

# Predicting Neurological Recovery with Canonical Autocorrelation Embeddings

Maria De-Arteaga<sup>1</sup>

Carnegie Mellon University

Machine Learning Department, School of Computer Science  
School of Public Policy, Heinz College

Committee: Artur Dubrawski (Chair), Gilles Clermont<sup>2</sup>, Alexandra Chouldechova

Heinz Second Paper

September 22, 2017

---

<sup>1</sup>Email: mdearte@cmu.edu. Current address: 4800 Forbes Ave. Pittsburgh, PA, USA

<sup>2</sup>University of Pittsburgh Medical Center

# Predicting Neurological Recovery with Canonical Autocorrelation Embeddings

Maria De-Arteaga

Committee: Artur Dubrawski (Chair), Gilles Clermont, Alexandra Chouldechova

## Abstract

In this work we present Canonical Autocorrelation Embeddings, a method for embedding sets of data points onto a space in which they are characterized in terms of their latent complex correlation structures, and where a proposed distance metric enables the comparison of such structures. This methodology is particularly fitting to tasks where each individual or object of study has a batch of data points associated to it, as in for instance patients for whom several vital signs or other health related parameters are recorded over time.

We apply this new methodology to characterize patterns of brain activity of comatose survivors of cardiac arrest, aiming to predict whether they would have a positive neurological recovery. Clinicians routinely face the ethically and emotionally charged decision of whether to continue life support for such patients or not. Both scenarios have potentially grave implications on patients and their close ones, so regardless of whether they believe they have enough information, clinicians are often forced to make a prediction. Our results show that we can identify with high confidence a substantial number of patients who are likely to have a good neurological outcome. Providing this information to support clinical decisions could motivate the continuation of life-sustaining therapies for patients whose data suggest it to be the right choice.

# Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
<b>2</b>	<b>Related work</b>	<b>5</b>
2.1	Component and Correlation Analysis . . . . .	5
2.2	Prognosis of Comatose Survivors of Cardiac Arrest . . . . .	7
<b>3</b>	<b>Data</b>	<b>8</b>
<b>4</b>	<b>Methodology</b>	<b>9</b>
4.1	Canonical Autocorrelation Analysis . . . . .	9
4.2	Non-linear and forbidden correlations . . . . .	12
4.3	Canonical Autocorrelation Embeddings . . . . .	13
4.4	K-Nearest Correlations . . . . .	14
<b>5</b>	<b>Experiments</b>	<b>15</b>
<b>6</b>	<b>Discussion</b>	<b>18</b>
<b>7</b>	<b>Conclusions and future work</b>	<b>21</b>
<b>A</b>	<b>EEG features</b>	<b>23</b>
<b>B</b>	<b>Proof: CAA distance metric</b>	<b>25</b>
<b>C</b>	<b>Principal angles and CAA</b>	<b>26</b>
<b>D</b>	<b>Acknowledgements</b>	<b>27</b>

## 1 Introduction

In many applications, machine learning is not intended as a tool to replace humans, but rather as an instrument to augment human knowledge and enable experts to make better, more informed decisions. Examples of such applications are common in public policy and medicine. In these cases, machine learning can be particularly useful in two ways: as a knowledge discovery tool and as a decision support device. In the first case, the goal is to uncover hidden structures in data that can provide experts with a better understanding of it. In the second case, interpretable algorithms that can accompany predictions with justifications are key to help experts make more informed decisions. In this work, we present Canonical Autocorrelation Embeddings, where the word *autocorrelation* refers to correlations existing between subsets of features of a single set. This work builds on previous research on Canonical Autocorrelation Analysis (CAA), which enables the automatic discovery of multiple-to-multiple correlation structures within a set of features. Through the introduction of a distance metric between CAA correlation structures, we are able to define a feature space embedding where each individual/object is represented by the set of its multivariate correlation structures.

This methodology is particularly fitting to tasks where each individual or object of study has a batch of data points associated to it, e.g., patients for whom several vital signs or multiple channels of brain activity are recorded over time. In the new feature space embedding, traditional machine learning algorithms that rely on distance metrics, such as clustering and K-Nearest Neighbors (K-nn), can be used, with the caveat that unlike traditional settings where each individual is represented by a single data point, in this case each individual is represented by a set of CAA structures in the newly defined space.

Cardiac arrest is deadly, and amongst patients who survive there is a high risk of brain injury. As the brain is deprived of oxygen during the cardiac arrest, many survivors are in a coma state, and predicting positive neurological recovery for this patients remains a challenging task for clinicians. Yet, prognosis is necessary, as clinicians need to constantly decide whether to continue providing life-sustaining therapy to patients. Currently, prognosis relies on observed patterns in EEG activity, and correlations are known to play an important role. Strong correlations between channels are indicative of a poor neurological state, and doctors have identified that correlations are relevant beyond what is evident to the human eye. Theta-coma and alpha-coma are examples of such cases in which the EEG recording seems to appears to show healthy variability between channels but, after some

preprocessing, it becomes evident that the brain activity is dominated by simple, strong, cyclic patterns. Given that correlation structures have been proven to be useful when estimating the chances of positive recovery of comatose patients, this paper aims at applying the proposed methodology to discover multiple-to-multiple correlation structures in each patient’s EEG activity, and using the resulting characterization to predict neurological outcomes. The goal of the project is to determine whether multivariate correlation structures can be helpful in predicting the positive recovery of comatose survivors of cardiac arrest. If successful, this could provide doctors with a prognosis tool that would enable them to better identify patients for whom life-sustaining therapy should be prolonged because they are likely to have a favorable recovery. Our results indicate that it is possible for our method to identify a portion of the patients that, if provided the appropriate care, will go on to have a positive recovery.

In the remainder of this paper, Section 2 presents a brief review of related work, both from the methodological perspective as well as from the medical one. Section 3 discusses the task and data in more detail. In Section 4 a new formulation and solution of the Canonical Autocorrelation Analysis optimization problem is proposed, which provides better theoretical guarantees, extends it to non-linear correlation structures and enables experts to prevent trivial correlations from obfuscating underlying signals. The proposed Canonical Autocorrelation Embeddings method is also described in detail in this section, as well as the use of a K-nn algorithm based on the discovered correlation structures. Section 5 contains the experimental results, Section 6 discusses our findings and Section 7 presents the conclusions and future work.

## 2 Related work

### 2.1 Component and Correlation Analysis

Canonical Correlation Analysis is a statistical method first introduced by [15], useful for exploring relationships between two sets of variables. It is used in machine learning, with applications to medicine, biology and finance, e.g., [11][6][22][23]. Sparse CCA, an  $\ell_1$  variant of CCA, was proposed by [23][24]. This method adds constraints to guarantee sparse solutions, which limits the number of features being correlated. Given two matrices  $X \in \mathbb{R}^{n \times p}$  and  $Y \in \mathbb{R}^{n \times q}$ , CCA aims to find linear combinations of their columns that maximize the correlation between them. Usually,  $X$  and  $Y$  are two matrix representations

for one set of objects, so that each matrix is using a different set of variables to describe them. Assuming  $X$  and  $Y$  have been standardized, the constrained optimization problem is shown in Equation 1. When  $c_1$  and  $c_2$  are small, solutions will be sparse and thus only a few features are correlated.

$$\begin{aligned} & \max_{u,v} u^T X^T Y v \\ & \|u\|_2^2 \leq 1, \|v\|_2^2 \leq 1 \quad \|u\|_1 \leq c_1, \|v\|_1 \leq c_2 \\ & \text{for } 0 \leq c_1 \leq 1, 0 \leq c_2 \leq 1 \end{aligned} \tag{1}$$

The extension of Sparse CCA for discovery of multivariate correlations within a single set of features to study brain imaging has been explored by [11] and [6]. Using the notion of autocorrelation, the authors attempt to find underlying components of functional magnetic resonance imaging (fMRI) and electroencephalogram (EEG), respectively, that have maximum autocorrelation. The types of data used in these work is ordered, both temporally and spatially. To find autocorrelations,  $X$  is defined as the original data matrix and  $Y$  is constructed as a translated version of  $X$ , such that  $Y_t = X_{t+1}$ .

