# Canonical Autocorrelation Analysis for Radiation Threat Detection

April 20, 2016

Maria De Arteaga[1]

Carnegie Mellon University

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

Committee: Artur Dubrawski (Chair), Daniel Nagin, Simon Labov[2]

**Abstract**

This paper introduces Canonical Autocorrelation Analysis (CAA), a method for finding multiple-to-multiple linear correlations within a single set of features. CAA can be useful when looking for hidden parsimonious structures of relationships in data, each involving only a small subset of all features. Correlations discovered by CAA are highly interpretable as they are formed by pairs of sparse linear combinations of the original features of the data. In addition, this results in useful visualizations of data. We show how CAA can be of use as a tool for anomaly detection when a structure of correlations characteristic of the expected data is not followed by anomalous data. In the case of radiation threat detection, this allows characterizing patterns of correlations between subset sets of bins of gamma-ray spectra known to represent benign background radiation, and flagging spectral measurements that do not follow the same patterns of correlations as anomalies potentially reflective of the presence of radiation threats. The resulting spectral anomaly detection technique performs substantially better than an unsupervised alternative prevalent in the domain, while providing valuable additional insights for threat analysis.

[1]Email: mdeartea@cmu.edu. Current address: 4800 Forbes Ave. Pittsburgh, PA, USA
[2]Lawrence Livermore National Laboratory

# Contents

# 1 Introduction

Ever since the invention of nuclear weapons, radiation threat detection has been a security priority around the globe. Even though the total number of weapons has declined since the Cold War, a continued investment in nuclear arsenal has increased the destruction capacity of existing warheads, thus the threats that characterized the Cold War are still a main concern for the international community [23].

Furthermore, a vast number of such weapons are tactical nuclear weapons, which are characterized by their incapability to inflict strategically decisive damage to the military or economy of the target, a trait that has kept them out of current nuclear arms control arrangements. Many of them are kept under dubious security standards, as is the case of many that are stored in remote, hard-to-defend locations, and to make matters worse, these are small and portable warheads that, although incapable of devastating a country's economy in a single blow, would cause large-scale destruction if used. All together, this makes up for a dangerous recipe that increases the risk of nuclear warheads falling in the hands of terrorists [15].

Robbery of stolen fissile material that can be used to build radiological devices, commonly known as"dirty bombs", is also a concern for governments [16]. This problem reached its peak after the collapse of the Soviet Union, when the risk of people who are unaware of the dangers of radioactive material getting hold of it was illustrated by the case of a man who died of radiation sickness after storing material stolen from a nuclear waste facility in a kitchen cabinet [10]. Such risks are still present; in 2015 international alerts were triggered after a container full of medical isotopes was stolen in Mexico, where two years earlier thieves accidentally got hold of a container with radioactive material used in medical equipment[20]. Even though the destruction capacity of such devices cannot be compared to that of a nuclear warhead [16], they could expose thousands of people to dangerous levels of radiation [2].

Thus, effectively monitoring borders to prevent smuggling of radioactive threats, as well as monitoring the interior for signatures of possible threats, are crucial needs for many countries, particularly for those at risk of being targeted by terrorist groups. This, however, is not an easy task. Radioactive materials are typically shielded, and the shielding can be engineered to make detection harder. In addition, faint sources of potentially harmful radiation can be hard to detect in scenes where intensity and spectral characteristics of benign background radiation vary significantly. This is often the case in man-made envi-

ronments, and so area searches for radiation threats throughout cities or inside buildings have to recognize and address this challenge.

An additional challenge comes from the fact that different types of threats follow different spectral patterns, and even if templates of some common threat types are available, relying on supervised analysis of field data is risky. Supervised detectors may fail to detect threats that were not present in the training data, or which were shielded in a particularly unexpected fashion. Therefore, efforts have been made to develop unsupervised methods that successfully detect threats without relying on source templates.

This paper presents a novel machine learning method - Canonical Autocorrelation Analysis - designed for finding multiple-to-multiple linear correlations within a single set of variables. The capability of automatically identifying multivariate structures of correlation make CAA naturally fitting in data-driven discovery, and it can be used as a building block for interpretable single-class learning and unsupervised-learning algorithms. Applying CAA for radiation threat detection enables us to characterize harmless radiation with a structure of correlations spanning sets of energy bins. Once this characterization is established, it can be used for spectral anomaly detection, as threats can be expected to not follow the same characterization as harmless ambience. We show in experiments that the ability of CAA to identify parsimonious subsets of features and later use it to model background radiation variability makes it more robust at threat detection than one of the most popular unsupervised methods used in the domain: a Principal Component Analysis (PCA) spectral anomaly detection method that considers all dimensions of spectra in linear combinations corresponding to subsequent principal components [24].

CAA is an extension of Canonical Correlation Analysis (CCA)[11]. CCA and its variant Sparse CCA[27] are useful tools for finding multivariate correlations between two sets of features. They work well when the features describing a task at hand are naturally divided into separate sets. For example, it can be used to find correlations between genes and diseases, when both types of information are available for a set of patients.

We believe that the capability of identifying relationships between sets of features is of general interest, even in cases when the natural or intuitive splits of features into separate subsets are not obvious. Therefore, we developed CAA, a method that discovers pairs of subsets of features that are maximally correlated, and just like CCA, CAA can identify multiple such pairs if the corresponding correlations exist. Figures 1a and 1b illustrate the different use cases of Sparse CCA and CAA, respectively. Sparse CCA finds multiple-to-multiple linear correlations between subsets of the features in a matrix $X \in \mathbb{R}^{n \times p}$ and

subsets of features in a matrix $Y \in \mathbb{R}^{n \times q}$, where the rows of both matrices correspond to the same items but the columns represent separate sets of variables, e.g., $X$ being genes and $Y$ being diseases. When the division of the features into two sets is not natural we have a single matrix $X \in \mathbb{R}^{n \times m}$ where all features are merged together, CAA then finds multiple-to-multiple correlations between disjoint subsets of these features.

$$\left[\begin{array}{cccc|ccc} x_{1,1} & \cdots & \cdots & x_{1,p} & y_{1,1} & \cdots & y_{1,q} \\ \vdots & \ddots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & \cdots & x_{n,p} & y_{n,1} & \cdots & y_{n,q} \end{array}\right]$$
$$\underbrace{\qquad\qquad}_{X} \qquad \underbrace{\qquad\qquad}_{Y}$$

**(a)** 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 $Y$.

$$\left[\begin{array}{ccc:ccc} x_{1,1} & \cdots & & \cdots & x_{1,m} \\ \vdots & \ddots & & \ddots & \vdots \\ x_{n,1} & \cdots & & \cdots & x_{n,m} \end{array}\right]$$
$$\underbrace{\qquad\qquad\qquad\qquad}_{X}$$

**(b)** 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 pairs of disjoint subsets of features in this matrix that yield strong correlations.

