Multi-View Relationships for Analytics and Inference

Eric Lei

August 2019
CMU-ML-19-112

Machine Learning Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee:
Artur Dubrawski, Chair
Barnabas Poczos
Mario Berges
Simon Labov

Submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

This research was supported by: Advanced Research Projects Agency-Energy award DEAR0000595; Air Force Research Laboratory awards FA87501220324 and FA87501420244; DARPA awards FA8750-17-2-0130 and FA8750-14-2-024; Defense Threat Reduction Agency award HDTRA11310026; Department of Energy contract DE-AC52-07NA27344; Department of Homeland Security awards 14DNARI00005, 16DNARI00025, 17DNERP00002, and 18DNARI000310100; Department of Defense awards HDTRA1-13-1-0026 and ST14291401; Lawrence Livermore National Security award DEAC5207NA27344; National Institutes of Health award R01GM117622; National Science Foundation awards 1320347 and IIS1320347; US Army award W911NF1820218; and grants from Foxconn Technology Group and Highmark Inc.

Keywords: Machine Learning, Statistics, Clustering, Classification, Anomaly Detection, Multi-View, Signal Processing, Gamma Source Detection, Medicine, Non-Intrusive Load Monitoring

Abstract

An interesting area of machine learning is methods for multi-view data, relational data whose features have been partitioned. Multi-view learning exploits relationships between views, giving it certain advantages over traditional single-view techniques, which may struggle to find these relationships or only learn them implicitly. These relationships are often especially salient in understanding the data or performing prediction. This work explores an underutilized approach in multi-view learning: to focus on multi-view relationships—the latent variables that govern relations between views—themselves as units of analysis. We investigate how this approach impacts analytics and inference in ways that standard multi-view and single-view learning cannot. We hypothesize that by ignoring relations between views or factoring them in only indirectly, standard approaches risk overlooking key structure. Accordingly, our goal is to investigate the extent multi-view relationships can be characterized and employed as units of analysis in descriptive analytics and inference. We present novel methods to do so, either using domain knowledge or by learning from data, which reveal structure that alternative methods do not or have competitive performance with the state of the art. Empirical results are presented in several application domains. First, we use domain knowledge to assume a known form for multi-view relationships in the task of gamma source detection. We aggregate the views by filtering their inferences collectively to perform classification. Second, we assume multi-view relationships are linear and learn them from data in a different approach toward gamma source detection. Our method detects anomalies when these relationships are disrupted. Third, we relax the assumption of linearity and propose a novel clustering method that finds cluster-wise linear relationships. This method discovers explanatory structure in a medical problem. Fourth, we extend this method to classification and demonstrate its competitive performance on a load monitoring problem.

Acknowledgments

The biggest acknowledgment goes to my advisor, Artur Dubrawski. I also owe Kyle Miller for practically being a second advisor. I'm grateful to my committee members, Barnabas Poczos, Mario Berges, and Simon Labov, and other primary co-authors, Karl Nelson, Michael Pinsky, and Rob MacLachlan. Thanks also to my collaborators in the Auton Lab and friends inside and outside CMU.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

A promising direction in machine learning is multi-view learning, which leverages different ways to observe a data-generating process in order to discover novel structure and enhance prediction. Formally, a relational dataset is *multi-view* if its features are partitioned into multiple observation-aligned subsets, known as *views*. This kind of natural partitioning exists in many common scenarios, such as:

Multiple sensors: Multiple sensors simultaneously observe related signals, where each sensor is a view. For example, gamma-ray spectrometers at separate but nearby locations search for a source in parallel. Their observed photons are related via source and background radiation.

Time series: Temporal relations can be represented by consecutive sliding windows. For example, each time step might have a window on either side. Each window is a view and represents either "before" or "after" the time step. The views are related via dynamics that determine change over time.

Multiple modalities: A subject can be observed through different mediums, where each medium is a view. For instance, a sentence can be spoken or written in the same language. Here the views would be related by semantic and language traits.

What the examples above—and practically all multi-view datasets—have in common is a set of latent variables that governs the relationship between views. These relationships often prove especially salient in understanding the data or performing inference. The exploitation of this structure by multi-view learning explains why it can exhibit advantages over single-view techniques, which may struggle to find these relationships or only learn them implicitly.

1.1 Main objective

1.1.1 Background

This work explores an underutilized approach in multi-view learning: to focus on multi-view relationships—the latent variables that govern relations between views—themselves as units of analysis. We investigate how this approach impacts analytics and inference in ways that standard multi-view and single-view learning cannot.

We present an illustrative example in Figs. 1.1–1.3. These figures show a multi-view dataset with three spatially overlapping clusters in two dimensions. Each dimension represents a projection, such as the top principal component, of one of two views. The ground truth is shown in Fig. 1.1. The result of common single-view clustering methods is shown in Fig. 1.2. Since these methods operate on spatial relationships, they fail to reveal the overlapping structure based on multi-view relationships. Next, the result of a popular multi-view clustering approach is shown in Fig. 1.3. Although this approach discovers overlapping clusters, it only finds two clusters, and they appear to be random samples from the original distribution. It finds clusters by applying a single-view approach to each view while enforcing agreement between the results. The issue with this procedure is that all information about the clusters is lost in this dataset if considered from only one view.

The failure of these methods broadly represents single-view and current multi-view methods: by ignoring relations between views or factoring them in only indirectly, they risk overlooking key multi-view structure. According to Liu et al. (2013), multi-view clustering strategies can usually be grouped into three categories. First, multiple views are integrated through the loss function, which includes the method in Fig. 1.3. Second, multi-view data are projected to a common subspace, in which any standard clustering algorithm is then applied. Third, a clustering solution is computed for each view individually, and then they are all fused to achieve a consensus. Every category, however, overlooks multi-view relationships. Even the second category of subspace learning, though most similar to the ideas we propose, ultimately operates on spatial relationships in some feature space by

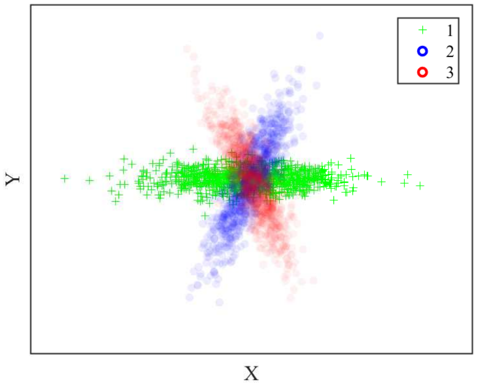


Figure 1.1: Ground truth clusters for an overlapping dataset, drawn translucently to illustrate overlap.

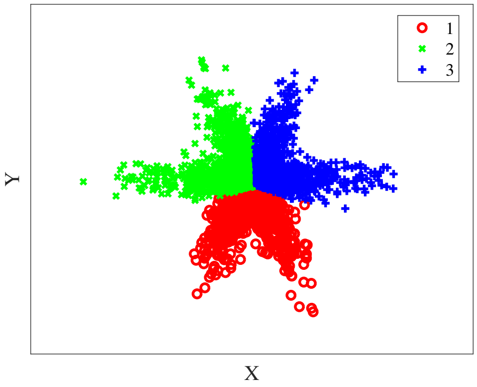


Figure 1.2: Clusters from k-means or spectral clustering for an overlapping dataset.

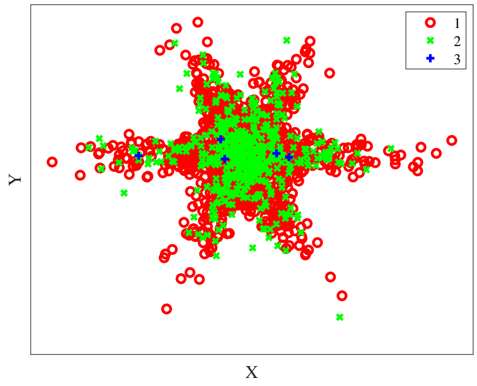


Figure 1.3: Clusters from modern multi-view clustering for an overlapping dataset.

spatially clustering data points in the end. These shortcomings generalize to almost all multi-view learning. In contrast, our methods define the inference task at hand through multi-view relationships, by, for example, identifying cluster variables as the relations themselves. In short, almost all current multi-view methods lack an additional layer of abstraction on top of views to truly separate themselves from single-view methods that project into a different feature space or apply a specific mode of regularization.

To provide this layer, we delve into three applied domains in which we showcase largely empirical results from novel methods. First, we investigate the problem of gamma source detection. A key research topic here is how to aggregate observations from different points in time and space, using known temporal and spacial relations between them to improve inference. Second, we examine physiological response to bodily trauma by analyzing blood pressure. We consider the inspiration and expiration phases of respiration as temporal views and learn the relationship between them to discover explanatory structure across both time and subjects. Third, we study how temporal views can improve inference in non-intrusive load monitoring. Our methodology attempts to exploit temporal dynamics often overlooked by standard methods.

One of these applications use an unconventional understanding of the notion *multi-view*, which raises an important point about the definition. In Ch. 2 we explore a task whose data are a multivariate time series. In our work it may be unclear at first how the task is multi-view. The natural approach would be to directly partition the features to obtain views, but instead we consider each time step itself as a different view. The rationale is that each observation comes from a different point in space, so they are different views of the same physical process. This idea reveals that the term *multi-view* is highly nebulous. If this line of thinking is generalized by treating observations of any dataset as views, an interesting argument can be made that almost any dataset can be considered multi-view. Consequently, we must clarify that we focus on multi-view datasets whose relationships between views are governed by an explainable latent process. This focus excludes the vast majority of arbitrary datasets because their observations are merely random samples from a population. Interestingly, however, it could include many non-stationary time series whose time steps are treated as views.

Furthermore, this thesis limits its scope to the development and evaluation of multi-view methods to investigate certain scientific hypotheses. These hypotheses explore unconventional applications of machine learning and do not have much crossover with popular problems and datasets in mainstream machine learning. For example, an interesting multi-view problem is to develop cross-lingual models of language, or word embeddings, using neural translation techniques (Ruder et al., 2017). If different languages are seen as views, they are related by semantic variables. It would be valuable to apply approaches in this work to this problem to show how they could complement popular methods. Multi-view tasks like this one are especially common in deep learning literature because of its suitability for modalities such as vision, audio, and text. Nonetheless, these topics only rarely intersect the scientific questions we study. We do our best to unite the two sides where relevant but mostly attempt to contribute in an orthogonal direction.

1.1.2 Hypothesis

This thesis investigates multi-view relationships as units of analysis in two ways. Part I assumes these relationships are known through domain knowledge to improve inference. This work is highly specialized to the task of gamma source detection, and we advance the state of the art in the aggregation of multiple views. Then Part II does not assume these relationships are known, so it learns them from data to perform clustering and classification. This idea has been previously addressed in statistics and machine learning but has received minimal attention. We make the earlier work more principled and show better empirical performance on gamma source detection, medicine, and load monitoring.

In the end, both approaches demonstrate the utility of multi-view relationships. This hypothesis is summarized by the thesis statement:

Thesis Statement: It is possible to characterize multi-view relationships and employ them as units of analysis in descriptive analytics and inference.

We present novel methods that characterize multi-view relationships, either using domain knowledge or by learning from data, and employ them as units of analysis. They reveal structure that alternative methods do not or have competitive empirical performance with the state of the art. This approach has received minimal attention in the past.

Moreover, this thesis promotes the underused concept of learning on properties of data distributions, particularly in Part II. Typically, machine learning operates on data points: a foundational condition is that similar points ought to have similar labels. This work, however, suggests to examine not individual points but multiple points as part of a distribution and then perform learning on properties of that distribution. The condition here is that similar distributions ought to have similar labels. For instance, our clustering method groups points together by which multi-view relationships they satisfy—relationships that can be considered properties of the data distribution. Thus, our work is one instance of learning on distributions. We aim to show that this kind of thinking can lead to learning methods that leverage novel characteristics of data. This concept has been studied in different ways previously, but we are among the first to apply it to multi-view learning.

1.2 Outline

In Part I, we cover known multi-view relationships in the task of gamma source detection. The problem of gamma source detection is to determine whether a potentially harmful source is present given counts of radioactive particles in the area. Every material, harmful or not, emits a spectrum of photons; this spectrum is usually a signature of the emitting material. The sum of photons from all sources is observed, but the target harmful source may or may not be present. Although many methods exist already, they often make assumptions that may prove unrealistic in practice. Ch. 2 proposes a method that handles the case when the assumption of informative training data is violated (Lei et al., 2017a). Exploiting smoothness in physical phenomena, this method may be considered multi-view if observations from different times and locations are considered different views, though admittedly this definition may be unconventional. Then Ch. 3 extends this method to aggregate observations from multiple sensors that simultaneously move on different paths, which can be regarded as separate views. The extension leverages the contemporaneous relationships between the sensors using domain knowledge. We show how the views can be aggregated by filtering their inferences collectively using structural information about their relationships.

In Part II, we address the problem of learning multi-view structure and apply it to descriptive analytics and inference. Ch. 4 proposes a method for gamma source detection when the target source template is only partially known (Lei et al., 2016). The method learns linear multi-view correlations and detects anomalies when these correlations are disrupted. Next, Ch. 5 extends this idea to nonlinear multi-view correlations by introducing a clustering method whose correlations are cluster-wise linear (Lei et al., 2017b, 2019). Standard clustering methods such as k-means and spectral clustering group observations together based on spatial relationships. However, another way to determine clusters is to instead consider correlations between observations. In a multi-view setting, we can cluster observations depending on the relationships between views. For example, it is possible that for some observations the features X and Y have positive correlation while for others it is negative. The chapter presents a novel method to accomplish this multi-view correlation clustering inspired by *Canonical Correlation Analysis* (CCA), a traditional statistical technique that somewhat resembles a two-view analogue of Principal Components Analysis. In short, CCA finds latent components that explain correlation between two views. Mathematically, it solves the optimization $\max_{u,v} \operatorname{Corr}(X^\top u, Y^\top v)$ where X and Y are random vectors and u and v are vectors of coefficients. Our clustering method discovers correlations similar to those in CCA that differ between clusters. We present empirical results on synthetic data and a dataset in medicine that demonstrate the utility of the method.

Then, Ch. 6 extends the multi-view clustering method to perform supervised classification by deciding class based on the cluster to which an observation belongs (Lei et al., 2017b, 2019). This method is demonstrated on the medical data and an electricity problem. This latter problem is called *non-intrusive load monitoring*, the task of identifying which appliances in a building are responsible for energy consumption. The total power consumption, which is the sum of consumptions of individual devices, is measured at the utility service entry and must be disaggregated. There are multiple ways to frame the problem, but here we frame it as a pipeline and consider the first two modules: event detection and event classification. Event detection is to identify time points at which an appliance is turned on or off; event classification is to identify the appliance at each event. Our classification method demonstrates superior results than state-of-the-art baselines on a common benchmarking dataset. Lastly, Ch. 7 summarizes key ideas and results and proposes future work.

Part I

Multi-view filtering using known multi-view relationships

Chapter 2

Gamma Source Detection by Simultaneous Estimation of Source Strength and Background

2.1 Introduction

This part of the thesis leverages multi-view relationships assuming they are known through domain knowledge. In particular, we build toward a method to aggregate multiple views by filtering. This specific chapter builds toward this goal by presenting a filtering method for a single view. The subsequent chapter explains how multiple views that each correspond to one of these filters can be combined through structural information.

Variation in local background has long been recognized as a principal challenge in gamma source detection (Ziock and Goldstein, 2002; Aage and Korsbech, 2003; Ziock and Nelson, 2007; Aucott et al., 2014; Bandstra et al., 2016). Background intensities and spectra vary with geology, composition of nearby buildings, amount of visible sky, cloud cover, humidity, and other atmospheric conditions (Bandstra et al., 2016). This variation may obfuscate detectability of especially less pronounced threatening sources. It can be particularly impactful in the context of mobile detectors as they move over heterogeneous environments. As a result, the data generating distributions for source detection tasks differ from commonly assumed Poisson and depend on application and environment (Bandstra et al., 2016). This challenge is often addressed by building statistical models of background variation designed to account for and suppress its influence. This approach is prevalent in source detection algorithms such as Spectral Anomaly Detection, Censored Energy Window, and matched filters (Nelson and Labov, 2009, 2010; Huggins et al., 2014; Tandon et al., 2016; Tandon, 2016). Modeling background variation requires training data from which background distributions can be estimated. Systems such as RadMAP (Bandstra et al., 2016) have been used to collect copious amounts of background data to enable studying natural variation of background radiation. However, during a source detection operation (the detection algorithm's test time), detection performance may suffer if the training data used to model background variation does not represent well the actual distribution. This effect can be quite limiting in practice when a detection algorithm needs to be used in new, unmapped yet environments. For example, the background inside a tunnel is probably much different from the outside.

To handle background variation, a state-of-the-art adaptive method is the Orthonormal Subspace Projection Matched Filter (RDAK) (Labov and Nelson, 2019), which learns and updates a representation of background. However, it requires a warm-up period of about 10 minutes before detection. This warm-up can be considered another type of training, and there are practical scenarios in which it would be infeasible, such as if

- Detection must begin as soon as possible, leaving no time for warm-up.
- The warm-up period could be contaminated by nuisance or threatening sources.
- The test background differs substantially from the warm-up.

These scenarios motivate our work. We propose a novel method that adapts to current local background and substantially reduces dependence of gamma source detection algorithms on relevant, comprehensive, and representative training data or warm-up. We demonstrate its effectiveness in scenarios when a training set is unavailable or when it differs from the background radiation distribution observed at the time of deployment. Assuming that background spectra vary smoothly over time and space, we apply filtering techniques to simultaneously track mean background spectrum and estimate source intensity via the Kalman filter (Kalman, 1960). We test two approaches to detection. First, we directly use source intensity estimated by the filter. Second we combine filtering with an existing detection method by substituting the dynamically estimated rates for the methods' static parameters. This work demonstrates this second approach on a classical matched filter for source detection (Huggins et al., 2014). We illustrate the performance of these approaches on a partially simulated roadside source detection task using gamma-ray spectrometers in two authentic datasets. In the first dataset, we alter training data to differ substantially from the test distribution in order to demonstrate the robustness of the method. The proposed method is empirically compared to a range of non-adaptive alternatives and demonstrates significant improvements in robustness. In the second dataset, we make no changes to the original background spectra, which contain more natural variation than the first. The method is compared to the state of the art, RDAK. It is demonstrated to perform much better when both methods have short but equal amounts of warm-up and when the warm-up differs significantly from test background.

In general, the presented results suggest that it is possible to adaptively, accurately, and simultaneously estimate the source intensity and background spectrum in real-time without prior accurate knowledge of the background radiation distribution, by exploiting smoothness in background rates and source intensity. The proposed approach lessens dependence of gamma source detection systems on costly, highly variable background radiation reference data for training capable detection models. Additionally, it establishes an alternative to RDAK in scenarios where warm-up is impractical.

2.2 Source detection task

This section describes foundations of the source detection task. The observed sets of gamma-ray spectrometer measurements can be expressed as time-ordered vectors of photon counts c_t, wherein each position corresponds to a range of photon energies (energy bin) and the photon counts are aggregated over a short period of time (integration time). We consider binned data with a one-second integration time, though the techniques used here could be easily modified for unbinned list-mode data streams. The count $c_{t,i}$ of particles observed in time interval t and energy bin i can be viewed as a Poisson random variable with rate $\lambda_i(t)$ (Bai et al., 2011), which can be decomposed into the sum of the background radiation rate and the source radiation rate. Examples of a radiation source spectrum, and the result of its injection into the background spectrum, are shown in Figure 2.1. Additionally, source *intensity* is inversely proportional to the square of distance from the detector

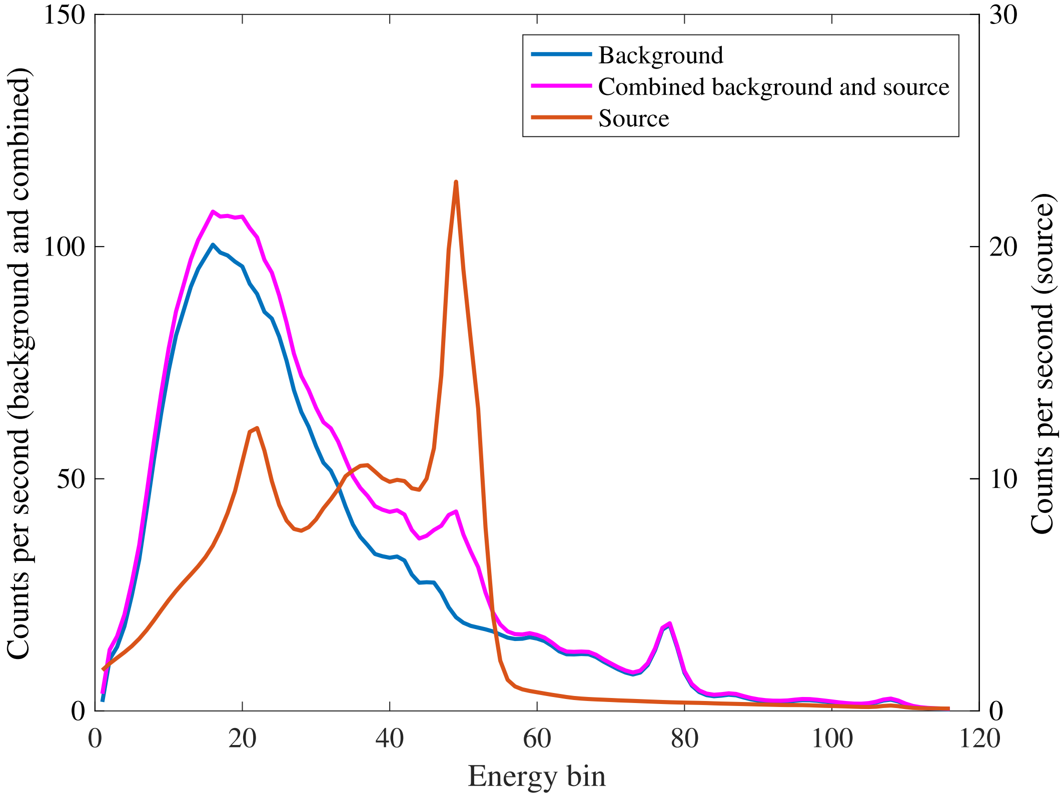


Figure 2.1: Example of a background radiation gamma-ray spectrum, source radiation spectrum, and the injected spectrum, which is the sum of background and source.

to the source (Miller et al., 2016). The higher the intensity, the more confident a detection method should be that the sought after source is present, keeping other factors constant such as relative distance between source to be detected and the sensor, shape of background radiation spectrum, and the extent of its variability.

Although the energy spectra are known for most harmful materials, this knowledge is often not sufficient when facing a noisy background environment due to the effects of shielding, and to the sensor's exposure to multiple emitting materials at varying distances in its proximity, and so the perception of the environment would vary depending on the spectrometer's position and over time. Consequently, the aggregate spectrum of the background sources is dynamic and generally not known exactly, although it is often approximated in practice as the mean of background spectra computed from a reference training set. The spectrum of the target source, on the other hand, can be assumed to be at least partially known by many detection methods (Nelson and Labov, 2012; Tandon, 2016). This is helped by the libraries constructed by domain experts: nuclear physicists have enumerated the spectra of a wide array of possible source types of interest, with various configurations of radioactive material and shielding (Nelson and Labov, 2012).

One requirement shared by all source detection methods listed above is a training set of background measurements. Typically some properties of the training background are learned, and when test measurements deviate from those properties, they are considered more likely to involve a non-background source to be detected. However, the training set could differ substantially from the test set, creating an opportunity to reduce dependence on the training set by updating the model with the test observations. To accomplish this, we propose an adaptive Kalman filter that exploits smoothness to simultaneously estimate local source strength and background spectrum.

2.3 Related work

2.3.1 Non-adaptive methods

Many non-adaptive methods have been proposed for source detection in the presence of noisy background. One fundamental approach is to treat the source detection as anomaly detection task in which one does not rely on information about the target source, but instead builds a model of expected background distribution to compare expected and measured spectra, and issue an alert when the observed discrepancy is substantial. Well-known instances of this approach, Spectral Anomaly Detection (SAD) methods are typically based on some variant of Principal Components Analysis (PCA) to learn the expected characteristics of variability of background radiation.

PCA-based anomaly detection (Nelson and Labov, 2012) utilizes a training set of background measurements. The top principal components of the background distribution are removed from consideration and new observations are scored by their reconstruction error versus the basis reduced to the remaining low-eigenvalue components. Conceptually, this approach learns the most important characteristics of the background presumably random variability and then subtracts them from the new measurements. If the remainder is small, then the measurement is adequately explained by background features. If the remainder is large, however, then the measurement is not explained well by background, which may indicate the presence of a novel source. A number of other Spectral Anomaly Detectors have been proposed as well. Aage and Korsbech (2003) suggest noise-adjusted singular value decomposition and a method of "stripping" away common background features. Runkle et al. (2006) fit a multivariate normal distribution to background in the top principal components space, and scored new observations by Mahalanobis distance. Du et al. (2010) suggest using k-nearest neighbors in the top principal components space of the background, given the availability of labeled training data. De-Arteaga et al. (2015) use Canonical Autocorrelation Analysis to discover multivariate correlation structures among subsets of bins of background spectra to create a null space model for Spectral Anomaly Detection task.

If the source spectrum template is known, in contrast to the anomaly detection scenario, more powerful classes of methods can be used. A popular type of method is the classical matched filter (MF) (Turin, 1960), which often achieves state-of-the-art results in radiation detection (Nelson and Labov, 2012). These filters attempt to intelligently measure the presence of a template in a signal. For instance, a standard MF that maximizes output signal-to-noise ratio takes as input a source template s and data matrix of background counts Y and learns the filter $h = \mathrm{Cov}(Y)^{-1}s$. Then a new observation y is scored as $f(y) = h^{\mathsf{T}}y$.

Another source-type-aware detection method is the Censored Energy Window (CEW) (Nelson and Labov, 2012). CEW finds a set of energy bins, called the energy window, in which a particular type of source is expected to be seen most clearly. A regression model is then fit using background data to predict the total counts within the window from the outer spectrum. New observations are scored according to the degree to which the predicted in-window background counts are exceeded.

Other detection methods also exist. In Anderson et al. (2008), an approach was presented that computes a ratio between observed and expected counts in key energy bins. This method may be viewed as a precursor to the Censored Energy Window Tandon (2016). An improved version of the ratio method was given by Pfund et al. (2016). A method was developed by Bai et al. (2011) and Kump et al. (2013) that uses the variable selection capability of LASSO regression Tibshirani (1996) when fitting the observed spectra to their expected values. Other types of statistical methods were surveyed by Fagan et al. (2012), including peak finding and discriminant analysis. The authors point out that Bayesian methods were largely untapped at the time and could improve performance by incorporating all sources of uncertainty. Subsequently, Tandon et al. (2016) developed Bayesian approaches for aggregating multiple observations and detection methods to achieve better detection.

2.3.2 Orthonormal Subspace Projection Matched Filter

The Orthonormal Subspace Projection Matched Filter (herein called RDAK after its software package) is an adaptive method that exhibits state-of-the-art detection performance (Labov and Nelson, 2019). The main intuition is that when given a new spectra, RDAK evaluates its discrepancy with previous background. The spectral shape of the discrepancy is then compared to the source template. Mathematically, it represents background with a basis B, similar to previous methods such as Spectral Anomaly Detection. Unlike them, however, B is updated adaptively when the current background does not fit. When RDAK has B at each step, it estimates a diagonal precision matrix W from the mean $\bar{y}$ of recent background counts as $W = (\mathrm{diag}(\bar{y}) + cI)^{-1}$, where $c > 0$ is a small constant and I is the identity matrix. The reason for W is to weight the energy bins to control for variance. Let s be the source template and Y a new observation. The score of RDAK is

$$RDAK(y) = \frac{s^\top W(I - B(B^\top WB)^{-1}B^\top W)y}{\sqrt{|y|_1}}.$$

RDAK requires a warm-up period to learn B before detection can begin. In our experiments, we find the ideal period to be 10 minutes.

2.4 Simultaneous estimation of source strength and background

All the source detection methods described above rely on training data, which could refer to warm-up, to learn characteristics of the background distribution. When the training set resembles the test set, the methods perform well. Nonetheless, in practice, there exist several scenarios in which training is infeasible. For example, there may not be time to train or warm-up, or there could be source contamination. Another example is if the test observations come from a much different distribution, which could happen realistically with step changes in background resulting from sudden changes in environment, such as passing through a tunnel or building. In these cases the methods above become less reliable. A final reason to avoid training is that it would be practically convenient for methods to function without setup.

Accordingly, we introduce a method to adaptively estimate source strength and background. We demonstrate how to limit dependence on training data and adapt to the test data by exploiting smoothness in source intensity and background rates. In particular, we simultaneously estimate source intensity and background spectrum using an adaptive Kalman filter (Mehra, 1970). From these estimates we propose two ways of source detection in the setting of a known source spectrum. First, we simply use the source strength estimated by the method as a detection score. Second, we improve existing methods by substituting the adaptively estimated rates for static background parameters. In this work, we select classical MF (not RDAK) to demonstrate this approach.

