Noise and Outlier Filtering in Heterogeneous Medical Data Sources

Noor Alaydie, Farshad Fotouhi, Chandan K. Reddy
Department of Computer Science
Wayne State University
Detroit, MI 48202, USA
{alaydie,fotouhi,creddy}@wayne.edu

Hamid Soltanian-Zadeh
CIPCE, University of Tehran
Tehran 14395-515, Iran
MIAL, Henry Ford Hospital
Detroit, MI 48202, USA
hamids@rad.hfh.edu

***Abstract*—There is a growing interest in studying the common features from multiple data sources. Fusing information that come from multiple heterogenous data sources promises to identify complex multivariate relationships among the heterogeneous sources. Such relationships can provide additional connectivity across the sources. A common way to analyze the relationships between a pair of data sources based on their correlation is canonical correlation analysis (CCA). CCA seeks for linear combinations of all variables from each data set with maximal correlation between the two linear combinations. However, the existence of non-informative data points and features makes it challenging for CCA to identify significant relationships among the examined data sets. In this paper, we propose a novel method, NOFA, Noise-Outliers Free Algorithm, that can be used to filter out the non-informative data points and features before applying the CCA. NOFA was applied to preprocess two epilepsy modalities, the MRI and neuropsychology, prior to applying CCA to find the association between them. The results show that the proposed method leads to interpretable results when noise plays a significant role in the acquisition of the data.**

***Index Terms*—canonical correlation analysis; regularization; noisy features; outliers; principal component analysis.**

I. Introduction

Recently, there has been an increasing interest in studying and extracting the common features from two sets of quantitative variables observed on the same experimental units [8]. Highlighting significant relationships between two sets of variables is important for many real-world applications [3].

In multivariate data analysis, generally, there are two standard techniques for extracting correlated features from two sets of variables: Partial Least Squares regression (PLS, [13], [6]) and canonical correlation analysis (CCA, [4]) [3]. PLS is appropriate when there is a dependency among the two sets of variables. In other words, PLS is suitable when one set of variables can be explained by the other set [3]. On the other hand, CCA is more suitable when the two sets of variables have symmetric role in the analysis, and the objective is to analyze the correlations between them [3], [2].

CCA is an old multidimensional explanatory statistical method that explores the sample correlations between two spaces of different dimensions and structure observed on the same experimental objects [4], [12]. More precisely, the main objective of CCA is to identify the linear combinations of all variables from each data set, such that the correlation between the two linear combinations is maximized [12], [8]. A necessary condition that is usually advocated to perform CCA is that $n \geq p + q + 1$ [3], where n is the number of observations and p and q are the number of variables in the first and second data sets, respectively.

One of the main weaknesses of CCA is that when the data set is noisy, the solution changes dramatically. CCA does not handle noisy data sets properly. The existence of non-informative data points and features makes it challenging to find meaningful relationships among the examined data sets using CCA.

In this paper, we propose a novel method, NOFA, Noise-Outliers Free Algorithm, to preprocess the data and remove non-informative features and data points based on principal component analysis (PCA). Next, CCA is applied on the cleansed version of the data sets to find the maximally correlated features. The proposed method is well-suited to a number of real life application. We provide the results of the experiments in a medical application. More specifically, we applied our method on two modalities of the epilepsy data set and we show that its interpretation evidences meaningful relationships between the filtered features. The paper is structured as follows: in the next section we describe the motivation for our method using a synthetic data set. In section III, we describe the proposed method, NOFA. In section IV, we present the results of our method. Section V concludes the paper.

II. Motivation

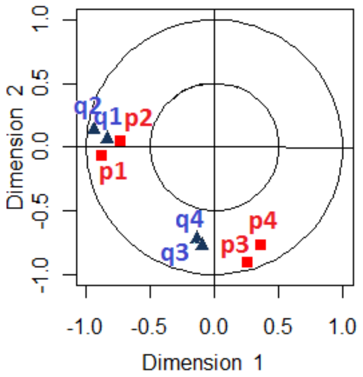
Outliers and noise are unavoidable in real life data sets obtained from numerous application domains. In the medical domain, noise plays a significant role in the acquisition of data. For example, medical data records naturally contain non-informative data points and features. CCA and its variations do not handle noisy data properly. This is illustrated in the following experiments. We generated two data sets with some correlated features measured on twenty objects. The first data set has six variables $(p_1, ..., p_6)$ and the second data set has six variables $(q_1, ..., q_6)$. The variables p_3 and p_4 from the first data set have local correlation. The variables q_1 and q_2 from the second data set have significant local correlation. Also, the variables q_3 and q_4 have strong local correlation. Variables p_1

TABLE I

SUMMARY OF THE SYNTHETIC DATA SET CORRELATIONS. LOCALLY CORRELATED MEANS THE CORRELATIONS EXIST IN A SINGLE DATA SET, WHILE GLOBALLY CORRELATED MEANS THE CORRELATIONS EXIST BETWEEN THE TWO DATA SETS.

