, all from Monash University, Australia. He joined Telstra Research Laboratories, in 1988, where he conducted research and development into artificial intelligence techniques for various telecommunication applications. In 2000, he joined the Department of Computing and Information Systems, University of Melbourne, Australia, where he is currently a professor. His research interests include using artificial intelligence for intrusion detection and sensor networks, and data mining techniques such as clustering and anomaly detection.

Ramamohanarao Kotagiri received the PhD degree from Monash University. He is currently a professor of computer science with the University of Melbourne. He served on the editorial boards of the *Computer Journal*. At present, he is on the editorial boards of *Universal Computer Science*, *Data Mining,* and *the International Very Large Data Bases* Journal. He was the program cochair for VLDB, PAKDD, DASFAA, and DOOD conferences. He is a steering committee member of the IEEE, ICDM, PAKDD, and DASFAA. He is a fellow of the Institute of Engineers Australia, a fellow of the Australian Academy Technological Sciences and Engineering, and a fellow of the Australian Academy of Science. He received the Distinguished Contribution Award in 2009 from the Computing Research and Education Association of Australasia. He is a member of the IEEE.

James Bailey received the PhD degree from the University of Melbourne, in 1998. He is currently an Australian research council future fellow in the Department of Computing and Information Systems, University of Melbourne. He has previously held appointments with Kings College, Birkbeck College, and the University of London. He has authored more than 100 articles. His research has consistently been supported through grants from the Australian Research Council. His research interests include the area of data mining and machine learning, with a focus on both fundamental topics such as contrast pattern mining and data clustering, as well as application aspects in areas such as health informatics and bioinformatics. He has received two Best Paper Awards. He serves on the program committees of many international conferences in data mining and is currently an associate editor of the *IEEE Transactions on Knowledge and Data Engineering* and *Knowledge and Information Systems*.

Marimuthu Palaniswami (S'84-M'87-SM'94-F'12) received the PhD degree from the University of Newcastle, Australia. He is currently a professor in the Department of Electrical and Electronic Engineering, the University of Melbourne. He has published more than 400 refereed research papers and leads one of the largest funded Australian Research Council programs. He has been a grants panel member for the US National Science Foundation, a steering committee member for the National Collaborative Research Infrastructure Strategy, Great Barrier Reef Ocean Observing System, Smart Environmental Monitoring and Analysis Technologies, and a board member for Information Technology and supervisory control and data acquisition companies. He has been funded (more than $40 million) to conduct research in sensor network, Internet of things, health, environmental, machine learning, and control areas. His research interests include sensor networks, IoT, machine learning, pattern recognition, and signal processing. He is a fellow of the IEEE.