**Figure 1:** Comparison between scenarios where Sparse CCA and CAA can be used.

Correlations retrieved by CAA can be seen as bidimensional projections where a scatter plot of the data follows a linear correlation and each axis corresponds to a sparse linear combination of the features, with both axes having disjoint feature support (eg. as in Figure 2b).

Extraction of informative projections has been tackled in the past [7][8]. Our work differs from those in two primary ways. First, each of our axes is defined by a linear combination of features, rather than a single feature, which enables the discovery of complex structures without loosing interpretability. Secondly, our method is unsupervised and its objective is to retrieve projections that contain multiple-to-multiple linear correlations, rather than focus on separating the classes.

Other methods for finding sparse representations of data comprised in a single matrix include the well-known Sparse Principal Component Analysis (Sparse PCA). However, CAA is fundamentally different from Sparse PCA. While CAA resembles Sparse PCA in the sense that it finds sparse representations of data contained in one matrix, Sparse PCA maximizes retained *variance* of data in one-dimensional projections, while CAA finds two-dimensional projections where *correlation* between two subsets of features is maximized. CAA specifically seeks projections composed by pairs of strongly correlated linear combinations of features, enabling discovery of hidden characteristic correlations in data, which cannot be easily found with other methods such as Sparse PCA. Section 4.2 goes into the details of the difference between the two methods from a theoretical perspective.

In the remainder of this paper, Section 2 contains a description of the data, Section 3 presents a brief review of related work, in Section 4 the proposed methods are described in detail, and Section 5 reviews the experiments when applying CAA to both synthetic and radiation threat detection data. Section 6 explores the applicability of the method to other domains, using the Breast Cancer Wisconsin benchmark data set. Finally, Section 7 concludes the paper and discusses future work directions.

## 2 Data description

Radiation is often characterized using gamma-ray spectra, which are typically represented as vectors of photon counts registered by the sensor at subsequent discrete and disjoint intervals of energy. These vectors, called in the application domain the energy spectra, become data points for analysis. In our research, 128 energy bins are used, thus each data point is a vector in $\mathbb{R}^{128}$.

There are two types of data used in this research: harmless background and threat-infected background.

- *Harmless background* Over a period of five consecutive days a truck drove around downtown Sacramento, California, with a double 4x16 NaI planar detector on its back. The data contains approximately 70,000 one-second observations collected enroute, that reflect background radiation as well as any nuisance sources.

- *Threat-injected background* Synthetic threat injections done at the Lawrence Livermore National Laboratory. Simulated data mimics 15 types of sources. For each

source, a data set of 10,000 observations is created by embedding synthetic threat signatures in harmless background data.

Once the model is trained using this data, it can be used on data collected by mobile detectors of radiation threat.

# 3 Literature review

## 3.1 Correlation analysis

Canonical Correlation Analysis is a statistical method useful for exploring relationships between two sets of variables. It is used in machine learning, with applications to multiple domains, including applications to medicine, biology and finance: [9], [6], [25] and [26].

A modified version of the algorithm was proposed by Witten et al. [26][27]. Sparse CCA, an $L_1$ variant of the original CCA, adds constraints to guarantee sparse solutions. This limits the number of features being correlated. Their formulation of the maximization problem also differs from the traditional CCA algorithm. We will use this version since it is suitable for our needs.

Principal Component Analysis (PCA) and its variant Sparse PCA are also related to our research. PCA and Sparse PCA find orthogonal projections of the data onto linear spaces where the variance of the data is maximized. This procedure aims at preserving as much variance as possible, while reducing dimensionality. The goal of CAA is different: it finds two-dimensional projections that emphasize correlation, where such correlations constitute interpretable patterns that are present in the data.

Additional work that shares the goal of our research is [9] and [6]. Using the notion of autocorrelation, they attempt to find underlying components of functional magnetic resonance imaging (fMRI) and electroencephalogram (EEG), respectively, that have maximum autocorrelation, and to do so they use CCA. The type of data they work with differs from ours in that their features are ordered, both temporally and spatially. To find autocorrelations, they take $X$ as the original data matrix and construct $Y$ as a translated version of $X$, such that $Y_t = X_{t+1}$. Since our data is not ordered we cannot follow the same procedure and must instead develop a new technique for finding autocorrelations.

### 3.2 Radiation source detection

Efforts to perform radiation threat detection in an unsupervised fashion tackle the problem as an anomaly detection task. Anomaly detection is the problem of finding patterns in data that do not follow the expected behavior[4].

One anomaly detection approach to detect radioactive sources is known as K-Sigma, which fits a parametric model assuming a Poisson distribution to the total gross photon counts of harmless radiation data [5]. This method takes as input the parameter $K$ and flags a data point if its total photon counts is over $K$ standard deviations away from the fitted model.

Unlike K-Sigma, which is based on total energy counts, spectral anomaly detection based on Principal Component Analysis (PCA) [14][24] takes into account photon counts for each bin. Using data from background radiation, it learns a PCA model. The first components capture the most variance, which in this case corresponds to what is considered "typical" variation for background. The algorithm drops the top components and projects new data on to the remaining PCA components, where variance should be low and therefore the observed discrepancy can be used as a threat score.

This PCA-based spectral anomaly detector, based on [17], works by calculating the magnitude of the residual after a background-subtracting projection. The background-subtracting is a strict projection onto the subspace spanned by the top few principal components of the covariance matrix. An alternative way of finding this projection is by a dilation modified projection where the correlation (not covariance) matrix is used to learn the low dimensional projection and then appropriate scaling of the measurement dimensions is performed before projection and scaled back after the projection. In any case, the transformation computes the estimated background contribution to a radiation measurement, assuming that the top few principal components represent expected typical background variation. After projection, the magnitude of the residual essentially provides the PCA-based spectral anomaly score, as it should be negligibly low for spectra consistent with training data distributions. Appendix B contains the PCA spectral anomaly detector algorithm.

# 4 Methodology

## 4.1 Canonical Autocorrelation Analysis

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.

We use the formulation of Sparse CCA given by [26]. Assuming $X$ and $Y$ have been standardized, the constrained optimization problem is:

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

When $c_1$ and $c_2$ are small, solutions will be sparse and thus only a few features are correlated.

Our goal is to find correlations within a single set of variables, which can be understood as having identical matrices $X$ and $Y$. Applying Sparse CCA when $X = Y$ results in trivial solutions $u = v$, corresponding to Sparse PCA solutions for $X$ [27].

The naive approach to finding correlations between disjoint subsets of features would consist of an exhaustive search, trying all possible ways of splitting the features into two groups and applying Sparse CCA. This may be computationally infeasible, as the number of times Sparse CCA would have to be applied is $O(\binom{m}{t})$, where $m$ is the number of variables of $X$ and $t = |\{u_i \neq 0\}|$ is determined by the constraints $c_1$ and $c_2$ that would be applied to the original matrix $X$ to obtain the desired level of sparseness.

