Missing Data and Matrix Factorization for an Electrodiagnostic Nerve Test Dataset

James Bell and Cody Rosevear

Abstract—Data samples from a 36-dimensional electrodiagnostic nerve test were reduced to three dimensions for visualization. Three methods of matrix factorization were used to reduce the data: principal components analysis, c-means clustering, and archetypal analysis. Principal components analysis was outperformed by c-means clustering by a factor of 5.0, and was outperformed by archetypal analysis by a factor of 4.42. When values were missing from the dataset, c-means clustering and archetypal analysis continued to outperform principal components analysis. In both cases, there was not a significant difference between the performance of c-means clustering and archetypal analysis.

Keywords—matrix completion, missing data, factor analysis, machine learning, nerves, diagnostic test.

1 Introduction

Matrix factorization is commonly used to fill in missing data and to provide additional insight into the existence of underlying factors. This study looks at data from a specific electrodiagnostic nerve test, using samples collected by the Clinical & Theoretical Neuroscience Laboratory at the University of Alberta (see Appendix A for more details about the test). The features of this nerve test are very difficult to interpret, in part because many of them are correlated, which makes it difficult to identify the underlying biological factors which change in the presence of neurological disorders. Dimensionality reduction would therefore be very beneficial to interpretation, especially if it reveals useful underlying features. Specifically, it would be beneficial to be able to plot the data in three dimensions while retaining much of the variance in the original data; 3D visualization would be helpful for the neuroscientists who use this test.

A secondary concern with this dataset is that values are occasionally missing from the recordings. Those missing values pose problems for many statistical methods, so a good algorithm will also be capable of filling in these missing values. Since matrix factorization naturally fills in missing values, it is a good fit for this problem.

Finally, since new results are added to the dataset on a regular basis, a key requirement for this algorithm is that it is able to operate on an individual data point, rather than recalculating factors for the entire dataset each time a new point is added. The algorithm must use a previously trained model to transform a newly added test result into the same 3D representation as the previous data. If the factors were recalculated, then the 3D visualization of the data would change; every previous data point would be plotted at a new location in the new representation. By retaining the factors, new points can be added to the existing visualization.

The current dataset consists of only 200 observations. That is too few to make strong generalizations about these algorithms or their effectiveness on this dataset. However, the goal of this study is to provide some pilot results and to set up a framework for further analysis once more results have been collected.

J. Bell is with the Neuroscience and Mental Health Institute and Department of Computing Science, University of Alberta, Edmonton, AB, Canada, e-mail: jbell1@ualberta.ca.
C. Rosevear is with the Department of Computing Science, University of Alberta, Edmonton, AB, Canada, e-mail: rosevear@ualberta.ca.

2 Background

Matrix factorization is a common method for dimensionality reduction, factor analysis, and filling in missing data. An initial n-by-m matrix X—in this case, 200 observations (n) by 36 features (m)—is factored into a matrix product

$$X = PF \,, \tag{1}$$

where P is an n-by-k matrix of k factors for each of the n participants, and F is a k-by-m matrix of k factors for each of the m features. Once the initial dataset, X, has been factored into PF, any new sample x (a 1-by-m row of features for a single participant) can be transformed into p, a 1-by-k row in P, by $p = xF^{+}$. This is useful for adding new samples without recalculating the factorization on the entire dataset. Those new entries in P can be converted back to a representation of x, $\hat{x}$, by

$$\hat{x} = xF^{+}F \,. \tag{2}$$

This is useful for filling missing values in new samples.

3 Experimental Design

The goal of this experiment was to quantitatively compare three algorithms to determine which one is the best at transforming new 36-dimensional samples of an electrodiagnostic test into a 3-dimensional representation suitable for graphing. The success of an algorithm was measured by how well it retained the original data and filled in missing values; the algorithm's reconstruction of a sample (equation 2) was compared to the known true value.

3.1 Data Organization

Each row of data contains 36 features. Most are continuous real values in the range of $(-\infty, \infty)$ (though they tend toward a much smaller range); a few features are discrete (e.g. age and sex). Many of the features are measurements taken from continuous graphs, so it could be possible to infer missing values and determine underlying factors based on relationships between them, but that is beyond the scope of this study. For more details about the specific features, see Appendix B.