Features	Description
(p_3, p_4)	locally correlated
(q_1, q_2)	locally correlated
(q_3, q_4)	locally correlated
(p_1, q_1)	globally correlated
(p_2, q_2)	globally correlated
(p_3, q_4)	globally correlated
p_5, p_6, q_5, q_6	noise features

Fig. 1. Variables plots for the first and second dimensions of CCA applied on synthetically generated data sets without adding the noisy features.



and q_1 have the strongest positive correlation from the two data sets, followed by the variables p_2 and q_2 and then the variables p_3 and q_4. The variables p_5, p_6, q_5 and q_6 are noisy variables. Table I summarizes the correlations exist in the data sets.

Figure 1 shows the results of applying CCA on the synthetic data sets without adding the noisy features. In this figure, we show the scatter plots of the initial variables from the two data sets. The most significant correlations are shown in the ring defined between the two circumferences of the inner and outer circles. Variables with a strong positive correlation are projected on the canonical variates plane in the same direction from the origin. On the other hand, variables with a strong negative correlation are projected in an opposite direction from the origin. As the figure shows, the variables p_1, p_2, q_1 and q_2 are projected in the same direction; which indicates a strong positive correlation. The figure also shows the significant local correlation between the variables q_3 and q_4. Moreover, the figure shows that the variables p_3 and p_4 have a correlation that is considered to be significant. And all of the variables p_3, p_4, q_3 and q_4 have also a significant correlation. In fact, since the variables p_1, p_2, q_1 and q_2 and the variables q_3 and q_4 are closer to each other than the variables p_3 and p_4, this indicates that the former correlations are stronger than the latter one.

As it is well-known, outliers in a dataset are not consistent with the rest of the data. In the second set of experiments, we generated a few randomly distributed outliers; a few data points that are arbitrarily far away from a sequence of identical data points. More specifically, we added two outliers to the variable p_1. Figure 2 shows the results for CCA after adding the outliers. Since p_1 has the outliers, CCA was not able to discover the correct relationship of p_1 with the other variables. As shown in Figure 2, p_1 is projected in the opposite direction from the origin of the variables p_2, q_1 and q_2 as opposed to the projection shown in Figure 1. Hence, CCA shows negative correlation of the variable p_1 with the variables p_2, q_1 and q_2.

Fig. 2. Variables plots for the first and second dimensions of CCA applied on synthetically generated data sets with the outliers added to p_1.

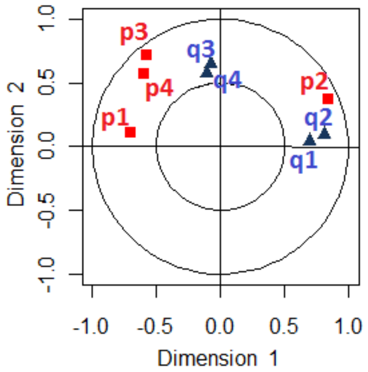
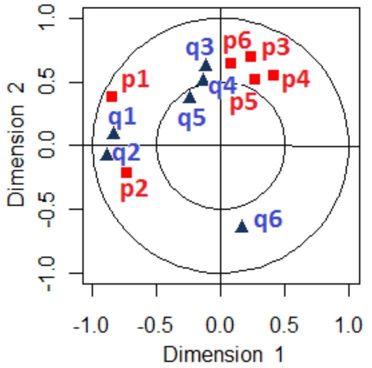


Fig. 3. Variables plots for the first and second dimensions of CCA applied on synthetically generated data sets with the noise features added to the two data sets.



In the third set of experiments, we added four gaussian noisy features, two features to each data set. Figure 3 shows the result of CCA after adding the noisy features, p_5, p_6, q_5 and q_6. As the figure shows, the solution is changed in a very obvious way when noisy features present in the data. Moreover, the noisy features show strong correlations either with each other or with other features. Such change in the results makes the results less informative. More specifically, the noisy features p_5 and p_6 are shown to have strong positive correlations with the variables p_3, p_4, q_3 and q_4. Moreover, the variables p_6 and q_6 show strong negative correlation. Since noisy features are meaningless, such correlations are not helpful and do not make any sense.