Canonical Autocorrelation Analysis (CAA) is a generalized approach to discovering multiple-to-multiple correlations within a set of features which we have developed previously [4]. Figure 1 illustrates the different use cases of Sparse CCA and CAA. That work has also described the use of such structures in data for anomaly detection. In this paper, we present a different formulation of the CAA optimization problem that provides better theoretical guarantees and allows us to extend the method to the discovery of non-linear correlations by incorporating penalties for disjoint support directly. Additionally, this new formulation also allows for the user to select sets within which correlations are forbidden, which is useful when e.g. trivial correlations should be avoided. More importantly, we introduce a distance metric between canonical autocorrelation structures, which gives substantially more power to CAA-based methodology, making it useful for various learning tasks, such as clustering and classification on datasets and distributions.

The comparison of correlation structures and principal components has been explored in the literature for decades. Most prominently, [18] discusses comparison of principal components between groups. To do so they propose a metric inspired by the concept of congruence coefficient [17], which is nothing but the cosine of the angle between the two  $p$ -dimensional vectors. Also related to our task is [10], where a metric between covariance matrices is proposed. The notion of a distance metric between canonical autocorrelation

$$\left[ \begin{array}{cccc|ccc}
x_{1,1} & \cdots & \cdots & x_{1,p} & y_{1,1} & \cdots & y_{1,q} \\
\vdots & \ddots & \ddots & \vdots & \vdots & \ddots & \vdots \\
x_{n,1} & \cdots & \cdots & x_{n,p} & y_{n,1} & \cdots & y_{n,q}
\end{array} \right]
\quad
\left[ \begin{array}{cc|c}
x_{1,1} & \cdots & x_{1,m} \\
\vdots & \ddots & \vdots \\
x_{n,1} & \cdots & x_{n,m}
\end{array} \right]$$

$X$                              $Y$                              $X$

**Figure 1:** Comparison between scenarios where Sparse CCA and CAA can be used. (**Left**) Sparse CCA set up:  $X$  and  $Y$  are two matrices where the rows correspond to the same items but the columns represent separate sets of variables. Sparse CCA finds sparse multiple-to-multiple linear correlations between subsets of the features in matrix  $X$  and subsets of features in matrix  $Y$ . (**Right**) CAA set up: In cases where there is no natural or intuitive division of the features into two sets, a possible division represented by the dotted line is no longer given. Instead, all features are part of one matrix  $X$ . CAA finds multiple-to-multiple correlations between subsets of features in this matrix.

structures differs from these in that CAA finds a factorization of the correlation matrix where each portion of the correlation matrix is expressed as the outer product of a pair of orthonormal vectors, which define a bi-dimensional space onto which the projected data follows a linear correlation. Section 4.3 discusses the proposed metric.

## 2.2 Prognosis of Comatose Survivors of Cardiac Arrest

In the United States, every year about 350.000 people suffer a sudden out-of-hospital cardiac arrest. The Cardiac Arrest Registry to Enhance Survival (CARES) National Summary Report indicates that for 2016 survival to hospital discharge was 10.8%, and survival with good outcome was 8.5% [19]. Moreover, a study considering all patients admitted to the intensive care unit in a general hospital in the United Kingdom between 1998 and 2003 after a cardiac arrest in the past 24 hours, shows that an estimated two-thirds of the patients dying after out-of-hospital cardiac arrest die due to neurological injury. Section 3 details the outcome statistics in the data set we work with.

The prevalence of cardiac arrest and the high risk of neurological injury amongst survivors has motivated decades of medical research focused on assessing and improving neurological outcomes for comatose survivors of cardiac arrest. Early prognosis of recovery has been explored [1, 8, 3] in an attempt to augment medical knowledge and provide deci-

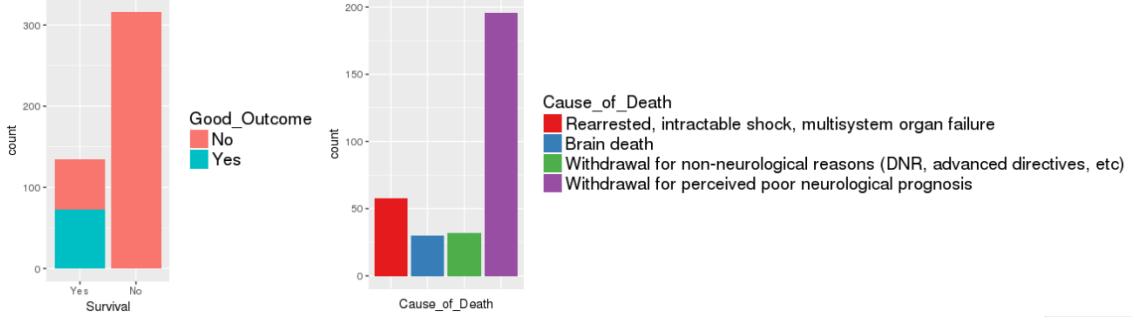
sion support systems for a decision that doctors continuously need to make in this cases: should life-sustaining therapies be continued? Meaningful neurological recoveries are rare, and treatment can be long, expensive and difficult for family members and caregivers [3], as a result, withdrawal and withholding of life support are practiced often [9, 12], even in countries where the legislation forbids it [9]. The impact of specific treatments has also been studied, with several researchers indicating that mild hypothermia treatment might mitigate brain damage [2, 16, 13]. Such research has also suggested that current practice leads to life-sustaining therapies been withdrawn too early [13], and there is a concern that “therapeutic nihilism” and inappropriate care withdrawal currently undermines postarrest critical care [7, 20].

Improving care and making better decisions requires more predictive power and a better understanding of the brain in the periods after cardiac arrest. Research on trajectory modeling using EEG signals [8] indicates that there are signals in the brain that would make it possible to improve prediction accuracy of positive neurological recovery. Amongst what is currently known by doctors regarding good/bad brain activity, correlations are a fundamental piece. Strongly correlated channels are a bad sign, but it has been found that not all negative patterns of correlation are visible to the naked eye, with theta coma being such an example. Motivated by this, our goal is to characterize patients in terms of their multivariate, non-linear structures of correlation and use the resulting featurization to predict their neurological outcome. In section 3 we describe the data in detail.

### 3 Data

The data used in this study corresponds to 451 comatose survivors of cardiac arrest, collected between 2010 and 2015. For each patient, featurized EEG signals at one-per-second resolution are available. The total number of features is 66 and they include artifact detector, seizure probability, amplitude-integrated EEG for the left and right hemisphere, spike detections, and suppression ratios, among other features that doctors have identified as informative. The feature generation and selection was performed by clinicians and the raw EEG channels are not available. The full list of features can be found in Appendix A.

For each patient it is known whether the patient survived. For those who lived it is known whether they had a good or bad outcome, although there is high variance as to what constitutes a bad outcome, as this may or may not refer to their neurological condition. For those who died the cause of death is available. Figure 2 shows this information in



**Figure 2:** Patient labels. Survival and outcome (Left), and cause of death (Right).

detail.

## 4 Methodology

### 4.1 Canonical Autocorrelation Analysis

In this section we present a new formulation and solution for Canonical Autocorrelation Analysis (CAA). Even though it is not radically different from the existing CAA formulation, it provides better guarantees and makes it possible to extend the method in key directions, as discussed in Section 4.2, substantially increasing the method’s descriptive and predictive power.