TABLE 1
Notation used for each of the splits of the dataset.

Matrix	Row	Description
X	x	The original 165 samples which are not missing any values.
X_{train}	x_{train}	The 125 samples which are used for training the algorithms.
X_{test}	x_{test}	The original 40 samples which are used to verify the performance of the algorithms, with no missing values.
X_{miss}	x_{miss}	The same 40 samples as X_{test}, but with some values deleted to assess the performance of the algorithms.

In order to quantitatively compare the results of the tested algorithms, any observations which are missing data were discarded. Since their true values are unknown, it would be impossible to draw any conclusions about the accuracy of the algorithms. Instead, only 165 of the original 200 observations were retained. These 165 observations were divided into two sets: training and test. The training set contained 125 observations, while the test set contained 40. Values were then deleted from the test set to simulate the missing data. See 1 for the notation used for each of these datasets.

To simulate the presence of missing values, values were randomly deleted from some columns of X_{miss}. This was based on the assumption that the data is missing at random. (Refer to Appendix A for more information about why that assumption is justified.) The feature most often missing, "Refractoriness at 2 ms (%)", was deleted from 30% of the observations, and the other features were deleted in relative proportions, as shown in Table 2. The prevalence of missing data in X_{test} was tripled compared to the original dataset in order to ensure the algorithms perform well in the presence of missing data. One additional constraint was also applied, based on known properties of the electrodiagnostic test: "Refractoriness at 2.5 ms (%)" was only deleted if "Refractoriness at 2 ms (%)" was also deleted.

No values were deleted from the training set because the goal was to construct the best possible feature matrix, F, which is the output of the algorithm. An F constructed without missing data is likely to perform better because

TABLE 2
Percent of observations that are missing for the four features that have missing data.

Feature Name	Percent Missing Originally	Percent Missing, Test
Hyperpolarization I/V slope	5.6%	17.5%
Refractoriness at 2 ms (%)	9.6%	30%
Refractoriness at 2.5 ms (%)	2.0%	6.25%
TEh(overshoot)	1.5%	4.7%

the trained factors will be based on correct values, rather than interpolations of missing values. Once that F was trained, it was used on the test set to determine the effectiveness of the algorithms on new data.

3.2 Evaluation

The algorithms were evaluated in two ways. First, which algorithm is most effective at retaining the original information after a projection down to three dimensions? Second, which algorithm is most effective at reconstructing missing data? In both cases, the 40 individual samples in the test set were compared to the algorithm's reconstruction of each sample (equation 2). The l2 norm of the difference provided a measure of the error for each sample. The first evaluation compared a reconstruction of x_{test} to x_{test}.

$$\text{Error}(x_{test}) = ||x_{test}F^{+}F - x_{test}||_2 \,. \tag{3}$$

The second evaluation compared a reconstruction of x_{miss} (that is, x_{test} with values deleted) to x_{test}.

$$\text{Error}(x_{miss}, x_{test}) = ||x_{miss}F^{+}F - x_{test}||_2 \,. \tag{4}$$

A successful algorithm will preserve most of the original data, but also fill in the missing values.

3.3 Algorithms

The algorithms used in this study were all chosen from those implemented by PyMF, a matrix factorization library written in Python [1]. This library was chosen because it provides over a dozen different algorithms for matrix factorization, with links to papers describing the features and performance of those algorithms. Some of the provided algorithms are only valid for non-negative matrix factorization (NMF). Since the electrodiagnostic nerve test data contains negative values, NMF algorithms are not an option. Of the remaining algorithms, principal component analysis, c-means clustering, and archetypal analysis were tested.

Principal component analysis (PCA) may be the most well-known method of matrix factorization. PCA weights each sample according to the eigenvectors of the dataset, and only the first k eigenvalues are retained. Since the eigenvalues are sorted by size, PCA is fairly effective at capturing variance in the dataset, but the factors are required to be orthogonal (since eigenvectors are orthogonal). This means it can be ineffective at highlighting factors that other methods can uncover, since factors are often not orthogonal. PCA's ubiquity and simplicity make it an excellent baseline algorithm for comparison.