We developed an algorithm capable of finding such autocorrelations by imposing an additional constraint in Equation 1 to prevent the model from correlating each variable with itself. Using the Lagrangian form of the added penalty, the problem can be written as follows:

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

This will penalize vectors $u$ and $v$ for having high values for the same component, which is precisely what we are trying to avoid. Through proper factorization we are able to solve this using Sparse CCA algorithm.

**Theorem 1.** *Solving Equation 2 is equivalent to solving Sparse CCA for the pair of matrices*

$$\hat{X} = [V(S^2 - \lambda I)]^T \text{ and } \hat{Y} = V^T.$$

*where the Singular Value Decomposition of* $X$ *is* $X = USV^T$ *and* $I$ *is the identity matrix.*

*Proof.* First, notice that

$$u^T X^T X v - \lambda u^T v = u^T (X^T X - \lambda I) v.$$

Therefore, the CAA problem can be written as:

$$\begin{aligned} & max_{u,v} u^T (X^T X - \lambda I) v \\ ||u||_2^2 \leq 1, ||v||_2^2 \leq 1 \quad & ||u||_1 \leq c_1, ||v||_1 \leq c_2 \end{aligned} \tag{3}$$

Finding the Singular Value Decomposition of $X$, we have:

$$\begin{aligned} & X & = & & USV^T \\ \Rightarrow \quad & X^T X & = & & VS^2V^T \\ \Rightarrow \quad & X^T X - \lambda I & = & & VS^2V^T - \lambda VV^T \\ & & = & & V(S^2 - \lambda I)V^T \end{aligned}$$

Now, setting $\hat{X} = [V(S^2 - \lambda I)]^T$ and $\hat{Y} = V^T$, the problem becomes:

$$\begin{aligned} & max_{u,v} u^T \hat{X}^T \hat{Y} v \\ ||u||_2^2 \leq 1, ||v||_2^2 \leq 1 \quad & ||u||_1 \leq c_1, ||v||_1 \leq c_2 \end{aligned} \tag{4}$$

This problem can be solved using Sparse CCA, and since the solutions obtained with this method are independent of the factorization of $\hat{X}^T \hat{Y}$, solving this is equivalent to solving the CAA maximization problem in Equation 3. □

Since the solution to CAA is obtained by applying Sparse CCA to newly defined matrices, the ability of Sparse CCA to retrieve multiple pairs of canonical vectors is inherited by CAA. Therefore, for each matrix $X$ we can find $k \leq m$ pairs of canonical vectors, where we will refer to the $i$th pair as $(u^{(i)}, v^{(i)})$, for $i = 1, ..., k$. CAA may return fewer pairs than $m$ when there are not enough vectors for which the data shows a substantial linear correlation when projected onto the planes defined by $u$ and $v$.

The maximization problem as formulated above is not convex with regard to $\lambda$ and it is sensitive to variations of this parameter. Our current approach to solving CAA performs a grid search and returns up to $m$ pairs of canonical vectors $u, v$. The grid search uses an evaluation metric for relative sparseness developed for this purpose. Vectors $u$ and $v$ are considered to have relative sparseness if components with high values in $u$ do not coincide with high valued components in $v$. This can be measured by the metric defined in Equation 5, where the case of $t(u, v) = 1$ corresponds to vectors $u$ and $v$ having strictly disjoint support.

$$t(u, v) = 1 - \sum_j |u_j v_j| \tag{5}$$

The $i$th CAA solution can therefore be obtained by finding the minimum value of $\lambda$ for which the corresponding CCA solution of $\hat{X}$ and $\hat{Y}$ have a relative sparseness of 1. Now, instead of applying Sparse CCA $\mathrm{O}(\binom{m}{s})$ times to find a pair of canonical variates, as in the naive approach, we apply it $\mathrm{O}(\lambda)$ times, where $\mathrm{O}(\lambda)$ refers to the resolution the grid search.

### 4.2 Relationship to Sparse PCA

As mentioned in Section 1, CAA and Sparse PCA have fundamentally different objectives, but given that Sparse CCA applied to identical matrices ($X = Y$) results in Sparse PCA components, it is worth taking a look at the details of how CAA and Sparse PCA differ. While Sparse PCA finds one-dimensional projections of data that maximize *variance* of data, CAA finds two-dimensional projections where *correlation* between the two sets is maximized. Furthermore, it is easy to see how the variables retrieved by CAA differ from those retrieved by Sparse PCA. As previously mentioned, applying Sparse CCA to matrices $X = Y$ results in Sparse PCA solutions $u = v$ [27]. Therefore, Sparse PCA can be written as

$$\begin{array}{rr} max_{u,v} u^T X^T X v \\ ||u||_2^2 \leq 1, ||v||_2^2 \leq 1 \quad ||u||_1 \leq c_1, ||v||_1 \leq c_2 \end{array} \tag{6}$$

In the following, we analyze how the objective values retrieved by CAA (Equation 2) and Sparse PCA (Equation 6) differ.

- Sparse PCA optimal criterion value retrieved:

$$\begin{aligned} u^T X^T X u &= (\textstyle\sum_{i \in P} u_i X_i)^T (\sum_{j \in P} u_j X_j) \\ &= \displaystyle\sum_{i \in P} \sum_{j \in P} u_i u_j X_i^T X_j \end{aligned}$$

  for $P = \{i | u_i \neq 0\}$

- CAA optimal criterion value retrieved:

$$u^T X^T X v - \lambda u^t v = (\sum_{i \in P_1} u_i X_i)^T (\sum_{i \in P_2} v_i X_i) - \lambda u^T v$$

  for $P_1 = \{i | u_i \neq 0\}$, $P_2 = \{i | v_i \neq 0\}$

  Provided that the grid search guarantees that the vectors $u$, $v$ in the solution are orthogonal, we can rewrite this as:

$$\begin{aligned} & (\sum_{i \in P_1} u_i X_i)^T (\sum_{j \in P_2} v_j X_j) \text{ s.t. } P_1 \cap P_2 = \emptyset \\ = \; & (\sum_{i \in P_1} \sum_{j \in P_2} u_i v_j X_i^T X_j) \text{ s.t. } P_1 \cap P_2 = \emptyset. \end{aligned}$$

Notice that in the case of Sparse PCA, all interactions between variables in the subset $P$ are considered, while CAA only considers interactions across two disjoint groups. Therefore, there are two types of interactions Sparse PCA considers that CAA does not: the variance of each variable and correlation/covariance between variables in the same subset. Note the first one is only relevant for Sparse PCA when using the covariance matrix, but the second is relevant both when Sparse PCA is applied to the correlation matrix or to the covariance matrix. As a result, CAA and Sparse PCA retrieve vectors that involve different subsets of features, and such subsets correspond to different types of structures in data.

### 4.3 CAA-based anomaly detection

How can the outcome of CAA be used once it has been applied to a matrix $X$? CAA produces several multiple-to-multiple linear correlation patterns. If the only goal is that of characterizing and understanding the data, one can analyze the coefficients in the canonical projections to understand which correlations are characteristic in a certain data set. In addition, such projections can be used as the basis of an anomaly detection method, which we introduce below.