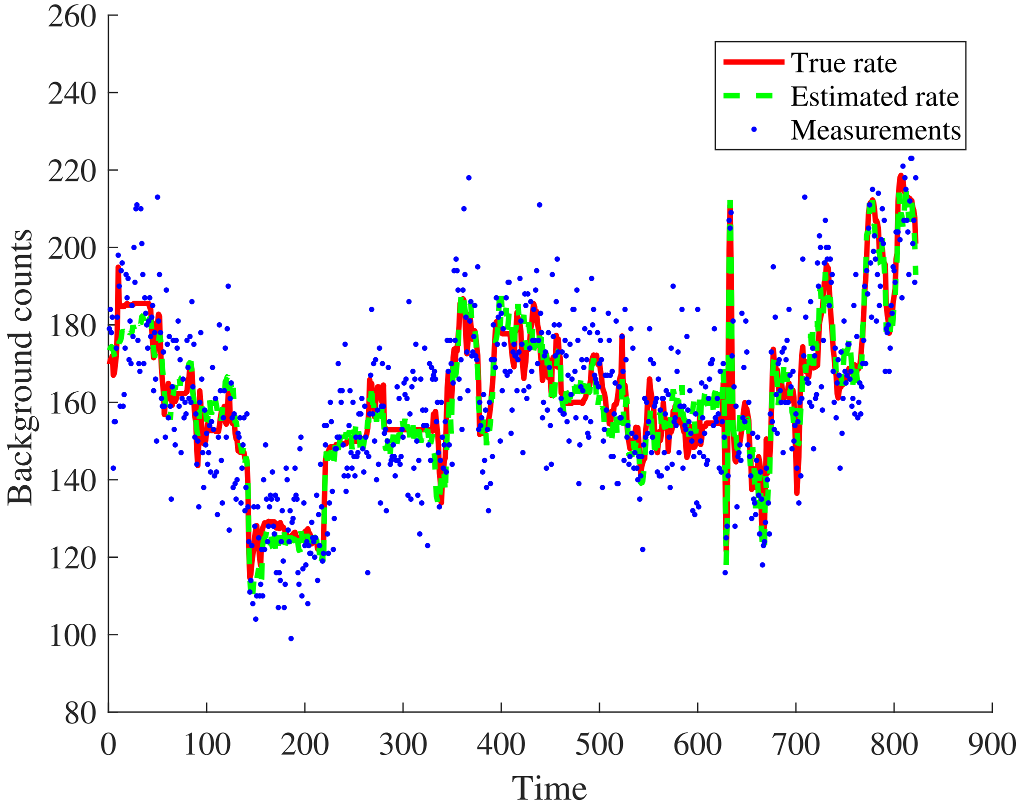


Figure 2.2: Using the Kalman filter to estimate background rates from noisy measurements in one energy bin over a sequence of 1-second spectral measurements.

2.4.1 Notation

We consider discrete stochastic processes indexed by $t \in \{1, \dots, T\}$. We use $a_{i:j}$ to refer to $\{a_i, a_{i+1}, \dots, a_j\}$ in a stochastic process $\{a_t\}$. Let $x^{(i)}$ be the i^{th} element of vector x. Let $x^{(i:j)}$ be a vector containing elements i through j (inclusive) of x. Let $\operatorname{diag}(z)$ be a function that maps a vector z to a diagonal matrix whose diagonal elements are given by z.

2.4.2 Kalman filter for radiation detection

First we explain the general Kalman filter, a widely used method for estimating an unobserved signal (state) of a linear dynamical system underlying an observed noisy discrete time series (Kalman, 1960). Implicitly assuming smoothness, it filters out statistical noise to obtain more precise estimates of a signal, such as in Fig. 2.2. Consequently, it is a natural candidate to model sequences of measured radiation spectra. Let $x_t \in \mathbb{R}^d$ be the state space and $y_t \in \mathbb{R}^p$ be the observation at time t. Let $A \in \mathbb{R}^{d \times d}$ and $C \in \mathbb{R}^{p \times d}$ be the transition and emission matrices respectively. Let $Q_t \in \mathbb{R}^{d \times d}$ and $R_t \in \mathbb{R}^{p \times p}$ be the covariances of process noise and measurement noise respectively. The parameters A, C, $\{Q_t\}$, and $\{R_t\}$ are assumed to be known. In this model we allow the covariances to be non-stationary, which is uncommon, but as long as they are non-random, all calculations are functionally identical to the stationary case. The linear dynamical system is defined as

$$x_{t+1} = Ax_t + w_t \qquad y_t = Cx_t + v_t$$

where $E(w_t) = 0$, $\operatorname{Var}(w_t) = Q$, $E(v_t) = 0$, and $\operatorname{Var}(v_t) = R$. We assume the $\{w_t\}$ and $\{v_t\}$ are independent. The Kalman filter gives the best linear minimum mean squared error estimator of x_t given $y_{1:t}$. Also, if the noise variables are Gaussian, then it is the optimal minimum mean squared error estimator (Shimkin, 2009). Update equations are given in A.

Now we discuss how the Kalman filter can be applied to simultaneously estimate source intensity and background rates. We dynamically estimate the background rates at each step. The state space $x_t \in \mathbb{R}^{d+1}$ contains the background rates in each of d energy bins in the first d elements. Furthermore, the source intensity is modeled in order to decouple its effect on the measured spectra from the background radiation. Source intensity can vary with distance and obfuscation by attenuating objects. As a result, we model source intensity in the last element $x_t^{(d+1)}$. The observation $y_t \in \mathbb{R}^d$ contains the measured photon counts at time t. The transition matrix A is set to the identity matrix because we have no *a priori* expectation as to how the background rates change. Thus, we propose the dynamical system given by

$$x_{t+1} = x_t + w_t \qquad y_t = x_t^{(1:d)} + x_t^{(d+1)} s + v_t$$

where $s \in \mathbb{R}^d$ is the target source spectrum. The second equation implies a form for the emission matrix C. The Kalman filter estimates the background rates $x_t^{(1:d)}$. The equations for updating the belief state after receiving a new measurement are provided in Appendix A. Note that since estimated rates can be non-positive in practice, we force them to be at least 0 after every iteration, which boosted performance in experiments.

The natural detection score of the Kalman filter method is simply the estimated intensity. We test this score in our experiments; however, we also aim to demonstrate how the adaptively estimated background rates can boost the performance of other methods that learn parameters from background. Here we use the standard matched filter (MF) as an example. Given a template s and data matrix B of background counts, the MF is $h = \text{Cov}(B)^{-1}s$. A new observation y is scored as $f(y) = h^{\mathsf{T}}y$. When the data B are unavailable or do not match the test distribution, the performance of MF may degrade. We can ameliorate this issue by estimating the background rates in real-time via the Kalman filter. We propose that we estimate $\text{Cov}(B)$ as $\text{diag}(x_t)$ where x_t is the estimated background at time t. This technique utilizes the Poisson maximum likelihood estimate of variance. It also assumes independence between energy bins, a simplifying assumption that reduces the amount of samples needed to a single point. Furthermore, since the estimated rates can be zero in practice, we add 10^{-3} to the estimated variances.

Next we explain how to set the remaining parameters, the transition noise covariances $\{Q_t\}$ and emission noise covariances $\{R_t\}$.

2.4.3 Adaptive covariance

Conventionally, the covariances of process and measurement noise are taken as input to the Kalman filter; in many scenarios, however, their values are unknown (Mehra, 1970). For instance, in the gamma source detection application it is possible that training data are scarce or differ from the test data, so it would be infeasible to estimate the covariances beforehand. When the covariances are estimated in real-time, the resulting method is known as an adaptive Kalman filter, and is well studied. Various approaches include Bayesian, maximum likelihood, covariance matching, and correlation techniques (Ottersten et al., 1998; Akesson et al., 2008; Bavdekar et al., 2011; Odelson et al., 2016). Most adaptive Kalman filters rely on estimators with desirable statistical properties such as asymptotic normality, unbiasedness, and consistency, yet these estimators are often computationally expensive. Here we propose a simple method for estimating the noise covariances by exploiting knowledge of the radiation domain. We do not show any statistical properties of the estimators but instead showcase their utility empirically. Although performance could be further improved by a more accurate and precise means of covariance estimation, the aim of this work is not to compare this method to other adaptive filters but only to show that it performs well compared to non-filtering approaches.

The method for computing the noise covariances is to use the previous state estimates as training data. In the process noise covariance Q_t, only the upper $d \times d$ block is estimated. For the remainder, we assume the background rates to be independent of source intensity, and the standard deviation of the intensity is set to 10. The exact value of this setting appeared insignificant in our experiments. For the upper $d \times d$ block, we first compute $\Delta\widehat{x}_i = \widehat{x}_i - \widehat{x}_{i-1}$ for $i = 1, \ldots, t-1$. We let $\widehat{Q}_0$ be the sample covariance of this variable.The next step is to apply Gaussian smoothing to $\widehat{Q}_0$. We selected a kernel length of 2 with a unit variance, but the smoothing hyperparameters were not found to be especially important. In our experiments, this process produced an effective, if potentially biased, estimator.

The measurement noise R_t can also be easily computed. Given the state x_t, the measurement noise $v_t = y_t - x_t$ is Poisson. In particular, $v_t^{(j)} \sim \text{Poisson}(x_t^{(j)})$. Furthermore, the elements are known to be conditionally independent. Therefore $R_t = \text{diag}(x_t)$. However, since $\widehat{x}_t$ is not known at time t, we instead estimate $\widehat{R}_t = \text{diag}(\widehat{x}_{t-1})$. In practice, this estimate performs better when divided by 2, which causes the estimated state to track the measurements more closely.

Additional design considerations reflect the challenge of unquantified uncertainty and filter burn-in. First, the approach for adaptively estimating the noise covariances described above is inconsistent with the assumptions of the method. While the conventional Kalman filter assumes Q_t and R_t are non-random, the estimators in this method depend on the output of the filter itself at the previous time steps, which is subject to random noise. Hence, when the filter estimates the uncertainty of its prediction, it misses the additional uncertainty contributed by the covariance estimators. This issue is solved in other approaches to adaptive filtering, but here we leave it as future work.

Second, we found it useful to enforce a filter burn-in period in our experiments. In particular, we assumed that measurements taken over the first L time steps do not contain source spectra, so we forced $\widehat{x}_t^{(d+1)} = 0$ when $t \leq L$ regardless of the output of the Kalman filter. The assumption is partially justified since a spectrometer can often stay in a zone known to have no source during the burn-in period. In the first dataset we set $L = 150$. When warm-up was used in the second dataset, we used half the warm-up as burnin. This approach allows the filter to calibrate to the correct noise covariances.

2.5 Experiments

This section includes experiments on single-pass roadside detection in two datasets.

2.5.1 Background with synthetic variation

Here we describe an experiment that simulated a source in the presence of heavily shifting background radiation.

Data

Our dataset was a collection of over 11,000 gamma-ray measurements recorded in one-second intervals by a sodium-iodide detector mounted on a vehicle moving around an urban area in and around Berkeley, CA, USA (Fig. 2.3), created as part of the RadMAP project (Bandstra et al., 2016). On average there were 10,000 photons counted per second, but this measure had significant variability (Fig. 2.4). Each measurement consisted of $d = 116$ quadratically spaced energy bins of photon counts. The measurements were assumed to be clean background data without presence of target or nuisance sources. Each measurement was annotated with the GPS position of the sensor vehicle and time. The background rates were estimated from the measurements and viewed as the true rates using GP estimation over observations within 10 meters of the current measurement being estimated (Miller et al., 2016). To facilitate source detection experiments, we synthetically injected this background

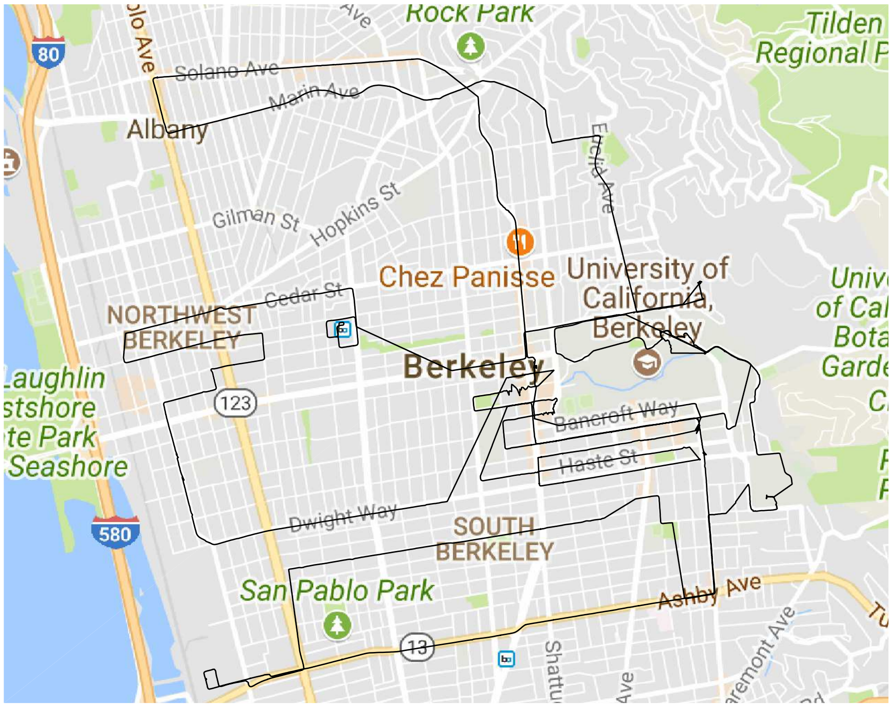


Figure 2.3: Path of the sensor.

data with signatures of a source spectrum corresponding to Americium-241, a nuclear waste isotope. Injections were generated as random samples from the Poisson rates determined by the source spectrum and assumed intensity.

Procedure

This experiment tested detection of a roadside source in a single pass. In each run of the experiment, a location for the synthetic source was randomly sampled from the set of points at most 100 meters from the path of the sensor. The test set was taken to be the data points where the sensor was at most 150 meters away from the source, along with all data points in between, forming a contiguous window. The window was extended on either side by 450 observations. The training set was taken to be the remaining data points.

To simulate different distributions between the training and test sets, we applied a

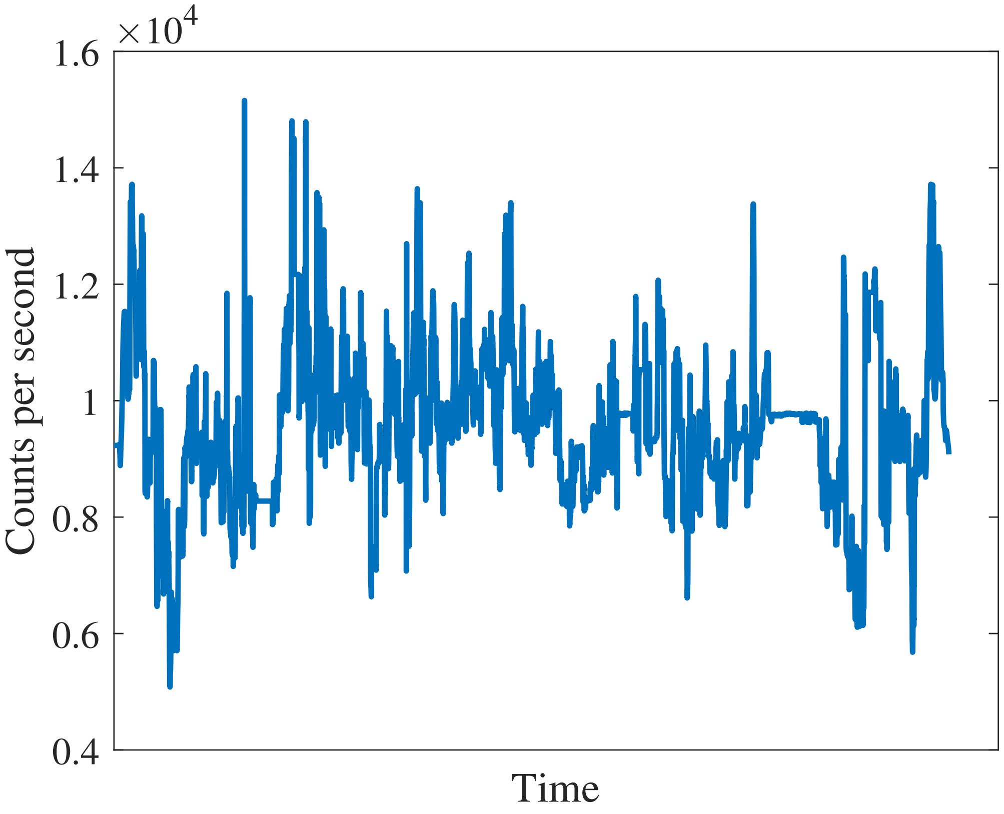


Figure 2.4: Fluctuation of total background counts per second over time.

shifting algorithm to each measurement in the training set (Fig. 2.5), which simulated gain drift (App. B). Photon counts in any particular energy bin were partially shifted to higher energy bins with a gain drift coefficient of 0.3. This shift led to Pearson correlation of roughly 50% between the mean training background spectrum and the test background spectra. Counts of photons yielded by the source were sampled and added to each observation in the test set according to the distance to the source.

The detection rule states that a source is predicted to be present if the maximum observation score is above a fixed decision threshold. All experimental runs were repeated identically except with no source injection, thereby producing negative examples.

Results

We compared performance between several methods:

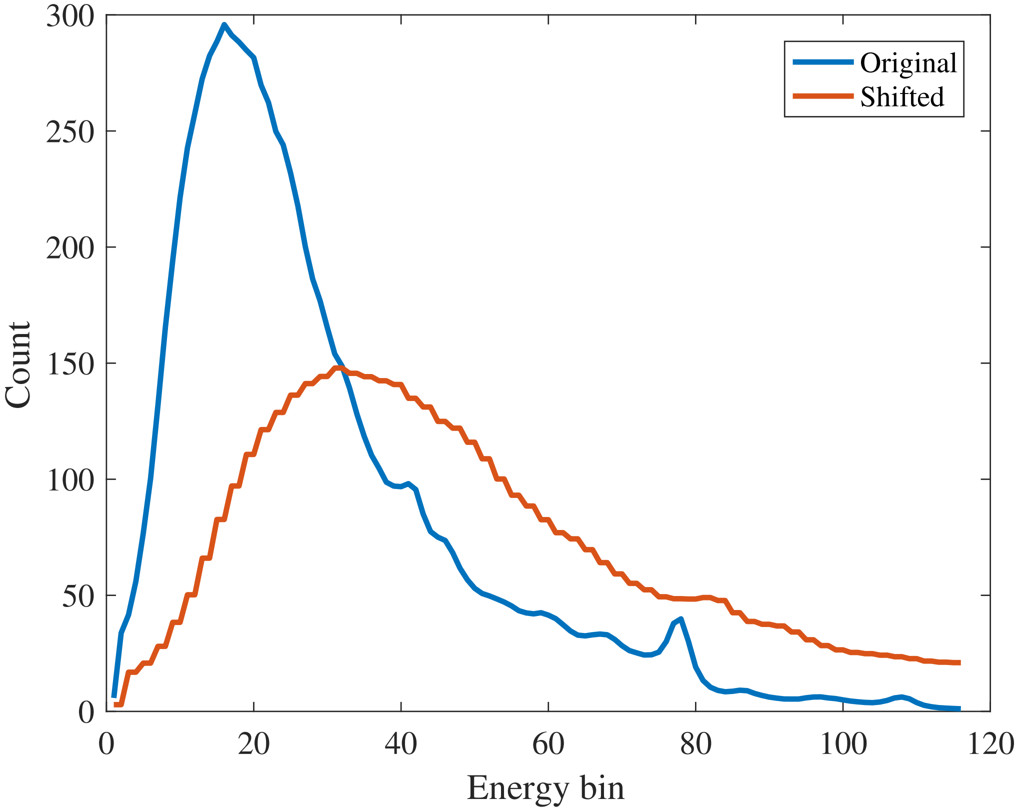


Figure 2.5: Mean background, showing training observations shifted to decrease correlation with test observations.

1. Oracle: Poisson likelihood ratio using perfect knowledge of the background rates and source intensity.
2. KF Intensity (KFI): Intensity estimated by the Kalman filter directly as score.
3. Matched Filter (MF): Matched filter calibrated on training data.
4. Adaptive Matched Filter (AMF): Matched filter combined with Kalman filter.

Figs. 2.6, 2.7, and 2.8 show the Receiver Operating Characteristics computed over 250 positive and negative runs with maximum source intensities of $M \in \{75, 125, 175\}$ photon counts per second (cps), obtained by varying the detection thresholds of the considered algorithms. The graphs show cumulative performance of the considered alternative models in terms of probability of source detection (denoted as TPR, true positive rate) as a function of false detection probability (denoted as

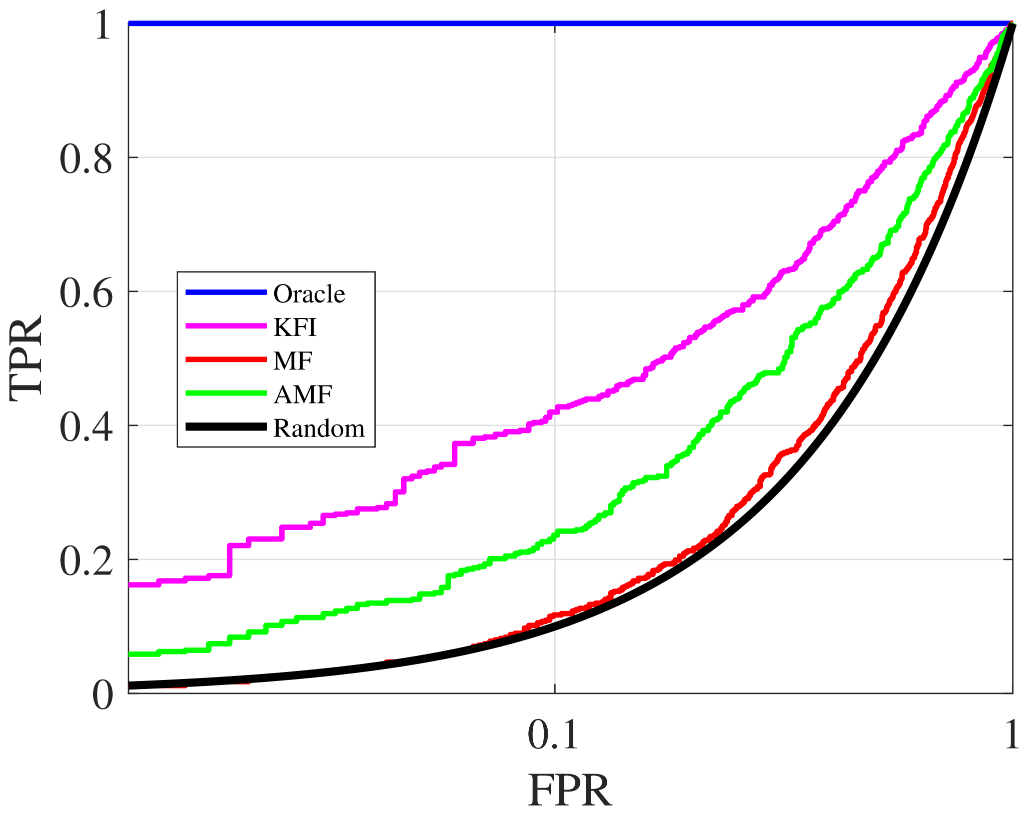


Figure 2.6: Receiver Operating Characteristics at source intensity of 75 photon counts per second.

FPR, false positive rate, shown in decimal logarithm scale). Furthermore, Table 2.1 provides statistics for the $M = 125$ case: Area-Under-Curve (AUC), TPR at 1% FPR, FPR at 50% TPR, and errors on estimated background rates and intensities for applicable methods. The background rate error is computed as the Kullback-Leibler divergence between the actual and predicted Poisson distributions, averaged over all observations and bins in the case where a source is present. Uncertainties of rate errors are extremely low and therefore omitted. The intensity error is the root-mean-square error between the actual and predicted source intensity in the case where a source is present. The uncertainties of these metrics can be found in Table 2.2.

For every M the best method was the Oracle, which was expected because it used more information than realistically available. We used these methods as a reference to calibrate upper-bound performance of the more practical alternatives.

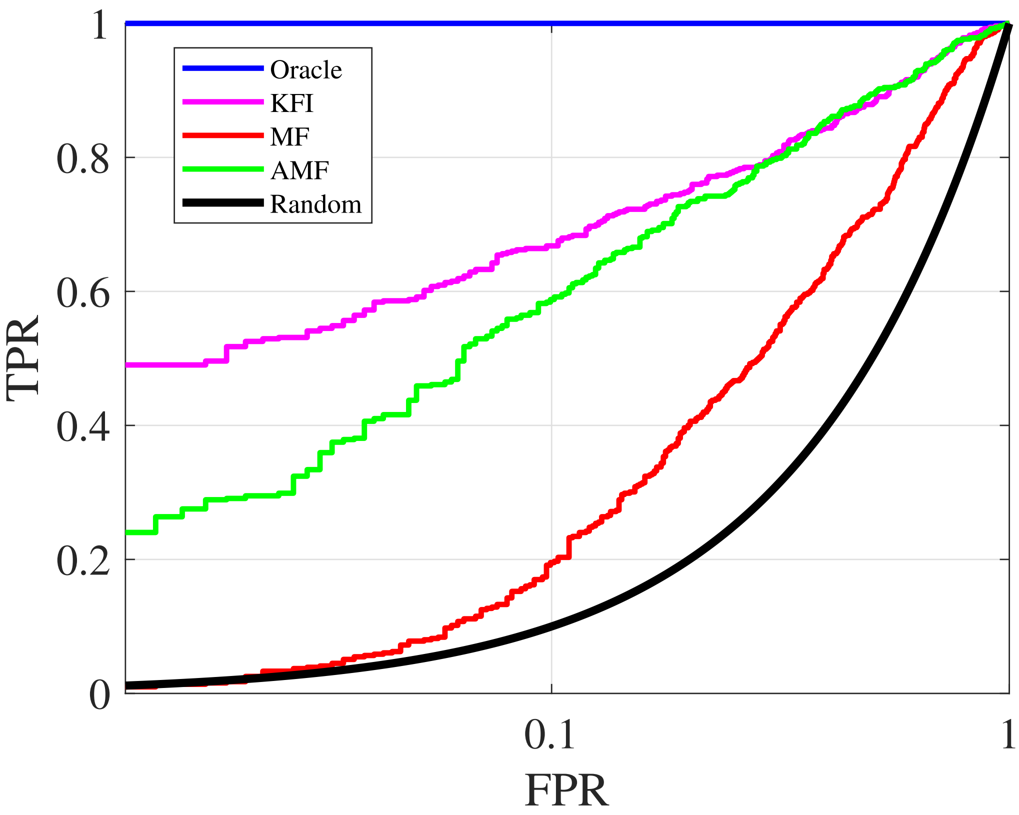


Figure 2.7: Receiver Operating Characteristics at source intensity of 125 photon counts per second.

The next best was one version of our proposed method, KFI, followed by the other version, AMF. The worst performer was MF. This advantage demonstrated the value of adaptive estimation of local background rates and source strength.

Fig. 2.9 visualizes the estimated background rates in an energy bin over a single trial of the experiment. It compares the actual rates to those estimated by the Kalman filter. In this trial, the sensor remained stationary near the source for some time. Visually, the Kalman filter tracked actual rates with minimal lag. Additionally, it overestimated the rate only slightly when the source intensity peaked.

Fig. 2.10 presents how the adaptive Kalman filter estimated the source intensity. Although there was substantial noise in the estimate when source intensity was zero, the estimated intensity detected the peak and plateau with minimal lag. This method was also able to detect the initial spike in source intensity, but this window lasted only briefly, and it was difficult to separate it from the substantial noise.

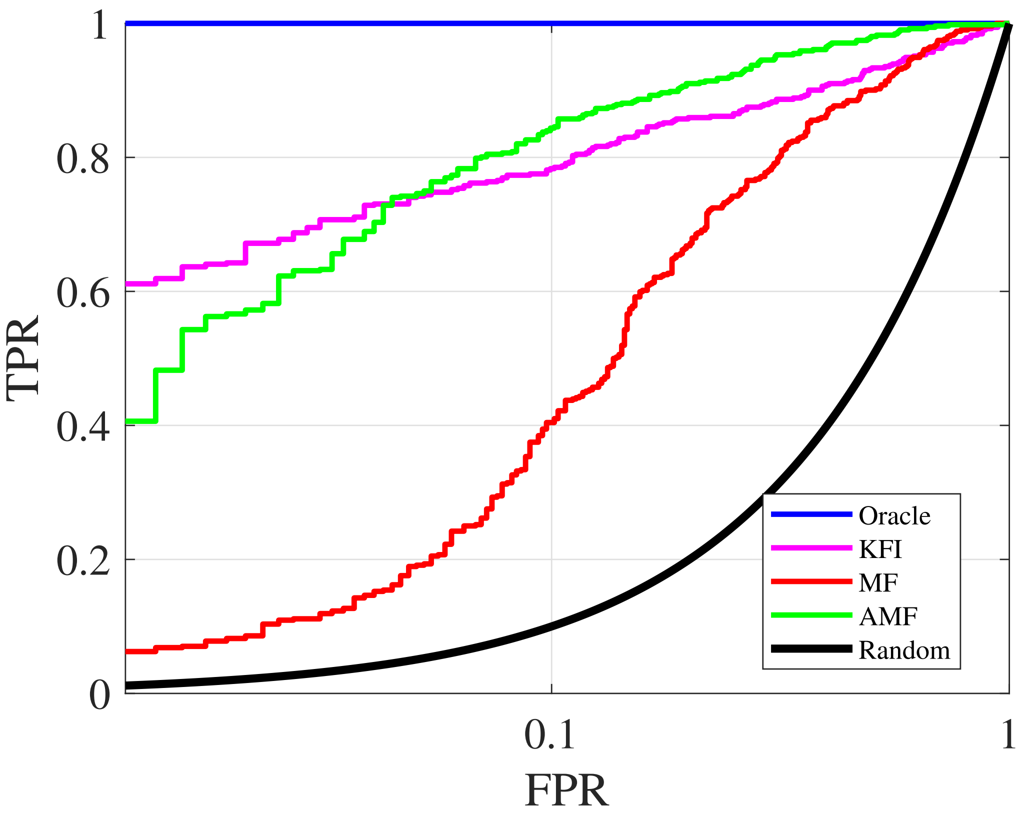


Figure 2.8: Receiver Operating Characteristics at source intensity of photon 175 counts per second.

2.5.2 Background with authentic variation

The previous experiment induced synthetic background variation, which may not reflect actual background conditions. Here we describe an experiment on data that contained substantially more background variation naturally.

Data and procedure

We used spectroscopic data from a major metropolitan area. [1] The average gross background count was about 400. We synthetically injected a source, Cobalt-60, an industrial isotope. The injections occurred at random roadside locations. Again, we tested single-pass detection. The passes included about 10 seconds before and after the source, and about 20,000 passes were simulated. This dataset included multiple

[1] Not publishable.

Table 2.1: Quantitative performance at source intensity of 125 photon counts per second. "Rate" refers to background rate estimation error in natural units. "Int." refers to source intensity estimation error. Its two columns correspond to units of counts per second and percentage of the maximum intensity respectively.