C-means clustering is based on k-means clustering. K-means clustering finds k points in the m-dimensional space around which the data can be clustered, and it associates each of the n rows with one of the k centers. When used as a matrix factorization method, k-means clustering collapses each data point to its nearest mean. This will result in poor performance. C-means clustering represents each sample as a linear combination of the k points instead of being constrained to match only the closest center, making it more effective for missing data and factor analysis problems.

Archetypal analysis (AA) is similar to c-means clustering, but AA chooses extreme data points ("archetypes") instead of centers. Like c-means clustering, each sample is the linear combination of the k archetypes.

The code for this study is available on GitHub [2].

3.4 Statistical Methods

When comparing two algorithms a t-test is a good test for significance. However, while the probability of a Type I error is low for a single t-test, the probability increases significantly as the number of t-tests increases, so a t-test is inadequate for a comparison between more than two groups. Instead, one-way ANOVA was used to reject the null hypothesis. Since the same samples were used for each algorithm, a repeated measures one-way ANOVA would be more relevant, but basic one-way ANOVA is adequate unless it fails to reject the null hypothesis. The one-way ANOVA was followed by a Games Howell post-hoc test [3], [4] to determine which differences were significantly different.

4 Results

Figure 1 shows the error introduced by reducing the dimension of the data from 36 to 3, for each of the three algorithms. There was a statistically significant difference between groups as determined by one-way ANOVA ($F(2, 117) = 2915.39, p = 1.54579e - 100$). PCA clearly introduces more error than c-means and AA, and a post-hoc analysis with the Games Howell test confirmed that there was not a statistically significant difference between c-means and AA. The error for PCA was greater than the error for c-means by a factor of 5.03, and a factor of 4.42 for AA.

Figure 2 shows the results with missing values. The samples with missing values are clearly visible in the figure; the samples without missing values are clustered closely around the mean, while samples with missing values spread far above the mean. In spite of that spread, it is clear that there was a statistically significant difference between groups ($F(2, 117) = 149.48, p = 5.95601e - 33$). The assumption of homogeneity of variances was clearly violated by those outliers, so the Games Howell post hoc test remains a good choice. Once again it indicated that there was not a significant difference between c-means and AA, but that PCA had 3.11 times more error than c-means and 2.91 times more error than AA.

The error for PCA did not increase very much after values were removed, but there was a substantial increase in the c-means and AA errors (see Table 3). For all algorithms, the standard deviation was significantly larger when values were missing.

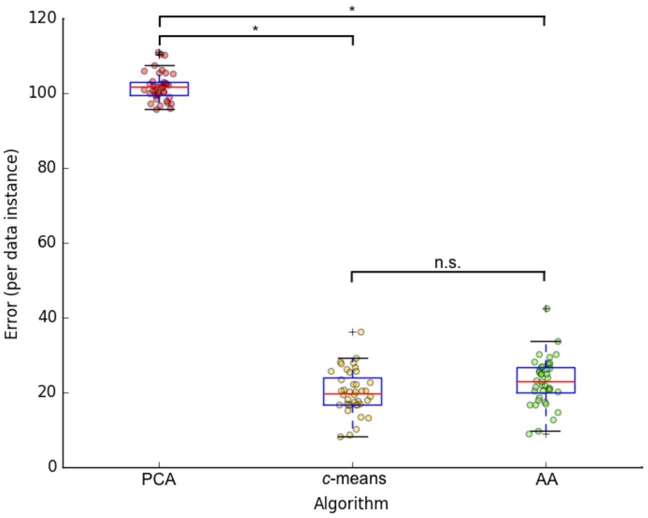

Fig. 1. The error introduced by reducing the dimension of the data from 36 to 3, for each of the three algorithms.

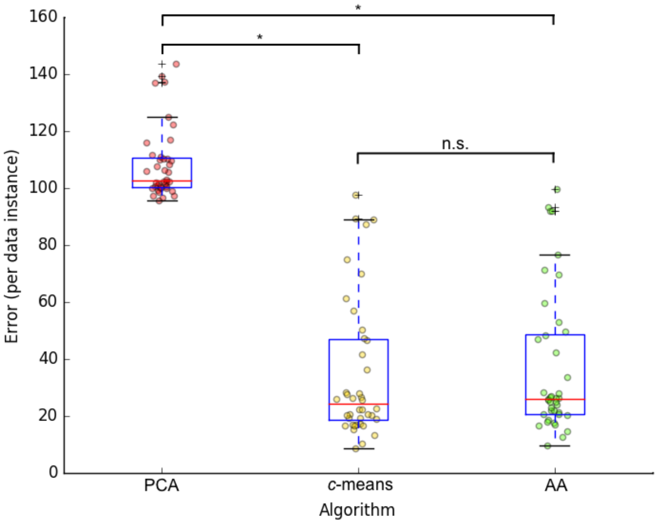

Fig. 2. The error resulting from attempting to reconstruct missing values by projecting down to three dimensions, for each of the three algorithms. The spread of data points above the mean shows the increased error for points that were missing values.