The goal of CAA is to find multivariate sparse correlations within a single set of variables. In the Sparse CCA framework, this could be understood as having identical matrices  $X$  and  $Y$ . Applying Sparse CCA when  $X = Y$  results in solutions  $u = v$ , corresponding to sparse PCA solutions for  $X$  [24]. The existing formulation of CAA [4] overcomes this issue by introducing a new constraint to induce disjoint support, which is shown in Equation 2.

$$\begin{aligned}
 & \max_{u,v} u^T X^T X v \\
 & \|u\|_2^2 \leq 1, \|v\|_2^2 \leq 1 \quad \|u\|_1 \leq c_1, \|v\|_1 \leq c_2 \\
 & u^T v = 0 \\
 & \text{for } 0 \leq c_1 \leq 1, 0 \leq c_2 \leq 1
 \end{aligned} \tag{2}$$

This formulation enforces that  $u$  and  $v$  must be orthogonal vectors. However, that is

not equivalent to having disjoint support, as there can be cases of orthogonal vectors with overlapping support. Hence, we incorporate the penalty for disjoint support directly, as shown in Equation 3.

$$\begin{aligned}
& \max_{u,v} u^T X^T X v \\
& \|u\|_2^2 \leq 1, \|v\|_2^2 \leq 1 \quad \|u\|_1 \leq c_1, \|v\|_1 \leq c_2 \\
& \sum_{i=1}^m |u_i v_i| = 0 \\
& \text{for } 0 \leq c_1 \leq 1, 0 \leq c_2 \leq 1
\end{aligned} \tag{3}$$

Understanding this as a new generalization of the PMD decomposition [24], the solution for CAA is analogous to that of other PMD-based approximations, although necessary adjustments have to be made to account for the additional constraint. Note that in the CAA optimization problem seen in Equation 3, the equality constraint can be seen as a weighted  $L_1$  penalty when either  $u$  or  $v$  are fixed. Replacing the equality constraint by an inequality constraint gives a biconvex problem, while resulting in the same solution. Therefore, we can solve it through alternate convex search [14], as shown in Algorithm 1.

---

**Algorithm 1:** CAA via alternate convex search

---

```

1 Initialize  $v$  s.t.  $\|v\|_2 = 1$ ;
2 repeat
3    $u \leftarrow \arg \max_u u^T X^T X v$ 
4   s.t.  $\|u\|_2^2 \leq 1, \|u\|_1 \leq c_1, \sum_{i=1}^m |u_i| |v_i| = 0$ 
5    $v \leftarrow \arg \max_v v^T X^T X u$ 
6   s.t.  $\|v\|_2^2 \leq 1, \|v\|_1 \leq c_2, \sum_{i=1}^m |u_i| |v_i| = 0$ 
7 until  $u, v$  converge;
8  $d \leftarrow u^T X^T X v$ ;

```

---

At each iteration, the resulting convex problem can be solved through the Karush-Kuhn-Tucker (KKT) conditions. Without loss of generality, assuming  $v$  is fixed and we are optimizing for  $u$ , the optimization problem in Lagrangian form can be written as formulated in Equation 4.

$$\begin{aligned} \min_u -u^T X^T X v + \sum_{i=1}^m (\lambda_1 |v_i|_1 + \lambda_2) |u_i|_1 \\ + \lambda_3 \|u\|_2^2 - \lambda_2 c_1 - \lambda_3 \end{aligned} \quad (4)$$

for  $0 \leq c_1 \leq 1, 0 \leq c_2 \leq 10 \leq \lambda_1, 0 \leq \lambda_2, 0 \leq \lambda_3$

The KKT conditions are:

- **Stationarity:**  $0 \in -X^T X v + 2\lambda_3 u + \Gamma$

for  $\Gamma_i = (\lambda_2 + \lambda_1 |v_i|) \operatorname{sgn}(u_i) \forall i = 1, \dots, m$

- **Complementary slackness:**

$$\lambda_1 \sum_{i=1}^m |u_i| |v_i| = 0; \quad \lambda_2 (\|u\|_1^2 - c_1) = 0; \quad \lambda_3 (\|u\|_2^2 - 1) = 0$$

- **Primal feasibility:**  $\sum_{i=1}^m |u_i| |v_i| \leq 0; \quad \|u\|_1^2 - c_1 \leq 0; \quad \|u\|_2^2 - 1 \leq 0$

- **Dual feasibility:**  $0 \leq \lambda_i, i = 1, 2, 3$

From complementary slackness and primal feasibility, either  $\lambda_3 = 0$  and  $\|u\|_2 \leq 1$ , or  $\lambda_3 > 0$  and  $\|u\|_2 = 1$ . Assuming  $\lambda_3 > 0$  and solving the stationarity condition, we obtain that for  $i = 1, \dots, m$ , Equation 5 holds, where  $S_\lambda(x)$  is the soft-thresholding operator.

$$2\lambda_3 u_i = S_{(\lambda_1 |v_i| + \lambda_2)}((X^T X v)_i) \quad (5)$$

From complementary slackness,  $\lambda_3$  must be such that  $\|u\|_2 = 1$ , therefore, Equation 6 is obtained.

$$u = \frac{S_{\Phi(v)}(X^T X v)}{\|S_{\Phi(v)}(X^T X v)\|_2^2} \quad (6)$$

$$\begin{aligned} \Phi(v, \lambda_1, \lambda_2) : \mathbb{R}^m &\longrightarrow \mathbb{R}^m \\ v_i &\longrightarrow \lambda_1 |v_i| + \lambda_2 \end{aligned}$$

Additionally,  $\lambda_1$  must be such that  $\sum_{i=1}^m |u_i| |v_i| = 0$ , which will be guaranteed by setting  $\lambda_1 = \max_i \frac{|(X^T X v)_i|}{|v_i|}$ . Finally, either  $\lambda_2 = 0$  results in a feasible solution, or  $\lambda_2$  is

chosen such that  $\|u\|_1 = c_1$ , which can be done through a binary search. The pseudo-code for solving the convex problems at each iteration of the alternate convex search is provided in Algorithm 2, where we solve for  $u$  without loss of generality.

---

**Algorithm 2:** CAA alternate convex search iteration via KKT conditions

---

```

1  $\lambda_1 = \max_i \frac{|(X^T X v)_i|}{|v_i|};$ 
2 if  $\|\frac{S_{\Phi(v\lambda_1,0)}(X^T X v)}{\|S_{\Phi(v\lambda_1,0)}(X^T X v)\|_2^2}\|_1 \leq c_1$  then
3   return  $u = \frac{S_{\Phi(v\lambda_1,0)}(X^T X v)}{\|S_{\Phi(v\lambda_1,0)}(X^T X v)\|_2^2}$ 
4 else
5   Binary search to find  $\lambda_2$  s.t.  $\|\frac{S_{\Phi(v\lambda_1,\lambda_2)}(X^T X v)}{\|S_{\Phi(v\lambda_1,\lambda_2)}(X^T X v)\|_2^2}\|_1 = c_1;$ 
6   return  $u = \frac{S_{\Phi(v\lambda_1,\lambda_2)}(X^T X v)}{\|S_{\Phi(v\lambda_1,\lambda_2)}(X^T X v)\|_2^2}$ 
7 end

```

---

To find multiple pairs of CAA canonical vectors, Algorithm 1 can be repeated iteratively, replacing  $X^T X$  with a matrix in which found correlations are removed, as shown in Equation 7.

$$X^T X - d(uv^T + vu^T) \quad (7)$$

## 4.2 Non-linear and forbidden correlations

In this section we propose a way of tackling two of the main current limitations of CAA by enabling the discovery of non-linear correlations and allowing practitioners to manually select pairwise correlations that should be avoided.