	AUC	TPR	FPR	Rate	Int.	Int. (%)
Oracle	100.0	100.0	0.0			
KFI	85.5	66.8	1.8	0.06	13.6	10.9
MF	65.9	19.5	28.3			
AMF	83.2	59.0	6.3			

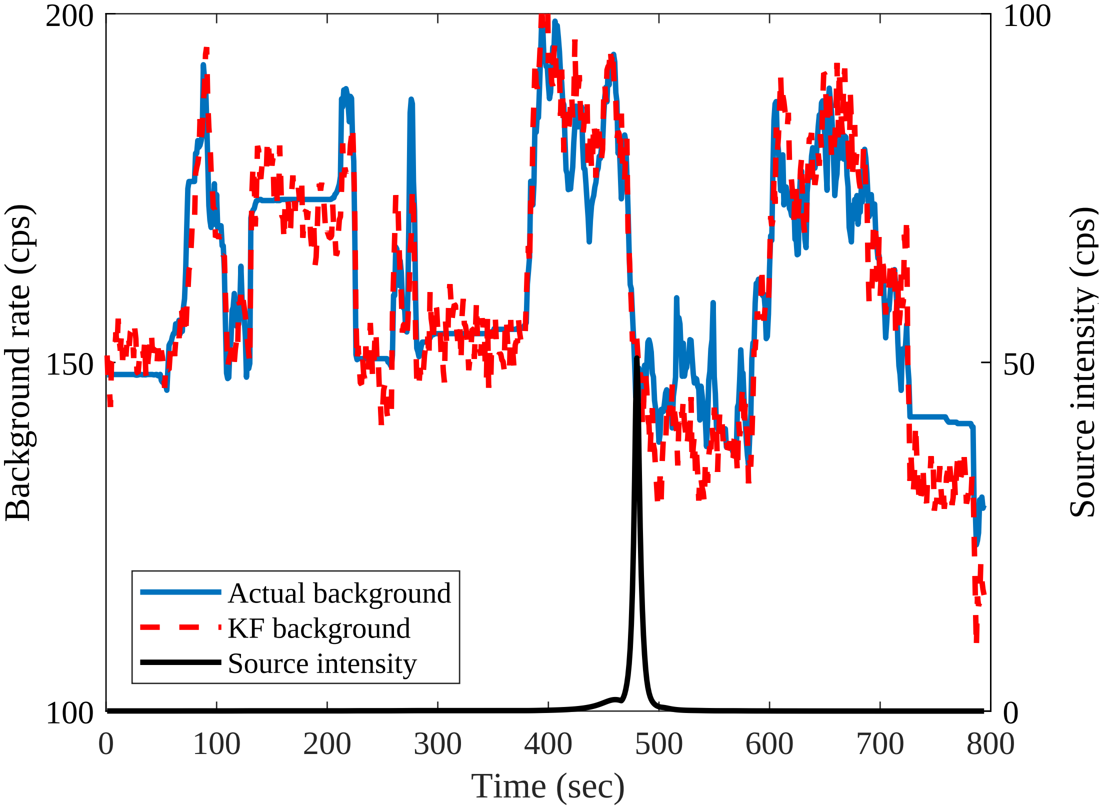


Figure 2.9: Actual background rates in the first energy bin (blue) are tracked very well by the KF method (red) despite a 175 gross cps source injection (black) containing about 150 cps in this bin.

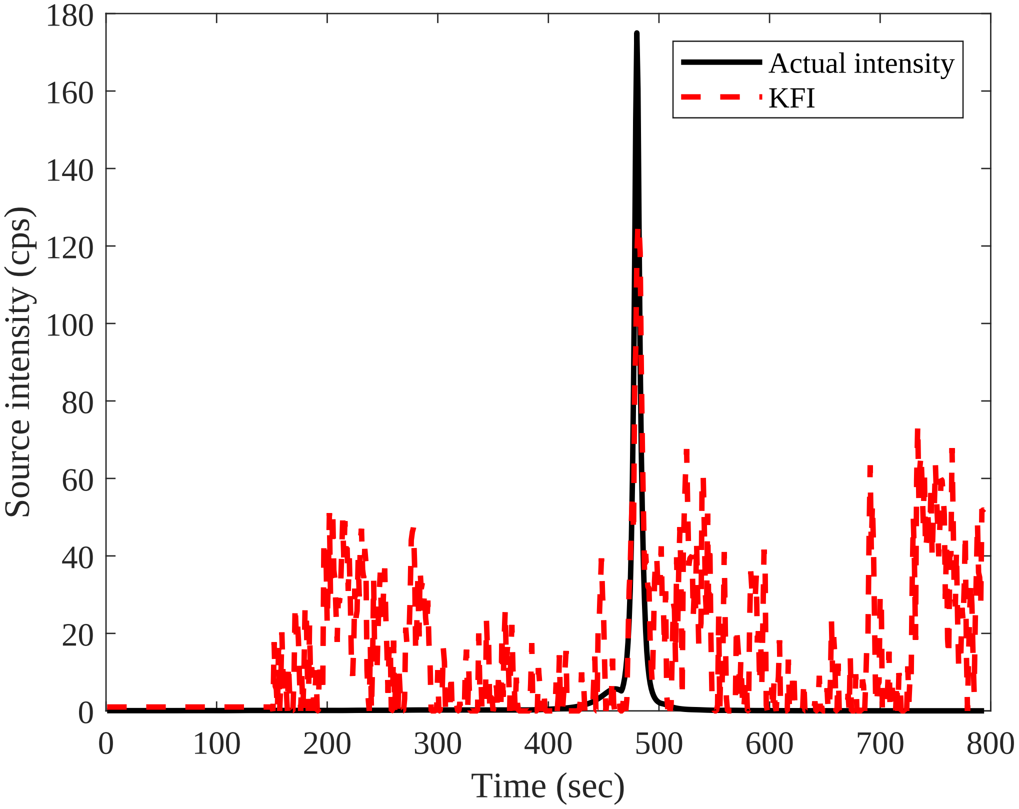


Figure 2.10: Source intensity estimated directly using KF method (red) over one trial with a 175 cps source (actual intensity shown in black).

Table 2.2: Uncertainty of quantitative performance at source intensity of 125 photon counts per second. Twice the standard error of values in Table 2.1.

	AUC	TPR	FPR	Int.	Int. (%)
KFI	4.4	5.9	1.7	0.3	0.3
MF	5.9	5.0	5.6		
AMF	4.7	6.2	3.0		

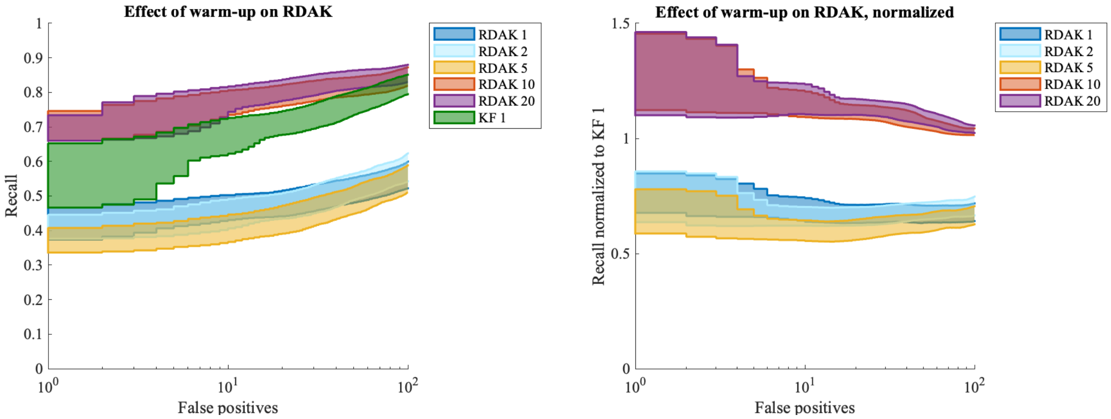


Figure 2.11: ROCs as warm-up varies for RDAK but is fixed for KF. Left: standard ROCs. Right: recall normalized by KF with 1 minute warm-up.

sensors in different parts of the area. All procedures carried over from the previous experiment unless otherwise mentioned.

Results

The first experiment varied the amount of warm-up given to KF and RDAK. Warm-up was taken from a disjoint time from the same sensor as used in the pass. Fig. 2.11 illustrates the performance when KF warm-up is fixed at 1 minute while RDAK is given varying amounts. The best method was RDAK with 10 or more minutes of warm-up, but it performed poorly with less than that. In the middle was KF with only 1 minute. Next, Fig. 2.12 illustrates performance when RDAK warm-up was fixed at 10 minutes and KF varied. Although KF with 0 warm-up performed poorly, it was largely insensitive to the amount of warm-up otherwise. In sum, RDAK was superior after running 10 minutes, while KF was better before that. This result suggests to create a switching scheme where KF is used while RDAK is warming up to improve overall detection performance. Interestingly, KF did not improve much with more warm-up, which probably reflected its limited memory as a

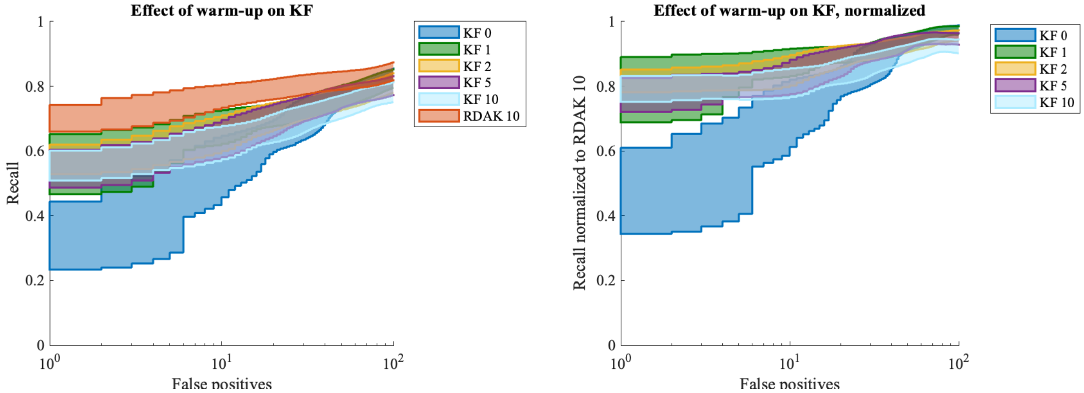


Figure 2.12: ROCs as warm-up varies for RDAK but is fixed for KF. Left: standard ROCs. Right: recall normalized by RDAK with 10 minutes warm-up.

smoothing algorithm. Nevertheless, there are potential problems with using warm-up in the first place. First, there may be no time to warm-up. Second, the warm-up could be contaminated by nuisance or threatening sources, which could drastically alter RDAK's basis. Third, the test background could have a severe mismatch with warm-up. An extreme example of this case would be a step change in background if, for instance, RDAK is warmed up inside a building and taken outside.

Thus, we simulated a step change between warm-up and test in the second experiment. Previously warm-up data were from the same sensor as the pass, whereas now we used distinct random sensors, an approach we called cross-training. Since sensors were from different places in time and space, the warm-up background was much less likely to match the test background. The results are shown in Fig. 2.13, which highlights a large dropoff in performance by RDAK when cross-trained. Meanwhile, KF was robust to the effects of the step change, signifying that it adapted better to large background fluctuations. This result indicated KF may be well-suited to environments with many sudden changes in background, such as a highway with many bridges and tunnels.

In the previous experiments, we only used a single source type. Therefore, we conducted a third experiment in which we tested the Minimum Detectable Amount of different injected source types. The strength of 15 different source types was adjusted until TPR was 50% at 8 false positives. We compared KF with 2 minutes of warm-up to RDAK with 10 minutes of warm-up (Tbl. 2.3). The results suggested that the difference between KF and RDAK was stable across different threats. On average RDAK outperforms KF because of its long warm-up period.

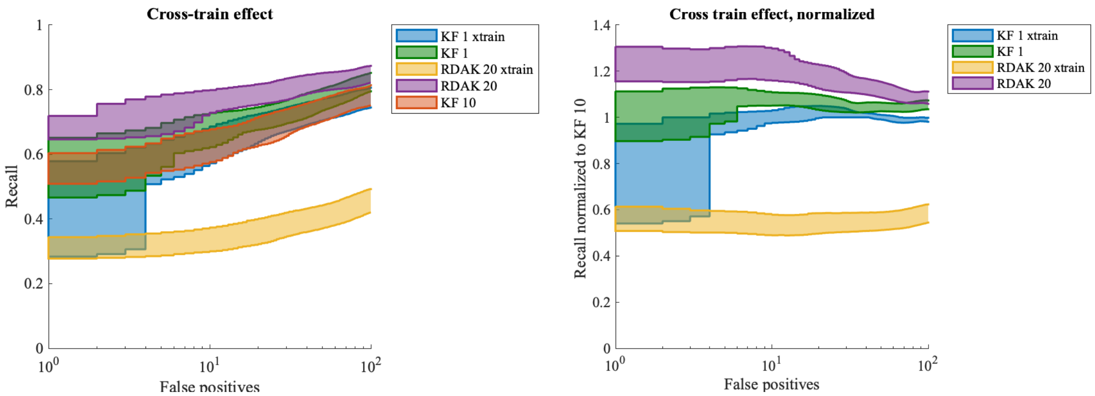


Figure 2.13: ROCs with and without cross-training. Left: standard ROCs. Right: recall normalized by KF with 10 minutes warm-up.

Table 2.3: Minimum Detectable Amount across 15 source types. Values show the source strength for KF with 2 minutes of warm-up divided by source strength for RDAK with 10 minutes of warm-up. Values over 100% indicate KF needs a brighter source to have equal performance to RDAK.

Source Type	KF MDA normalized by RDAK MDA
Pu_1	70%
Co60_2	87%
Co60_3	107%
Co60_1	116%
Cs137_3	123%
Cs137_2	128%
Cs137_1	133%
HEU_3	165%
HEU_1	172%
HEU_2	187%
Am241	188%
HEU_4	189%
Pu_3	305%
Pu_4	362%
Pu_2	365%
Mean	180%

2.6 Conclusion

A method based on the Kalman filter was proposed for the gamma source detection problem in the presence of noisy and unknown background radiation. Background and source radiation were modeled as a linear dynamical system, allowing the filter to exploit smoothness to simultaneously estimate source intensity and background rates, thereby separating source from background. We proposed two ways to convert the filter's estimates to detection scores. First, the intensity can be used directly. Second, the background rates may be inserted into previous methods to substitute their static parameters with local estimates. We showcased this second approach using a standard matched filter. The advantage of the presented method is that it does not assume an informative and comprehensive training set of background radiation data is available yet remains adaptive to changes in background, although it does assume that the source's spectrum template is known. The new method was empirically demonstrated to perform better than realistic alternatives on RadMAP data when a mismatch was induced between training and test data in both classification performance and quality of the estimates of source intensity and background rates. Additionally, it was evaluated against a state-of-the-art adaptive algorithm on unaltered background from another dataset. These experiments showcased better performance with extremely short warm-up periods and robustness against dramatic changes in background. These results suggest our method is a preferred alternative in many practical scenarios.

To more rigorously evaluate the method, an option would be to marginalize over multiple source types. Typically the correct source template is unknown, and it is assumed to be one among a library of them. Methods like KF and RDAK are run on each template and the scores aggregated, often by a simple maximum. In our work we only used the correct source template, which would reduce false positives across all methods. To better simulate practical conditions, it would be ideal to include marginalization over a realistic source library.

Another extension of this method might be to track weights for a background basis rather than the full spectrum of background rates because a basis can efficiently represent background characteristics with less variance in its predictions. Ideally the basis would be computed adaptively, perhaps in the same manner as RDAK, which would make our method perform better with more warm-up. Another approach to computing the basis, albeit non-adaptive, is given by Bilton et al. (2019), using a nonnegative matrix factorization that minimizes Poisson rather than Gaussian loss. It is possible that this method could applied to a rolling window of background to compute an adaptive basis.

An additional intellectual opportunity created by this work is to integrate RDAK and the Kalman filter. We previously remarked that we could switch between them. However, there could be a more principled way to combine them. For example, RDAK employs its own algorithm to adaptively estimate background, but it is possible that this algorithm could be replaced by the Kalman filter.

Lastly, the work here applied to radiation with relatively high counts per bin. This condition allowed the Kalman filter to perform near optimally because the Poisson distribution can be approximated by the normal. However, there are many datasets in which counts are too low. Here, the Kalman filter could break down. A future avenue would be to test modifications of this work for low count conditions. One simple change would be to apply an Anscombe transform (Anscombe, 1948) to change counts to have an approximately normal distribution. Another would be to replace the Kalman filter with a particle filter, which can handle non-Gaussian cases.

Chapter 3

Multi-view filtering using known multi-view relationships

3.1 Introduction

This chapter extends the previous one to multiple views by exploiting domain knowledge about the multi-view relationship to improve inference. We consider a similar problem setting as the previous chapter but with multiple sensors at separate but nearby locations, a common practical scenario (Tandon, 2016). For instance, multiple vehicles could scan the area around the same building. Here, sensors would be exposed to different background radiation and source intensity. Our objective is to investigate how multiple views can be combined using structural information about their temporal and spatial relationships to enhance prediction. We model this relationship with physical domain knowledge to refine our inferences, aggregating multiple simultaneous sensors by a novel multi-view filtering method tailored to the domain. This approach allows us to identify a source of interest in an adaptive manner similar to the Kalman filter method but with greater expected detection performance. The first advantage of our method compared to other aggregation methods is a lesser dependence on the amount of data. The second advantage is that if sensor passes are simultaneous, we leverage all collective information gathered to refine inference in real-time. Our work assumes at most one source is present, the source template is known and stationary, any source is isotropic and stationary in position, and no occlusions exist.

3.2 Bayesian Aggregation

Our method extends Bayesian Aggregation (BA), a state-of-the-art detection method that can flexibly incorporate many variables such as multiple sensors (Tandon, 2016). BA spatially aggregates detection scores $\{x_i\}_{i=1}^T$ to compute probabilities of different source hypotheses. Let the *luminosity* of a point source refer to the total observed counts per second from the source by a sensor that coincides in space. In its simplest form, BA tracks a null hypothesis H_0 that no source is present as well as k alternate hypotheses $H_{1,L}, \ldots, H_{k,L}$ that a source of luminosity I_k and fixed type is at a particular location L. A planar grid is placed over the area with the null and alternate hypotheses at each lattice point. Then Bayes' theorem is used to update hypothesis probabilities upon new observations. An important assumption is that x_i and x_j are independent if $i \neq j$. The final score is the highest likelihood ratio of nonzero source intensity to zero source intensity,

$$\max_{k,L} \frac{\prod_{i=1}^T \Pr(x_i | H_{k,L})}{\prod_{i=1}^T \Pr(x_i | H_0)}.$$

Multiple sensors are not explicitly modeled but naturally incorporated by dynamic updates at any point in time and space.

BA requires two training sets. The first set is used to train a model of background radiation variability, called the *detector*, such as Matched Filter or Censored Energy Window, to get the x_i variables from raw observed spectra. The second set is used to learn probability distributions $\Pr(x|H)$ over the scores for each hypothesis given the expected exposure to source radiation, a function of distance, velocity, and duration of measurement. These distribution can be estimated nonparametrically, such as with a kernel density estimator. Alternatively, if scores are given by an affine function of the observed spectra, the distributions can be estimated parametrically by assuming a normal distribution. The distribution can be directly fit for the null hypothesis, and for each hypothesis the distribution can be derived in closed-form.

3.3 Aggregation of sensors by multi-view filtering

3.3.1 Motivation

Although BA performs well empirically, it requires a large quantity of training data to achieve confident predictions (Tandon, 2016). This drawback motivates a need for a detection method with reduced dependence on training. Additionally, although BA can flexibly incorporate inferences from multiple detectors at different sensors, if the series of inferences are simultaneous, it misses the opportunity to share information between sensors to refine detection in real-time. We propose to use BA to aggregate information from all sensors at every time step; furthermore, we use the aggregated information to form a feedback loop by informing each detector of BA's inferences about potential sources. This strategy allows sensors to indirectly communicate information through a central hub.

This new method, which we name the Bayesian Aggregation Filter (BAF), extends BA. It can be summarized by the following changes to BA:

1. By leveraging physical relationships, BAF requires little to no training data, in contrast to BA's requirement for two training sets. To enable this feature, BAF uses the Kalman filter (KF) from Ch. 2 for the detector at each sensor to skip the detector training step. The KF intensity is the detector score. Then it applies a physical model (Eqn. 3.1) to skip the distribution training step.

2. At each time step, BAF strengthens the detector at each sensor by feeding it the collective inference of all sensors from the previous time step, unlike BA, which does not exploit temporal or multi-view structure across detector scores. To enable this feature, BAF applies BA at each time step to the KF estimates of intensity so far, computing the most likely source hypothesis according to likelihood ratio. Then the expected intensity at each sensor is computed according to a physical model (Eqn. 3.1). Lastly, BAF modifies the KF detectors to incorporate the previous step's expected intensity as an observed variable.

The key difference between BA and BAF is that BA assumes observations are independent, whereas our method contrarily assumes dependencies exist across both time and sensors.

The most important practical assumptions of this work are that each sensor makes a single pass by the source, as in Fig. 3.1, and that these passes occur simultaneously.

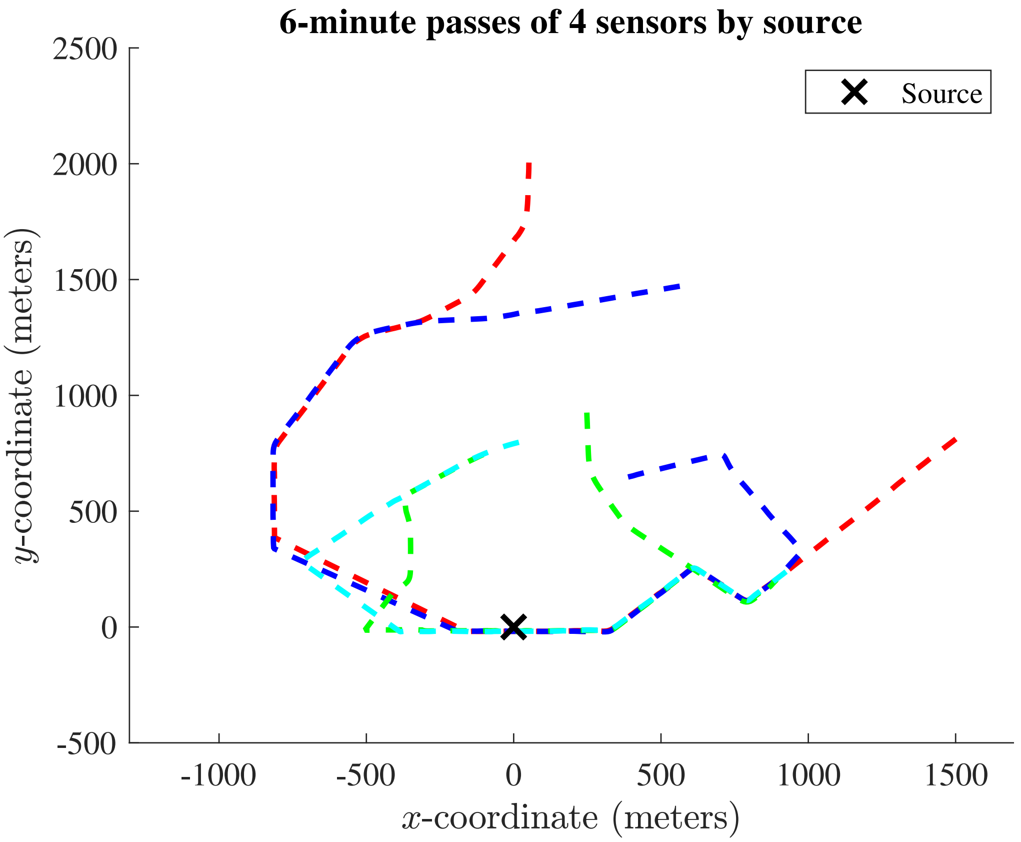


Figure 3.1: Path of four sensors near the same source over six minutes.

After all, our goal is to utilize contemporaneous multi-view structure; it would be less interesting to aggregate sensor passes that intersect in space but not in time because there would be no such structure. However, we are careful not to exploit the exact knowledge that the source is detectable to all sensors simultaneously, which would be infeasible to know in practice. Instead, our method leverages contemporaneous structure to keep all detectors up-to-date with their collective knowledge. Also, we assume all sensors are identical spheres. We ignore that sensor velocity and other factors affect source intensity, assuming source exposure is only a function of distance.

3.3.2 Technical details

In the remainder of this section, we describe the technical details of BAF, whose pseudocode is given by Alg. 1. In this pseudocode, $x_{i:j} = \{x_i, \ldots, x_j\}$ and $x_{i:j,m:n} = \{x_{u,v}\}$ where $u = 1, \ldots, j$ and $v = m, \ldots, n$. According to Eqn. 3.1, if luminosity

Algorithm 1 Bayesian Aggregation Filter

```
procedure BAF(source template $s$, sequence of spectra $\{Y_{t,1:K}\}_{t=1}^{\infty}$ from $K$ sensors, coordinates of sensor paths)
    $H_0 \leftarrow 0$ luminosity source at arbitrary coordinates
    for $t = 1, \ldots$ do
        for $k = 1, \ldots, K$ do
            Compute expected intensities $\bar{I}_{1:t,k}$ at sensor $k$ from $H_{t-1}$
            Estimate intensities $\gamma_{1:t,k}$ with Kalman filter on $\{Y_{1:t,k}, \bar{I}_{1:t,k}\}$
        Compute $H_t$ using BA on $\gamma_{1:t,1:K}$
```

and location of a source is presumed known, we can compute the expected source intensity at a sensor at time t as $I(r_t)$, where r_t is the distance between source and sensor. Now we explain the first change to BA (1). To get a distribution over a KF estimate of intensity $\hat{I}_t$, we use a normal distribution with mean $\mu_t = I(r_t)$ and variance $\sigma_t^2 = \widehat{\mathrm{Var}}(I_t)$ from the KF estimate of state variance. Since luminosity and location are given by each hypothesis, the score probabilities can be computed in this manner. The normal distribution is selected to match the KF assumption of normality.

Next we explain the second change to BA (2). At time t, after BA selects a source hypothesis H_{t-1}, the expected intensity to each sensor is computed. It can be shown that a sensor a distance of r away from a source observes an intensity $I(r)$ of approximately (Miller et al., 2016; Tandon, 2016)

$$I(r) \propto 1/r^2. \tag{3.1}$$

However, since the hypothesis can change at every iteration, we compute not just intensity at t but at every step τ so far, $\{I(r_\tau)\}_{\tau=1}^{t}$. These values are incorporated as an observed variable in the KF, which is run over all observations $\tau = 1, \ldots, t$ at each sensor. The observation model simply states that the observed intensity ought to match the estimated intensity in expectation. Mathematically,

$$I(r_\tau) = \hat{I}_\tau + \epsilon_\tau$$

where $\epsilon_\tau \sim N(0, \sigma^2)$ and $\sigma^2 = 100$ is a hyperparameter that we set by hand to approximately match the scale of the variance of $\hat{I}_\tau$. In our experiments, this hyperparameter does not impact the method much unless multiple orders of magnitude away from 100. Separately, since the estimates $\hat{I}_\tau$ change at every step t, we run BA from scratch on the new estimates. In total, both KF and BA is restarted at each step, which results in $O(T^2)$ runtime, where T is the total number of steps in the pass. This runtime is a significant slowdown from all other methods, which are all $O(T)$ to our knowledge.

A final design choice is the technique to compute the final score that aggregates all passes. In BA, this score is typically the maximum likelihood ratio over hypotheses. In BAF, however, our experiments showed that this approach performed no better than BA. Therefore, we choose a different technique: we take the maximum over each KF detector's final estimated intensities; then we average the top two values. In our experiments this approach performed best out of many other aggregation techniques such as mean, median, maximum, and minimum over sensors. Intuitively, there is a tradeoff to using using values from more sensors. One one hand, it decreases the impact of outliers, but on the other hand, it increases the impact of ordinary stochastic variance in the KF intensity, especially in the sensors farther away because they do not observe as high of intensities in the first place. Thus, the choice of two as the number of sensors used may represent a point of parity between the two effects.

3.4 Experiments

3.4.1 Procedure

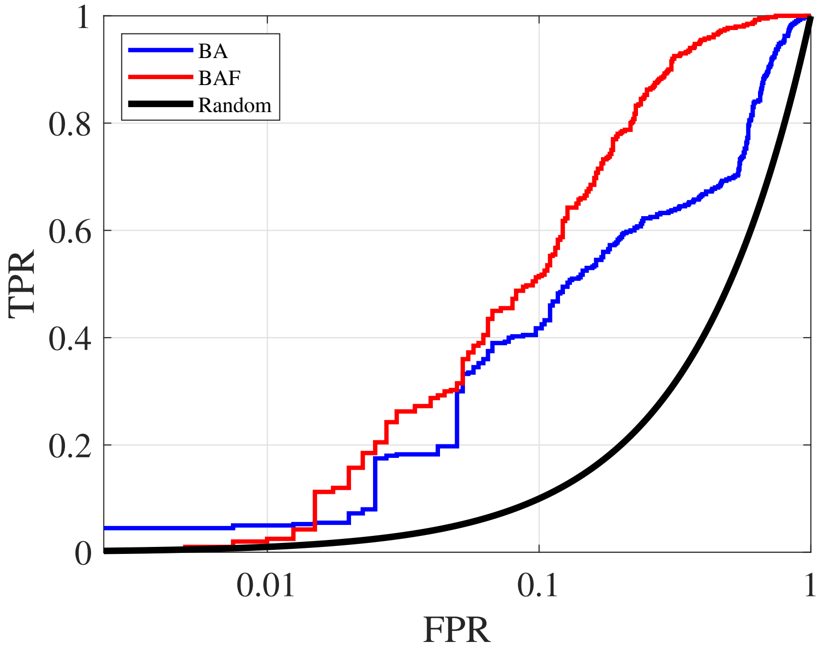
We test our methods on the same dataset as in the second experiment of Ch. 2, which contains spectroscopic data from a major metropolitan area recorded by vehicle-mounted NaI detectors. We use a subset of 6.6 hours with an integration window of one second, yielding about 24,000 observations. Our procedure is identical to Ch. 2 except for some changes. Since simultaneous nearby passes by multiple sensors are rare, we simulate these scenarios by multiple disjoint passes of the same sensor at different times. This simulation only lacks exact fidelity because of temporal variation of background, which should not be too impactful since passes are separated by hours at most. KF is ran without any warm-up period. We vary the number M of passes between 1 and 4 and constrain all passes to come within at least 20 meters of the randomly placed source. Then BA is trained on all data outside the passes, half for each training set. The outside data are also used to compute mean gross counts, which is given to BP. In all methods the hypothesis space was spatially restricted to the intersection of a 100 meter envelope around each sensor path. The source luminosity is computed by setting the maximum source intensity at any sensor to 84 cps, which corresponds to SNR of 4. Each pass contains three minutes before passing the source and three minutes after. In addition, we selected a source template that corresponds to Cobalt-57, a medical and industrial isotope, although simple experiments showed the choice of template did not affect results significantly. Lastly, all hyperparameter tuning was conducted on a separate fold of the data that was entirely disjoint in time and almost disjoint in space.