The main objective of the proposed method is to remove the outliers data points and noisy features prior to the application of the CCA method. Noise plays a significant role in preventing the discovery of meaningful correlations. Hence, getting a cleaned copy of the data is considered crucial before applying any further processing.

III. NOFA: The proposed Noise-Outliers Free Algorithm

In this section, we present the details of NOFA, the Noisy-Outliers Free Algorithm. The following notation is used through the discussion of the method. G_1 and G_2 are two groups of measurements, with ranks M x P and M x Q, respectively, observed on the same set of objects, N. The i^{th} column of G_1 is denoted as $p^{(i)}$, likewise, the i^{th} column of G_2 is denoted as $q^{(i)}$. The k^{th} row of G_1 or G_2 is denoted as O_k.

Our main objective is to identify and remove the non-informative data points and features. We aim to remove the outliers and noisy features so that CCA leads to more meaningful results. The cleansing process is achieved using PCA as the basic stone in the whole algorithm. In fact, PCA is applied in an elegant way on each data set. The objective of PCA is to find a new set of dimensions that captures the variability in the data. PCA transforms the set of correlated features into a smaller set of uncorrelated variables, called principal components. The first principal component is chosen to capture as much of the variability in the data as possible. The second principal component, which is orthogonal to the first principal component, captures as much of the remaining variability as possible and so on [10]. PCA identifies the strongest patterns in the data. Since noise has weaker patterns, the use of PCA can eliminate much of the noise.

Algorithm 1 shows the basic steps in the filtering process. Initially, the data sets are normalized so that each data set has zero mean and one standard deviation. Generally, there are two main steps to clean the data; filtering out the non-informative features i.e., the noisy features, and filtering out the non-informative data points i.e., the outliers data objects. The output of the first step is the input to the second step, which is the noisy-free data set. Since our ultimate goal is to find correlated features from two sets of variables measured on the same set of data objects, we need to check whether the same data object is still present, in both data sets after the filtration process or not. If a given object does not survive in one data set i.e., considered as an outlier and removed, this object will be removed also from the second data set and that object has to be analyzed no further. Filtering out the noisy features is done next followed by filtering out the outliers data objects. Finally, a consistent view is constructed using the cleaned copies of the data sets.

A. Noise removal stage

Algorithm 2 shows the process for removing the noisy features. The process begins by fitting linear models (in the Algorithm, it is called FLM) using PCA for every possible feature pairs. In order to find a good line that models the data, we use segments of the first principal component that are bounded by 3σ/2 from the mean of the whole data set (see Algorithm 4). The sum of the square distances, for all of the data points, is computed from the linear model that is fitted for each feature pairs. In fact, we divided the sum of square distance for each linear model by $\sqrt{2}$, we call the new value $dist(p^{(i)}, p^{(j)})$. Once the distance, $dist$, values are available for all of the feature pairs, the β threshold is computed as the following: $\beta = \frac{\sum_{i=1}^{n} dist(p^{(i)},p^{(j)})}{n} - \alpha * std(dist)$ where n is the number of elements of $dist$ and std is the standard deviation. The parameter α controls the contribution of the standard deviation of $dist$ to the β threshold. Note that β is automatically set after getting the sum of square distances for all feature pairs.

Algorithm 1 $NOFA$

```
Input: $G_1$ and $G_2$: two data set of sizes $M \times P$ and $M \times Q$, respectively, observed on the same set of objects, $M$.
$\eta$: outliers threshold that controls the percentage of the data points to be considered as outliers.
$\alpha$: a parameter that controls the contribution of standard deviation in the beta threshold.
$\epsilon$: noise threshold that controls the percentage of the data points to be considered as noisy features.
Output: $Clean\_G_1 = K \times L$: the preprocessed data set of K objects and L features, where $K \leq M$ and $L \leq P$.
$Clean\_G_2 = K \times L$: the preprocessed data set of K objects and V features, where $K \leq M$ and $V \leq Q$.
Algorithm:
Normalize $G_1$ and $G_2$
$G_1^{'} = Remove\_Noisy\_Features(G_1, \alpha, \epsilon)$
$G_1^{''} = Remove\_Outliers(G_1^{'}, \eta)$
$G_2^{'} = Remove\_Noisy\_Features(G_2, \alpha, \epsilon)$
$G_2^{''} = Remove\_Outliers(G_2^{'}, \eta)$
$Clean\_G_1 = \cup\{G_1^{''}(O_i) : \forall O_i \in G_1^{''}\{\exists y \in G_2^{''} \wedge y = O_i\}\}$
$Clean\_G_2 = \cup\{G_2^{''}(O_i) : \forall O_i \in G_2^{''}\{\exists d \in G_1^{''} \wedge d = O_i\}\}$
```