When looking for latent structures of correlation in data, it is desirable to allow for the discovery of non-linear correlations. This is particularly true in the application presented in this paper, where the original data corresponds to waveforms. Given the Weierstrass approximation theorem [5], the most trivial way of doing so is by extending the feature space with powers of the original features. However, the current CAA formulation would likely result in a feature being correlated with transformations of itself. Similarly, a current limitation of the application of CAA is the fact that often datasets contain sets of features that are trivially correlated. For example, in the data we are working with in this project several features correspond to basic statistics of the amplitude integrated EEG.

To overcome both these limitations, we modify the optimization problem to extend the disjoint support to sets of features. Assuming each feature  $x_i$  has a subset  $S_i$  of feature indices associated to it, which indicates which features should not be correlated with  $x_i$ , the resulting optimization problem is presented in Equation 8.

$$\begin{aligned}
& \max_{u,v} u^T X^T X v \\
& \|u\|_2^2 \leq 1, \|v\|_2^2 \leq 1 \quad \|u\|_1 \leq c_1, \|v\|_1 \leq c_2 \\
& \sum_{i=1}^m \sum_{j \in S_i} |u_i v_i| = 0 \\
& \text{for } 0 \leq c_1 \leq 1, 0 \leq c_2 \leq 1
\end{aligned} \tag{8}$$

The new constraint for disjoint support can still be understood as a weighted-L<sub>1</sub> penalty at each iteration of the biconvex algorithm. Hence, the optimization algorithm can still be solved in the way presented in Section 4.1, with the only difference that the parameters of the soft-thresholding operator will change.

### 4.3 Canonical Autocorrelation Embeddings

CAA allows us to find bidimensional projections where the data follows a linear distribution. Each axis of these projections corresponds to a linear combination of the original features, i.e., each CAA canonical space is defined by a pair of vectors  $u, v \in \mathbb{R}^m$ .

Since the correlations discovered by CAA are defined by pairs of vectors in  $\mathbb{R}^m$ , we can measure the distance between two CAA canonical spaces in terms of the Euler angles defining the transformation from one pair of axes to the other. Given that measuring the angle between two vectors is equivalent to measuring the arc between them, and that  $\|u^{(i)}\|_2 = \|v^{(i)}\|_2 = 1 \forall i$ , the distance between two CAA canonical spaces  $C_1$  and  $C_2$  can be defined as shown in Equation 9. Note that the minimum is simply finding the best of two possible ways of doing the transformation.

$$d(C_1, C_2) = \min(\|u_1 - u_2\|_2 + \|v_1 - v_2\|_2, \|u_1 - v_2\|_2 + \|v_1 - u_2\|_2) \tag{9}$$

It is easy to prove that this function satisfies the necessary conditions for well-defined distance, see Appendix B for the proof. Moreover, if two CAA canonical spaces represent the same correlation structure, the vectors defining them must be equal. This stems from

the fact that such correlation structure would take the form of a matrix  $Co \in \mathbb{R}^{m \times m}$ . Therefore, Equation 10 can be seen as a system of linear equations with at most one solution.

$$Co = uv^T \quad (10)$$

Even though we believe this is a good distance metric that captures what we desire to measure, we do not claim this is the only nor the best such metric, and the exploration of better suited metrics could yield even better results. Appendix C contains a short discussion of why principal angles, one of the metrics most commonly used to measure distance between subspaces and which naturally comes to mind in this setting, is not well-suited for this task.

#### 4.4 K-Nearest Correlations

Having a distance metric enables us to use a wide range of machine learning algorithms, such as clustering and K-nn. However, there is a caveat that must be taken into account. While in most traditional cases each data point in the metric space represents an object of study, e.g., a patient, in our case each individual has a set of CAA correlation structures associated to it, which means that each object/individual is represented by multiple points in our embedded space.

This setting can be incorporated into the K-nn framework by calculating the class probability for each correlation structure through the votes of their  $k$  nearest neighbors, and then aggregating over all correlations associated to an object using log-odds, as shown in Equation 11, where  $n_{p,i,j}$  denotes the class label of the  $j$ th neighbor of the  $i$ th correlation of patient  $p$ .

$$\begin{aligned} q_i &= \frac{\sum_{j=1}^k n_{p,i,j}}{k} \\ \hat{y}_p &= \log\left(\prod_{i=1}^{m_p} \frac{q_i}{1 - q_i}\right) \end{aligned} \quad (11)$$

However, it is likely that some type of correlation structures will be common to both classes, while others are discriminative. To reduce noise and allow for those discriminative correlations to lead the decision, we incorporate a threshold  $t$ , so that log-odds are only

calculated over those correlation structures with a class probability that is discriminative enough, as shown in Equation 12. Incorporating this threshold also enhances interpretability, as it reduces the number of structures that are used for making a prediction, making it easier for practitioners to understand which correlation structures are relevant for the task at hand. The parameters  $k$ , indicating the number of neighbors, and  $t$  can be tuned through cross-validation.

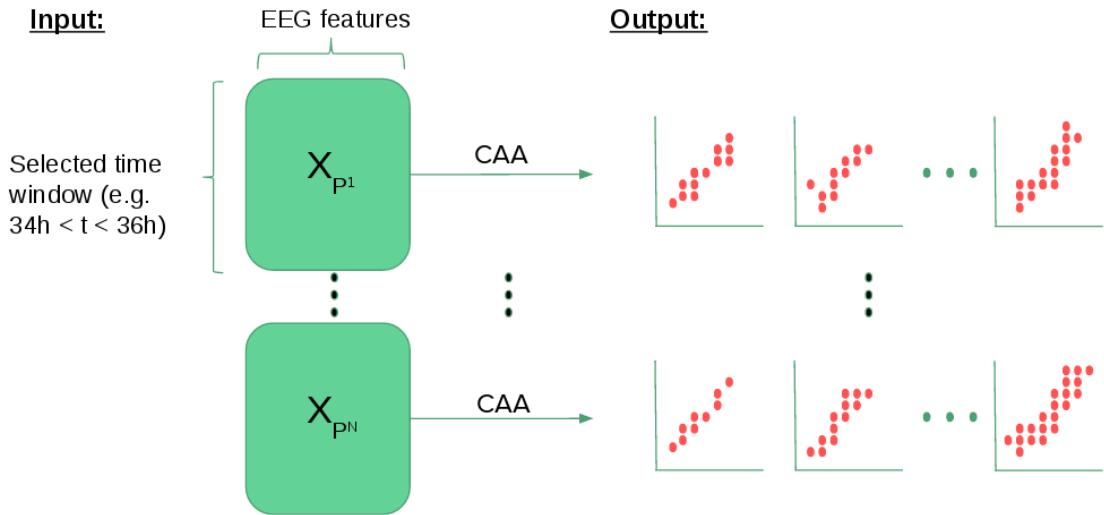
$$\hat{y}_p = \log\left(\prod_{i=1}^{m_p} \mathbb{I}_{(|q_i - 0.5| > t)} \frac{q_i}{1 - q_i}\right) \quad (12)$$

## 5 Experiments

Our goal is to improve care given to comatose survivors of cardiac arrest through a decision support system that can help improve reliability of clinical prognosis. To do so, we propose a way of characterizing patients through their latent multivariate structures of correlation, and using the resulting featurization to build a predictive model.