3.4.2 Results

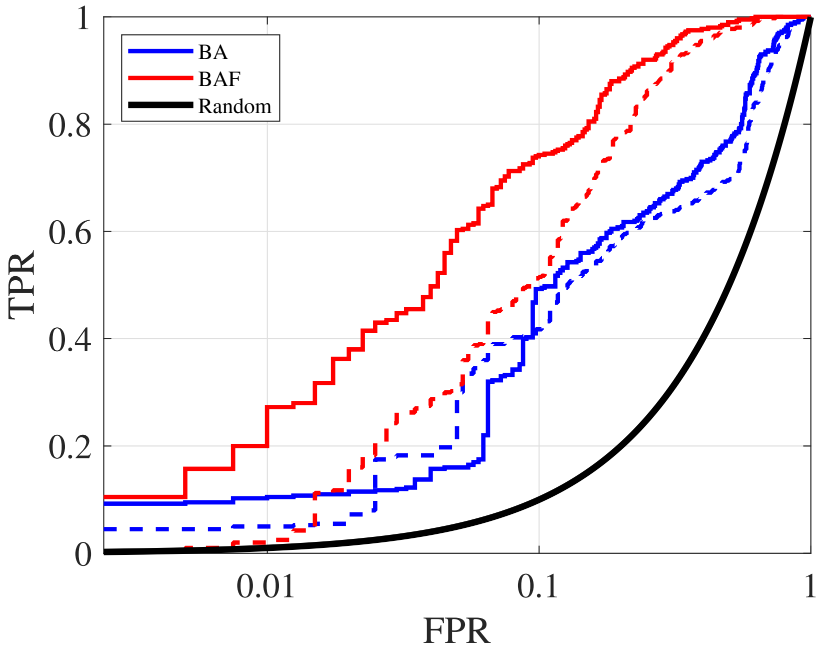
We compared BAF to BA, using a classical Matched Filter (Turin, 1960) as BA's detector. ROCs are displayed in Fig. 3.2. Also, certain points on the ROC are illustrated in Fig. 3.3, in particular TPR at 1% FPR and FPR at 50% TPR. Overall the results demonstrate that BAF has comparable performance to BA at one sensor but enjoys a significant advantage with multiple sensors. Also, BP exhibits close to nil detection power. Furthermore, BAF performance hugely improves with more than one sensor but does not differ between values above one. In contrast, BA's TPR at 1% FPR consistently improves with the number of sensors, although the same does not hold for the FPR at 50% TPR.

The relative performance of BA and BAF matches our expectations. First, they perform comparably with only one sensor. This result makes sense because no multi-view structure exists for BAF to exploit. Accordingly, BAF performs much better than BA with more sensors. This difference reflects the advantage BAF creates by disseminating collective inferences to each sensor at every step. It is less intuitive that BAF does not improve with more than two sensors. A plausible explanation is that the third and fourth sensors observe intensities that are too low to be detectable within background noise. Thus, no detection power is added. In fact, the background noise may be worsening performance. This result suggests that it can be detrimental to add views to a multi-view approach if their SNR is too low. Unlike BAF, however, BA marginally improved in TPR at low FPR with more sensors. This trend may indicate that BA is more robust at finding signal in highly noisy observations.

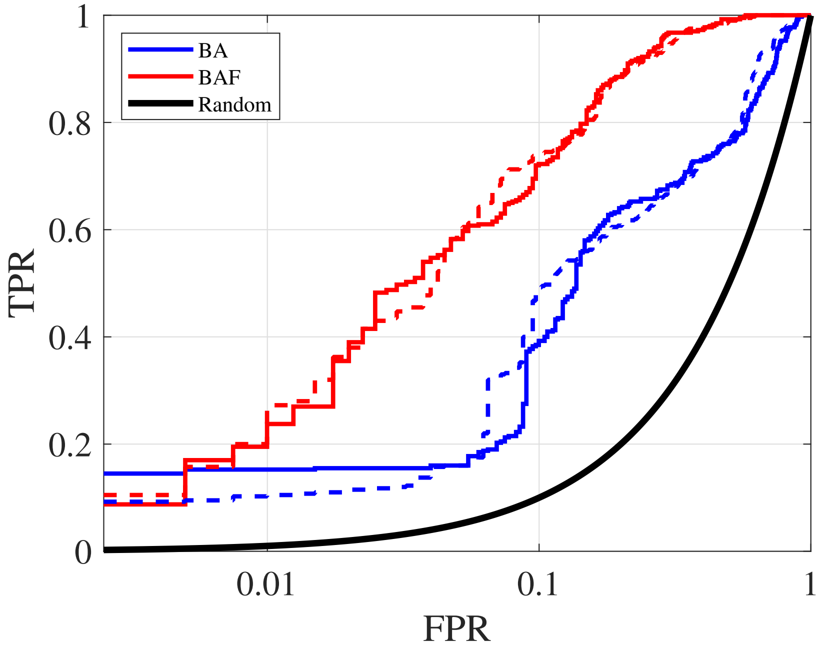
Additionally, we present an example of estimated intensities at each sensor from a single trial in Fig. 3.4. The figures shows collective inferences in both injected and background runs from BAF as well as individual inferences in a background



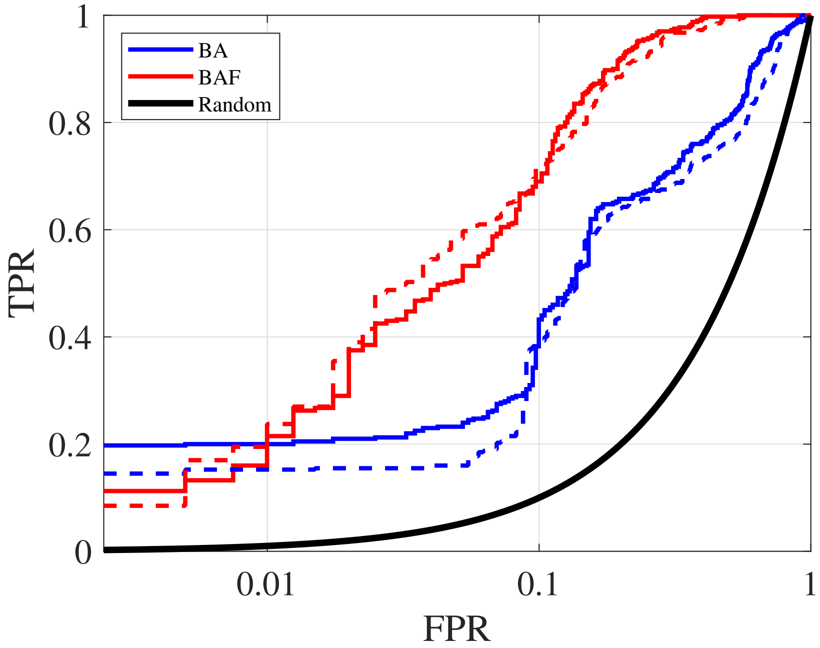
(a) ROC with 1 sensor.



(b) ROC with 2 sensors. Dashed lines indicate results from previous number of sensors.

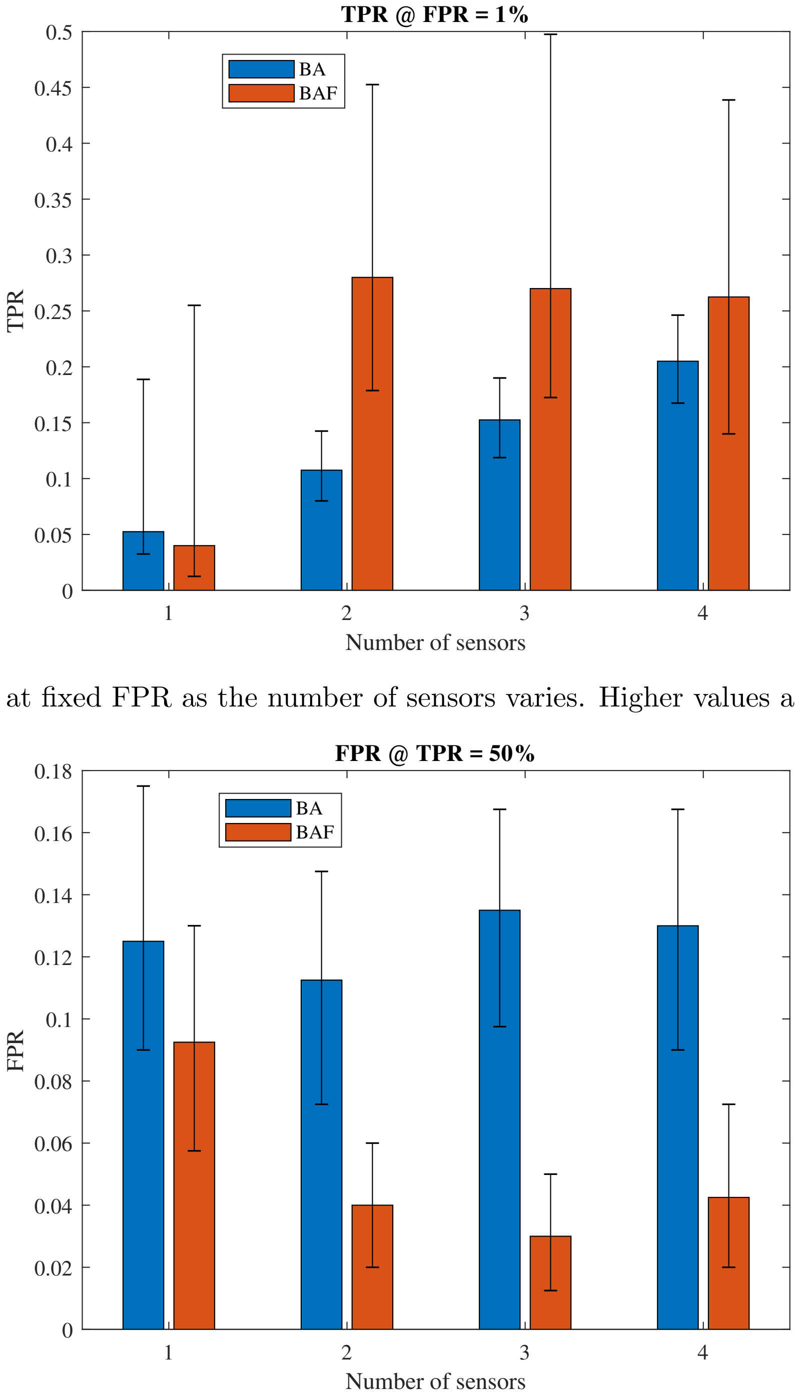


(c) ROC with 3 sensors. Dashed lines indicate results from previous number of sensors.



(d) ROC with 4 sensors. Dashed lines indicate results from previous number of sensors.

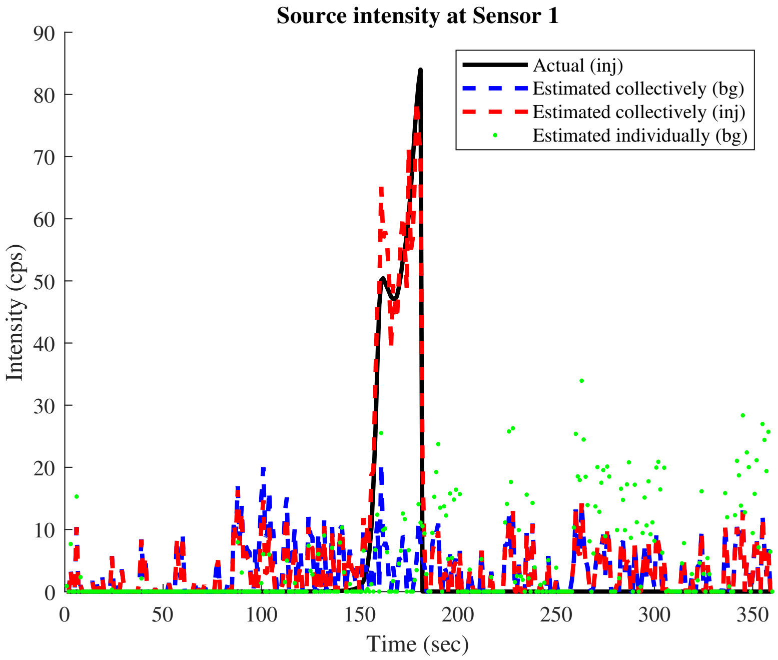
Figure 3.2: ROCs with multiple sensors.



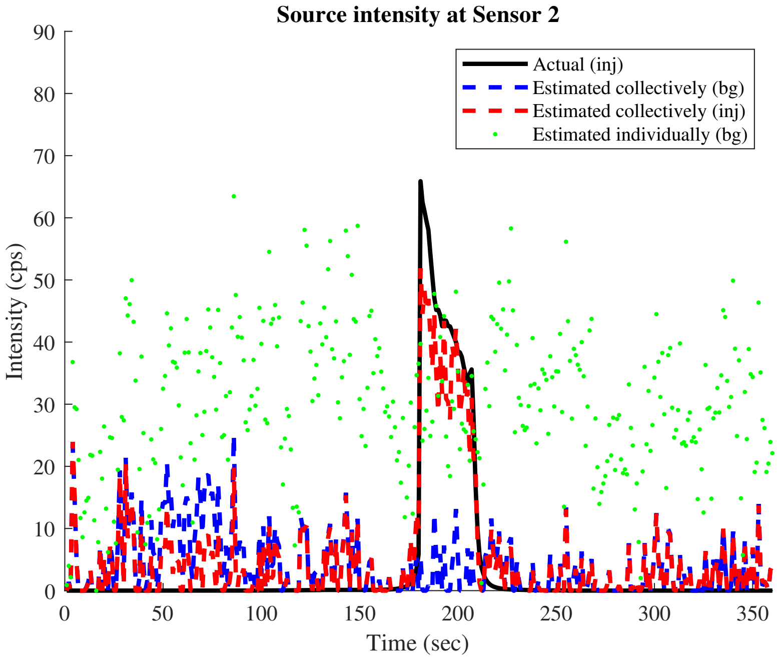
(a) TPR at fixed FPR as the number of sensors varies. Higher values are desirable.

(b) FPR at fixed TPR as the number of sensors varies. Lower values are desirable.

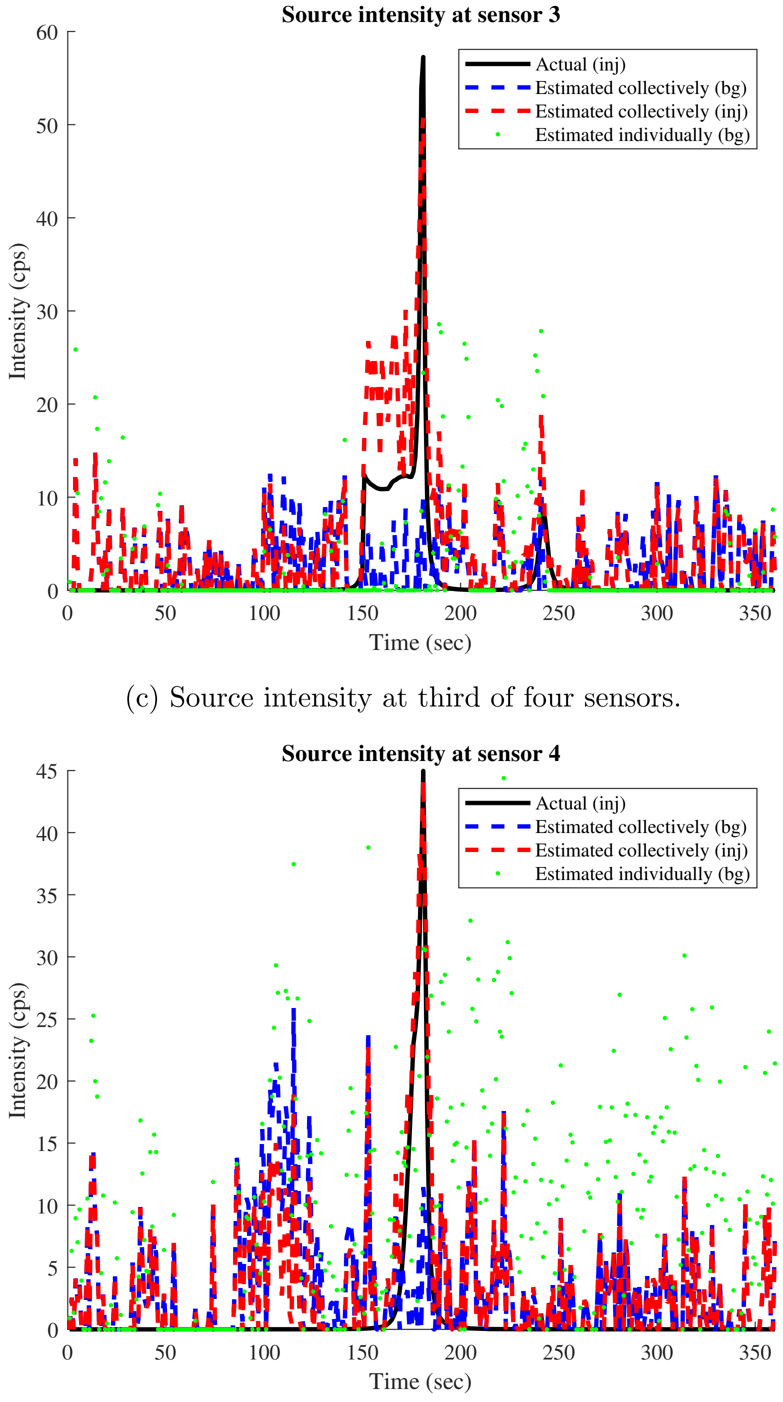
Figure 3.3: Statistics from ROCs as the number of sensors varies. Error bars show 95% confidence intervals from bootstrapped sensor passes.



(a) Source intensity at first of four sensors.



(b) Source intensity at second of four sensors.



(c) Source intensity at third of four sensors.

(d) Source intensity at fourth of four sensors.

Figure 3.4: Source intensities at different sensors, comparing collective to individual inferences.

run from standard single-view KF. The goal is to illustrate how collective inference greatly enhances estimates by suppressing noise. In Figs. 3.4b and 3.4d, for example, individual estimates in a background run reach values as high as the actual injected intensity. These errors result from the uncertainty in KF. Yet when the estimates are made with collective information, these false peaks decrease significantly or vanish altogether. This improvement results from the hypothesis selected by BA using inferences from all sensors, which informs each KF of what intensity to expect.

3.5 Conclusion

We proposed a novel method for gamma source detection by aggregating observations from multiple sensors. The method has two principal advantages over the state of the art, Bayesian Aggregation (BA). First, it does not require any training data. Second, it leverages contemporaneous multi-view structure between sensors to enhance the quality of predictions. Empirically it performs as well as or better than BA, a notable feat given the lack of training. The method employs individual Kalman filters for each sensor, which are then synthesized by BA to track location and strength of the source. It sends its inferences back to the individual sensors to enhance their predictions in a feedback loop. This method illustrates how multi-view relationships can be explicitly modeled by structural information to inform inference.

An interesting direction of future work is to utilize not only the intensity from the Kalman filter but also the background rates. We propose to refine each sensor's background estimates by smoothing across sensors using spatial information. More precisely, we could use a kernel smoother on background estimates from every sensor at each time step. The smoothed background would be incorporated in the observed variables of the Kalman filter in the same manner as the refined source intensity from BA. By leveraging the spatial smoothness of background, this approach would augment the method with an orthogonal multi-view relationship to what we use here.

Part II

Learning multi-view relationships

Chapter 4

Robust Gamma Source Detection with Incomplete Source Information and Gain Drift

4.1 Introduction

This part of the thesis investigates how to learn multi-view relationships when they are unknown, in contrast to the previous part. This particular chapter assumes the relationships are linear and demonstrates how to use them for prediction. We present a method that utilizes linear relationships to perform anomaly detection in the gamma source detection problem.

Many effective statistical methods, including Spectral Anomaly Detection, Censored Energy Window, and Matched Filter (Tandon, 2016), rely on reliable knowledge about spectral shapes of target sources and/or background spectra and their variability patterns. There has been little investigation on effects of error in this information, which may be flawed due to differences in sensor hardware used to collect training data and at time of deployment (e.g. miscalibration or gain drift) or uncertainty in source composition or shielding. Here we explore the effects of low quality information on existing methods and propose a new robust alternative.

The proposed method extends the Censored Energy Window approach using Canonical Correlation Analysis (Hotelling, 1936). We simulate incomplete and inaccurate information by removing a target source from a known source library that is utilized by the method, thereby simulating a situation in which a previously unknown target source is encountered. The method is applied to an authentic radiation dataset and demonstrated to significantly improve performance over alternatives with incomplete or inaccurate source information. Additionally, we evaluate the impact of imperfect information due to miscalibration (gain drift) on different methods. Our method is shown to outperform the baseline methods. In general, these results suggest that our extension of the Censored Energy Window method is robust to distortions in spectral information. For a summary of related work, please see Ch. 2.3.

The proposed method extends the known Censored Energy Window approach using Canonical Correlation Analysis (Hotelling, 1936). To validate our method empirically, we emulate incomplete and inaccurate information by removing a target source from a library of known source types that is utilized by the method, which models a situation in which a target source of a previously unknown design is encountered. The method is applied to an authentic spectroscopic dataset and demonstrated to significantly improve performance over alternatives in scenarios involving incomplete or inaccurate source information. Additionally, we evaluate the impact of imperfect information due to miscalibration (gain drift) on different reference methods. Our approach is shown to outperform these baselines. In general, our results suggest that the proposed extension of the Censored Energy Window approach can be robust to distortions in spectral information.

4.1.1 Censored Energy Window

Introduced by (Nelson and Labov, 2012), Censored Energy Window (CEW) assumes partial knowledge of the sought-after source spectrum. It takes as input a set of (not necessarily contiguous) energy bins, referred to as the energy window, in which the target source is expected to be seen most clearly. If the source spectrum is known, this corresponds to the signal-to-noise ratio-maximizing energy window for the given background radiation variation pattern. CEW learns a linear relationship between out-of-window energy bins and the total in-window photon count, in the absence of a source. If the in-window counts significantly exceed the predicted value, then a source is likely present. If the source and background spectra are known, the optimal energy window can be computed by Algorithm 2 in App. C (Miller et al., 2016). In practice, if the source spectrum is known, the mean background spectrum learned from training data is often used to determine the window.

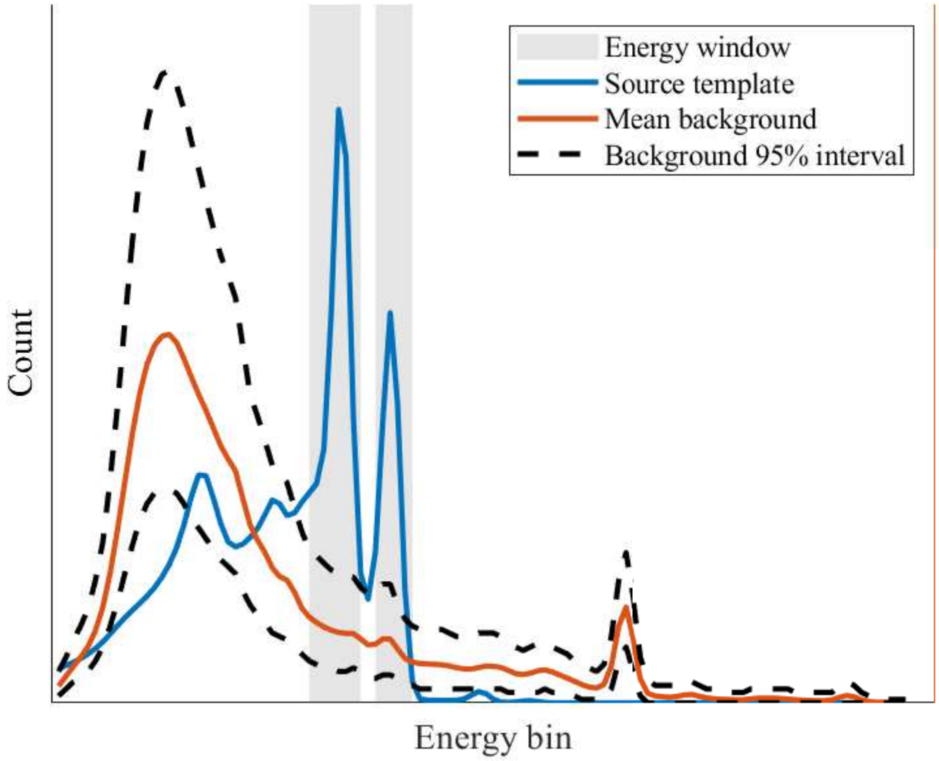


Figure 4.1: Example of an energy window and mean background. The blue line shows a source template, and the shaded gray area is the (non-contiguous) energy window found for it to maximize the SNR vs. mean background template. The orange line shows the mean background on a different scale. The dashed black lines show the empirical 95% interval about the median.

A multiple linear regression model is then fit to predict the sum of counts inside the window from the vector of counts outside of it. The source detection score is then defined as a difference between the predicted and the actually observed particle count in the window. Figure 4.2 illustrates the processing flow of the CEW method. The linear regression step may use ridge regression with a small regularization parameter to alleviate overfitting and collinearity. As a last step not shown in Fig. 4.2, the in-window sum of counts $\hat{y}$ is normalized. Since the score quantity can be modeled as a Poisson random variable, the Anscombe transform (Anscombe, 1948) $A(x) = 2\sqrt{x + 3/8}$ can be used to transform the in-window counts to an (approximately) normally distributed random variable with unit variance. This stabilizes the variance

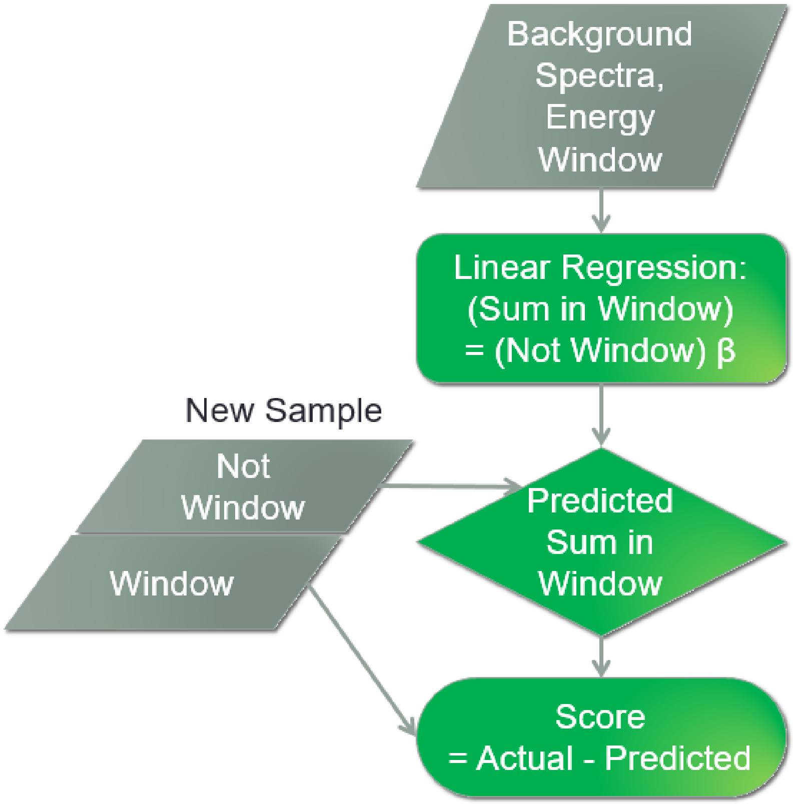

Background Spectra, Energy Window
Linear Regression: (Sum in Window) = (Not Window) β
New Sample
Not Window
Window
Predicted Sum in Window
Score = Actual - Predicted

Figure 4.2: The CEW detection algorithm. The sum of in-window counts is predicted from the out-of-window counts via linear regression. The prediction error is the score. (Normalization step not shown.)

of the photon counts, especially if they are low. The final score is a (function of) $A(\hat{y}) - A(y)$.