If the data are scattered in the feature spaces for a particular pair of features, then the $dist$ value for that pair will be greater than β threshold. We flag such a pair for further analysis as there is a chance for one or both of the features to be noisy features. However, if the principal curve for that pair of features is good, then the computed $dist$ value will be less than the β threshold. Hence, such pairs are considered as successfully passing the noisy features test. After checking the $dist$ values for all of the feature pairs and flagging the suspect pairs, we count the frequency of each feature in the suspect pairs. Those features that have their frequencies greater than the user specified threshold (ϵ), are considered as noisy ones and hence are removed from the data set. Through this approach, the noisy features are discarded and do not proceed for the correlation analysis part. It is worthy to mention that the user has the control over the percentage of features to be considered as noisy features.

```
Algorithm 2 $Remove\_Noisy\_Features$
  Input: $D = M \times N$: data set of M objects and N features.
  $\alpha$: a parameter that controls the contribution of standard deviation to the beta threshold.
  $\epsilon$: noise threshold.
  Output: $D^{'} = M \times L$: the preprocessed data set of $M$ objects and $L$ features, where $L \leq N$.
  Algorithm:
  P=features($D$)
  for each pair $p^{(i)}$ and $p^{(j)} \in P$ do
      $[sqr\_dist, proj\_Ind] = FLM([p^{(i)}, p^{(j)}])$
      $dist[(p^{(i)}, p^{(j)})] = \frac{\sum sqr\_dist}{\sqrt{2}}$
  end for
  $\beta = \frac{\sum_{k=1}^{n} dist[(p^{(i)}, p^{(j)})]}{n} - \alpha * std(p)$
  $\mathring{f} = \{(p^{(i)}, p^{(j)}) : dist(p^{(i)}, p^{(j)}) > \beta, \forall (p^{(i)}, p^{(j)}) \in P\}$
  for each $p^{(k)} \in P$ do
      $count(p^{(k)}) = |\mathring{f}(p^{(k)}, p^{(l)})|, \forall p^{(l)} \in P$
  end for
  $CNF = \{p^{(i)} : count(p^{(i)}) > \epsilon, \forall p^{(i)} \in P\}$
  $D^{'} = \{D - CNF\}$
```

B. Outliers removal stage

The next step is identifying and removing outliers from the remaining non-noisy informative features. Algorithm 3 describes the pseudocode for recognizing and removing the outliers. Basically, a linear model is fitted once for all of the features using PCA. As a result, the distance of each data point from the fitted linear model, (sqr_dist), is computed. Next, data points that have sqr_dist greater than a threshold (η) are considered as outliers and removed from the data set.

```
Algorithm 3 $Remove\_Outliers$
  Input: $D = M \times L$: data set of $M$ objects and $L$ features.
  $\eta$: Outliers threshold.
  Output: $D^{''} = K \times L$: the preprocessed data set of $K$ objects and $L$ features, where $K \leq M$ and $L \leq N$.
  Algorithm:
  $[sqr\_dist, proj\_Ind] = FLM(D)$
  $\tilde{f} = \{d_i : sqr\_dist(d_i) > \eta, \forall d_i \in D\}$
  $D^{''} = \{D - \tilde{f}\}$
```

```
Algorithm 4 $FLM$
  Input: $X$: feature data set of dimensionality $n$ x $p$
  Output: $Optimal\_Linear\_Model$ fitted to the data.
  $sqr\_dist$: projection distance of the data.
  $proj\_Ind$: projection indices of the data points on the fitted linear model.
  Algorithm:
  $X\_1^{st}\_PC = First\_Principal\_Component(X)$
  $Optimal\_Linear\_Model = PC\_Segment(X\_1^{st}\_PC)$
  $sqr\_dist = Projection(X, Optimal\_Linear\_Model)$
  $proj\_Ind = Rank\_Data(X, Optimal\_Linear\_Model)$
```