5 Discussion

Principal components analysis is not as effective as c-means clustering and archetypal analysis for retaining data integrity after compression down to three dimensions. It is less

TABLE 3
Increase in error and standard deviation after values were deleted.

Algorithm	Increase in Error	Increase in Standard Deviation
PCA	6.4%	227%
c-means	72%	327%
AA	62%	297%

susceptible to an increase in error due to missing values, but that is not enough to make up for its relatively poor performance. The poor performance of PCA may be due to overfitting. PCA will always produce the best error on the training set because it is composed of the eigenvectors with maximal variance, but that might result in it not generalizing as well to the test set.

c-means and AA are equally effective at dimensionality reduction for this electrodiagnostic test data. Even though they perform 62-72% worse with missing data, they are still reasonable methods for presenting this high-dimensional data in a dimension suitable for visual analysis. The similarity in their performance is unsurprising because they follow similar approaches. Both algorithms select k points (in this case, three) and describe the remaining samples in relation to those points. While they are likely to select different points, they are much more similar to one another than they are to PCA.

6 Future Work

There are many other algorithms for matrix factorization, including simplex volume maximization, semi-non-negative matrix factorization, CUR decomposition, and compact matrix decomposition. Testing all of them was outside the scope of this study, but future work could compare their performance to the ones tested in this study. There are even more algorithms available for factorizing matrices without any negative values, such as standard non-negative matrix factorization, convex non-negative matrix factorization, and convex-hull non-negative matrix factorization. Those algorithms could also be tested if the input data were first transformed to be non-negative (e.g. by scaling or exponentiation).

In this study, the choice of k was fixed for 3D visualization. A future study could instead determine the optimal k to preserve a balance between reconstructing complete samples (which is maximized with zero error when $k = 36$) and the ability to reconstruct missing values (which requires a smaller k). A validation step could be used to determine the optimal number of factors, and the test set would then verify that result. The rate of missing values could also be adjusted to consider its impact on the choice of k.

With additional data, future work could confirm if these results hold more generally. Data could also be included for various neurological disorders, such as multiple sclerosis (MS), chemotherapy-induced nerve damage, or amyotrophic lateral sclerosis (ALS) to see if the 3D scatter plot reveals clustering. Adding data for disorders could also show if the factor matrix F for healthy nerves can generalize to unhealthy nerves.

7 Conclusion

The goal of this study was to find the algorithm that is best at filling missing values while preserving known values for an electrodiagnostic nerve test dataset. The algorithm was required to perform well even after the data was compressed into a 3-dimensional space for visualization, and it was required to operate on individual samples without any changes to the previously calculated underlying factors. Principal components analysis performed significantly worse than both c-means clustering and archetypal analysis, but there was no significant difference between c-means and AA. The algorithms all performed worse in the presence of missing values, but c-means and AA still performed better with missing values than PCA without any missing values. In conclusion, c-means clustering and archetypal analysis are equally suitable algorithms for 3D visualization of this dataset.