CAA can be applied to a set ($X \in \mathbb{R}^{n \times m}$) of data points that are assumed to not be anomalous. The result will be $k \leq m$ pairs of canonical vectors, where we will refer to the $i$th pair as $(u^{(i)}, v^{(i)})$, for $i = 1, ..., k$. Each of these vector pairs maps the data into a new bi-dimensional space, where the x-axis corresponds to ${u^{(i)}}^t X^t$ and the y-axis corresponds to $Xv^{(i)}$. We define the projection of the data onto the $i$th canonical space as

$$X_i^{proj} = ({u^{(i)}}^t X^t, Xv^{(i)})$$

We expect top canonical projections to yield scatter plots in which the training data shows a strong diagonal tendency if the underlying correlations indeed exist. We can characterize the resulting distribution by fitting a parametric density model (e.g. bivariate Gaussian) or using a non-parametric density model (e.g. kernel density estimation),

$$X_i^{proj} \sim F_i(\boldsymbol{\theta}_i)$$

This yields a characterization of the non-anomalous data points by:

$$\begin{cases} \text{Canonical vectors } (u^{(i)}, v^{(i)}) & for \quad i = 1, ..., k \\ \text{Distributions } F_i \text{ parameterized by } \boldsymbol{\theta}_i & for \quad i = 1, ..., k \end{cases}$$

Given a new data point $x$, it can be projected onto the $k$ canonical projections and a score of anomalousness can be computed for each projection using the likelihood (Equation 7).

$$s_i(x) = \mathbb{P}(x|\boldsymbol{\theta}_i) \tag{7}$$

Finally, a cumulative single score can be computed using an aggregation metric $M(\cdot)$, where the choice of this metric depends on the particular application (e.g., minimum or product), as shown in Equation 8.

$$S(x) = M_{\{i=1:k\}}(s_i(x)) \tag{8}$$

For the experiments in this paper, we assume that the training data follows a bivariate Gaussian in each of the canonical projections, i.e,

$$X_i^{proj} \sim \mathcal{N}(\mu_i, \Sigma_i).$$

Therefore, the data set can be characterized by:

$$\begin{cases} (u^{(i)}, v^{(i)}) & for \quad i = 1, ..., k \quad k \leq n \\ (\mu_i, \Sigma_i) & for \quad i = 1, ..., k \end{cases}$$

For each $i$ we obtain a characterization of the training data that involves multiple features. A new data point $x$ can be simultaneously mapped onto all $k$ canonical spaces, and given the assumption of a bivariate Gaussian, we can use Mahalanobis distance metric as an equivalent to the likelihood. Therefore, our score $s_i(x)$ is given by

$$\begin{aligned} s_i(x) &= D_{M_i}(x_i^{proj}) \\ \text{where } x_i^{proj} &= ({u^{(i)}}^T x^T, x v^{(i)}) \\ D_{M_i}(x_i^{proj}) &= \sqrt{(x_i^{proj} - \mu_i)\Sigma_i^{-1}(x_i^{proj} - \mu_i)} \end{aligned} \tag{9}$$

Note $D_{M_i}(\cdot)$ is the Mahalanobis distance from $x_i^{proj}$ to $N(\mu_i, \sigma_i)$, where $x$ is the current observation.

If the new data point follows the same correlation patterns as the training data, all of the Mahalanobis distances computed for it should be small. It can be expected that a data point that is anomalous in the CAA sense would not match that behavior. It will likely fail to follow one or multiple of these characterizations, which will result in one or multiple large Mahalanobis distances. To marginalize the resulting distribution of scores into a total score $S(x)$, one conceivable option is maximization, as in Equation 10.

$$S(x) = \max_{i=1,...,k} D_{M_i}(x^{proj}) \tag{10}$$

Maximization is only one of many possible ways to aggregate scores from multiple CAA projections. We have found it effective in our threat detection application because

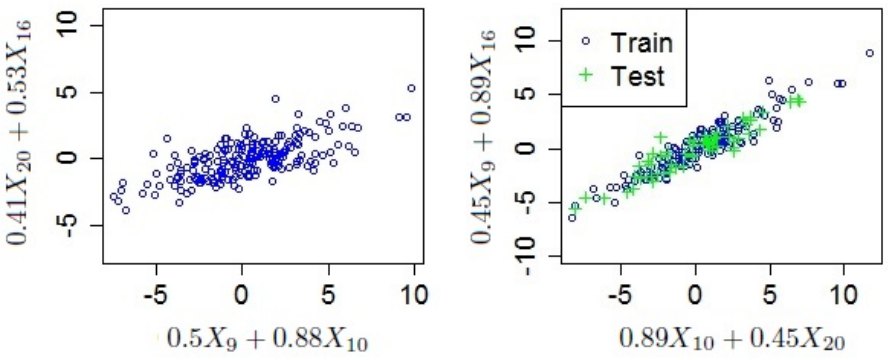
it is typically sufficient for a gamma-ray spectrum measurement to substantially deviate from the expectation in only a few energy bins to warrant attention. However, in other applications alternative forms of $S(x)$ may be more relevant and effective.

We calibrate the threat detection threshold following [12], by assuming a particular rate of nuisance positives in training data (2-5%).

# 5 Experiments

## 5.1 Synthetic data

The first experiment aims to illustrate how known correlations can be successfully retrieved by CAA.




**(a)** Synthetic correlation **(b)** Retrieved correlation

**Figure 2:** Comparison between a synthetic correlation pattern and the correlation pattern retrieved by CAA. Equations have the form of $k_i X[,i] + k_j X[,j]$, where $X[,i]$, $X[,j]$ are the $i$th and $j$th columns of X and $k_i$, $k_j$ are the linear combination coefficients.

A Gaussian bivariate distribution is generated with an assumed mean and covariance, and 200 data points are sampled from it. A matrix $X$ of dimensions $200 \times 20$ is created such that there exist sparse vectors $u, v$, each with two non-zero components, for which $(u^T X^T, Xv)$ correspond to the previously generated Gaussian. Next, 70% of data is used to train a CAA model and the rest is used for testing. Figure 2a contains a scatter plot of the data sampled from the Gaussian, where the axes indicate the linear combinations of

columns of X that map the original data onto the Gaussian. Figure 2b shows the projection of both training and testing data onto the space determined by the first pair of canonical vectors retrieved by CAA, where the equations on the axes correspond to the correlation they establish. Note that the method is able to successfully identify the four features of the data for which multiple-to-multiple linear correlation exists, even though they are not grouped identically as in the original design.

### 5.2 Radiation threat detection

The radiation data used in our experiments is featurized into 128 disjoint energy bins, and reflects photon counts obtained from gamma-ray spectrometer measurements. There are 20,000 records available for harmless background data, and 10,000 records for each of 15 types of threat-injected data, to simulate various radiological threats.