4.2 Robust method for incomplete information

When methods like CEW and Matched Filter (MF) are given accurate information about target sources, they can perform well. In practice, however, the source spectrum could be uncertain due to different configurations of material, shielding, and environmental factors (Tandon, 2016). These methods can then underperform when their input information is incomplete or inaccurate. In CEW for example, the energy window may be misspecified and the observed in-window counts might not show a sufficiently strong signal when a source is present to facilitate its detection. To remedy these effects, we propose using a Canonical Correlation Analysis (CCA)

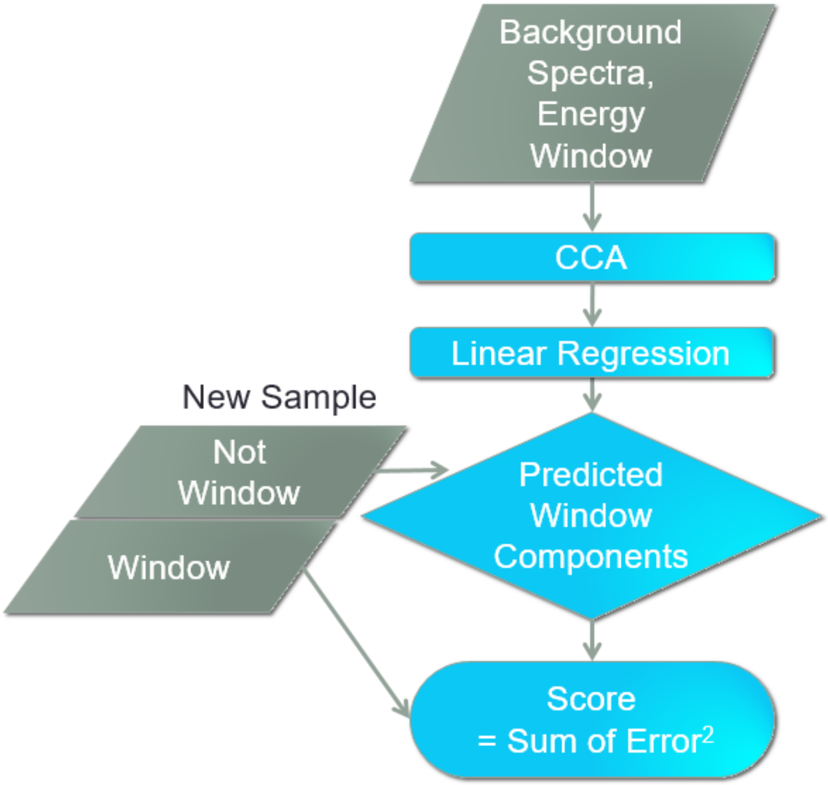


Figure 4.3: The CCA detection algorithm. The canonical variates from in-window counts are predicted from the canonical variates from out-of-window counts via linear regression. The sum of squared prediction errors is the score. (Normalization step not shown.)

based detection. This method discovers a source in the presence of background noise while tolerating imperfect knowledge of the source template. It could be useful in practical applications when the characteristics of the sought-after sources are not precisely known a priori.

4.2.1 Background: Canonical correlation analysis

In this section we summarize canonical correlation analysis (CCA), a useful starting point for understanding the proposed methods. CCA analyzes cross-covariance between two sets of variables that have aligned observations. By performing CCA, one can understand how much variance in the sets can be explained by common factors. It finds linear projections from each view into a shared latent space such that the projections have maximal correlation. According to Bach and Jordan (2006), the canonical, or latent, variables can be considered the basis of a generative Gaussian model for the observed views. These canonical variables often have some practical meaning, such as a certain combination of genes that corresponds to a combination of phenotypes Witten and Tibshirani (2009). CCA, or more generally component analysis, can therefore be used to analyze complex datasets in an interpretable fashion.

We mathematically define CCA. Let $X \in \mathbb{R}^{d_X}$ and $Y \in \mathbb{R}^{d_Y}$ be random vectors. Without loss of generality, assume $\mathrm{E}[X] = \mathrm{E}[Y] = 0$. Then CCA for the m-th component solves the problem

$$\begin{aligned} \max_{u \in \mathbb{R}^{d_X}, v \in \mathbb{R}^{d_Y}} \quad & \mathrm{Corr}(X^\top u, Y^\top v) \\ \text{subject to} \quad & \mathrm{Cov}(Xu, Xu_i) = \mathrm{Cov}(Yv, Yv_i) = 0, \\ & i = 1, \ldots, m-1. \end{aligned} \tag{4.1}$$

Define $\Sigma_{XY} = \mathrm{Cov}(X, Y)$, $\Sigma_{XX} = \mathrm{Cov}(X)$, and $\Sigma_{YY} = \mathrm{Cov}(Y)$. This optimization is non-convex, but it has a closed-form solution Hardoon et al. (2004): u and v are the respective m-th largest eigenvectors of

$$A = \Sigma_{XX}^{-1} \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{XY}^\top,$$

$$B = \Sigma_{YY}^{-1} \Sigma_{XY}^\top \Sigma_{XX}^{-1} \Sigma_{XY}.$$

4.2.2 Robust Censored Energy Window via CCA detection

Our method takes as input a collection of background measurements and an energy window. This window may be taken as a set of (not necessarily consecutive) energy bins that collectively maximize the signal-to-noise ratio (SNR) for a particular source template of interest, as in Algorithm 2, but it can be arbitrary, and we vary its alignment with the properties of the target source spectra in our experiments. The first step is to apply CCA to the training set composed of reference background spectra measurements. Here, X and Y correspond to photon counts in the spectrum bins inside and outside of the chosen energy window respectively. Next, for each pair (u, v) of weights found, we fit a simple linear regression of $X^\top u$ on $Y^\top v$. We compute the residuals for all samples and fit a univariate normal distribution to them. Then given a new sample to classify, we compute the regression residuals and find their z-scores for each pair (u, v). The sign of each score is not meaningful, unlike CEW in which only an elevated observation suggests presence of a source. Thus, we square the z-scores and compute the sum of squares as the final source detection score. In the same way that CEW finds a single-output linear relationship between photon counts inside and outside the energy window, CCA identifies and concurrently leverages multiple arbitrary multiple-to-multiple linear relationships between bins of measured spectra. Figure 4.3 illustrates the information processing flow of the method.

4.3 Experiments

We demonstrate the utility of the CCA-based detection algorithm using authentic field data in which imperfect information is simulated in three experiments. First, we corrupt a predefined source spectrum template. Second, we remove the true template from a library of known sources. Third, we corrupt measurements by inducing miscalibration of the detector.

4.3.1 Data

Our data consists of over 86,000 gamma ray measurements recorded using a one-second integration time using a sodium-iodide detector mounted on a vehicle moving around an urban area in Baltimore, MD, USA. On average, it measured 2,600 background photon counts per second. Each raw measurement was projected onto $d = 116$ linearly spaced energy bins. The resulting dataset was partitioned into training and test subsets with the first 60% as training. Along with the background measurements, we used a library of 67 source templates corresponding to high-fidelity source simulations applying a range of configurations of radioactive material and shielding (Nelson and Labov, 2012). The templates were normalized to each produce 100 photons from the source per second. Each template was independently sampled for every positive test set observation. These simulated source signatures were added to the field observations to form synthetic positive measurements. This process resulted in a labeled set of negative (no source injected) and positive (source signatures injected) data, separately for each source template in the library.

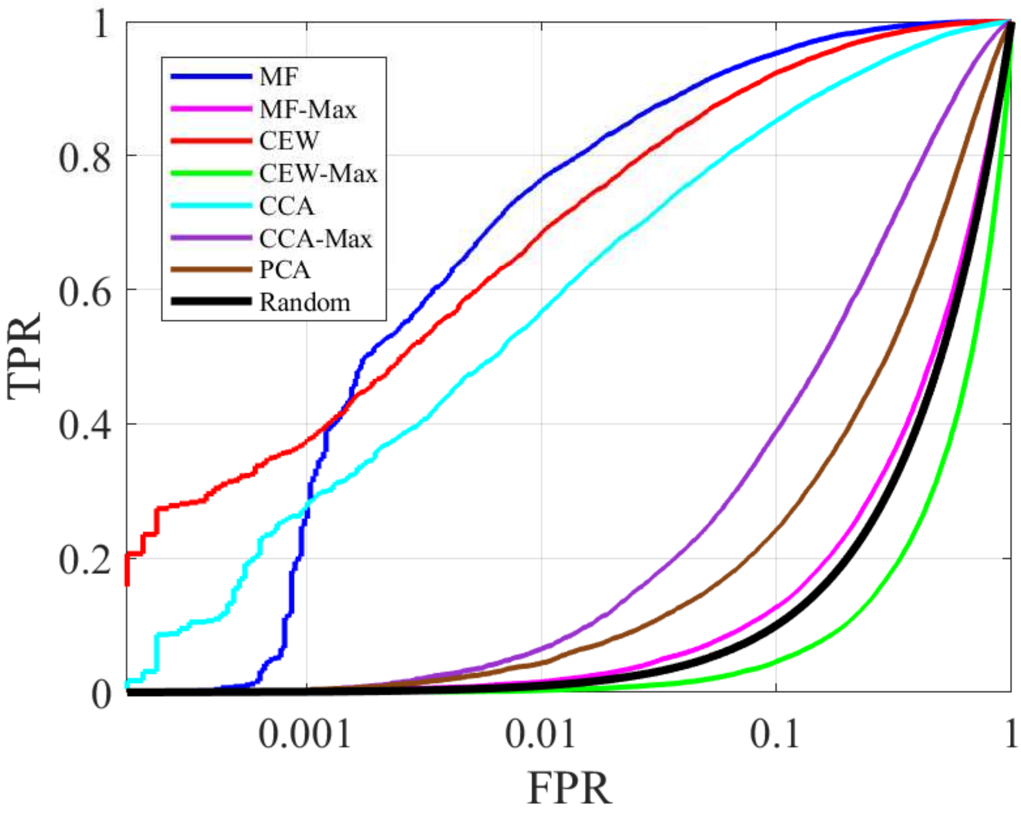


Figure 4.4: Example of methods' performance on a single source template. MF (blue), CEW (red), and CCA (cyan) perform best when the source template is known. When it is unknown, CCA-Max (purple) performs better than MF-Max (magenta), CEW-Max (green), and PCA (brown).

4.3.2 Incomplete source information

The task was binary classification of each observation as either containing a source or not. We compared CCA detection with MF and CEW given perfect source information. These methods benefit from strictly more information and thus approximate an upper-bound on expected detection performance. We also compared to the source-unaware PCA method. Additionally, we let MF, CEW, and CCA address imperfect information by marginalizing over the source library and taking the maximum score (CEW-Max, MF-Max, and CCA-Max).

Imperfect energy window

In our first experiment, we directly measured the impact of energy window quality. Note that among methods we consider, only CEW and CCA use energy windows. We quantify detection performance of CEW-Max and CCA as the energy window changed from an accurate, source-specific window to a common window that maximizes average SNR across all considered types of sources in the library, but not specifically tailored to any of them. To emulate this, a convex combination was taken between the target source source template and the average source template with alternative weights of 0, 0.25, 0.5, 0.75, and 1, where 0 corresponds to the simple average and 1 to the true target. The energy window was computed for the interpolated spectrum. Detection performance was measured by the false positive rate (FPR) at a fixed true positive rate (TPR) of 50%, and TPR at a fixed FPR of 1%. The results, were averaged across all sources in the library, taking each as the target in turn. These average detection performances are displayed in Figures 4.5b and 4.5a, along with the performance of the source-optimal MF and PCA-based SAD. With fully source-specific windows, CCA performed slightly but insignificantly worse on average than CEW. With a common, source-type-nonspecific window, CCA enjoyed large, statistically significant advantages in both metrics, boosting resistance to false positives at 50% detection probability by a substantial margin.

Unknown source template

In our second experiment, we simulated a lack of information. We removed a group of mutually similar sources from the library and used one of them as the target for detection. To do so, we clustered the source template library using K-means algorithm with $K = 10$. For MF-Max, CEW-Max, and CCA, we tested them for each source template by removing its cluster from the dataset and marginalizing over the remainder. Figure 4.4 shows an example ROC obtained for the considered methods using a source template including Cobalt-57, an industrial isotope. The distribution of FPR and TPR over source templates is displayed in Figure 4.6. By TPR (Figure 4.6a), it is difficult to distinguish between PCA, CCA, and MF-Max. By FPR (Figure 4.6b), however, the CCA method usually performs better than MF-Max, CEW-Max, and PCA.

Gain drift

In our third experiment, we simulate miscalibration of the sensor by gain drift in test data. The gain drift coefficient was varied between 0 and 0.1 by increments of 0.01, resulting in transformed spectra such as those depicted in Figure 2.5. This range was selected as an approximation of drift that might be encountered in practice. To simulate gain drift, we employed an algorithm that effectively transforms a spectrum to a different definition of energy bins. Given photon counts data, the energy frequencies of the current bin boundaries, and the energy frequencies of the new bin boundaries, the algorithm adapts the counts data to the new bins. When the new boundaries come in-between the old ones, counts are linearly interpolated. In this experiment, the CCA method was compared to MF (non-marginalized version), CEW, and PCA. The TPR and FPR metrics of the methods observed at each coefficient, averaged over sources, are displayed in Figure 4.8. While CCA was not initially the best, it obtained an advantage that became more pronounced as gain drift increased, especially in terms of lower false detection rates. In Figure 4.9, we show the performance of the methods at 0.1 gain drift along with paired t-test statistics demonstrating that the CCA method can be superior with statistical significance when facing sensor miscalibration issues.

In the remainder of this section, we discuss the relationship between CEW and CCA. Fig. 4.10 displays the similarity between CEW scores and each component of CCA scores at 10% gain drift. The first component is almost perfectly correlated to CEW, while the others differ substantially. This result suggests CCA finds multiple uncorrelated relationships and is therefore more robust to disturbances in any particular one.

In Fig. 4.11, we illustrate the hypothesis that CCA performs better than CEW because it utilizes multiple relationships. The figure shows the observed spectra projected onto four pairs of canonical component projections sorted by the decreasing order of their correlation with mean background spectrum. The blue contours show the kernel density estimated distribution of the background data on which CCA was learned. The trendlines indicate the linear relationship between the views in each canonical projection. In green and red we highlight a small set of test spectra. The green are background while the red have the same background but have added source injections. As expected, the green points generally fit the trendlines, while the red are farther away. Longer lengths of the residual to the line contribute more to the

CCA detection score, so these plots show how source template injection results in higher scores. Consider, for example, the observation marked by a triangle. In the first component, the residual actually shrinks, lowering the contribution to score. In the other components, the residual grows significantly as expected. This example demonstrates how different components are more sensitive to injection. In particular, the first component, which is almost the same as CEW, contains minimal signal. On the other hand, the change in score can be seen much more clearly in the other components. This result explains why multiple components offer benefits over CEW.

Lastly, in Fig. 4.12a and 4.12b, we show detection performance of CEW and CCA with related methods at 10% gain drift. The leftmost method is Gross Counts (GC), the sum of counts in all bins. The next is Energy Window Gross Counts (EW-GC), the sum of counts in the energy window. The next is Univariate CEW (Uni-CEW), CEW in which the out-of-window counts are summed and treated like a single bin. Altogether, these five methods form a spectrum that operates on accumulated counts in different ranges. On the left is the simplest, which uses only one weighting of bins—uniform—and on the right is the most complex, which uses multiple learned weightings in two groups to capture maximal structure in the background spectra. This spectrum of complexity largely aligns with detection performance.

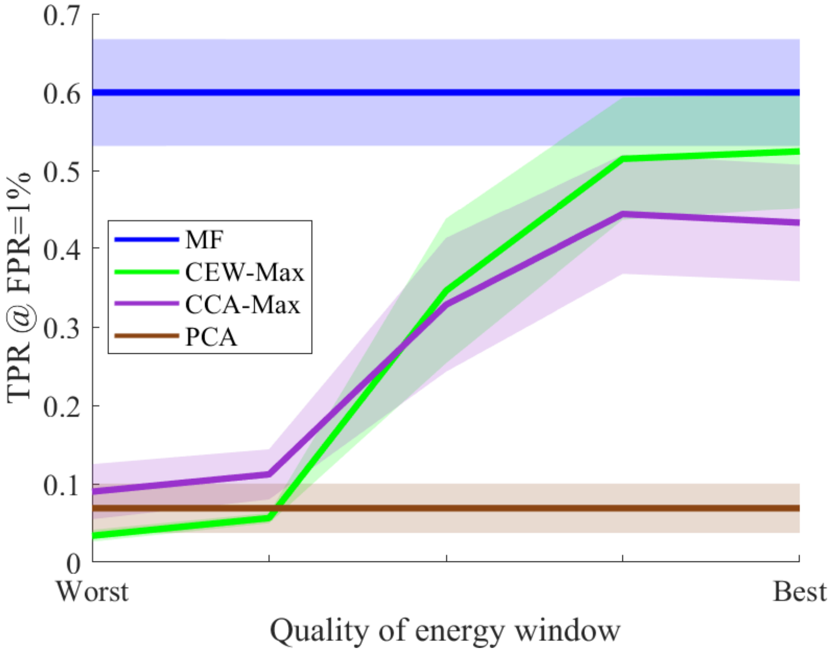
4.4 Conclusion

We have introduced a new method, CCA detection, for processing gamma-ray spectral measurements and applied it to the task of detecting gamma-ray sources in noisy background radiation environments. This new method is intended to enhance practical utility of radiation detection systems by making them more robust to various imperfections of the application settings, such as limited knowledge of the sought-after source designs, or sensor miscalibration, that frequently undermine the performance of radiation threat detection systems in practical application scenarios.

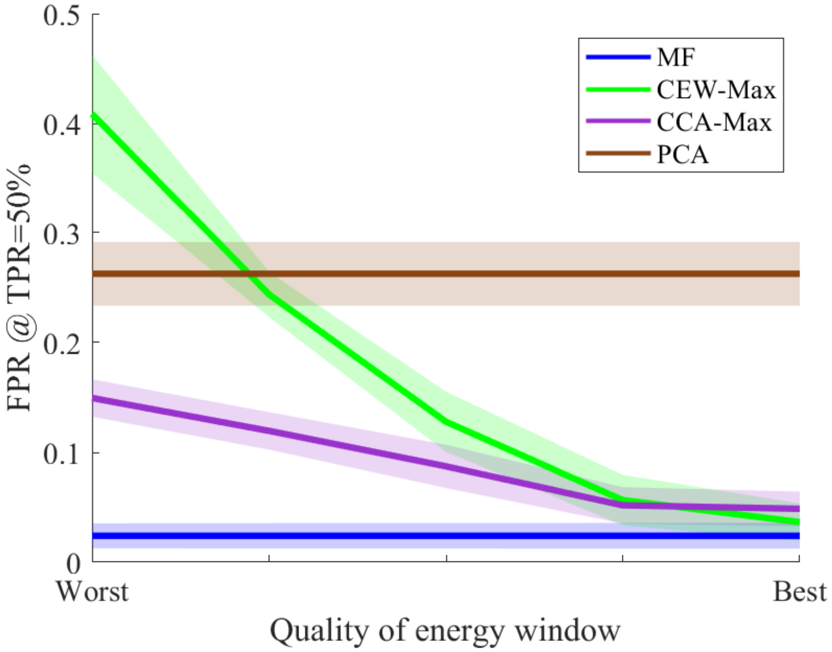
Our experiments demonstrate that CCA detection can be more robust to poor knowledge of the target source template in the task of gamma source detection. When the source is unknown to the reference library, CCA detection usually performs better than relevant alternatives including PCA-based Spectral Anomaly Detection, CEW-Max, and MF-Max. Likewise, when the energy window is suboptimally computed—corresponding to a situation in which the target source template is unknown or inaccurately represented—CCA detection performs better than CEW-Max by a significant margin. We hypothesize that this robustness advantage over the CEW method stems from the fact that whereas CEW finds one relationship between in-window and out-of-window energy bins, CCA detection finds multiple such relationships, which when considered jointly, reduce the impact of one or a few of them becoming uninformative when the window definition becomes less accurate. Intuitively, by taking a simple sum of photon counts in the window, CEW only utilizes a fraction of the spectral information available. By comparison, CCA detection utilizes much more information by finding more than one relationship. Furthermore, a plausible explanation for why CCA can be more robust than MF is that it simply allows weaker knowledge of the source—an energy window instead of the exact template. Intriguingly, CCA also performs better than the completely source independent PCA detection, even if the knowledge about the source is imperfect. In contrast, MF and CEW, which depend much more on exact source information, tend to be worse than PCA in these situations.

Thus, CCA detection may represent a superior trade-off between robustness to imperfect source information and capacity to leverage that information, and as such it can serve as a universal tool whose applicability spans the full range of problem configurations from the perfect knowledge of the targeted source design, to the complete ignorance about such design. The same reasons can also help explain why the CCA method can perform better in the presence of sensor gain drift.

The work in this chapter showcases the utility of learning linear multi-view relationships as discriminative factors in classification. The next step is to extend to nonlinear relationships.

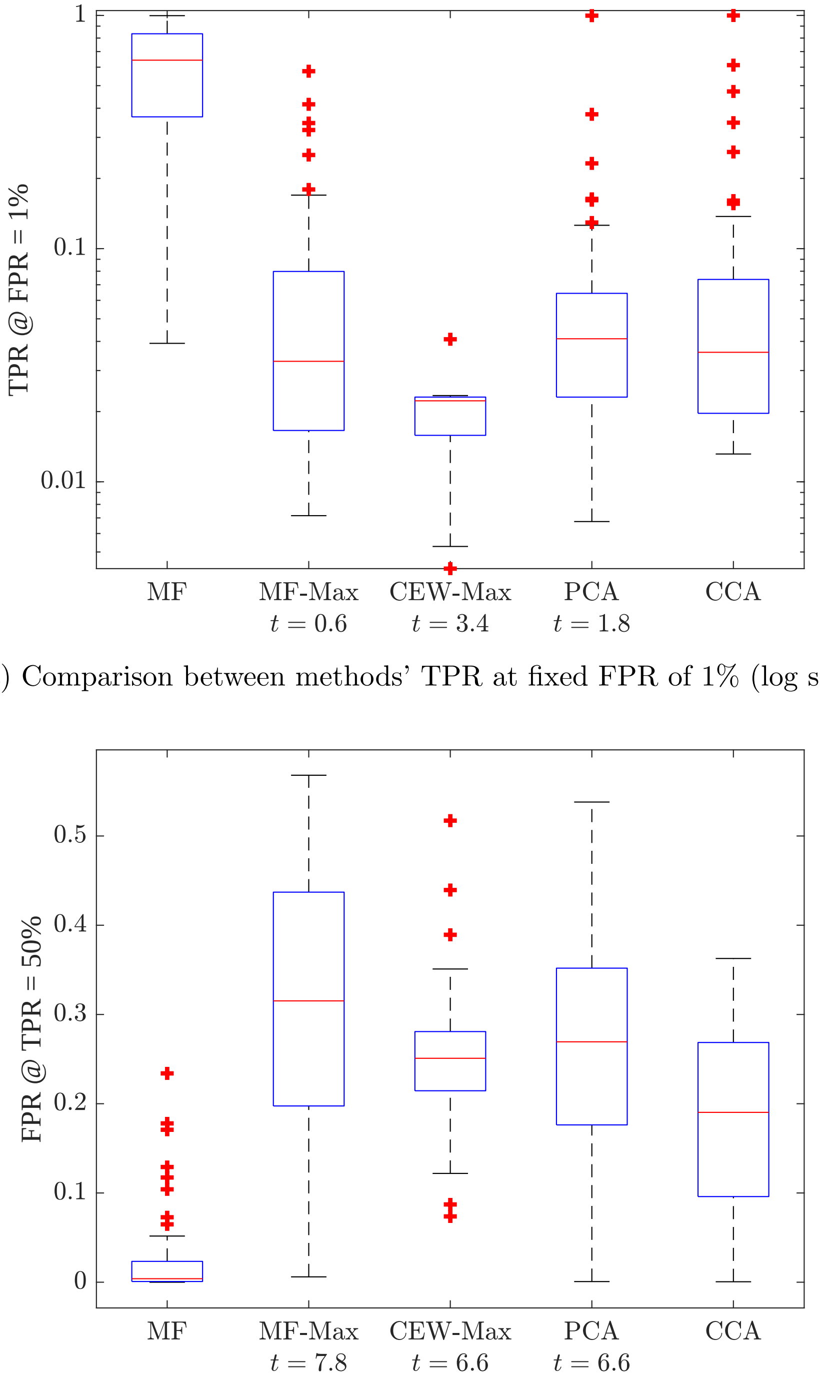


(a) Comparison between methods' TPR at fixed FPR of 1%.



(b) Comparison between methods' FPR at fixed TPR of 50%.

Figure 4.5: Comparison between methods as energy windows change from common to source-specific. Values are averaged over all available source templates, and bands correspond to two standard errors on each side.



(a) Comparison between methods' TPR at fixed FPR of 1% (log scale).

(b) Comparison between methods' FPR at fixed TPR of 50%.

Figure 4.6: Comparison between methods when target source's cluster is missing from the library. Box plots show distribution over source templates. MF, which knows the true target source, is included as a reference.

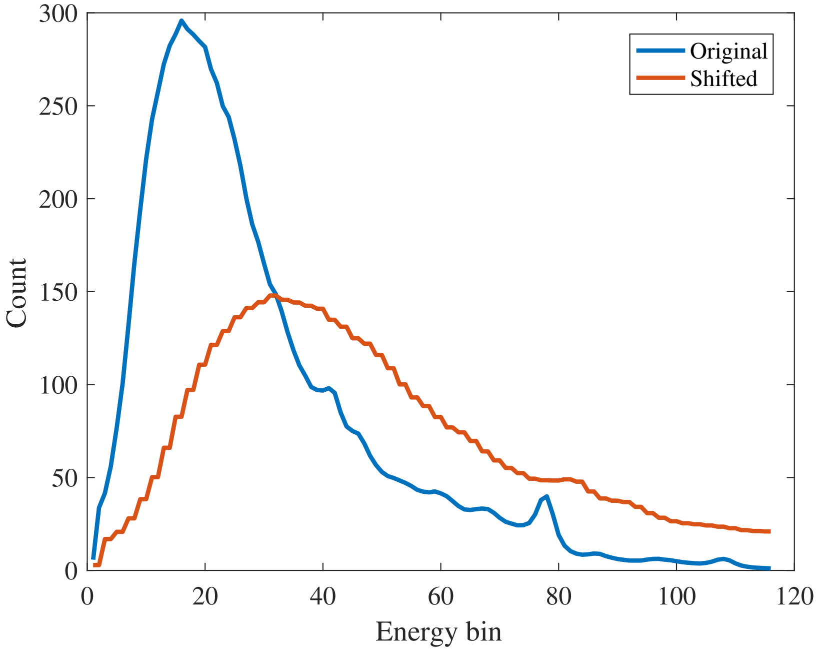
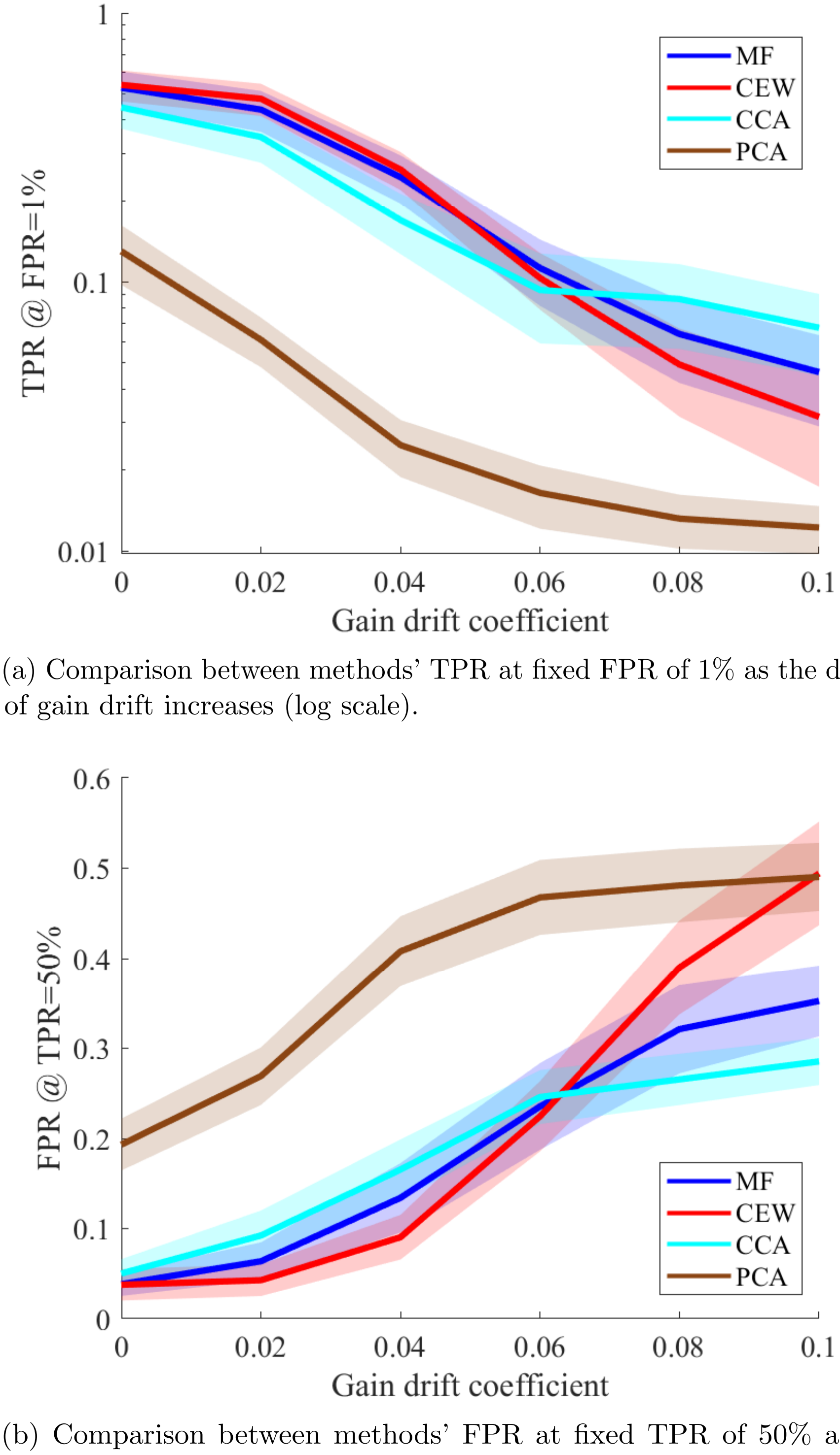


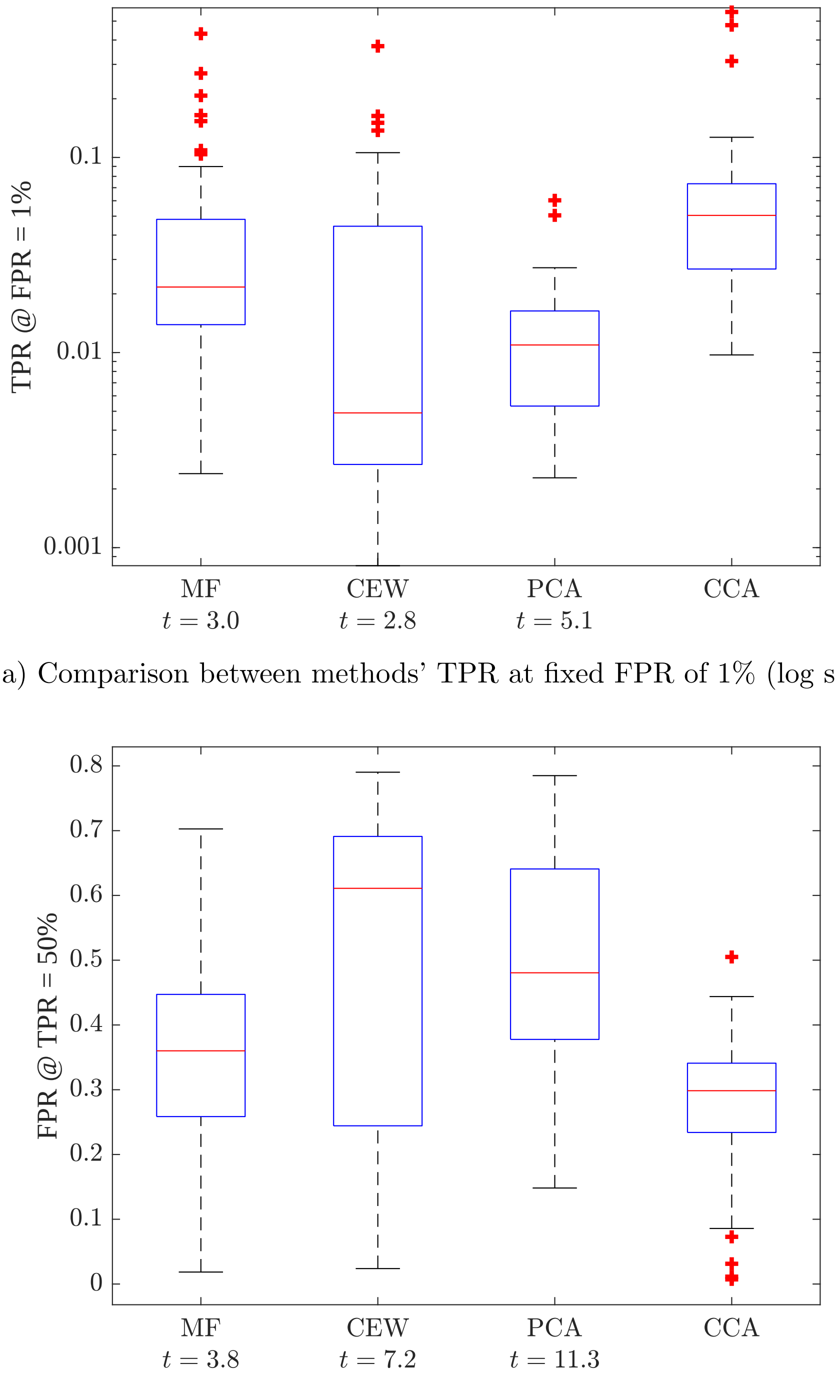
Figure 4.7: Example of a gain drift-shifted spectrum with a coefficient of 0.1.