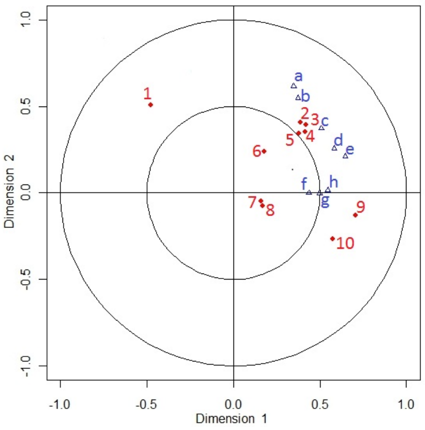


Fig. 4. The variables plot for the epilepsy data set using the first two dimensions obtained from CCA. The red circles are the MRI features, where 1 refers to volume ratio, 2 refers to intensity average, 3 refers to intensity standard deviation and $4 - 10$ refer to wavelet texture features. The blue triangles are the neuropsychology features, where a refers to Rey-Osterreith non-verbal memory immediate, b refers to Rey-Osterreith non-verbal memory delayed, c refers to verbal IQ, d refers to full scale IQ, e refers to non-verbal IQ, f refers to Boston Naming Test, g refers to Wechsler verbal memory immediate and g refers to Wechsler verbal memory delayed.

IV. Results

In this section, we present the results of NOFA applied on epilepsy data set. We evaluate the effectiveness of NOFA by applying CCA after preprocess the data set using NOFA.

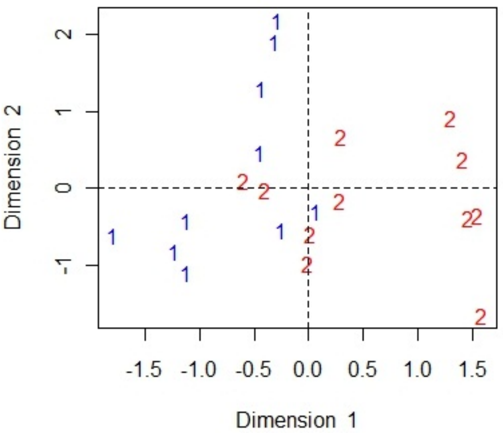
A. Epilepsy data set

The data set was generated from a study performed at Henry Ford Health Systems on mesial Temporal Lobe epileptic patients (mTLE). Temporal lobe epilepsy is a form of focal epilepsy in which the patients experience recurrent seizures arising from one or both of the brain's hippocampus, an inner aspect of the medial temporal lobe [5]. Approximately, 65% of epilepsy patients become free of seizures using anti-epileptic medications [7]. Another $8 - 10\%$ benefit from surgical treatment [7].

A retrospective study on twenty eight unilateral mesial temporal lobe epilepsies with Engel class *Ia* outcomes were undertaken. Engel class *Ia* is used to describe patients who rendered without seizures postoperatively. By setting class to *Ia*, we establish a genuine criterion for laterality. The following two sets of variables were acquired:

- MRI based features: Three sets of features were extracted from each hippocampal region of interest (*ROI*): mean and standard deviation of the FLAIR MR signal intensity, wavelet transform-derived energy and volumetry. A ratio of the measured values of the two hippocampi for each feature is used to express the final value of the feature. The ratio is taken for normalization purposes and to avoid the problem of variance in FLAIR signal intensity from case-to-case and scan-to-scan.
- Neuropsychology features: neuropsychological testing measures simple and complex verbal and visual memory

Fig. 5. The individuals plot for the epilepsy data set using the first two dimensions obtained from CCA. 1 means Left side lateralization and 2 means right side lateralization



[11]. One of the ideas that affect the pre-surgical decision making for TLE is the segregation of verbal and non-verbal forms of memory completely and their localization to left or right hippocampi [9].

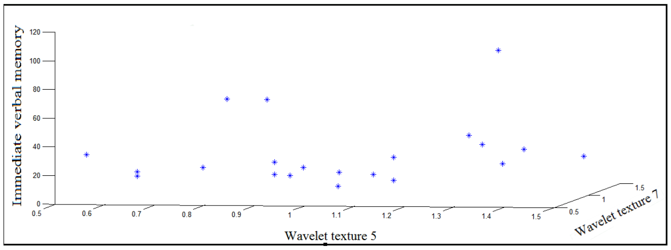
B. Discussion

The goal of the study is to find the correlated features in each set of variables and to highlight the significant correlated features from both sets. In this study, there is a need to find the associations between neuropsychological testing results and the FLAIR/T1 MRI results in order to be more accurate about the laterality of this disorder. As it is known, medical data has noise acquired with the data. So, we need to detect and eliminate the noise. Figure 4 shows the variables plot for the epilepsy data set after applying the preprocessing stage. In this figure, we show only those correlations that are considered to be significant. Figure 5 shows the individuals plot, where 1 means left side lateralization and 2 means right side lateralization. As it is shown in the figure, the first dimension of CCA was able to separate the data points in more than 80% accuracy. Actually, we got two instances having right side lateralization that were not separated correctly. While there is one case with left side lateralization that was misclassified, although that point was close to the separation line.