The first fundamental decision to make is what data and labels to use for training. As it can be seen in Figure 2, the main cause of death for patients in our data set is withdrawal of life-sustaining therapies for perceived poor neurological prognosis. However, as mentioned in Section 2, it is believed that in many cases treatment is been withdrawn too early, hence including this data in our training would have a high risk of introducing bias, as the model could learn and replicate the mistakes doctors are making. Considering this and the fact that our goal is to predict neurological outcome, we train our model using those patients who lived, making our target label whether they had a good or a bad outcome.

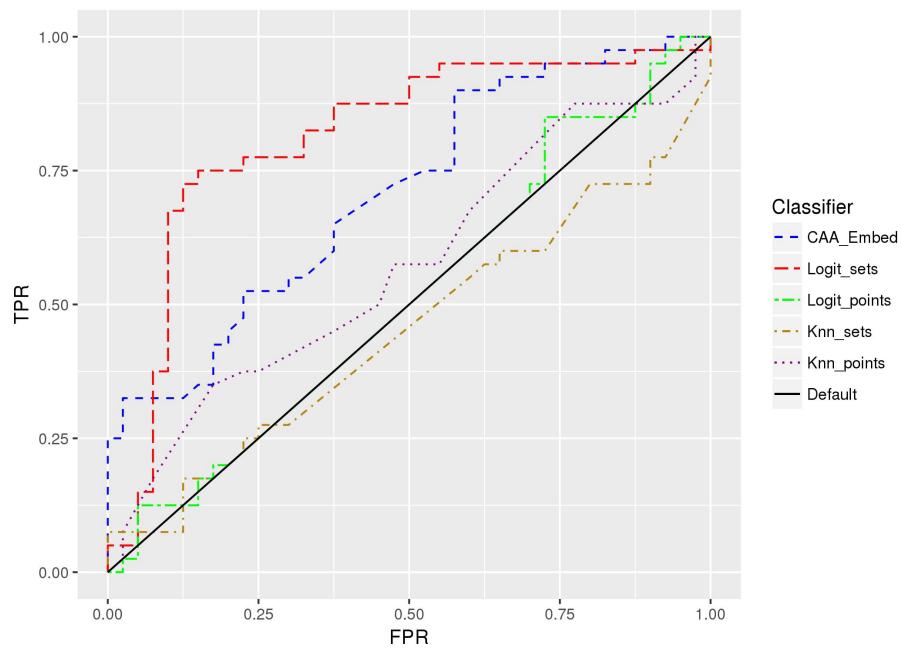
For each patient, their entire stay is recorded, with lengths of stay varying from less than an hour to more than 50 hours. We select those patients who stayed at least 36 hours under monitoring, and use CAA to characterize a period of time of two hours after 34 hours of monitoring. The specific question the proposed model answers is: can the correlations present over a period of time of two hours after 36 hours of monitoring be predictive of whether the patient will have a good neurological outcome? The reason why only two hours are considered is because it can be expected that the state of the patient fluctuates during their state, therefore periods of time that are two long might obfuscate important patterns of correlation. Figure 3 illustrates the characterization of patients' EEG with CAA.



**Figure 3:** Diagram illustrating CAA patient characterization using EEG features as input data.

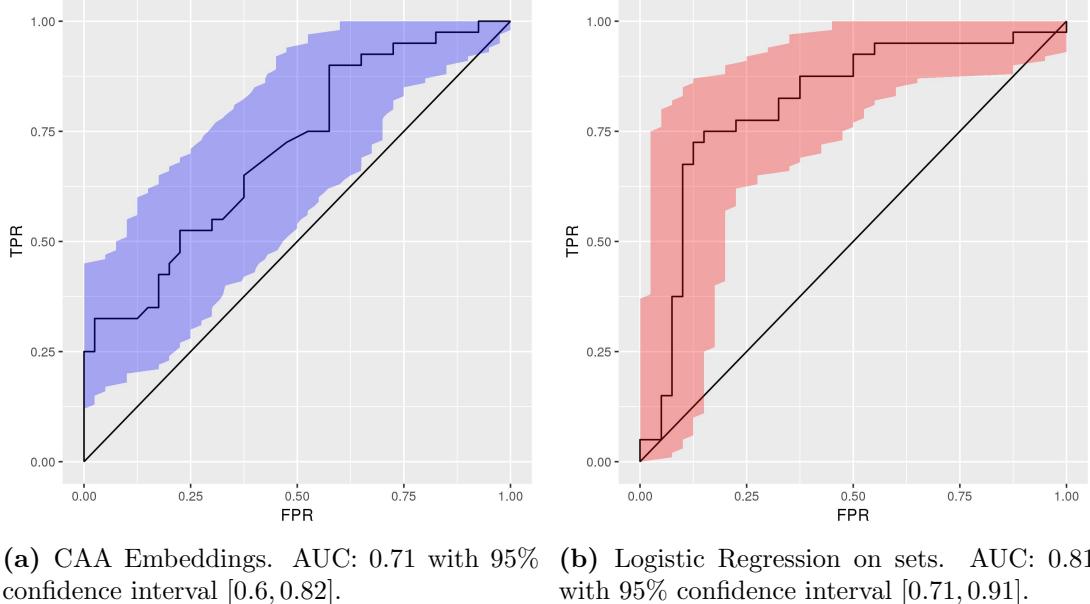
To avoid spurious correlations from misleading the model, only discovered correlations with an  $R^2 > 0.25$  are considered. Moreover, given that beyond a certain threshold the distance metric is not informative since the correlations are no longer related, we can assume our set of correlations to be not-complete and only take into account those distances smaller than  $\sqrt{2}$ , a threshold that corresponds to a  $90^\circ$  rotation over one axis. Using the resulting distance matrix we apply K-Nearest Correlations on the space of CAA Embeddings. The prediction results obtained through 10-fold cross-validation, tuning parameters  $k$  and  $t$  through 10-fold cross-validation as well, are presented in Figures 4, 5a, 6a, 7a.

We also fit a Logistic Regression with Lasso regularization [21] to predict recovery 36 hours after cardiac arrest. Given that Logistic Regression is not suited for sets, but rather takes as input individual data points, we explore two avenues. The first approach consists on taking the last data point after 36 hours of monitoring, that is, the recording at one time step. For the remainder of the paper, we will refer to this approach as *Logistic regression on points*. In the second approach, in an attempt to provide the model with information pertaining the two hours used by the Canonical Autocorrelation Embedding model, we calculate quintiles for each feature over this two hours, and provide them as features to the model. For the remainder of the paper, we will refer to this approach as *Logistic regression on sets*. The choice of the  $\lambda$  parameter is made through 10-fold cross-validation. The



**Figure 4:** ROC curves displaying false positive rate in  $x$ -axis and true positive rate in  $y$ -axis showing performance of CAA Embeddings, Logistic Regression on sets, Logistic Regression on points, K-nn on sets and K-nn on points.