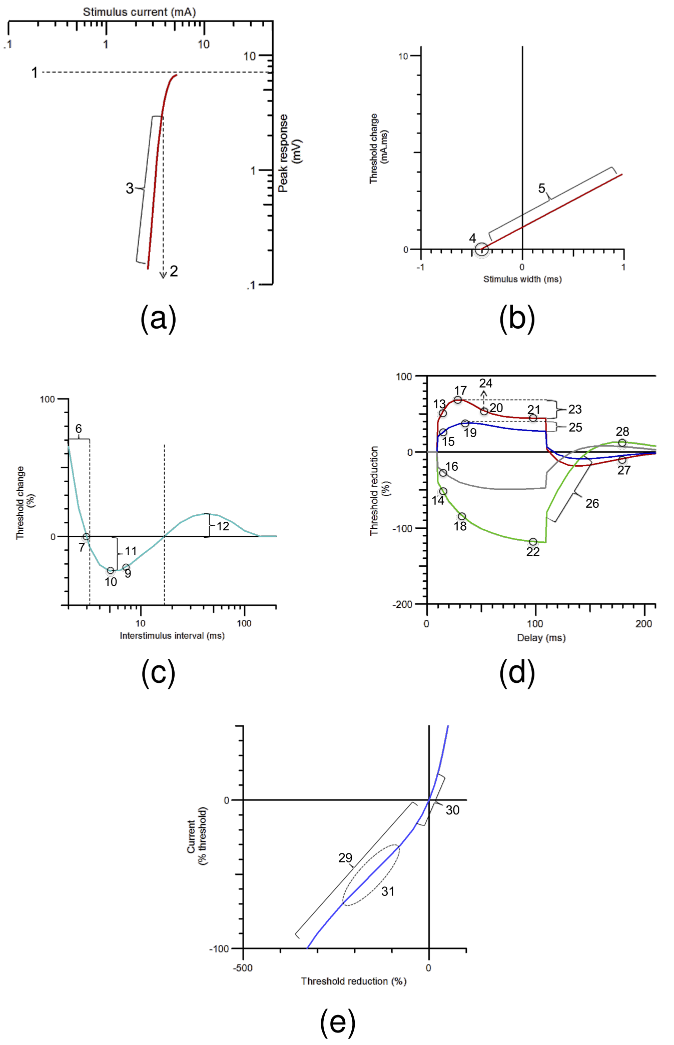

Fig. 3. The five plots produced by the Nerve Excitability Test. The numerical labels on the plots identify some of the 36 features that were analyzed in this study. The five plots are (a) Stimulus-Response Curve, (b) Charge Duration, (c) Recovery Cycle, (d) Threshold Electrotonus, and (e) Current-Voltage Relationship.

Acknowledgments

We thank Martha White and Kelvin Jones for providing guidance in the design of the project. We thank Touqir Sajed, Farzane Aminmansour, and Katie Tran for providing feedback. The Clinical & Theoretical Neuroscience Laboratory provided access to the private dataset used for this study.

Appendix A
Electrodiagnostic Nerve Test

Electrodiagnostic tools are used to measure the bioelectric properties of nerve and muscle tissue. The Nerve Excitability Test is one such tool used in both research and clinical settings. It has been used to study the way age, sex, and BMI affect the bioelectric properties of human motor axons [5], [6], [7], [8]; clinically it has been used with many disorders [9]. However, the Nerve Excitability Test is not yet common clinically because few clinicians have the expertise to interpret the multiple waveforms and over 30 discrete measures it produces.

The test takes twenty minutes to administer and produces waveforms of current and voltage. Out of those waveforms, certain measurements are used to construct graphs showing key relationships with neurological meaning (see figure 3). From those graphs, neuroscientists have identified a few dozen important features, such as the slope of a certain portion, the y-value at a specific x-value, or the maximum y-value. Those few dozen features are commonly used to compare neurological disorders.

Some values have obvious constraints. For example, the slope of a section of a graph will be either positive or negative, with a known mean. The value at a given point might be any real value, or it might always be positive or negative. Age and sex are covariants and are obviously constrained; age is usually recorded as an integer, and sex is usually recorded as a binary class. In this study, these constraints were not taken into account, other than to avoid non-negative matrix factorization techniques.

The four features with missing data are “Refractoriness at 2 ms (%)” (label 8 in figure 3c), “Refractoriness at 2.5 ms (%)” (label 7 in figure 3c), “Hyperpolarization I/V slope” (label 29 in figure 3e), and “TEh(overshoot)” (label 28 in figure 3d). Refractoriness is a measure of how quickly a nerve can respond to a stimulus after responding to a previous stimulus. It is possible that missing refractoriness data indicates the nerve has such a slow response that it could not respond even after 2 or 2.5 ms, which would not be missing at random, but it is much more likely that the data is missing due to signal noise, which classifies this as missing at random. The electrical charge used to stimulate the nerve has a much larger magnitude than the measured response, so it sometimes obliterates the desired signal, resulting in lost data at 2 ms and occasionally 2.5 ms. Hyperpolarization I/V slope can be missing for similar reasons: if the first point or two in the Recovery Cycle plot (figure 3e) is missing, the slope for the entire segment cannot be calculated. TEh(overshoot) is the maximum value in a certain region of one of the Threshold Electrotonus plots (figure 3c). That segment might not have the expected concave shape due to noise, resulting in a missing value for TEh(overshoot). Missing values for all of these four factors can most likely be considered to be missing at random.