(a) Comparison between methods' TPR at fixed FPR of 1% as the degree of gain drift increases (log scale).

(b) Comparison between methods' FPR at fixed TPR of 50% as the degree of gain drift increases.

Figure 4.8: Comparison between methods as the degree of gain drift increases. Values are averaged over all available source templates, and bands correspond to two standard errors.



(a) Comparison between methods' TPR at fixed FPR of 1% (log scale).

(b) Comparison between methods' FPR at fixed TPR of 50%.

Figure 4.9: Comparison between methods when gain drift is fixed at 0.1. Box plots show distribution of results over library of source templates.

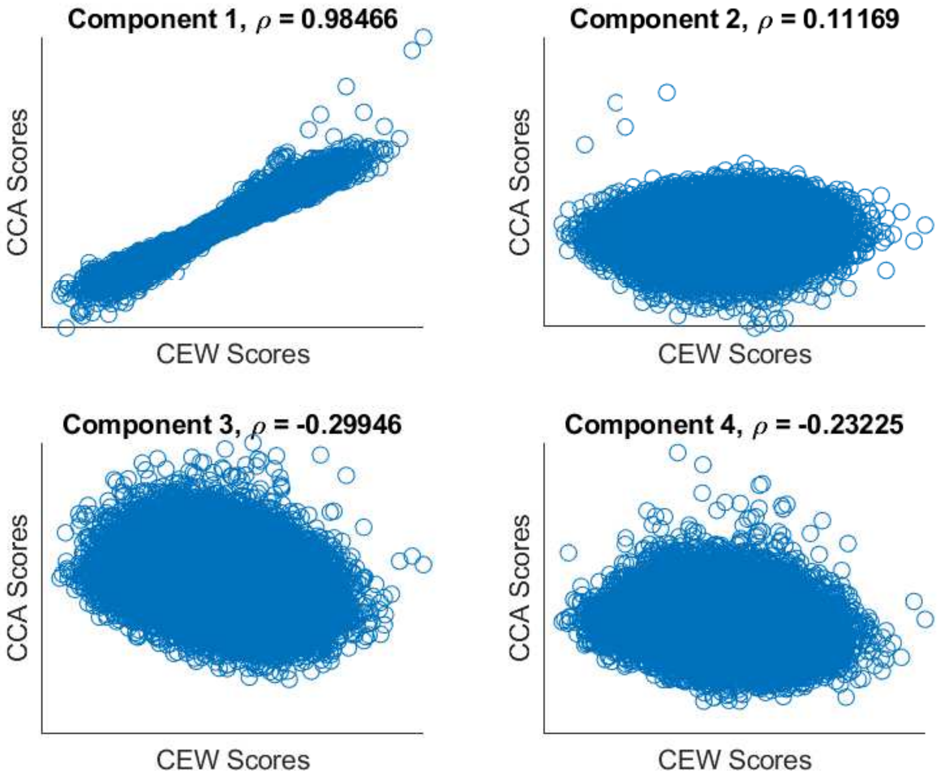


Figure 4.10: Comparison of CEW scores and scores in each CCA component. The first component is almost equivalent to CEW.

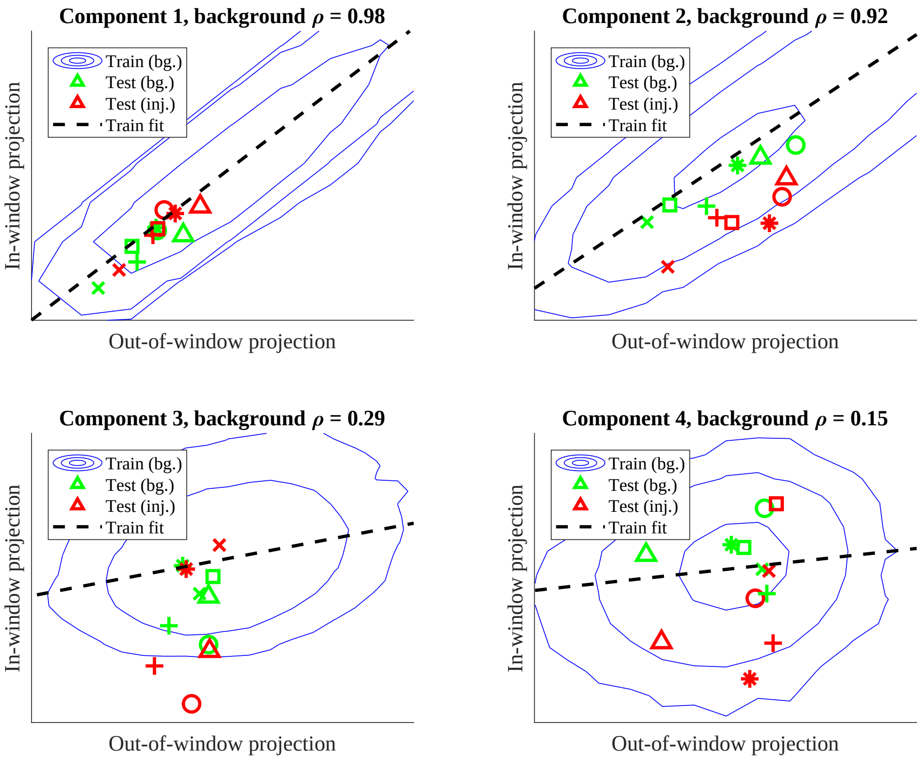
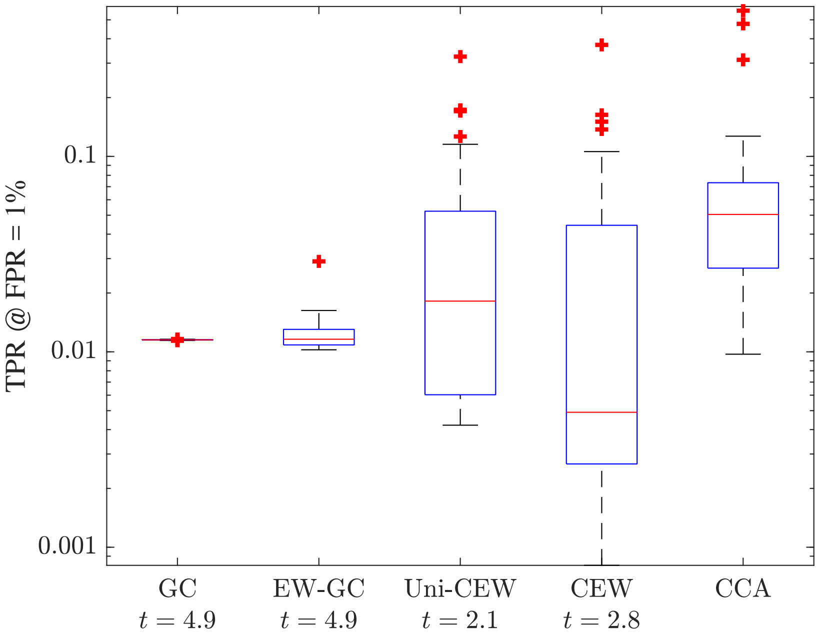
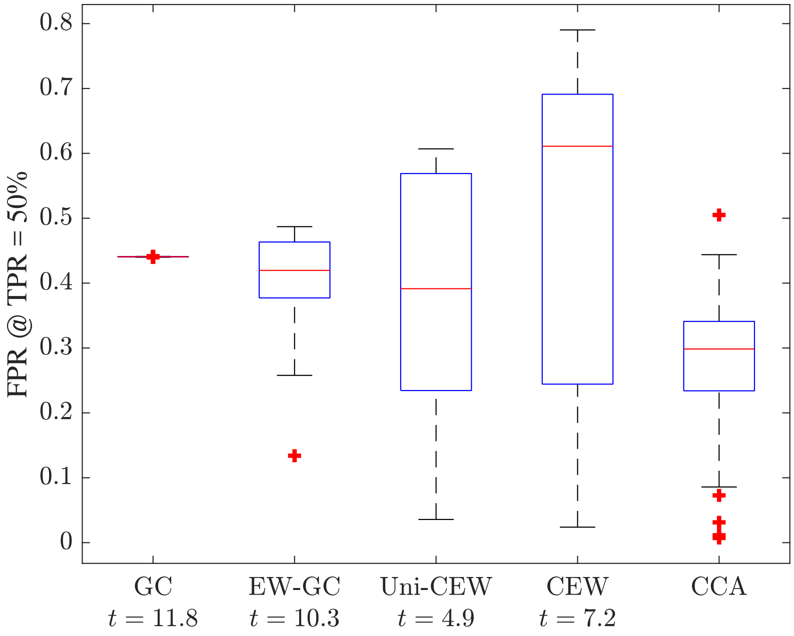


Figure 4.11: Background and injected test samples and how they fit the learned correlations from background. Blue contours show levels of the kernel density estimated distribution of training background at 0.1, 0.01, and 0.001. Green and red symbols show test observations before and after injection. The impact of injection is small in individual components but is noticeable when all components are considered together.



(a) TPR as detection methods progress in complexity.



(b) FPR as detection methods progress in complexity.

Figure 4.12: Comparison between methods when gain drift is fixed at 0.1. Box plots show distribution of results over the library of source templates.

Chapter 5

Clustering by Multi-View Relationships

5.1 Introduction

In this chapter, we ask how we can learn and perform inference on multi-view relationships, but unlike the previous chapter, we move our attention to nonlinear multi-view relationships.It is often interesting to analyze the correlation between two views using Canonical Correlation Analysis (CCA) Hotelling (1936), which finds linear relationships. A more detailed explanation can be found in 4.2. In many practical scenarios, however, the relationships may be nonlinear. One way to represent such structure is cluster-wise linearity. That is, different subsets of observations may have distinct patterns of correlation; important canonical variables might differ between subsets of observations. For instance, certain subpopulations might express a gene combination differently, or distinct subsets of subjects might have a different physiological response to medical trauma. Additionally, there may be globally nonlinear correlation structure in the data that may be approximated by cluster-wise linear models.

To allow discovery of such structures, we propose a method called Canonical Least Squares (CLS) clustering for dense, continuous data. A single CLS model can be regarded as a multiple-to-multiple correlation model that finds latent variables to connect inputs and outputs, somewhat like CCA. The proposed approach, however, also identifies a clustering of observations, which may be useful when different correlation structures appear in different subsets of the data. This approach can be considered a form of correlation clustering, a class of clustering methods that groups observations based on their correlation patterns Zimek (2009). We also introduce a supervised classification method that relies on CLS clustering. Practical benefits of these methods stem from their capability to find interpretable structure in the data to explain their predictions. In the simplest case, discussed in this paper, the correlation structures found in each cluster are linear, which aids interpretation, and the classification score has a gradient that is straightforward to compute and interpret.

To illustrate our approach, consider the two-dimensional dataset in Fig. 5.1. It contains three spatially overlapping Gaussian clusters with different covariance structures. When k-means or Gaussian kernel spectral clustering is employed, the resulting clusters are displayed in Fig. 5.2. As expected, they are contiguous in space, but they do not match the ground truth because of the overlap in data. In contrast, Fig. 5.4 illustrates the clusters learned by CLS clustering. They closely resemble the ground truth despite the overlapping data, though the learned clusters do not overlap. This example demonstrates that there are certain problems in which the data have an interesting structure that cannot be discovered by straightforward methods. This idea motivates our approach, distinguished from more common methods by the ability to account for such more unusual, but prevalent in practice, structures.

5.2 Related work

5.2.1 Multi-view clustering

There has been substantial past work on multi-view clustering. Multi-view versions of k-means and Expectation Maximization were considered by Bickel and Scheffer (2004) and found to outperform the single-view counterparts. A method by Chaudhuri et al. (2009) uses CCA to find the subspace spanned by the means of mixture components. However, their work assumes views are conditionally uncorrelated given the component. This is essentially the opposite of our work, which directly exploits these correlations. Kumar et al. (2011) propose a multi-view spectral clustering. This framework employs co-regularization to enforce agreement between spectral clusterings in different views. Another line of work by Nie et al. (2011) and Wang et al. (2013) approaches clustering as a regression-like problem of fitting the data to cluster membership probabilities. Wang et al. (2013) apply structured sparsity to weight features in different views by their importance.

According to Liu et al. (2013), multi-view clustering strategies can usually be grouped into three categories. First, multiple views are integrated through the loss function. Second, multi-view data are projected to a common subspace, in which any clustering algorithm is then applied. Third, a clustering solution is computed for each view individually, and then they are all fused to achieve a consensus. Recent work has heavily focused on Liu et al. (2013)'s work on the first strategy of clustering by non-negative matrix factorization (NMF). Their idea is to seek a common latent factor by regularizing similarity between a matrix factorization of each view. Most current articles propose extensions of this work. For example, Zhao et al. (2017) propose a hierarchical NMF with graph regularization that incrementally groups points from the same class closer together in each layer. In addition, Zong et al. (2017) propose a multi-manifold regularized NMF that better preserves local geometric structure of the multi-view space. However, almost all previous multi-view clustering strategies identify clusters from spatial relationships; clusters are defined as sets of points near each other in some space—either in each view or a combined space—but the relationships between views are not an explicit factor in determining clusters.

5.2.2 Single-view correlation clustering

Zimek (2009) considers the problem of clustering data based on patterns of correlation when the variables are not partitioned into two groups. Unlike CLS or CCA, however, this work assumes a single view; the correlation refers to correlation between all the variables, not just between two sets. The paper presents a diverse body of algorithms for the task.

5.2.3 Cluster-wise linear

Späth (1982) introduces a method for clustering the observations in a single-output regression dataset. Like k-means, this method is greedy and iterative and alternates between two steps. Given cluster labels, it fits a linear regression to each cluster. Given regression coefficients, it assigns each observation to the cluster whose regression residual is the smallest for that observation. It is simple to show that this method is a special case of CLS clustering in which the regression inputs are one set of variables and the regression output by itself is the other set—i.e., one view is univariate.

5.2.4 Dependency seeking clustering

An interesting approach to correlation clustering is explored by Klami and Kaski (2008) and Rey and Roth (2012). Klami and Kaski (2008) establish a probabilistic generative modeling framework to allow Bayesian inference. They do so by proposing a model of probabilistic families for finding dependency and give a general clustering algorithm for this family. CCA is shown to be a special case. A key assumption is that a linearly transformed Gaussian latent variable produces the variation in the data. However, there may be severe model mismatch when this assumption was violated. To remedy this behavior, Rey and Roth (2012) deploy a copula mixture model to the framework, enabling them to model mixtures of CCA, similar to the clustering setup in this work. A Bayesian clustering algorithm is proposed and shown to perform well on synthetic and real datasets. However, a disadvantage of this approach is that it requires a prior distribution to be specified for every feature, which could require substantial tuning to avoid mismatch.

5.2.5 Correlational spectral clustering

In Blaschko and Lampert (2008), a multi-view clustering method is proposed based on kernel CCA (KCCA). It simply runs KCCA on the kernel matrices of two views and then runs k-means on the latent variables in one view. The authors state that this method generalizes spectral clustering to arbitrary kernels and paired data. A notable distinction from our work is that KCCA cluster assignments for test observations depend on only one view. For example, if it were run on the data in Fig. 5.1, it would find clusters divided along vertical lines. Hence, the clusters have limited dependence on correlation between views.

5.3 Canonical least squares

In this section we develop our method for correlation clustering called Canonical Least Squares (CLS) clustering. We then describe how it can serve as the basis of supervised classification. Recall Canonical Correlation Analysis (CCA), the multi-view method summarized in Ch. 4.2.1. A standard reformulation Hardoon et al. (2004) of the objective function in (4.1) relevant to our method is

$$\min_{u \in \mathbb{R}^{d_X}, v \in \mathbb{R}^{d_Y}} \mathrm{E}\left[\|Xu - Yv\|_2^2\right]. \tag{5.1}$$

Like CCA, CLS takes sets of variables X and Y and produces up to $m \leq d_X \wedge d_Y \wedge \operatorname{rank}(X^\mathsf{T} Y)$ pairs of vectors (u, v) such that the components $X^\mathsf{T} u$ and $Y^\mathsf{T} v$ have some kind of relationship. Unlike CCA, this relationship is not of maximum correlation but of least squared error. It functions best on dense, continuous data.

5.3.1 First components

First consider only the top pair of components ($m = 1$). We redefine $X \in \mathbb{R}^{n \times d_X}$ and $Y \in \mathbb{R}^{n \times d_Y}$ as centered data matrices. Then (5.1) becomes

$$\begin{aligned} \min_{u \in \mathbb{R}^{d_X}, v \in \mathbb{R}^{d_Y}} \quad & \|Xu - Yv\|_2^2 \\ \text{subject to} \quad & u^\mathsf{T} X^\mathsf{T} X u = v^\mathsf{T} Y^\mathsf{T} Y v = 1. \end{aligned}$$

We propose the following modification, which has the same objective but different constraints:

$$\begin{aligned} \min_{u \in \mathbb{R}^{d_X}, v \in \mathbb{R}^{d_Y}} \quad & \|Xu - Yv\|_2^2 \\ \text{subject to} \quad & v^\mathsf{T} v = 1. \end{aligned} \tag{5.2}$$

This optimization has a positive semidefinite objective but quadratic constraints. We denote (5.2) CLS (for the first component). One major difference from CCA is the lack of X or Y in the constraints. This difference enables CLS to form the building block of a clustering method with a well-defined optimization procedure, as will soon be explained. The other difference is the lack of u in the constraints. When only v is constrained, the problem generalizes ordinary least squares, which does not constrain the coefficients of the independent variables, to multiple outputs.

Next we present the solution to (5.2). First let v be fixed. The problem becomes ordinary least squares in u, yielding

$$u = (X^\mathsf{T} X)^{-1} X^\mathsf{T} Y v.$$

Let $H = I - X(X^\mathsf{T}X)^{-1}X^\mathsf{T}$, a symmetric idempotent matrix. After substituting for u, the problem in v is given by

$$\min_{v \in \mathbb{R}^{d_Y}} \|HYv\|_2^2 \qquad \text{subject to} \quad v^\mathsf{T}v = 1.$$

This problem resembles PCA except with a minimum instead of maximum. The solution v is the eigenvector with the lowest eigenvalue of $Y^\mathsf{T}H^\mathsf{T}HY = Y^\mathsf{T}HY$. The complexity of this routine is $O(n(d_X + d_Y)^2 + (d_X + d_Y)^3)$.

5.3.2 Multiple components

In CCA, subsequent canonical variables are uncorrelated with each other. After changing these constraints to be independent of the data, we are left with simple orthogonality constraints between vectors of coefficients. The generalization of (5.2) to m components is then

$$\begin{aligned} \min_{\substack{U \in \mathbb{R}^{d_X \times m} \\ V \in \mathbb{R}^{d_Y \times m}}} \quad & \|XU - YV\|_\mathcal{F}^2 \\ \text{subject to} \quad & V^\mathsf{T}V = I. \end{aligned} \tag{5.3}$$

Again, this problem has a positive semidefinite objective but quadratic constraints. It is difficult to solve analytically because all components must be found simultaneously. We instead choose an easier suboptimal solution: let V be the eigenvectors corresponding to the m lowest eigenvalues from the solution to (5.2), and compute U accordingly. This solution corresponds to greedily solving for each component sequentially under orthogonality. The computational runtime of this algorithm is $O(n(d_X + d_Y)^2)$. It is an interesting tangent to juxtapose this procedure with Principal Components Analysis (PCA), which solves a similar problem

$$\max_{W \in \mathbb{R}^{d \times d}} \|ZW\|_\mathcal{F}^2 \quad \text{subject to} \quad W^\mathsf{T}W = I$$

where $Z \in \mathbb{R}^{n \times d}$ is a centered data matrix. In PCA, the greedy eigenvector solution is optimal because of the orthogonality constraints between full vectors of coefficients. In CLS, however, only the vectors v_i must be orthogonal, rendering the greedy solution suboptimal.

Separately, in the special case that $m = \min\{d_X, d_Y\}$, then U or V is an orthogonal matrix, so CLS reduces to ordinary least squares on the columns of X or Y respectively.

5.4 CLS clustering

So far we have presented how to change CCA, a multiple correlation problem, to CLS. On its own, CLS is probably uninteresting, but it becomes relevant in the context of clustering. Our proposed CLS clustering algorithm takes matrices X and Y, a number k of clusters, and a number m of components. Let $X^{(i)}$ and $Y^{(i)}$ denote X and Y with rows sub-sampled to those in cluster i. Let the coefficients corresponding to that cluster be $U^{(i)}$ and $V^{(i)}$. To find cluster labels for each data point, we iterate the following steps until convergence:

- **CLS step** Given cluster labels, for each cluster $i = 1, \ldots, k$: run CLS (5.3) on $X^{(i)}$ and $Y^{(i)}$ to find $U^{(i)}$ and $V^{(i)}$.
- **Labeling step** Given CLS coefficients $U^{(i)}$ and $V^{i)}$, for each observation (x_ℓ, y_ℓ), $\ell = 1, \ldots, n$: assign it to

$$\text{argmin}_i \, \|y_\ell^\mathsf{T} V^{(i)} - x_\ell^\mathsf{T} U^{(i)}\|_2^2.$$

This procedure takes a block coordinate-wise iterative approach, resembling Expectation-Maximization Dempster et al. (1977), to solving the overall optimization problem

$$\begin{aligned} \sum_i \min_{\substack{U^{(i)} \in \mathbb{R}^{d_1 \times m} \\ V^{(i)} \in \mathbb{R}^{d_2 \times m}}} & \|R^{(i)}(XU^{(i)} - YV^{(i)})\|_\mathcal{F}^2 \\ \text{subject to} \quad & V^{(i)\mathsf{T}} V^{(i)} = I, \; i = 1, \ldots, k, \end{aligned} \tag{5.4}$$

where $R^{(i)}$ is a length n diagonal matrix whose ℓ-th diagonal element is the binary indicator of whether observation ℓ is assigned to cluster i.

A convergence guarantee exists when $m = 1$, i.e., when only the first pair of components is used. The CLS step optimizes over the u_i's and v_i's, while the labeling step optimizes over the $R^{(i)}$'s. Thus the objective is non-increasing at every step, so convergence is guaranteed. If $m > 1$, an exact solution to CLS would also guarantee monotonicity, but since a greedy approximation is used, monotonicity is not guaranteed. Nevertheless, we have found the objective function to usually behave monotonic empirically.

Furthermore, in some applications it is helpful to add must-link constraints that designate sets of points that must appear in the same cluster. These constraints can be encoded by assigning each set of points to the cluster that minimizes the sum of squared errors over the points.

In addition, an analogous clustering algorithm was proposed by Fern and Friedl (2005) called CCA clustering. While CCA maximizes correlation between variables in the latent space, CLS minimizes the squared error. These objectives are similar, but CLS can find components with weaker correlation and smaller residuals, which is not necessarily an advantage or disadvantage. However, CLS clustering solves one important issue with CCA clustering. Recall that the CCA optimization had constraints dependent on data,

$$u^\mathsf{T} X^\mathsf{T} X u = v^\mathsf{T} Y^\mathsf{T} Y v = 1.$$

As a result, when cluster assignments change, the constraints for each cluster's CCA problem change as well. To be consistent, the search space for cluster assignments would also have to satisfy those constraints, but this requirement is infeasible. Consequently, there is no reason for the CCA clustering algorithm to improve its solution at each iteration. By removing the dependence on data in constraints, CLS clustering avoids this problem and therefore permits a more well-behaved optimization routine. Indeed, we conducted simple simulations on synthetic Gaussian data and found that CCA clustering often finds poor solutions whereas CCA clustering finds reasonable solutions much more often.

5.5 Practicalities

Intercept An intercept should be incorporated in CLS clustering by augmenting X with a column of 1's.

Data scale CCA is affine-invariant with respect to X and Y. However, CLS is sensitive to scaling because it uses Euclidean distance, similar to k-means. Therefore, we recommend normalizing the column variance in preprocessing.

Initialization Like in all greedy iterative algorithms similar to k-means, random initialization over many runs improves the chance of CLS clustering to reach a robust solution.

5.6 Synthetic data experiment

We deployed different clustering methods on a medium-sized synthetic dataset. The dataset consisted of 10 equally sized clusters of 1,000 points each. Each cluster was sampled from a different multivariate Gaussian in $\mathbb{R}^{100}$ centered at the origin with covariance drawn from a Wishart distribution. The first 50 features composed one view while rest composed the other view.

Table 5.1: Cluster Quality on Synthetic Data

	CLS	CCA	k-means	SC
ARI	.99 ± .01	.94 ± .02	.005 ± .003	.000± .000

We computed the Adjusted Rand Index (ARI) relative to the true cluster labels Yeung and Ruzzo (2001). We compared CLS clustering to CCA clustering, k-means, and Gaussian kernel spectral clustering (SC). The simulation was run 50 times. Table 5.1 displays the average ARI and two times the standard error. The best performer was CLS clustering. The next best was CCA clustering, which usually produced solutions with ARI of either about 1.00 or .89 possibly because of convergence issues. Lastly, k-means and SC performed very poorly because they searched for isolated L_2 clusters structure while the true clusters overlapped.

5.7 Bleeding experiment

We used a medical dataset to conduct three sets of experiments corresponding to different levels of supervision for CLS clustering.

5.7.1 Background

We considered a dataset in which we attempted to detect the presence of bleeding and other conditions by monitoring central venous pressure (CVP), the blood pressure measured invasively in the central veins close to the heart, or estimated from indirect less invasive measurements. It helps quantify right atrial pressure and can be used as an estimate of right ventricular preload. Predictive tasks based on CVP have been the subject of several studies in the medical literature (Michard and Teboul, 2000; Pinsky and Payen, 2005; Kumar et al., 2004; Marik and Cavallazzi, 2013; Damman et al., 2009; Boyd et al., 2011). Here we investigate the CVP signal within a controlled setting by attempting to classify CVP waveforms as indicative of an active bleeding episode vs. periods of no-bleeding. We show how CLS can make predictions as well as automate the discovery of insights of potential clinical interest. CLS clusters can be interpreted as clinical phenotypes characterizing patients' pre-bleeding or post-bleeding responses. Also, the relationship of bleeding with inspiration and expiration can be interpreted in terms of the original CVP waveforms.

5.7.2 Data description

The data were collected from an experiment in which healthy pigs were subjected to controlled bleeding. The experimental procedure was similar to that in Pinsky (1984). Thirty-eight Yorkshire pigs were anesthetized, instrumented with catheters, and allowed to stabilize for 30 minutes. Then they were bled at a constant rate of 20 mL/min. Their CVP was monitored for 25 minutes before bleeding and 25 minutes after its onset. Two CVP waveforms (Fig. 5.5) were extracted from each respiration cycle, one from the top of the inspiration phase of breathing and the other from the bottom of expiration. The respiration cycles lasted 5.2 seconds each on average, resulting in an average of 556 observations per pig over the 50 minutes of observation. Thirteen features were extracted from each waveform as averages and ratios between different characteristic points of the CVP waveform, landmarks used commonly in clinical analysis. These features included differences between peaks and troughs such as the height SA between points S and A as well as ratios of ranges such as VR over CQ (Fig. 5.5).