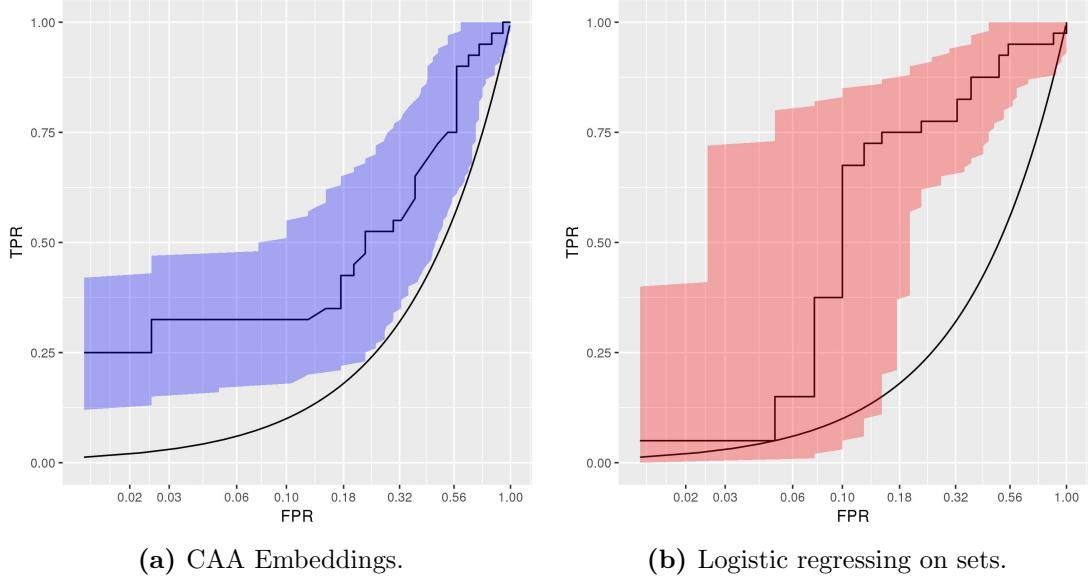
results are included in Figures 4, 5b, 6b, 7b. Furthermore, to be able to better assess the role of the CAA Embeddings, we also include results for K-nn using Euclidean distance taking the same inputs as the Logistic Regression models, which we will refer to as *K-nn on points* and *K-nn on sets*, respectively.



**Figure 5:** ROC curves with 95% confidence intervals displaying false positive rate in  $x$ -axis and true positive rate in  $y$ -axis.

## 6 Discussion

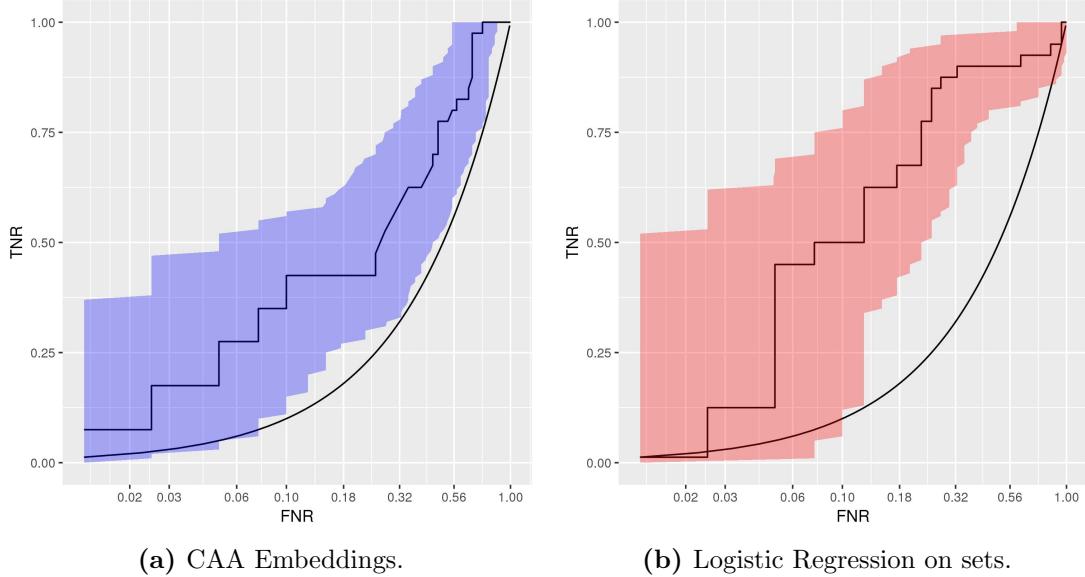
The results presented in Section 5 show that the proposed methodology has predictive power, and the comparison to K-nn using Euclidean distance on points and set-aggregated features highlights the role of CAA Embeddings. The proposed methodology has an Area Under the Curve (AUC) of 0.71 with a 95% confidence interval [0.6, 0.82], while Logistic Regression with Lasso regularization has an AUC of 0.81 with a 95% confidence interval [0.71, 0.91]. From this, it can be concluded that the Logistic Regression approach has overall better performance. However, in applications like the one explored in this paper, machine learning would not be used to automate a process. Rather, it would be used as a decision support system that only makes a recommendation when the algorithm has



**Figure 6:** ROC curve with 95% confidence intervals displaying false positive rate in  $x$ -axis and true positive rate in  $y$ -axis, with  $x$ -axis in log-scale to emphasize area of low-false positive rate.

high confidence on its prediction. In this circumstances, it is important to observe the performance at low false positive rates and false negative rates, given that these are the thresholds that would be used in practice. The ROC curves with the  $x$ -axis in logarithmic scale to emphasize the low false positive and low false negative rates are shown in Figure 6 and Figure 7, respectively. The performance of CAA Embeddings at low false positive rates is promising, while the performance of Logistic Regression at both low false positives and low false negative rates is not above random. Hence, the apparent good performance of Logistic Regression reflected in its AUC is misleading in this case, as the model could not be used for clinical decision support. Meanwhile, the proposed model can identify with high confidence patients who will likely go on to have a good neurological outcome.

Even though several voices within the medical community advocate for maintaining life-sustaining therapies for at least 72 hours [7, 20], the burden associated to continuing life-support for patients who will not have a positive neurological recovery still leads doctors to end treatment when they do not see signs of positive outcomes. Hence, being able to identify with high confidence patients that will recover with good outcome has the potential to save lives. Figure 6a shows that we can identify 25% of the patients that will recover with



**Figure 7:** ROC curves with 95% confidence intervals displaying false negative rate in  $x$ -axis and true negative rate in  $y$ -axis, with  $x$ -axis in log-scale to emphasize area of low-false negative rate.

little chance of incurring in false positives. And even if we are conservative and consider the lower bound of the confidence interval, that would correspond to 12.5% of patients that go on to recover.

It is hard to evaluate the immediate medical impact of these findings in the absence of clinical context. To appropriately estimate the potential impact of such a decision support system in terms of lives saved, it would be necessary to compare against physicians' assessments. Therefore, the results found are promising, but at this point there are no direct clinical implications. Our model would only be useful if it had the potential to change the prior beliefs a clinician has regarding the outcome of a patient, a comparison that would require access to clinicians' prognosis throughout the patients' stay.

From the ROC curve in Figures 6a and 7a, it can be seen that it is easier for the model to determine if patients will have a positive neurological recovery than if they will not. However, this must be taken with a grain of salt, and we cannot yet conclude that correlation structures in the brain are more useful to predict positive than negative outcomes. The available labels denote positive/negative outcome, but this is not limited to neurological activity. A patient could have a positive neurological recovery but have other types of

complications, which would result in a *bad outcome* label. Meanwhile, the positive recovery label is much more homogeneous and is sure to indicate positive neurological recovery (as well as positive recovery in other arrays). The fact that some patients who had positive neurological recovery could be labeled as having a bad outcome might be adding noise, and it is possible that a cleaner dataset would increase predictive power for those patients who will not go on to have a good neurological outcome.

## 7 Conclusions and future work

Cardiac arrest is a leading cause of death around the world, coma after cardiac arrest is common, and good neurological recovery is rare. Everyday, clinicians are tasked with making a prediction that determines whether they will continue life-sustaining therapies for these patients or not. Motivated by the emphasis the clinicians place on potential informativeness of the correlation structures in EEG data, we have proposed a way to characterize and compare patients based on the latent structures of multivariate non-linear correlations, and use such information to predict positive neurological recovery. To do so, we have extended Canonical Autocorrelation Analysis to allow for the discovery of non-linear correlations and for the enforcement of disjoint support for sets of features in order to prevent potentially strong but trivial correlations from obfuscating parsimonious but important structures. Additionally, we have proposed Canonical Autocorrelation Embeddings to enable the comparison of discovered correlation structures, making it possible to leverage the power of machine learning methodologies that rely on the use of distance metrics.