As it was previously explained, PCA-based spectral anomaly detection assumes that the top few principal components represent the expected envelope of background variation, and uses the residual after removing these top components as a spectral anomaly score. In the case of CAA, multiple-to-multiple combinations of energy bins that are well correlated provide a characterization model for background radiation. This model can be used as the basis of the anomaly detection method described in Section 4.3, which identifies threats when radiation spectra depart from the expected patterns of correlation.

For evaluation purposes, the CAA-based anomaly detector is compared to the PCA-based anomaly detector introduced in Section 3.2. Additionally, a Sparse PCA anomaly detector was designed and implemented for comparison. The algorithm is analogous to the PCA alternative, with some minor modifications that enable the exchange of PCA for Sparse PCA in the pipeline. The algorithm is described in detail in Appendix C.

All three models were trained using 10,000 background records, and the resulting models were evaluated on 15 types of radiation threats. Each of the 15 test sets contained 10,000 samples of injected data corresponding to a particular threat type, combined with the remaining 10,000 background records. Three performance comparison metrics were used: area under the ROC curve (AUC), recall at a fixed low false discovery rate, and false discovery rate at a fixed recall rate of 50%. Figure 5 and Tables 1 and 2 summarize the results, and Appendix A contains ROC curves for all 15 threat types used in the experiments, with the false positive rate axis in logarithmic scale to enhance view at low false positive rates, where most applications tend to reside. For ten of these fifteen threats CAA

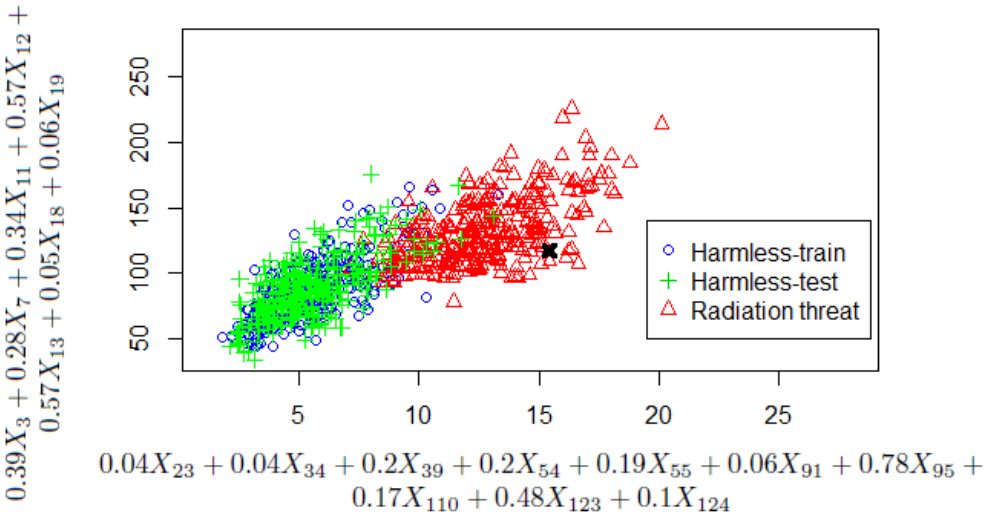



**Figure 3:** Projection of background radiation and threat-injected background radiation measurements onto space determined by one pair of CAA canonical vectors. The equations indicate the multiple-to-multiple linear correlation that defines this projection. Data point labeled with black $x$ corresponds to the individual threat case analyzed in Figure 4.

performs significantly better than PCA and Sparse PCA according to all three performance metrics, and only for one type of threat is another method significantly better than CAA according to all metrics. This is threat type $A$, where PCA performs best, which can be potentially explained by the fact that sparsity is apparently not very useful in this case and a larger number of bins is necessary to detect this particular type of threat. When the algorithms are applied to all threats combined into a single batch, CAA is by far better than the other two competitors.

Figure 3 shows an example of the mapping onto the space determined by a pair of CAA canonical vectors $(u, v)$. The data points corresponding to background radiation, as well as those corresponding to a particular threat, are mapped onto this projection. This particular pair $(u, v)$ is used most often to score the data points belonging to this particular type of threat (listed as type "L" in table labels), meaning the one where the maximum Mahalanobis distance to the Gaussian characterizing background radiation is found most often, as defined in Equation 11.

$$(u, v) = \arg\max_{u_i, v_i} D_{M_i}((u_i^T x^T, x v_i))|_{i=1}^{k} \tag{11}$$

As Figure 3 shows, in this example the threat-injected data distribution visibly diverges from the test set distribution of benign data. For this threat type, CAA model achieves the AUC of 0.995, while PCA-based detector has the AUC of 0.821.

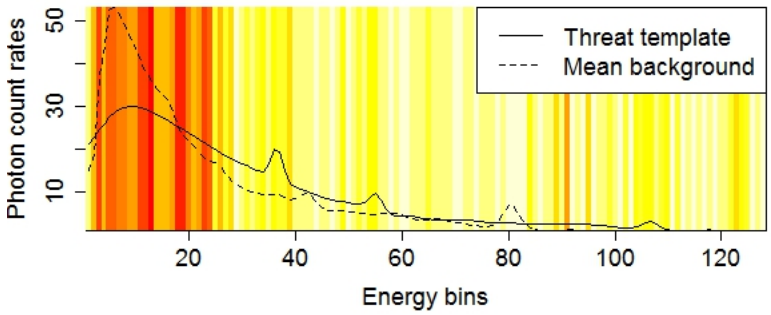
| Th | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | **all** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| cor. | .72 | .79 | .79 | .81 | .87 | .87 | .88 | .88 | .88 | .91 | .92 | .92 | .94 | .94 | .96 | NA |
| caa | 7 | **4*** | **4*** | **14*** | **7*** | **7*** | **2** | **15*** | **100** | **14*** | **87*** | **14** | **20*** | **13*** | 5 | **21*** |
| pca | **59*** | 1 | 1 | 1 | 2 | 2 | 1 | 3 | **100** | 7 | 8 | 13 | 2 | 1 | 4 | 14 |
| Spca | 6 | 1 | 1 | 2 | 3 | 3 | 1 | 4 | **100** | 5 | 9 | 6 | 2 | 2 | **6** | 10 |

**Table 1:** Performance of CAA, PCA and Sparse PCA in terms of recall rate (given in %) at fixed false discovery rate of 0.01. Radiation threat types are ordered according to the strength of the correlation between mean background spectrum and threat template. Asterisks mark cases when the winning method performs significantly better than the second-best.