APPENDIX B
FEATURES OF THE DATA

Feature Name	Data Type
Stimulus (mA), 50% max	real number
Strength-Duration (ms)	real number
Rheobase (mA)	real number
Stimulus-Response Slope	real number
Peak Response (mv)	real number
Resting I/V Slope	real number
Minimum I/V Slope	real number
Temperature (°C)	positive integer
RRP (ms)	real number
TEh(90-100ms)	real number
TEd(10-20ms)	real number
Superexcitability (%)	real number
Subexcitability (%)	real number
Age (Years)	positive integer
Sex (Male/Female)	binary
Latency (ms)	real number
TEd(40-60ms)	real number
TEd(90-100ms)	real number
TEh(10-20ms)	real number
TEd(Undershoot)	real number
TEh(Overshoot)	real number
TEd(peak)	real number
S2 Accommodation	real number
Accom. Half-Time (ms)	real number
Hyperpol. I/V Slope	real number
Refractoriness, 2.5ms (%)	real number
TEh(20-40ms)	real number
TEh(Slope 101-140ms)	real number
Refractoriness, 2ms (%)	real number
Superexcitability, 7ms (%)	real number
Superexcitability, 5ms (%)	real number
TEd20(Peak)	real number
TEd40(Accom)	real number
TEd20(10-20ms)	real number
TEh20(10-20ms)	real number
Nerve (Arm/Leg)	binary

REFERENCES

[1] C. Thurau, “PyMF - Python matrix factorization module,” https://github.com/cthurau/pymf, 2014, accessed: 2017-11-27.

[2] C. Rosevear and J. Bell, “Code for missing data and matrix factorization for an electrodiagnostic nerve test dataset,” https://github.com/ZerkTheMighty/551Project, 2017, accessed: 2017-12-08.

[3] P. A. Games and J. F. Howell, "Pairwise multiple comparison procedures with unequal N's and/or variances: A Monte Carlo study," *Journal of Educational Statistics*, vol. 1, no. 2, pp. 113–125, 1976. [Online]. Available: http://www.jstor.org/stable/1164979

[4] A. Trujillo-Ortiz and R. Hernandez-Walls, "GHtest: Games-Howell's approximate test of equality of means from normal population when variances are heterogeneous. a MATLAB file." http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=3676&objectType=FILE, 2003, accessed: 2017-12-07.

[5] S. Jankelowitz, P. McNulty, and D. Burke, "Changes in measures of motor axon excitability with age," *Clinical Neurophysiology*, vol. 118, no. 6, pp. 1397 – 1404, 2007.

[6] J. S. Bae, S. Sawai, S. Misawa, K. Kanai, S. Isose, K. Shibuya, and S. Kuwabara, "Effects of age on excitability properties in human motor axons," *Clinical Neurophysiology*, vol. 119, no. 10, pp. 2282 – 2286, 2008.

[7] J. C. McHugh, R. B. Reilly, and S. Connolly, "Examining the effects of age, sex, and body mass index on normative median motor nerve excitability measurements," *Clinical Neurophysiology*, vol. 122, no. 10, pp. 2081 – 2088, 2011.

[8] I. Casanova, A. Diaz, S. Pinto, and M. de Carvalho, "Motor excitability measurements: The influence of gender, body mass index, age and temperature in healthy controls," *Neurophysiologie Clinique/Clinical Neurophysiology*, vol. 44, no. 2, pp. 213 – 218, 2014.

[9] M. C. Kiernan and C. S. Y. Lin, "Chapter 15 - nerve excitability: A clinical translation," in *Aminoff's Electrodiagnosis in Clinical Neurology*, 6th ed., M. J. Aminoff, Ed. London: W.B. Saunders, 2012, pp. 345 – 365.