Future work would require training and validating the model with more data. It is reasonable to believe that there is high variability across patients, hence more data of more subjects is crucial to ensure reproducibility of results. It would also be desirable to obtain better quality labels for the negative outcomes, as at this point bad outcomes of different kinds are conflated under a single label. Additionally, taking into account the results presented in the literature, the power of the model could be enhanced by incorporating trajectory modeling, rather than using only two hours of data for training the models in the way that currently ignores their sequential structure.

Applying Canonical Autocorrelation Embeddings to the raw EEG channels rather than the aggregated featurization of data would also be interesting to explore once that data becomes available to us. Our motivation to characterize brain activity with CAA comes

from the importance clinicians give to correlations. However, the correlations they know to be informative are across raw EGG channel measurements, and it is likely that at the current level of data aggregation a big portion of the information is obfuscated.

Finally, there is a source of bias in our model given that the data with which we train it is limited to those patients for whom life-sustaining therapies were continued. If patients for whom treatment was stopped early are significantly different from those in our training set, which is very likely the case, our model could perform poorly on this faction of the population. Tackling this selective labeling problem in a way that allows us to incorporate doctors' knowledge while not reproducing their mistakes is a key phase in the path to successfully using machine learning to save human lives.

## A EEG features

Below is the complete list of features available and used in this study.

Feature	Details
Artifact Intensity	Muscle
Artifact Intensity	Chew
Artifact Intensity	V-Eye
Artifact Intensity	L-Eye
Artifact Detector	
Seizure Probability	Max
aEEG, Left Hemisphere	Min
aEEG, Left Hemisphere	Median
aEEG, Left Hemisphere	Q75%
aEEG, Left Hemisphere	Q25%

aEEG, Right Hemisphere	Max
aEEG, Right Hemisphere	Min
aEEG, Right Hemisphere	Median
aEEG, Right Hemisphere	Q75%
aEEG, Right Hemisphere	Q25%
aEEG+(custom filt.)(LFF0.16sec, HFF(off), custom aEEG 2-20 512), Left Hem.	Max
aEEG+(custom filt.)(LFF0.16sec, HFF(off), custom aEEG 2-20 512), Left Hem.	Min
aEEG+(custom filt.)(LFF0.16sec, HFF(off), custom aEEG 2-20 512), Left Hem.	Median
aEEG+(custom filt.)(LFF0.16sec, HFF(off), custom aEEG 2-20 512), Left Hem.	Q75%
aEEG+(custom filt.)(LFF0.16sec, HFF(off), custom aEEG 2-20 512), Left Hem.	Q25%
aEEG+(custom filt.)(LFF0.16sec, HFF(off), custom aEEG 2-20 512), Right Hem.	Max
aEEG+(custom filt.)(LFF0.16sec, HFF(off), custom aEEG 2-20 512), Right Hem.	Min
aEEG+(custom filt.)(LFF0.16sec, HFF(off), custom aEEG 2-20 512), Right Hem.	Median
aEEG+(custom filt.)(LFF0.16sec, HFF(off), custom aEEG 2-20 512), Right Hem.	Q75%
aEEG+(custom filt.)(LFF0.16sec, HFF(off), custom aEEG 2-20 512), Right Hem.	Q25%
PeakEnvelope, 1 - 20 Hz, Left Hemisphere	
PeakEnvelope, 1 - 20 Hz, Right Hemisphere	
Spike Detections	
Suppression Ratio, Left Hemisphere	
Suppression Ratio, Right Hemisphere	
FFT Power, 1 - 4 Hz, Left Hemisphere	
FFT Power, 1 - 4 Hz, Right Hemisphere	
FFT Power, 4 - 8 Hz, Left Hemisphere	
FFT Power, 4 - 8 Hz, Right Hemisphere	
FFT Power, 8 - 13 Hz, Left Hemisphere	
FFT Power, 8 - 13 Hz, Right Hemisphere	
FFT Power, 13 - 20 Hz, Left Hemisphere	
FFT Power, 13 - 20 Hz, Right Hemisphere	
FFT Alpha/Delta, 8-13/1-4 Hz, Left Hemisphere	
FFT Alpha/Delta, 8-13/1-4 Hz, Right Hemisphere	
Rhythmicity Spectrogram, Left Hemisphere	1-4Hz
Rhythmicity Spectrogram, Left Hemisphere	4-8Hz
Rhythmicity Spectrogram, Left Hemisphere	8-13Hz
Rhythmicity Spectrogram, Left Hemisphere	13-20Hz
Rhythmicity Spectrogram, Right Hemisphere	1-4Hz
Rhythmicity Spectrogram, Right Hemisphere	4-8Hz
Rhythmicity Spectrogram, Right Hemisphere	8-13Hz
Rhythmicity Spectrogram, Right Hemisphere	13-20Hz

## B Proof: CAA distance metric

In this section we prove that the metric defined to measure the distance between CAA canonical spaces satisfies the necessary conditions to be a well-defined distance.

$$d(C_1, C_2) = \min(||u_1 - u_2||_2 + ||v_1 - v_2||_2, ||u_1 - v_2||_2 + ||v_1 - u_2||_2) \quad (13)$$

- Non-negativity: stems directly from the non-negativity of the  $\ell_2$  norm, together with the fact that the set of non-negative real numbers is closed under the summation and minimum operations.
- Identity:

$$\begin{aligned} 0 &= \min(||u_1 - u_2||_2 + ||v_1 - v_2||_2, ||u_1 - v_2||_2 + ||v_1 - u_2||_2) \\ \Leftrightarrow 0 &= ||u_1 - u_2||_2 + ||v_1 - v_2||_2 \vee 0 = ||u_1 - v_2||_2 + ||v_1 - u_2||_2 \\ &\Leftrightarrow (0 = ||u_1 - u_2||_2 \wedge 0 = ||v_1 - v_2||_2) \\ &\quad \vee (0 = ||u_1 - v_2||_2 \wedge 0 = ||v_1 - u_2||_2) \\ &\Leftrightarrow (u_1 = u_2 \wedge v_1 = v_2) \\ &\quad \vee (u_1 = v_2 \wedge v_1 = u_2) \end{aligned}$$

Given that we are dealing with these as non-ordered pairs,  $d(C_1, C_2) = 0 \Leftrightarrow C_1 = C_2$ .

- Symmetry: Stems directly from the fact that we define  $C_1$  and  $C_2$  as non-ordered pairs, hence the definition of the distance for each is exactly the same.
- Triangle inequality: The triangle inequality comes as a result of the triangle inequality of the  $\ell_2$  norm. We want to show that

$$d(C_1, C_3) \leq d(C_1, C_2) + d(C_2, C_3)$$

$$\begin{aligned} d(C_1, C_3) &\leq ||u_1 - u_3||_2 + ||v_1 - v_3||_2 \\ &= ||u_1 - u_3 + u_2 - u_2||_2 + ||v_1 - v_3 + v_2 - v_2||_2 \\ &\leq ||u_1 - u_2||_2 + ||u_2 - u_3||_2 + ||v_1 - v_2||_2 + ||v_2 - v_3||_2 \\ &= ||u_1 - u_2||_2 + ||v_1 - v_2||_2 + ||u_2 - u_3||_2 + ||v_2 - v_3||_2 \end{aligned}$$

Through an analogous process,

$$\begin{aligned} d(C_1, C_3) &\leq \|u_1 - u_3 + v_2 - v_2\|_2 + \|v_1 - v_3 + u_2 - u_2\|_2 \\ &\leq \|u_1 - v_2\|_2 + \|v_1 - u_2\|_2 + \|v_2 - u_3\|_2 + \|u_2 - v_3\|_2 \end{aligned}$$