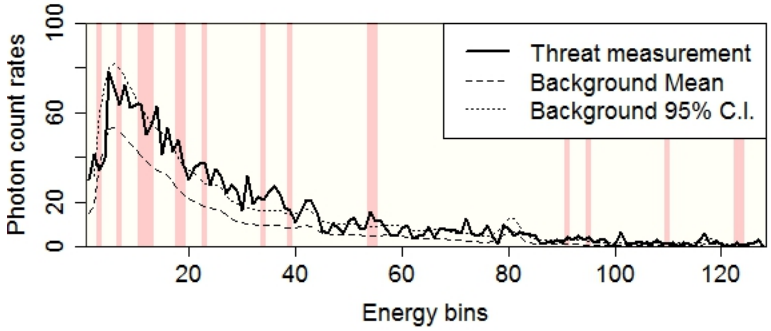
| Th | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | **all** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| cor. | .72 | .79 | .79 | .81 | .87 | .87 | .88 | .88 | .88 | .91 | .92 | .92 | .94 | .94 | .96 | NA |
| caa | 11 | **21*** | **21*** | **6*** | **13*** | **13*** | **41** | **6*** | **0** | **7*** | **0*** | **7*** | **4*** | **7*** | 18 | **8*** |
| pca | **1*** | 43 | 39 | 33 | 42 | 41 | 53 | 33 | **0** | 23 | 12 | 15 | 28 | 32 | 42 | 27 |
| Spca | 14 | 46 | 49 | 34 | 27 | 27 | 43 | 20 | **0** | 17 | 9 | 15 | 36 | 35 | **16*** | 23 |

**Table 2:** Performance of CAA, PCA and Sparse PCA in terms of false discovery rate (given in %) at a fixed recall of 50%. Radiation threat types are ordered according to the strength of the correlation between mean background spectrum and threat template. Asterisks mark cases when the winning method performs significantly better than the second-best.

In addition to its good empirical performance, the proposed method yields readily interpretable outputs. When a spectral measurement is identified as a possible threat, the energy bins on which it fails to follow background patterns can be pointed out. This has two main advantages: first, when analyzing an individual data point the user knows which energy bins the algorithm used to make its decision, for easy adjudication of the results (Figure 4b. Secondly, when applied to a batch of data associated to a particular threat




**(a)** Threat batch: Heat map indicates the frequencies with which bins are used by CAA to flag one particular threat type, together with mean background spectra and spectral template for that threat.




**(b)** Adjudication of an individual measurement: Radiation spectrum the method correctly labels as inclusive of threat signatures compared to background radiation distribution. Colored bins were used to flag measurement as anomalous, corresponding to the support of the CAA canonical vectors that define the projection where the maximum Mahalanobis distance to the baseline distribution was found. This corresponds to CAA projection shown in Figure 3, where this individual measurement is labeled with a black $x$.

**Figure 4:** Visualization of energy bins that are used to label gamma-ray measurements as likely inclusive of threats.

type, it is possible to identify the bins on which the threat's appearance systematically differs from the background behavior, providing the way to characterize this type of threat (4a).

Figure 4 shows the frequency with which energy bins are used to identify anomalies. The top plot shows the usage frequency of bins for a threat-infused data batch associated to the

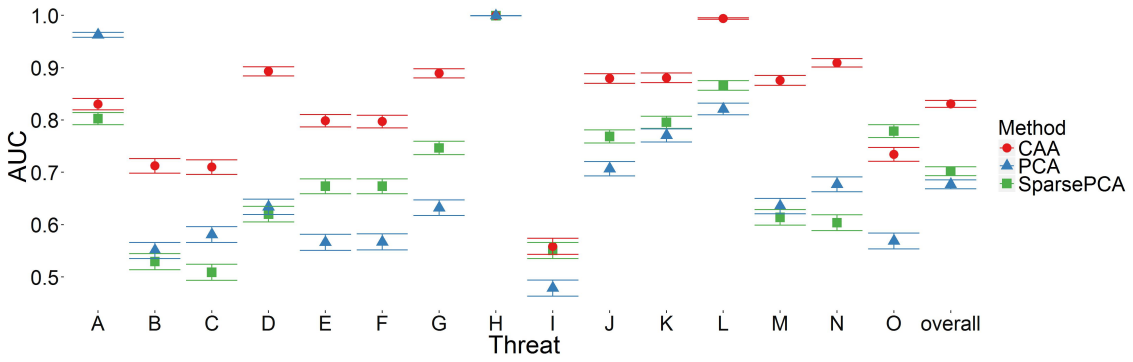



**Figure 5:** AUC and confidence intervals for CAA, PCA and Sparse PCA applied to detecting radiation threats of various types. Radiation threat types are ordered according to correlation between mean background spectrum and threat template spectrum. The right-most column shows performance when all threat types are combined in one batch.

threat type used in Figure 3. The bottom plot shows an example of a radiation spectrum the method correctly labels as representative of a threat, mean of the background radiation spectra used for training, and colors the energy bins that were used to label the data as anomalous. This individual threat measurement is labeled with a black $x$ in Figure 3. It is interesting to see that even though the method is fully unsupervised, such bins correspond to spikes in the injected threat template.

## 6 Application to other domains

Even though our development of CAA method is motivated by the radiation threat detection task, it can be applied to many other domains. Such an example is breast cancer diagnosis.

Breast cancer diagnosis presents similar challenges to the task discussed above. We use the Wisconsin Breast Cancer data set [28]. It contains 699 cases of tumors, 683 after removing those with missing values, labeled as either malignant (239) or benign (444). Each sample is described with 9 numeric variables that take on values between 1 and 10. Previous work has been done with this data to build predictive models that distinguish malignant from benign tumors [19][13][1][18][21][29]. These approaches rely on supervised classifiers. The use of single-class approaches to detect malicious tumors may however be beneficial, given that some types of cancer are fairly rare and different from more typical

cases. There has been research showing that cancer can follow multiple developmental pathways in asynchronous orders [3], and even these studies are not known to be complete, so there is much motivation to anticipate at least some cancerous tumors that do not follow characteristics of the most common types of cancer, but are still identifiable as deviating from typical patterns in benign data in some way. We therefore expect an unsupervised approach, such as CAA, to be a contender with many supervised (and therefore more informed about frequent types of cancer) methods, while potentially being more reliable than supervised methods at detecting types of cancer that were not well represented in the training set.

Hence, we approach this problem as an anomaly detection task, and train CAA using benign cases. We then use the CAA-based anomaly detection method to score both malignant and benign test cases. To set the decision threshold, we set $\rho = 0.05$. This means we assume 5% of our training data to be likely anomalous and choose the threshold accordingly. As a comparison, we also apply the PCA-based anomaly detection approach previously described in Section 3.2, the Sparse PCA-based anomaly detector developed for comparison purposes in this paper, and a one-class SVM classifier.

We compare the four methods, as well as previous results reported in literature that use supervised learning approaches. All our experiments were performed using 10-fold cross-validation, and we report both mean accuracy and its standard deviation. In the case of one-class SVM, which is very sensitive to the choice of parameters, we perform 10-fold cross-validation for parameter tuning inside the benign-class training set in each fold to find the optimal values of the kernel width, as well as the parameter which determines the amount of training points that are considered outliers. The results are presented in Table 3.