5.7.3 Blood loss known exactly

In the first experiment, the exact amount of blood loss was assumed to be known as zero before bleeding and linear at a rate of 20 mL/min after the onset of induced bleeding. CLS clustering was run to cluster respiration cycles with all CVP features as the input view and blood loss as the output view. The data from all 38 pigs were concatenated. Since the output view was univariate, the method corresponded to cluster-wise linear regression (Späth, 1982). The purpose of this experiment was to learn clusters that corresponded to bleeding status by directly incorporating bleeding information. We tried $k = 4$ and $m = 4$. These values were selected by hand to optimize visual quality of the clusters. The cluster assignments are shown in Fig. 5.6. Each row represents a subject, while each column represents a time step. The clusters are color-coded. The clusters were largely contiguous in time, even though there was no such constraint in the method. One cluster corresponded to no bleeding, but the bleeding period was separated into three separate phases. This result confirms the hypothesis that the chosen parameterization of the CVP waveform carries in its structure the information about the bleeding status of the subject and is to some extent informative of the amount of blood lost.

5.7.4 Blood loss unknown

In the previous setting, the amount of blood loss is unknown in practice, but this information is required at test time to cluster new observations. To simulate a more practical environment, the second experiment was to deploy CLS clustering on only CVP features without knowing blood loss. The two views were inspiration and expiration, which are both known at test time. Observations from the same pig were constrained to belong to the same cluster during training. The pigs were partitioned into training and test sets of 25 and 13 subjects respectively. Data from all training pigs were concatenated to learn the cluster model. The purpose of this experiment was to examine the qualitative performance of this method in more realistic conditions. We chose $k = 4$ and $m = 4$ to match the previous experiment.

The cluster assignments are shown in Fig. 5.8. The clusters were still mostly time-contiguous, although they appeared significantly noisier than before. This result was expected because critical information, the amount of bleeding, was excluded. Two clusters corresponded predominantly to either no bleeding or bleeding, but the two other clusters identified two small groups of pigs that appeared distinctive before the onset of bleeding. This suggests that our method discovered diversity within the subjects' physiology in the stable state, later confirmed via independent review. Crucially, the clusters still corresponded to bleeding status even though no information about bleeding was known. For comparison, Fig. 5.7 displays the clusters from k-means on the concatenated views. The clusters are time contiguous and smoother than CLS clusters, which is expected because observations close in time are probably close in L_2 distance. Overall, the two approaches find similar cluster structures. However, there are important differences. Most notably, our method finds a much clearer divide between no bleeding and bleeding, such as in Pigs 1, 3, and 11. In addition, Pig 4 is interesting because k-means does not differentiate between its bleeding and no-bleeding states, whereas for CLS clustering that distinction is more clear. These results indicate multi-view relationships are more powerful than spatial relationships for discovering temporal structure in the physiological response to bleeding. Furthermore, it is plausible that k-means is underfitting the clusters, while the noise in the CLS clusters indicates that our method may be further from underfitting (and closer to overfitting).

Additionally, Fig. 5.9 shows for comparison the clusters from CCA clustering on inspiration and expiration views. The same constraints and number of clusters and components were used as in CLS clustering, yet the clusters do not appear much time-contiguous and do not correspond with bleeding status. Also, we clustered the same data using spectral clustering (Von Luxburg, 2007) using a Gaussian kernel to produce the affinity matrix, but over 98% of data points were assigned to the same cluster.

Fig. 5.10 shows the four latent components over time of one pig in a particular cluster. The components appear almost constant before bleeding and become more or less linear after bleeding. This behavior is interesting because it resembles the amount of blood loss over time, even though this information was excluded from training. Although this example is from one cluster, it is also representative of all other clusters.

5.7.5 Discussion

Clinical relevance Discussed in Sec. 6.3.

Quantitative evaluation of correlation clusters To our knowledge, there is no consensus in correlation clustering community on a framework for quantitative evaluation. Ideally, our method's cluster quality would be numerically compared on real datasets to alternatives such as k-means, spectral clustering, correlational spectral clustering (Blaschko and Lampert, 2008), CCA clustering (Fern and Friedl, 2005), and copula-based dependency-seeking clustering (Rey and Roth, 2012). However, there are a couple underlying issues with such comparisons. First, the former three alternatives are based on L_2 distance to some extent, which makes them improper comparisons because they identify fundamentally different cluster variables. The second issue is the shortage of ground truth in public datasets for correlation clusters. Consequently, in correlation clustering literature, it is common practice to perform qualitative evaluation of clusters rather than quantitative (Fern and Friedl, 2005; Rey and Roth, 2012; Zimek, 2009). We do the same in this work, following our best intuition.

Bleeding clusters Figs. 5.6 and 5.8 highlight an interesting pattern. For many pigs, there was a dominant cluster before bleeding, but when bleeding started, a different cluster took over. This new cluster typically only held observations from the first ten or fewer minutes after bleeding. Afterward, other clusters became dominant. One interpretation is that the physiological response to bleeding changed as the induced stress escalated. There may have been an initial compensation surge, followed by a more systemic response mediated through autonomic nervous control which could also change in its modality as a function of escalating stress. This hypothesis may be supported by Fig. 5.10, which shows that the immediate onset of bleeding corresponded to a spike in latent variables. This pattern is an example of how the interpretable structure of CLS clustering can lend itself to finding practical insights. In addition, our method identified diversity

Appropriate types of data In elided experiments we tested CLS clustering on many different types of synthetic data, which we summarize here. The most appropriate type was found to be continuous values with no missing data. Some distributions on which it performed included Gaussian, log-normal, and uniform. On missing or sparse data, the method requires additional treatment such as imputation before or during the learning process.

Soft clustering We developed a soft clustering extension of CLS based on ideas in Hathaway and Bezdek (1993). One way to view this extension is that cluster probabilities of an observation are regularized toward a uniform distribution over clusters. In the soft version, the optimization is much smoother, resulting in more consistent solutions over different runs. In the applications shown in this paper, however, it was outperformed by the hard version, even though the assignment step in the hard version is highly non-smooth. A potential avenue for future work would be to analyze this and other theoretical optimization properties of the method.

Spectral interpretation Recall that the solution v for the first component of CLS was given by the last eigenvector of $Z \equiv Y^{\mathsf{T}}(I - X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}})Y$. Let $\Sigma_{xx} = X^{\mathsf{T}}X$, $\Sigma xy = X^{\mathsf{T}}Y$, and $\Sigma_{yy} = Y^{\mathsf{T}}Y$. Assuming the data are centered, these variables are covariance and cross-covariance matrices of X and Y. Then $Z = \Sigma_{yy} - \Sigma_{xy}^{\mathsf{T}}\Sigma_{xx}^{-1}\Sigma_{xy}$ is the Schur complement of the covariance matrix of the joint distribution of X and Y. If this joint distribution is multivariate normal, then Z is the conditional covariance of Y given X. Hence CLS can be interpreted as finding the direction of minimum variance in Y given X. When Y has less variance after controlling for its relationship with X, it is easier to find a better linear fit with X. CLS clustering is similar in this regard to correlation clustering methods by Zimek (2009), which also leverage eigenvectors of lower variance.

5.8 Conclusion

This work considered the problem of discovering interpretable structures in complex datasets. In particular, we proposed a method to learn correlation clusters for multi-view data, where important relationships between the views are discovered. The method was tested on CVP waveform datasets of induced bleeding and was demonstrated to find interesting structure. Although we demonstrated the potential utility of the proposed method on the task of real-time monitoring of surgical patients, it can be useful in a wide range of multi-view problems in clinical and biological engineering applications, wherever distinct multi-modal structures of relationships between views of data can reveal operationally useful information.

A useful extension of this method would be to incorporate sparseness, most likely via an L_1 penalty in the CLS objective. Although it is not difficult to preserve the analytic solution with this penalty on one view, it is an open problem to do so while penalizing both views.

Another possibility for future work would be to augment the CVP featurization with systematic features. It is potentially problematic that the current featurization depends on synchronicity of inspiration and expiration by pairing them as views because the phases often do not align well. A solution would be to create features using harmonic or wavelet transforms. Rather than extracting a single waveform from each phase, the transforms would be applied to each entire phase. This change would reduce the noise in features and would complement the existing featurization.

Another future direction in clustering would be to generalize to arbitrary nonlinear relationships, not only cluster-wise linear ones. One approach would be to kernelize CLS similar to Kernel CCA (Akaho, 2006). Alternatively, nonlinear models have become highly popular with the advent of deep learning. It would be intriguing to extend our framework with a technique such as Deep CCA (Andrew et al., 2013) to replace CLS.

This chapter showed how multi-view relationships can be learned nonlinearly to perform unsupervised learning. This result raises the question of whether we can do the same for supervised learning, the subject of the following chapter.

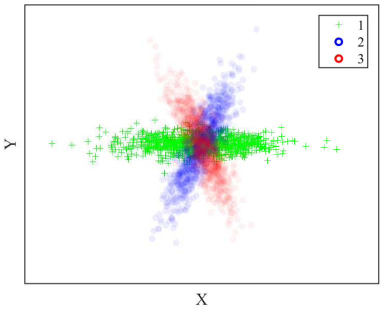


Figure 5.1: Ground truth clusters for an overlapping Gaussian dataset, drawn translucently to illustrate overlap.

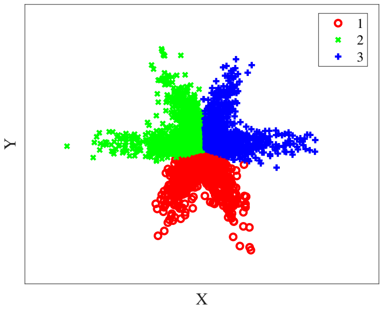


Figure 5.2: Clusters from k-means or spectral clustering for an overlapping Gaussian dataset.

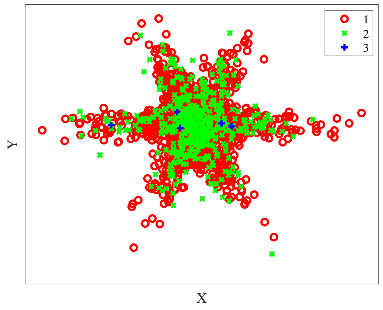


Figure 5.3: Clusters from deep non-negative matrix factorization Zhao et al. (2017) for an overlapping Gaussian dataset.

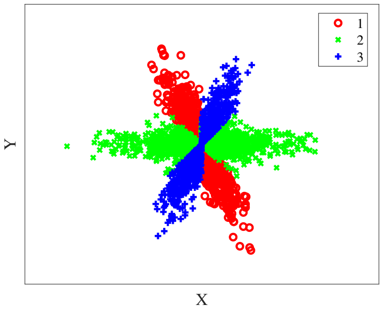


Figure 5.4: CLS clusters for an overlapping Gaussian dataset.

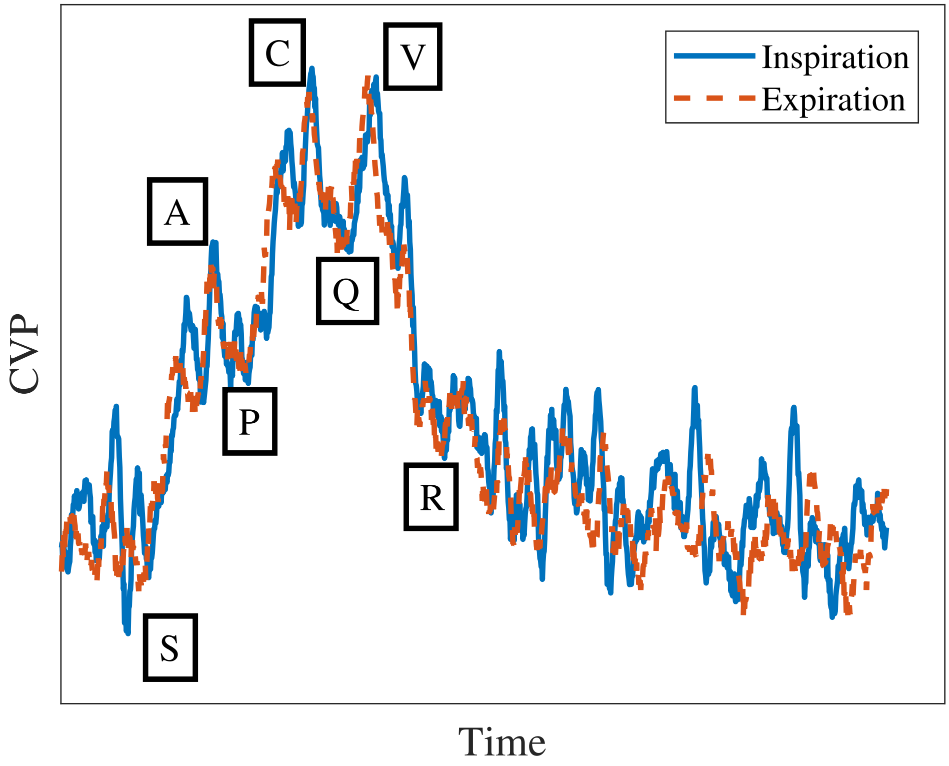


Figure 5.5: An example of a central venous pressure waveform for inspiration and expiration phases of breathing cycle, along with labeled key points.

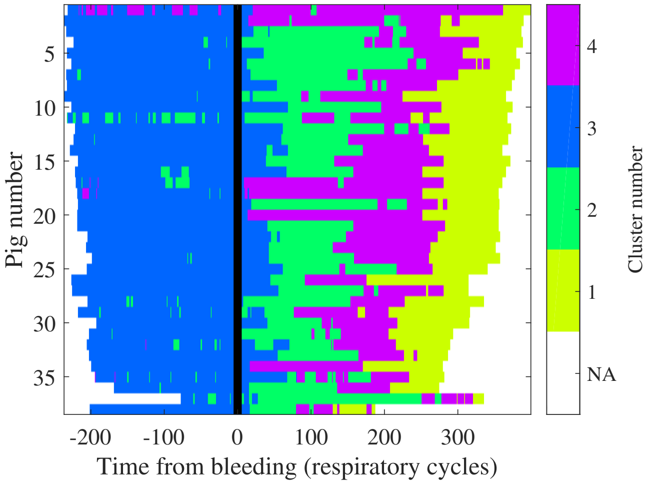


Figure 5.6: CLS cluster assignments with 4 clusters when blood loss is known.

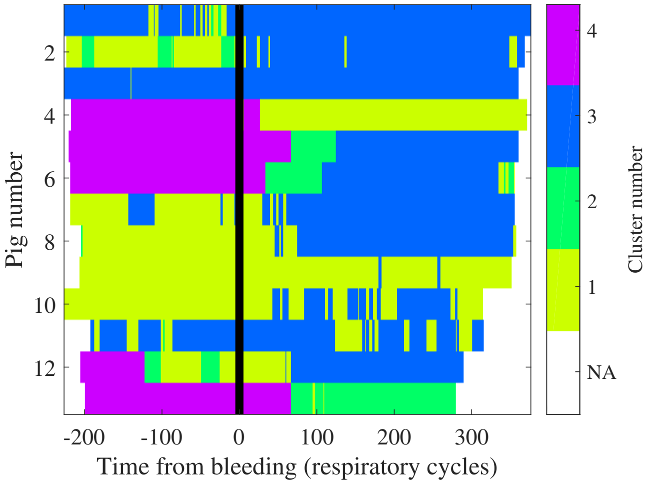


Figure 5.7: Cluster assignments from k-means with 4 clusters when blood loss is unknown.

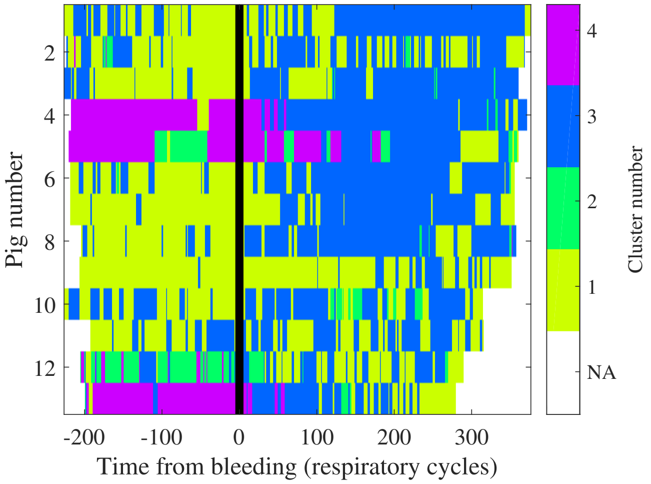


Figure 5.8: CLS cluster assignments with 4 clusters when blood loss is unknown.

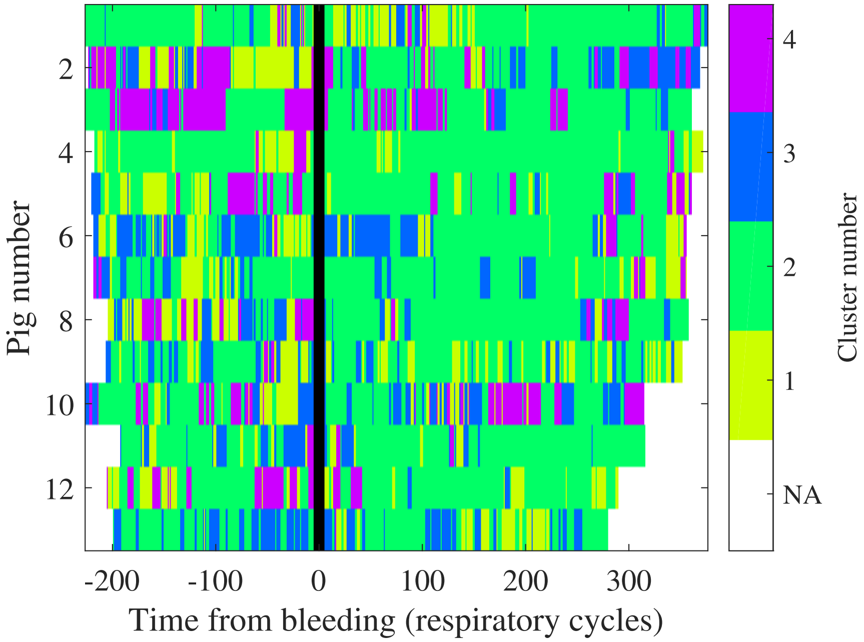


Figure 5.9: CCA cluster assignments with 4 clusters when blood loss is unknown.

.

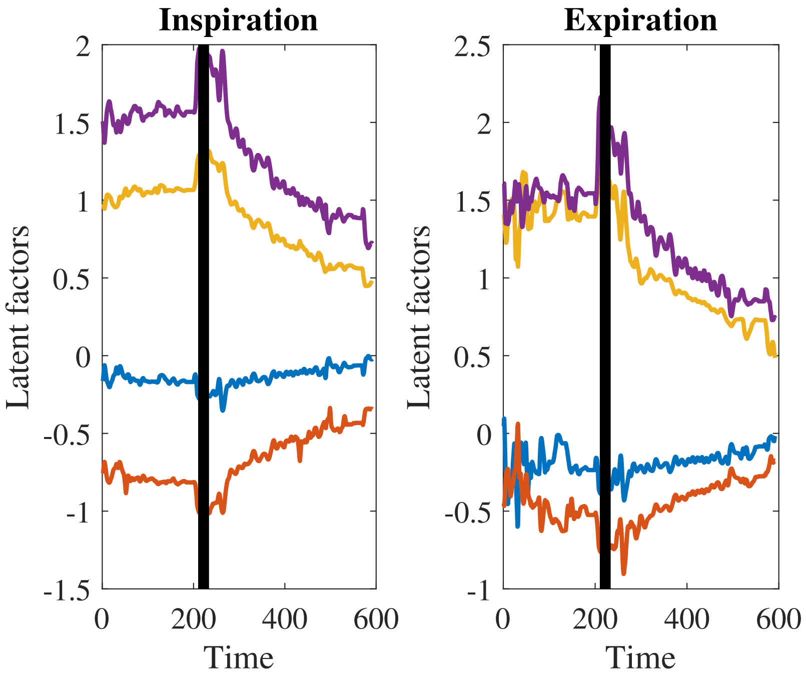


Figure 5.10: CLS latent components over time for one cluster when blood loss is unknown.

Chapter 6

Classification by Multi-View Relationships

6.1 Introduction

This chapter is an extension of Ch. 5, which focused on descriptive analytics. Here we switch to prediction—classification in particular. We aim to use the same idea of nonlinear multi-view correlations, but now we hypothesize that each class has its own distinct structure. The intuition is that we perform comparisons similar to the nearest-neighbor algorithm but with multi-view relationship features rather than spatial features. This approach generalizes the previous anomaly detection method in Ch. 4. The importance of this chapter is that it proposes a unique multi-view approach to classification based on multi-view relationships rather than the agreement between views of spatial relationships. Additionally, it offers a quantitative way to evaluate mixtures of linear relationships, our choice to represent globally nonlinear relationships.

6.2 Method

It is straightforward to build a supervised classification method on top of CLS clusters. First CLS clusters are learned independently on each class. Then new points are scored for each class according to the best fitting (lowest scoring) cluster in that class's fitted model. More formally, the score of point (x, y) in cluster i is given by

$$-\frac{1}{2}\|x^\mathsf{T} U^{(i)} - y^\mathsf{T} V^{(i)}\|_2^2.$$

The minimum of these scores is taken over the clusters in a given class to produce the score for that class. The final classification is the class that has the best fitting cluster overall. In addition, the procedure can be run many times with different random initializations and the scores averaged, which would make this classifier an ensemble method.

One possible drawback of this method is that by learning models on different classes independently, it does not maximize separation between classes. However, anomaly detection does not either. Indeed, this method can be considered a nonlinear multiclass generalization of the anomaly detection method in Ch. 4. This intuition may justify why it performed well in experiments.

Additionally, in many applications, it is interesting to examine only two of the learned clusters and ask how to decide which of them a new observation should belong to. It is possible to derive a locally linear model of the relevant factors, which should be readily interpretable. Of course, we can only interpret a single weak learner from the ensemble, not the entire ensemble at once, but this difficulty is shared by all ensemble methods. Now, let the observation be given by (x, y) and let z be the vertical concatenation of x and y. Let the two clusters of interest have coefficients $U^{(i)}$ and $V^{(i)}$, where $i \in \{0, 1\}$, and let $W^{(i)}$ be the vertical concatenation of $U^{(i)}$ and $-V^{(i)}$. The loss in cluster i is then $z^\mathsf{T} W^{(i)} W^{(i)\mathsf{T}} z/2$. The classification score between the two clusters is

$$\frac{1}{2} z^\mathsf{T} (W^{(0)} W^{(0)^\mathsf{T}} - W^{(1)} W^{(1)^\mathsf{T}}) z$$

where a higher score indicates membership in cluster 1. We determine the effect of a small change in any individual feature by computing the gradient,

$$(W^{(0)} W^{(0)^\mathsf{T}} - W^{(1)} W^{(1)^\mathsf{T}}) z.$$

6.3 Experiment on bleeding data

In the unsupervised setting, it was difficult to obtain quantitative measures without ground truth. Thus, the third experiment was to run CLS classification. The classification task was to decide whether an observation came from before or after the

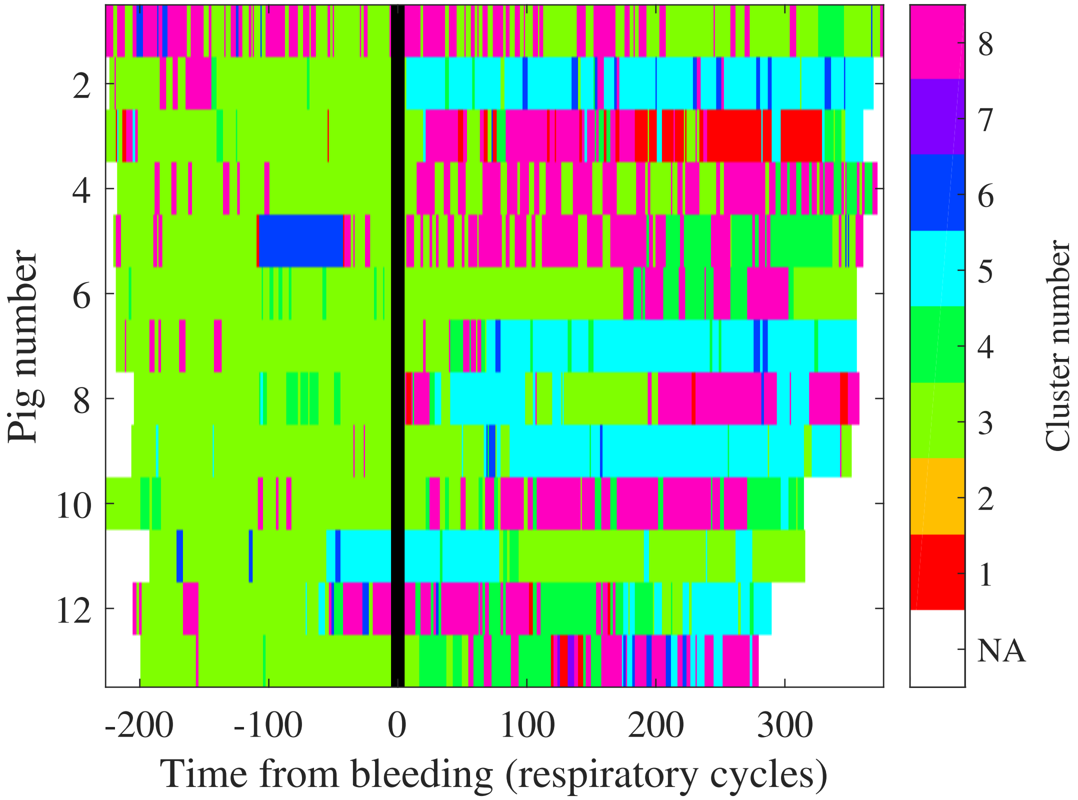


Figure 6.1: CLS cluster assignments with 3 non-bleeding (1-3) and 5 bleeding (4-8) clusters when blood loss is known as a binary label.

onset of bleeding. The binary label is required at training time but is not needed to classify or cluster unseen observations, so this form of supervision is more practical than the first. Under the same training/test split, data from all training pigs were concatenated to learn the cluster model. Leave-one-subject-out cross-validation was employed to select the hyperparameters k and m for non-bleeding and bleeding models. Following the classification algorithm from the previous section, CLS clusters were learned separately on the two classes. Observations from the same pig were constrained to belong to the same cluster during training. The classification scores on a left-out pig were used to compute the area under the receiver operating characteristic curve (AUC), true positive rate (TPR) at a false positive rate (FPR) of 10% and 1%, and FPR at a TPR of 50%. Hyperparameters were selected by optimizing the AUC. We chose 3 clusters with 6 components each for pre-bleeding and 5 clusters with 7 components each for post-bleeding.

Table 6.1: Bleeding Classification Performance

	Single cluster CLS	Final CLS	Random forest
AUC	.701 ± .128	.862 ± .064	.891 ± .075
TPR @ .10 FPR	.468 ± .185	.674 ± .145	.762 ± .167
TPR @ .01 FPR	.222 ± .134	.501 ± .185	.610 ± .210
FPR @ .50 TPR	.239 ± .152	.064 ± .055	.073 ± .075

Fig. 6.1 illustrates the resulting cluster assignments. Similarly to the previous experiment, the clusters are mostly time contiguous but contain substantial noise. There is one predominant cluster for no bleeding and several for bleeding. Intriguingly, the cluster structure does not appear too similar to the previous experiment and has more differentiation between subjects during the bleeding period, suggesting some extent of individualization of the response to harmful effects of bleeding across subjects.

Table 6.1 shows performance metrics of the final model on test data. It also gives results from a model that learns only one cluster on each class. The sizable gap in results demonstrates the benefit of searching for correlations that exist in subsets of the data, as opposite to a global correlation model identifiable in the whole set. The table includes results from a random forest classifier (Breiman, 2001) with 100 trees trained on the combined views. The random forest performs best in most metrics, but its advantage vs. CLS is not statistically significant. This result is acceptable since CLS enables detailed yet interpretable view of discovered structures in data while its performance metrics remain within the confidence interval of otherwise powerful random forest classifier. The explainability of CLS results will be discussed later in this section and in Sec. 5.7.5.

To understand the model's decisions, we used the method involving the gradient derived in the previous section on weak learners from the pre- and post-bleeding ensembles. We checked the score that determined whether a certain pig belonged to cluster 1 or 4, where cluster 1 was pre-bleeding and cluster 4 was post-bleeding. We computed the gradient of the score on a pre-bleeding observation. The results are displayed in Fig. 6.2. The original waveform of the observation is plotted on the left. According to the gradient, the most major changes that would make the observation closer to a bleeding waveform were shortening the lengths SA and AP during

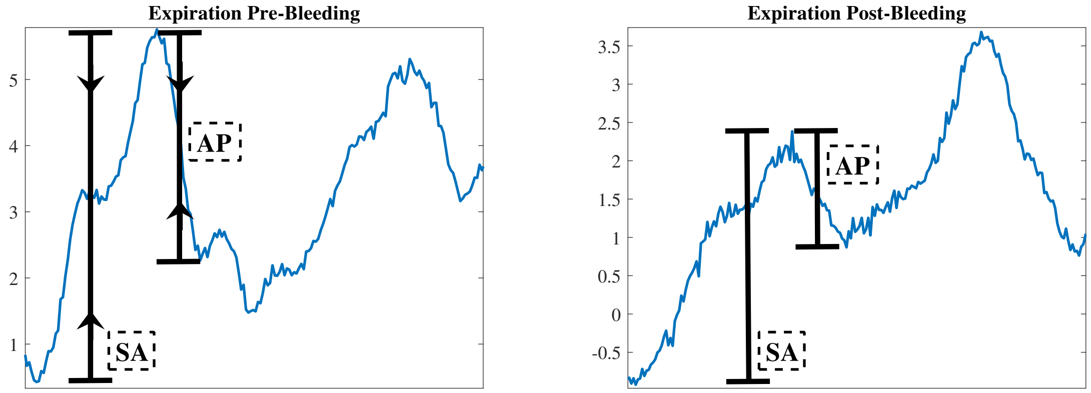


Figure 6.2: A pair of CVP waveforms from expiration before and after bleeding. The impact of certain features has been labeled on the pre-bleeding side. Arrows indicate lengths that must decrease to appear more like a waveform from after bleeding.

expiration. Correspondingly, the figure shows on the right an expiration waveform from soon after the onset of bleeding. The two characteristic waveform parameters have shrunk dramatically, and bumps and dips at A and S have substantially diminished.