Additionally, the following is also true:

$$d(C_1, C_3) \leq \|u_1 - v_3\|_2 + \|v_1 - u_3\|_2$$

Therefore, through analogous reasoning, we derive the following two sets of inequalities:

$$\begin{aligned} d(C_1, C_3) &\leq \|u_1 - v_3 + u_2 - u_2\|_2 + \|v_1 - u_3 + v_2 - v_2\|_2 \\ &\leq \|u_1 - u_2\|_2 + \|v_1 - v_2\|_2 + \|u_2 - v_3\|_2 + \|v_2 - u_3\|_2 \end{aligned}$$

$$\begin{aligned} d(C_1, C_3) &\leq \|u_1 - v_3 + v_2 - v_2\|_2 + \|v_1 - u_3 + u_2 - u_2\|_2 \\ &\leq \|u_1 - v_2\|_2 + \|v_1 - u_2\|_2 + \|v_2 - v_3\|_2 + \|u_2 - u_3\|_2 \end{aligned}$$

The four inequalities we have derived span the four possible cases for  $d(C_1, C_2) + d(C_2, C_3)$ , which concludes our proof.

## C Principal angles and CAA

Although principal angles might initially seem like a good alternative to measure distances between CAA canonical spaces, note that this is not a viable option. Even though each pair of vectors defining a CAA canonical space constitute an orthonormal basis of a subspace, two orthogonal basis defining the same subspace do not represent the same correlation structure. This can be derived from the fact that, as shown in Section 4.3, two different pairs of vectors cannot represent the same correlation structure. It is also easy to under-

stand why this would not be the case with a simple counterexample in  $\mathbb{R}^3$ . Consider the following two pairs of vectors:

$$\begin{cases} u_1 = (1, 0, 0) \\ v_1 = (0, 1, 0) \end{cases} \quad \begin{cases} u_2 = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0) \\ v_2 = (\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0) \end{cases}$$

Even though they are both orthonormal bases of the same subspace,  $u_1 v_1^T \neq u_2 v_2^T$ .

## D Acknowledgements

This work has been partially supported by DARPA under awards FA8750-12-2-0324 and FA8750-17-2-0130.

## References

- [1] Bassetti, C.; Bomio, F.; Mathis, J.; and Hess, C. W. 1996. Early prognosis in coma after cardiac arrest: a prospective clinical, electrophysiological, and biochemical study of 60 patients. *Journal of Neurology, Neurosurgery & Psychiatry* 61(6):610–615.
- [2] Bernard, S. A.; Gray, T. W.; Buist, M. D.; Jones, B. M.; Silvester, W.; Gutteridge, G.; and Smith, K. 2002. Treatment of comatose survivors of out-of-hospital cardiac arrest with induced hypothermia. *New England Journal of Medicine* 346(8):557–563.
- [3] Booth, C. M.; Boone, R. H.; Tomlinson, G.; and Detsky, A. S. 2004. Is this patient dead, vegetative, or severely neurologically impaired?: assessing outcome for comatose survivors of cardiac arrest. *Jama* 291(7):870–879.
- [4] De-Arteaga, M.; Dubrawski, A.; and Huggins, P. 2016. Canonical Autocorrelation Analysis for Radiation Threat Detection. *Heinz College First Paper / Data Analysis Project Machine Learning Department, Carnegie Mellon University*.
- [5] De Branges, L. 1959. The stone-weierstrass theorem. *Proceedings of the American Mathematical Society* 10(5):822–824.
- [6] De Clercq, W.; Vergult, A.; Vanrumste, B.; Van Paesschen, W.; and Van Huffel, S. 2006. Canonical correlation analysis applied to remove muscle artifacts from the electroencephalogram. *Biomedical Engineering, IEEE Transactions on* 53(12):2583–2587.

- [7] Elmer, J., and Callaway, C. W. 2017. The brain after cardiac arrest. In *Seminars in neurology*, volume 37, 019–024. Thieme Medical Publishers.
- [8] Elmer, J.; Gianakas, J. J.; Rittenberger, J. C.; Baldwin, M. E.; Faro, J.; Plummer, C.; Shutter, L. A.; Wassel, C. L.; Callaway, C. W.; Fabio, A.; et al. 2016. Group-based trajectory modeling of suppression ratio after cardiac arrest. *Neurocritical care* 25(3):415–423.
- [9] Ferrand, E.; Robert, R.; Ingrand, P.; Lemaire, F.; et al. 2001. Withholding and withdrawal of life support in intensive-care units in france: a prospective survey. *The Lancet* 357(9249):9–14.
- [10] Förstner, W., and Moonen, B. 2003. A metric for covariance matrices. In *Geodesy-The Challenge of the 3rd Millennium*. Springer. 299–309.
- [11] Friman, O.; Borga, M.; Lundberg, P.; and Knutsson, H. 2002. Exploratory fmri analysis by autocorrelation maximization. *NeuroImage* 16(2):454–464.
- [12] Geocadin, R.; Buitrago, M.; Torbey, M.; Chandra-Strobos, N.; Williams, M.; and Kaplan, P. 2006. Neurologic prognosis and withdrawal of life support after resuscitation from cardiac arrest. *Neurology* 67(1):105–108.
- [13] Gold, B.; Puertas, L.; Davis, S. P.; Metzger, A.; Yannopoulos, D.; Oakes, D. A.; Lick, C. J.; Gillquist, D. L.; Holm, S. Y. O.; Olsen, J. D.; et al. 2014. Awakening after cardiac arrest and post resuscitation hypothermia: are we pulling the plug too early? *Resuscitation* 85(2):211–214.
- [14] Gorski, J.; Pfeuffer, F.; and Klamroth, K. 2007. Biconvex sets and optimization with biconvex functions: a survey and extensions. *Mathematical Methods of Operations Research* 66(3):373–407.
- [15] Hotelling, H. 1936. Relations between two sets of variates. *Biometrika* 321–377.
- [16] Hypothermia after Cardiac Arrest Study Group. 2002. Mild therapeutic hypothermia to improve the neurologic outcome after cardiac arrest. *N Engl J Med* 346(5):549–556.
- [17] Korth, B., and Tucker, L. R. 1976. Procrustes matching by congruence coefficients. *Psychometrika* 41(4):531–535.

- [18] Krzanowski, W. 1979. Between-groups comparison of principal components. *Journal of the American Statistical Association* 74(367):703–707.
- [19] Mozaffarian, D.; Benjamin, E. J.; Go, A. S.; Arnett, D. K.; Blaha, M. J.; Cushman, M.; Das, S. R.; de Ferranti, S.; Després, J.-P.; Fullerton, H. J.; et al. 2016. Heart disease and stroke statistics2016 update. *Circulation* 133(4):e38–e360.
- [20] Mulder, M.; Gibbs, H. G.; Smith, S. W.; Dhaliwal, R.; Scott, N. L.; Sprenkle, M. D.; and Geocadin, R. G. 2014. Awakening and withdrawal of life-sustaining treatment in cardiac arrest survivors treated with therapeutic hypothermia. *Critical care medicine* 42(12):2493.
- [21] Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288.
- [22] Todros, K., and Hero, A. 2012. Measure transformed canonical correlation analysis with application to financial data. In *Sensor Array and Multichannel Signal Processing Workshop (SAM), 2012 IEEE 7th*, 361–364. IEEE.
- [23] Witten, D. M., and Tibshirani, R. J. 2009. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical applications in genetics and molecular biology* 8(1):1–27.
- [24] Witten, D. M.; Tibshirani, R.; and Hastie, T. 2009. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* kxp008.