Performance of CAA is comparable to that of Sparse PCA in our experiments and to that of supervised methods reported in the literature, and it is better than both one-class SVM and the PCA-based alternatives. One of the advantages of a single-class method is that it can be expected to be more reliable than a supervised method when tasked to detect types of cancer that were not present in the training set. The fact that CAA is fully interpretable is also an advantage, as domain experts can see which features make a sample appear anomalous. CAA yields multiple-to-multiple linear correlation scatterplot in which the current data does not conform to the expected behavior, and in this particular case correlations do not involve more than 4 to 5 tumor features at one time, facilitating interpretability. Finally, classification criteria that are easily interpretable by a domain

| Type | Method | Reference | Accuracy ± sd (%) | |
|---|---|---|---|---|
| Supervised | LS-SVM | [18] | 98.53 | N/A |
| Supervised | K-SVM | [29] | 97.38 | N/A |
| Supervised | LLWNN | [21] | 97.20 | N/A |
| Supervised | Supervised fuzzy clustering | [1] | 95.57 | N/A |
| Supervised | NEFCLASS | [13] | 95.06 | N/A |
| Supervised | C4.5 | [19] | 94.74 | N/A |
| Single-class | Sparse PCA | | 96.33 ± 2.53 | |
| Single-class | **CAA** | | 94.29 ± 2.7 | |
| Single-class | one-class SVM | | 86.84 ± 4.2 | |
| Single-class | PCA | | 81.06 ± 5.82 | |
| Most common value | default | | 65 | |

**Table 3:** Comparison of classification accuracy on the Wisconsin Breast Cancer Data Set. Experiments were conducted using 10-fold cross-validation. Standard deviation is not available for the results cited from the literature.

expert may lead to new qualitative insights of how different types of cancer manifest in data.

## 7 Conclusions and future work

This paper presents Canonical Autocorrelation Analysis (CAA), a new method that automatically finds subsets of features of data that form strong multiple-to-multiple linear correlations. It has been shown how CAA can be useful at anomaly detection tasks which involve multivariate numeric data where detecting changes in the structure of correlations may be of interest. CAA-based anomaly detector was applied to the task of radiation threat detection, where it has proven to be successful. In addition to this, CAA can be a valuable tool in many scenarios where in spite of an existing division between "non-anomalous" and "anomalous", the "anomalous" class is composed of multiple sub-classes, many of which might be unknown or underrepresented in the training data, but are still important to detect. To illustrate how CAA can be applied to other domains, breast cancer diagnosis on a publicly available dataset was performed, where it has proven to be competitive to supervised methods, while being potentially better at detecting unusual types of cancer that are not present or that are underrepresented in the training data.

The models built by CAA are readily interpretable. The canonical projections are

sparse, and even though the anomaly detection method analyzes every projection found, often it is sufficient to consider only one of them to adjudicate a query data as possibly anomalous, and each test case may invoke a different canonical pair as the most useful for handling it. Interestingly, these most useful projections are usually also the ones which provide the most interpretable intuition on why and how the particular query does not seem to fit the reference distribution. Consider that each canonical projection spans the axes of a 2-dimensional Cartesian system which happen to be sparse and easy to interpret linear combinations of the native features of data, and it should be clear why CAA may be particularly useful in applications where machine learning is used to aid humans in their decision making.

The next steps for this research are multiple. In the radiation threat detection domain, further validation against diverse field data sets will be performed, different criteria for identifying the optimal level of sparsity will be explored, and the output of CAA will be used to inform supervised (threat-type aware) methods that rely on energy windows for detection.

The next steps towards further development of methodology include the extention of CAA towards supervised learning and enabling discovery of non-linear relationships via kernelization.

## 8 Acknowledgements

This work has been partially supported by DNDO under competitively awarded grant 2010-DN-077-ARI040-02, by DTRA under award HDTRA1-13-1-0026, by the NSF under award 1320347, and DARPA under awards FA8750-12-2-0324 and FA8750-14-2-0244.

## References

[1] Abonyi, J., and Szeifert, F. 2003. Supervised fuzzy clustering for the identification of fuzzy classifiers. *Pattern Recognition Letters* 24(14):2195–2207.

[2] against Nuclear Weapons, S. P. "nuclear terrorism".

[3] Almendro, V.; Kim, H. J.; Cheng, Y.-K.; Gönen, M.; Itzkovitz, S.; Argani, P.; van Oudenaarden, A.; Sukumar, S.; Michor, F.; and Polyak, K. 2014. Genetic and phenotypic diversity in breast tumor metastases. *Cancer research* 74(5):1338–1348.

[4] Chandola, V.; Banerjee, A.; and Kumar, V. 2009. Anomaly detection: A survey. *ACM computing surveys (CSUR)* 41(3):15.

[5] Chandy, K. M.; Bunn, J.; and Liu, A. 2010. Models and algorithms for radiation detection. In *Modeling and Simulation Workshop for Homeland Security*, 1–6.

[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] El-Arini, K.; Moore, A. W.; and Liu, T. 2006. Autonomous visualization. In *Knowledge Discovery in Databases: PKDD 2006*. Springer. 495–502.

[8] Fiterau, M., and Dubrawski, A. 2012. Projection retrieval for classification. In *Advances in Neural Information Processing Systems*, 3023–3031.

[9] Friman, O.; Borga, M.; Lundberg, P.; and Knutsson, H. 2002. Exploratory fmri analysis by autocorrelation maximization. *NeuroImage* 16(2):454–464.

[10] Holdstock, D., and Waterston, L. 2000. Nuclear weapons, a continuing threat to health. *The Lancet* 355(9214):1544–1547.

[11] Hotelling, H. 1936. Relations between two sets of variates. *Biometrika* 321–377.

[12] Lazarevic, A.; Ertöz, L.; Kumar, V.; Ozgur, A.; and Srivastava, J. 2003. A comparative study of anomaly detection schemes in network intrusion detection. In *SDM*, 25–36. SIAM.

[13] Nauck, D., and Kruse, R. 1999. Obtaining interpretable fuzzy classification rules from medical data. *Artificial intelligence in medicine* 16(2):149–169.

[14] Nelson, K., and Labov, S. 2009. Detection and alarming with sords unimaged data: Background data analysis. Technical report, Lawrence Livermore National Laboratory.

[15] Nichols, T.; Stuart, D.; and McCausland, J. D. 2012. Tactical nuclear weapons and nato. Technical report, DTIC Document.

[16] Nrc: Fact sheet on dirty bomb, howpublished = `http://www.nrc.gov/reading-rm/doc-collections/fact-sheets/fs-dirty-bombs.html`, note = Accessed: 2015-10-21.

[17] Parra, L.; Deco, G.; and Miesbach, S. 1996. Statistical independence and novelty detection with information preserving nonlinear maps. *Neural Computation* 8(2):260–269.

[18] Polat, K., and Güneş, S. 2007. Breast cancer diagnosis using least square support vector machine. *Digital Signal Processing* 17(4):694–701.