Figure 6 shows an example of the outliers that are detected and removed by NOFA algorithm. In this Figure, we show three features, two from the MRI data set and one from the neuropsychology data set. In fact, these three features (and others) are found to be significantly correlated using CCA after preprocessing the data sets using NOFA (see Figure 4). The outliers are visually detected in this Figure and automatically detected and removed using NOFA.

The complexity of PCA for a matrix of size $M \times N$, where M is the number of objects and N is the number of features is given by $O(MN^2 + N^3)$ [1]. In NOFA, since each time we are applying PCA on only a couple of features, the time complexity for PCA will be reduced to $O(M)$. Hence, the overall complexity for NOFA is $O(MN^2)$.

Fig. 6. Original three features from the epilepsy data set. Two of them, wavelet texture 5 and wavelet texture 7, are from the MRI study, while the third one, Immediate Verbal Memory, is from the neuropsychology study.



V. Conclusion

In this paper, we have proposed a preprocessing algorithm, NOFA, that can be used to remove the non-informative data points and features, that are most likely outliers and noisy features, before applying the CCA method. The preprocessing algorithm utilizes the principal component analysis method. The need for the proposed preprocessing step tends to be relevant to a number and a variety of real life applications where the noise may be acquired frequently with the data, such as the medical domain.

References

[1] Cheng T. Chu, Sang K. Kim, Yi A. Lin, Yuanyuan Yu, Gary R. Bradski, Andrew Y. Ng, and Kunle Olukotun, *Map-reduce for machine learning on multicore*, NIPS (Bernhard Schölkopf, John C. Platt, and Thomas Hoffman, eds.), MIT Press, 2006, pp. 281–288.

[2] R. Gittins, *Canonical analysis : a review with applications in ecology*, Springer, 1985.

[3] I. Gonzlez, S. Djean, P.G.P. Martin, O. Gonalves, P. Besse, and A. Baccini, *Highlighting relationships between heterogeneous biological data through graphical displays based on regularized canonical correlation analysis*, Journal of Biological Systems **17** (2009), no. 2, 173–199.

[4] H. Hotelling, *Relations between two sets of variates*, Biometrika **28** (1936), 321377.

[5] David Y Ko, *Temporal lobe epilepsy*, eMedicine Neurology (2009).

[6] K. A. L Cao, P.G.P. Martin, C. Robert-Grani, and P. Besse, *Sparse canonical methods for biological data integration: application to a cross-platform study*, BMC Bioinformatics **10** (2009).

[7] Florian Mormann, Ralph G. Andrzejak, Christian E. Elger, and Klaus Lehnertz, *Seizure prediction: the long and winding road*, Brain Journal **130** (2007), no. 2, 314–333.

[8] Elena Parkhomenko, David Tritchler, and Joseph Beyene, *Sparse canonical correlation analysis with application to genomic data integration*, Statistical Applications in Genetics and Molecular Biology **8** (2009), 1–34.

[9] Michael M. Saling, *Verbal memory in mesial temporal lobe epilepsy: beyond material specificity*, Brain: A journal of Neurology **132** (2009), no. 3, 570–582.

[10] Lindsay Smith, *A tutorial on principal components analysis*, Tech. report, Feb 2002.

[11] G. Vingerhoetsa, K. Miattona, M.and Vonckb, R. Seurincka, and P. Boonb, *Memory performance during the intracarotid amobarbital procedure and neuropsychological assessment in medial temporal lobe epilepsy: The limits of material specificity*, Epilepsy & Behavior **8** (2006), no. 2, 422–428.

[12] A. Wiesel, M. Kliger, and A. O. Hero, *A greedy approach to sparse canonical correlation analysis*, submitted to ArXiv, http://arxiv.org/abs/0801.2748 (2008).

[13] H. Wold, *Estimation of principal components and related models by iterative least squares. in multivariate analysis*, New York: Academic Press, 1966.