[19] Quinlan, J. R. 1996. Improved use of continuous attributes in c4. 5. *Journal of artificial intelligence research* 77–90.

[20] Reals, T. 2015. Radioactive material stolen in mexico.

[21] Senapati, M. R.; Mohanty, A. K.; Dash, S.; and Dash, P. K. 2013. Local linear wavelet neural network for breast cancer recognition. *Neural Computing and Applications* 22(1):125–131.

[22] Sjöstrand, K.; Clemmensen, L. H.; Larsen, R.; and Ersbøll, B. 2012. Spasm: A matlab toolbox for sparse statistical modeling. *Journal of Statistical Software Accepted for publication*.

[23] Smith, A. 2015. Hiroshima 70th anniversary: What to know about nuclear weapons in 2015.

[24] Tandon, P. 2015. *Bayesian Aggregation of Evidence For Detection and Characterization of Patterns in Multiple Noisy Observations*. Ph.D. Dissertation, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA.

[25] 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.

[26] 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.

[27] 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.

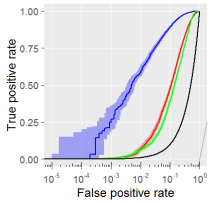
[28] Wolberg, W. H., and Mangasarian, O. L. 1990. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proceedings of the national academy of sciences* 87(23):9193–9196.

[29] Zheng, B.; Yoon, S. W.; and Lam, S. S. 2014. Breast cancer diagnosis based on feature extraction using a hybrid of k-means and support vector machine algorithms. *Expert Systems with Applications* 41(4):1476–1482.

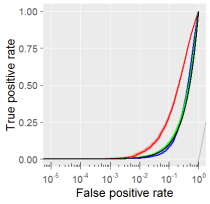
[30] Zou, H.; Hastie, T.; and Tibshirani, R. 2006. Sparse principal component analysis. *Journal of computational and graphical statistics* 15(2):265–286.

# A Radiation threat detection: ROC curves

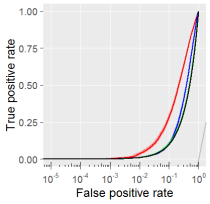
Receiver operating characteristic (ROC) curves for all 15 types of threats for which CAA was tested are shown below. All PCA, Sparse PCA and CAA were trained using benign background radiation, and the resulting model was evaluated on similar background data inclusive of signatures of 15 different types of threats. In the ROC curves, the false positive rate axis is shown in logarithmic scale, to enhance view at low false positive rates.



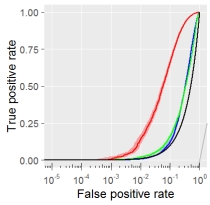

**(a)** Threat A



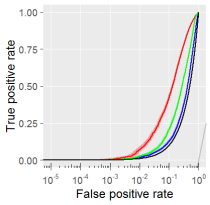

**(b)** Threat B



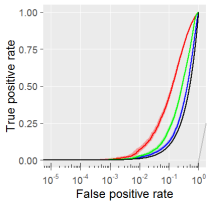

**(c)** Threat C



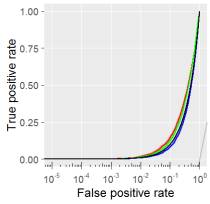

**(d)** Threat D



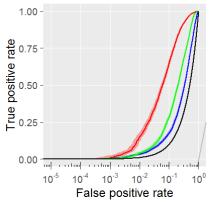

**(e)** Threat E



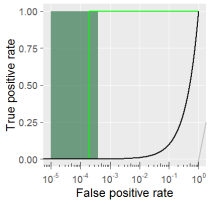

**(f)** Threat F



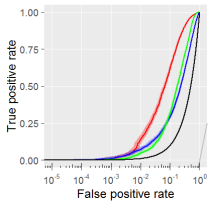

**(g)** Threat G



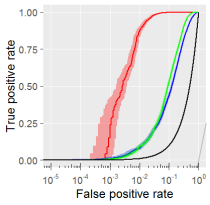

**(h)** Threat H



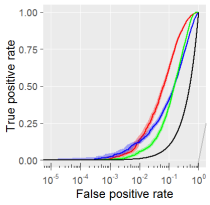

**(i)** Threat I



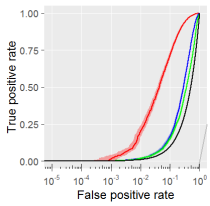

**(j)** Threat J



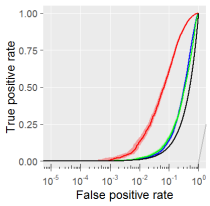

**(k)** Threat K



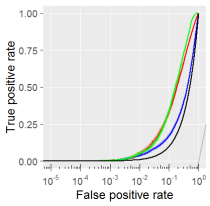

**(l)** Threat L



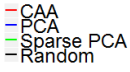

**(m)** Threat M




**(n)** Threat N




**(o)** Threat O



## B PCA spectral anomaly detector

The PCA spectral anomaly detector used for comparison purposes in this paper is frequently used in the radiation threat detection domain, and the description provided in this abstract can also be found in [24]. The algorithm first filters the energy data and performs smoothing via a 10s rolling window. It then computes the special covariance matrix shown in Equation 12, where we assume the background data is a matrix $X \in \mathbb{R}^{n \times q}$. This covariance matrix retains 0.01 of the mean, instead of fully centering the data.

$$\begin{aligned} \Sigma &= \frac{XX^T}{q} - 0.99mm^T \\ m_j &= \sum_{i=1}^{n} X_{i,j} \end{aligned} \tag{12}$$

The correlation matrix $C = A\Sigma A$ is later calculated, where $A$ is the design matrix

$$A = diag(\frac{1}{diag(\Sigma + 1)})$$

Finally, the Singular Value Decomposition is performed on the correlation matrix and the basis matrix $T$ is created as

$$T = I_q - A^{-1}U_{PC}U_{PC}^T A$$

where $U_{PC}$ contains the top principal component eigenvectors. Finally, the residuals can be obtained as $\sigma = || \, TX_{test}^T \, ||_2$.

## C Sparse PCA spectral anomaly detector

Even though Sparse PCA is not generally used in the radiation threat detection domain, a Sparse PCA spectral anomaly detector was designed and implemented for comparison purposes in this paper. As in the case of the PCA-based approach, the algorithm first filters the energy data and performs smoothing via a 10s rolling window. It then normalizes the data to have mean of 0.01 the original mean and and Euclidean length of each column equal to 1, where each column corresponds to a variable.

Top sparse principal components, $U_{Spc}$, are extracted using the algorithm in [22], which is itself based on the formulation done in [30].

The basis matrix $T$ is created as

$$T = I_q - U_{Spc}U_{Spc}^T$$

Given a new data set $X_{test}$, it is first normalized using the mean and Euclidean length of the training data, and then residuals are obtained as $\sigma = || \, TX_{test}^T \, ||_2$.