Clinical relevance Many physiologic factors interact to define a given CVP or its mean change during the ventilatory cycle making these measures insensitive to changes in effective circulating blood volume as bleeding occurs except at the extremes, where such monitoring is not needed. Importantly, as depicted in Fig. 6.2, the dynamical waveform changes at end-expiration in the CVP waveform features compared to end-inspiration are very informative of dynamic changes in volume status, even if the absolute CVP values are not. This is relevant to bedside monitoring of critically ill patients for several reasons. First, CVP monitoring is common in critically ill patients because central venous catheters safely deliver fluid and drugs that cannot safely be infused by a peripheral source. Thus, its monitoring is readily available. Second, although absolute CVP values may be inaccurate for technical reasons of zero reference values, the pressure waveform datasets remain accurate, allowing their featurization for CLS and other machine learning applications that until now have been underutilized. And finally, early identification of occult bleeding would allow earlier corrective therapies to minimize or prevent hypovolemia associated tissue hypoperfusion. Such earlier interventions would markedly reduce hypoperfusion related morbidities, like acute kidney injury, ileus and secondary wound infection.

Comparison to random forest In the above experiments we compared the classification performance of the proposed method to a random forest (Breiman, 2001). Our goal was to illustrate that CLS classification was pragmatically close to state-of-the-art of methods used in clinical settings, even if just slightly worse. The advantage of the proposed method is not intended to be classification performance but rather its interpretable structure. Although the individual decision trees in a random forest are somewhat interpretable, a major difference are the types of problems for which they are suited. The proposed method is more suited for multi-view data that are hypothesized to have interesting correlations between the views, especially when those correlations differ between subsets of observations, as random forests do not incorporate any clustering mechanism.

6.4 Application to non-intrusive load monitoring

6.4.1 Introduction

This chapter covers an additional application of our classification methodology. We continue to use our methodology to investigate temporal structure of signals but in a new application. The previous source separation problem covered here, radiation detection, included exactly two kinds of signals, the source and background. Furthermore, information is often known about the source. It is interesting to consider a different source separation problem that comes with less structure. We turn to the task of non-intrusive load monitoring (NILM), in which there is an arbitrary unknown number of unknown source types. We characterize change points in an aggregated signal by relationships between past and present windows of time.

As demand grows, it is increasingly important to conserve energy for financial and environmental purposes. This problem is addressed statistically in NILM, a problem about discovering which appliances in a building are responsible for energy consumption. The aggregated power consumption is measured at the utility service entry and must be disaggregated into consumptions of individual appliances. This information could then detect anomalous consumption patterns, allow homeowners to compare their patterns to each other, help electrical utilities refine their load profiles to better forecast loads, and analyze load flexibility of different buildings. In the event-based workflow of NILM (Hart, 1992), the process is often broken down into at least three steps. First, time steps are identified at which an appliance is switched on or off. Second, the identified events are classified according to which appliance was involved. Third, the exact power usage is computed for each appliance.

In this work, we focus on the first step, event detection, and the second step, event classification (also known as load identification). In time series literature, when an abrupt change must be detected, the problem is called change detection. This domain has inspired the current state-of-the-art of NILM event detection, the chi-squared goodness-of-fit (GOF) test, which compares two consecutive windows of power to check if an event occurred between them (Jin et al., 2011b). This method makes questionable assumptions. To address these problems, we propose a novel approach that applies our multi-view classification methodology to harmonic features derived from power. The views in this problem represent times before an event and times after, so the relationship between views is temporal change. We hypothesize that our method can characterize this change in a highly discriminative fashion. Separately, detected events must be classified by appliance. Many methods have been explored in the past; see Zoha et al. (2012) for a survey. The most relevant of these simply apply a standard classifier to features derived from power. We, however, demonstrate that a more powerful classifier can be trained with our multi-view methodology under the same intuition as event detection. We consider these two problems independently, which is standard in the literature albeit impractical.

6.4.2 Related work

There is an interesting collection of literature on the event detection step of event-based NILM. Most methods are based on probabilistic models. In Shaw et al. (2002), the authors draw from change detection literature to propose generalized likelihood ratio (GLR), which checks if a time step is an event by evaluating whether it came from the window immediately before or after under an assumed distribution, typically Gaussian. A voting scheme is applied to smooth the scores (Anderson et al., 2012). GLR is often used as a baseline. Another popular baseline is the previously mentioned GOF (Jin et al., 2011b,a), often referred to as state-of-the-art. Recently, De Baets et al. (2017) argue that GOF is too sensitive to variance in base load power consumption and requires extensive parameter optimization. Accordingly, they propose a modified version of GOF, which they show to have greater performance with a higher base load. Additionally, a recent paper by Pereira and Larsys (2017) proposes to use GLR but replace the voting scheme with a locator algorithm for the extrema of scores. Another category of method is based on heuristics. For example, in Meehan et al. (2014), an event is detected by checking whether the absolute difference of power between two points exceeds a pre-defined threshold.

A significant amount of research has been done on event classification too. Sadeghianpourhamami et al. (2017) provides a survey on different techniques for feature selection. Many techniques process the P-Q plane, the series of real power (P) and reactive power (Q), using harmonic analysis. Some techniques process raw waveforms, and others VI trajectories, the series of voltage (V) and current (I). In terms of classification techniques, there have been a variety of approaches, partially surveyed by Zoha et al. (2012). Figueiredo et al. (2011) try typical classifiers, Support Vector Machine and k-Nearest Neighbors. Lin et al. (2011) apply a classifier based on Fuzzy C-Means to exploit cluster structure in crest factors. Recently, research has focused on deep learning. Chang et al. (2014) employ wavelet transforms to featurize power waveforms and feed them to a neural network for classification. Several other deep neural architectures are tested by Kelly and Knottenbelt (2015). More recent work has included Convolutional Neural Networks (De Baets et al., 2018; de Paiva Penha and Garcez Castro, 2018).

6.4.3 Baselines

GOF is referred to as the state-of-the-art in event detection (Jin et al., 2011b). Given a time series of power, it checks whether an event has occurred in a window by comparing the window to a previous window. A hypothesis test is performed to analyze whether the windows come from the same probability distribution. With a window size of n, let $\{p_i\}_1^n$ and $\{q_i\}_1^n$ be consecutive windows of power. The test statistic is

$$\sum_i \frac{(p_i - q_i)^2}{p_i},$$

which follows a χ^2_{n-1}-distribution under a Gaussian assumption on power. If the test statistic is high, there is more evidence to believe that the distributions differ and an event occurred in the second window.

The voting scheme in Anderson et al. (2012) can be applied to GOF. A window of length w_{vote} is moved over a series of scores. A vote is given to the highest score in each window. This procedure greatly reduces false positives at the expense of some true positives.

In event classification, there is no single state-of-the-art method, but many standard classifiers have competitive performance (Zoha et al., 2012). We employ a random forest on the same harmonic features as in event detection. Preliminary experiments showed k-NN, SVM, and multinomial regression probably held no advantage over random forest.

6.4.4 Proposed approach

We hypothesize CLS classification can outperform GOF in event detection. Although GOF may be considered a basic multi-view method where each window is a view, it potentially oversimplifies the multi-view structure of the problem because it only considers pairs of corresponding time steps in order. The i-th observation of a window is only compared to the i-th of the other; however, there is no physical reason for these observations to correspond. There is room for improvement if this constraint is relaxed so that multiple relationships between observations are considered. Furthermore, GOF explicitly assumes that power is normally distributed, but in real datasets this does not appear to always hold. To address these drawbacks, we replace GOF with the CLS classification method in Ch. 6. By characterizing events and non-events by correlation patterns between the windows, our method learns much more complex multi-view structure. It also does not make any assumptions about the probability distribution of power. Since GOF is applied to just power, a univariate feature, we restructure the data to be multivariate by computing harmonic features over the windows using the Fourier transform, a common featurization (Sadeghianpourhamami et al., 2017). We then select the top $n_{PC} = 5$ principal components in each window. After the classifier is run, the voting scheme is applied to the scores. Note that the harmonic features are linear combinations of power, meaning that a Gaussian assumption propagates. Thus, GOF can be applied by summing the test statistics over channels.

Furthermore, we conjecture that CLS classification can also outperform random forest in event classification. Although random forest makes no obviously problematic assumptions, it is inherently single-view; its constituent trees split on individual features at a time without analyzing how they relate to other features.

By explicitly characterizing temporal change as a multi-view relationship, we believe our method can readily capture dependencies often missed by single-view methods. Additionally, rather than use simple CLS classification, we run with many different initializations and average the class scores across models, building an ensemble of classifiers. This technique reduces variance and empirically improved performance in this task.

6.4.5 Experiments

The dataset is BLUED (Filip, 2011), a collection of high-frequency electricity data in one building. The dataset is commonly used for benchmarking. It contains measurements of current and voltage on two phases A and B at 12kHz from which we compute active power. Here we select phase B because it is has many appliances that frequently overlap. Ground truth events are provided. We compute the discrete Fourier transform (DFT) from 12kHz power, but we evaluate the methods at 60Hz in testing.

Event detection

We compare our method to GOF using both 12kHz power and DFT. Positive samples are from time steps containing events, while negatives are one-second blocks separated by three seconds on either end of each event. Results are shown in Fig. 6.3 and Tbl. 6.2. Our method performs better at all levels of TPR with statistical significance, where randomness is over model initialization. This example shows how our method outperforms a multi-view alternative because of its less restrictive assumptions.

Table 6.2: Comparison of FPR of event detection methods on BLUED dataset.

TPR	80%	85%	90%	95%	98%
GOF on power	.10%	.15%	.50%	.81%	5.46%
CLS on DFT	**.08%**	**.09%**	**.26%**	**.67%**	**3.92%**

Event classification

We compare our method to random forest on harmonic features. We subsample the events to the most common appliances because many appliances have very

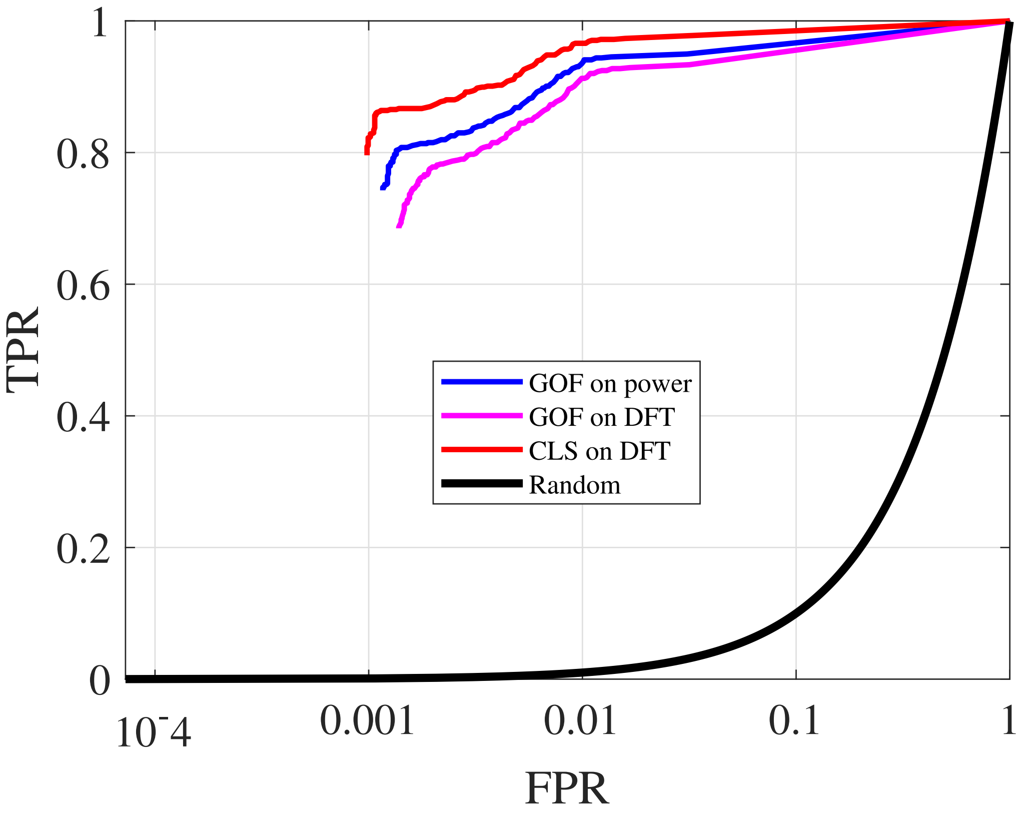


Figure 6.3: ROCs of event detection methods on BLUED dataset.

few samples. Classes are defined by both appliance type and direction of the switch—increase or decrease in total power. We use the M top classes, where $M \in [4 \mathinner{..} 10]$. Since we consider event classification in a vacuum, as is standard in the literature (Zoha et al., 2012), the data lack complications such as false positive events and unknown classes in test data. Results are shown in Fig. 6.4 and Tbl. 6.3. Our method performs better in F_1, precision, and recall (uniformly averaged over class) at every value of M with statistical significance, where randomness is over model initialization and subsampling. This example reveals how the multi-view nature of our method provides an advantage over single-view methods.

Next we present a case analysis of a mistake made by random forest but not by CLS classification. In Fig. 6.5 there are features of three examples with five features in each view. The examples (a) and (b) are from the same class while (c) is from a different class. CLS correctly classifies both (a) and (b), but random forest misclassifies (a) as the same class as (c). The features of (a) and (b) are similar and

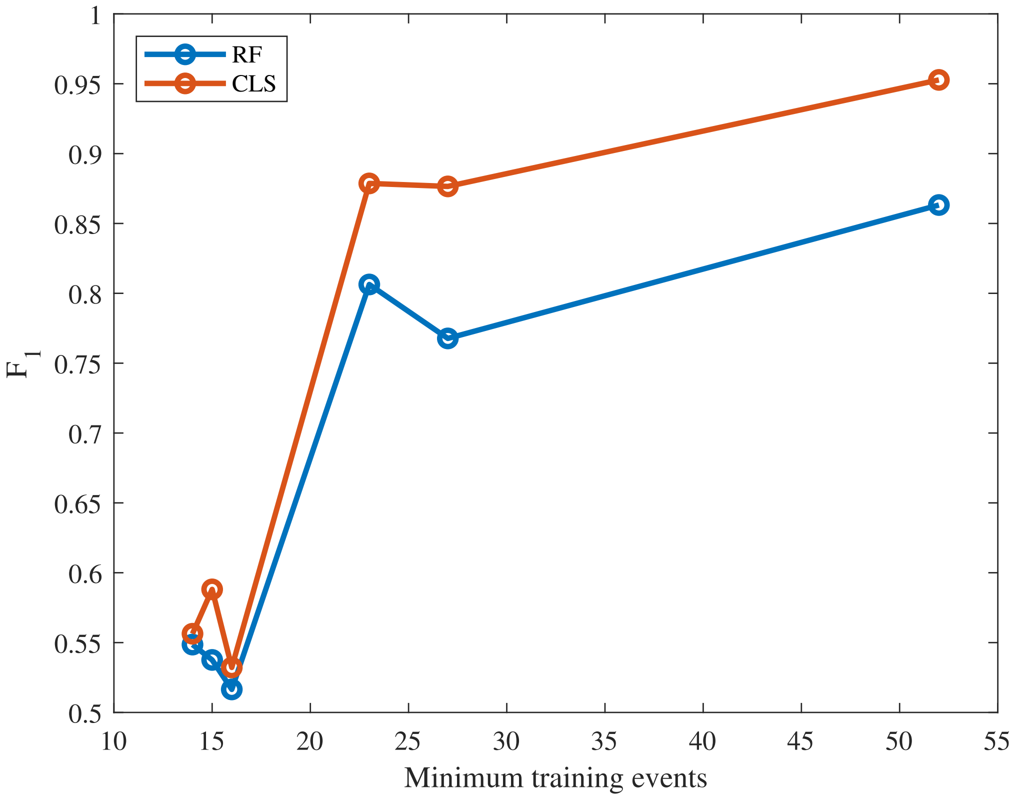


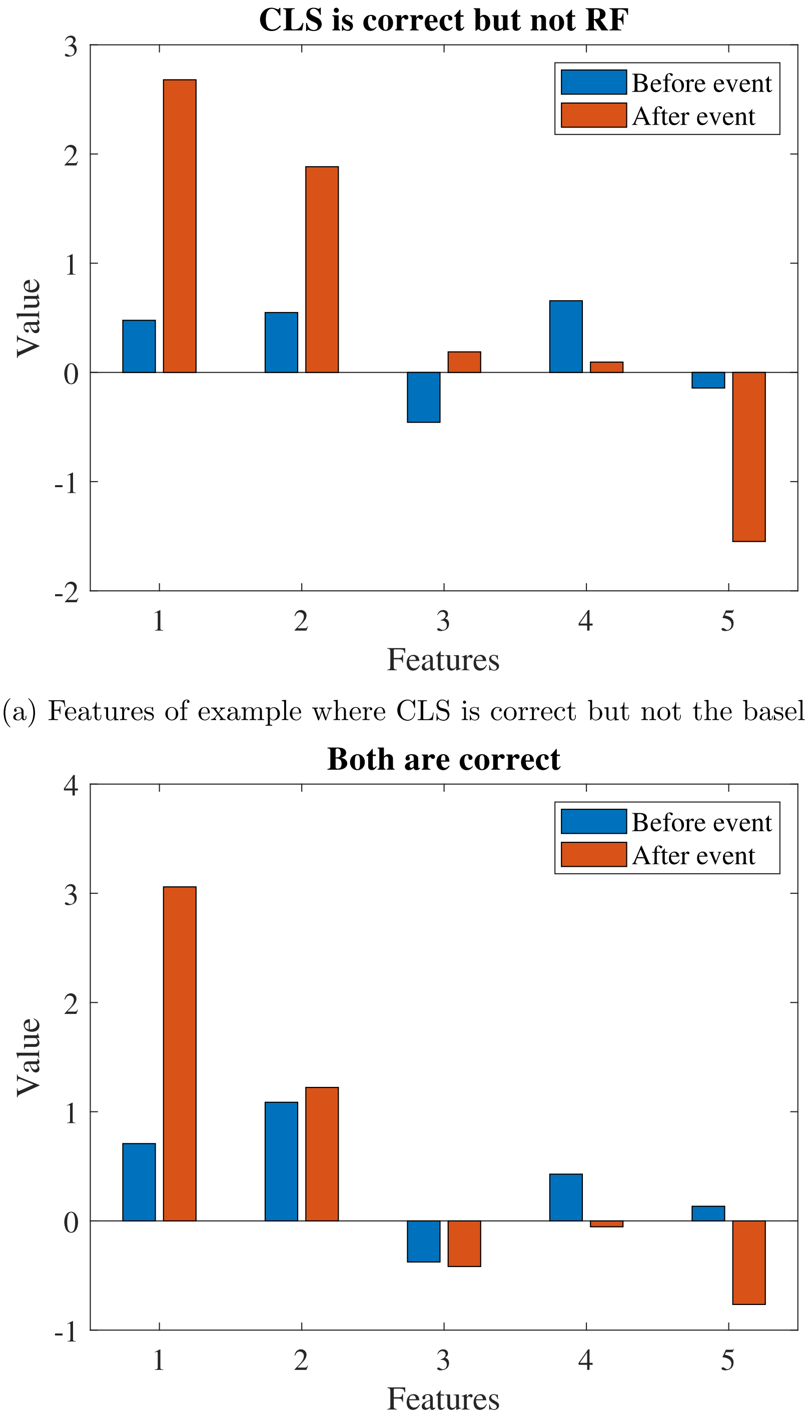
Figure 6.4: Comparison of methods' F_1 as the number of samples to include a class grows.

have the same multi-view relationship, which allows CLS to classify them properly. However, the value of Feature 2 in the "After" view differs from about 2 to about 1. It turns out that many decision trees in the random forest quickly classify examples with this value of Feature 2 as the class of (c) because it is an effective rule for that class, as demonstrated by Fig. 6.5c. The random forest can make this mistake because it only looks at individual features and not at the relationship between views.

6.4.6 Discussion

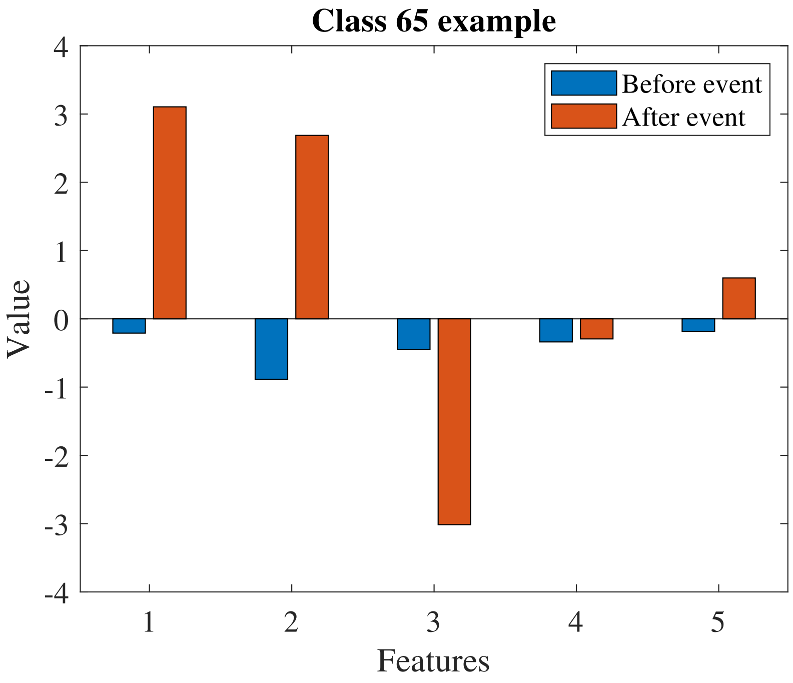
Performance of our method

We have demonstrated our method of CLS classification to outperform baselines in event detection and classification. The method performs better than the multi-view baseline in detection because it learns much more sophisticated multi-view structure.



(a) Features of example where CLS is correct but not the baseline.

(b) Features of example where both CLS and the baseline are correct.



(c) Features of example in class that the baseline mistakenly assigned the example above in Fig. 6.5b.

Figure 6.5: Features of three examples, with five features in each view. The top two are the same class and appear similar. The bottom is a different class and is only similar in the first two features of the ”After” view.

It also performs better than the single-view baseline in classification because it can identify useful changes in temporal structure through multi-view analysis. As detection and classification comprise core components of event-based NILM, these results suggest that our method might help advance the state of the art. Furthermore, this work supports the hypothesis that classification can be approached by using multi-view relationships as a unit of analysis to leverage structure that other methods struggle to find.

The results highlight the advantage of a multi-view treatment of event classification. Our method models dynamics of temporal change through views of "before" and "after," a well-suited treatment for frequently sharp or step changes when appliances are switched. This intuition inspires potential augmentations to random forest to improve its ability to exploit these dynamics. For example, each decision tree could be constrained to consider features that alternate between views or to always select the corresponding feature from the other view. In addition, the random forest could be given a features that represent multi-view relationships, such as the CLS projections in each cluster.

End-to-end evaluation

In a practical scenario detection and classification should be evaluated end-to-end. The ideal evaluation would be to reconstruct a binary signal of whether each device is on or off and treat it as a series of classification tasks. However, in this work we consider the problems independently because numerous practical issues arise that we consider out-of-scope. First, many devices have state spaces more complex than

Table 6.3: Classification performance as the number of events to include a class grows. Asterisk (*) denotes statistically significant advantage. Note F_1 is not tested.

Min. training events	Num. classes	RF F_1	**CLS $\mathbf{F_1}$**	RF Precision	**CLS Precision**	RF Recall	**CLS Recall**
14	10	55	**56**	56	**66***	55	**56**
15	9	54	**59**	55	**70***	53	**58**
16	8	52	**53**	53	**60**	51	**54**
23	6	81	**88**	82	**89***	80	**89***
27	5	77	**88**	78	**87***	76	**90***
52	4	86	**95**	89	**97***	85	**94***

just on or off, such as a fan with multiple speeds. Many common appliances have this problem. Moreover, these states are not labeled in the data, so one would need to propose an approach to estimate the states. Second, detection will produce false positives that are then fed to classification, and true positives may be unknown devices. A method must be designed to purge these anomalies; we believe a good approach is to insert an independent anomaly detector after classification for each class, but there may be better ways. Third, the ON and OFF events predicted by classification do not always come in pairs, which would be expected in appliances with only two states. A simple fix would be to only consider adjacent pairs of ON and OFF events, but doing so would often ignore many events that are correctly detected but misclassified. A rigorous treatment for this problem is suggested in Giri and Bergés (2017). Due to the scope and difficulty of these three issues, we opt to leave end-to-end evaluation as future work.

VI trajectory

Our multi-view approach to NILM focused on temporal views. Nonetheless, there may be other viable approaches. According to Sadeghianpourhamami et al. (2017), some researchers have used features based on the P-Q plane or VI trajectory. These features would be naturally multi-view. For example, the VI trajectory is circular for phase shifted power and current. We could imagine that different combinations of appliance states could result in different circles. Then a multi-view clustering method might be an ideal candidate to identify these states.

6.5 Conclusion

Like the previous chapter, this work considered the problem of discovering multi-view structure in complex datasets. We proposed a classification method that learns correlation clusters for multi-view data on each class individually and makes classifications based on which model fits better. The method was tested on CVP waveform datasets of induced bleeding and in NILM on event detection and classification. Experiments demonstrated it to perform well because it leveraged relations between views. In essence, these results illustrate that cluster-wise nonlinear multi-view relationships can be employed as discriminative factors in classification.

In the future it would be relevant to investigate why our method performs in terms of the original power signal. For instance, it would be interesting to check if some events corresponded to certain changes in power that were easier or harder to detect with our method. One way to do so would be to perform a gradient analysis on the classification score to measure the impact of different features, which correspond to principal components of the Fourier transform. The components could then be projected back to the original feature space. Alternatively, an interpretable classifier could be fit to the binary variable of whether our method is correct while random forest is not. Its decisions could then be interpreted in terms of features of the original power.

Additionally, we could attempt to try other distribution tests to replace the baseline GOF test such as Maximum Mean Discrepancy (Borgwardt et al., 2006), a distribution-free test that does not make the same problematic assumptions.

Another interesting future avenue would be to obtain theoretical results about the performance of this method. In the empirical results here, we explained how our method leveraged multi-view structure, but we did not attempt to understand the underlying reason our method could perform better than baselines rather than just by operating on different statistical properties. A challenging but important research topic would be to quantify the strength or discriminativeness of multi-view relationships to theoretically bound the effectiveness of this method.

Chapter 7

Conclusion

In multi-view learning, common approaches lack a layer of abstraction to truly separate themselves from single-view methods. To attempt to provide this layer, this thesis initially postulated that it is possible to characterize multi-view relationships and employ them as units of analysis in descriptive analytics and inference. To this end, our work proposed machine learning algorithms based on these multi-view relationships, which we demonstrated to discover novel structure or have competitive empirical performance with the state of the art. We tested our methods on a variety of domains with significant practical ramifications.

7.1 Contributions

Now we summarize our key contributions. This thesis was divided into two parts. In Part I considered a specific application in which the relationship between views was known from domain knowledge. This work was complemented by Part II, which showed how to learn unknown multi-view relationships from data.

Starting off Part I, Ch. 2 introduced the problem of gamma source detection, where the principal statistical challenge was low signal-to-noise ratio. We raised concerns with the standard assumption of training data or a warm-up period: source contamination, need for immediate detection, and mismatch between training and background. These issues could severely degrade performance in practical scenarios with commonly used methods. Consequently, we proposed a method based on the Kalman filter to reduce dependence on learning from training or warm-up data. The filter exploited smoothness in radiation dynamics to simultaneously estimate source intensity and background. Experiments demonstrated that our method was robust to background variation and low amounts of training. Thus, this method could improve detection performance in a variety of situations in which training is undesirable.

That work was extended in Ch. 3, which brought multiple views into the problem through multiple sensors. Because the location over time of each sensor was known, there existed a multi-view relationship between them that involved their relative proximities to the source. Our goal was to leverage this relationship, a task ignored by the state-of-the-art baseline, Bayesian Aggregation (BA). Although BA already offered a way to aggregate observations, it did not utilize the information that certain observations corresponded to different sensors. Instead, it assumed that each observation was independent, a likely false assumption. To leverage their contemporaneous multi-view relationships, we constructed a multi-view filter, named the Bayesian Aggregation Filter, based on our previous Kalman filter detector. The filter shared information between sensors at every time step to refine their inferences. The information was shared by utilizing relationships between views known from domain knowledge. This method had equal or superior empirical performance to the BA, all with little to no training data, rendering it a practical and effective choice in multi-sensor situations.

To complement the ideas in the first part, Part II considered the case in which multi-view relationships were unknown, establishing how to learn and perform analytics and inference on them. It began by modeling and applying linear relationships in Ch. 4. These relationships were computed by Canonical Correlation Analysis (CCA). We used them to devise an anomaly detection method, which we evaluated on the gamma source detection task. The experiments simulated imperfect source information, violating a standard assumption to model a practical scenario, as well as sensor miscalibration. The problem was treated as multi-view because of energy windows, energy ranges in which a source was expected to be seen most clearly. Energy windows were considered more robust to imperfect information, so they were a natural fit for a multi-view environment. The results demonstrated that detection utilizing multiple linear multi-view relationships was more robust to imperfect information than baselines. These scenarios corresponded to a variety of practical situations. This study also demonstrated that multi-view relationships could effectively serve as discriminative factors in classification.

Then we extended our approach toward multi-view analysis to nonlinear relationships in Ch. 5 because there could be many datasets in which a linear model of relationships would not suffice. We considered unsupervised learning based on these relationships. Our intuition was that different subsets of observations could exhibit distinct linear relationships. To generalize CCA nonlinearly, we modeled nonlinearity as clusters of linear relationships, intuitively similar to a mixture model of latent variable models. We proposed the Canonical Least Squares (CLS) Clustering algorithm to learn these clusters and relationships simultaneously. The key characteristic of this approach was that it defined cluster variables by linear multi-view relationships rather than spatial relationships. Thus, the clusters it discovered could potentially be radically different from almost all other clustering methods. This work was evaluated quantitatively on synthetic data and demonstrated good performance. In addition, we employed it on a medical dataset and discovered qualitatively useful insights about physiological behavior. In particular, our work was possibly the first to showcase the utility of a type of blood pressure in characterization of hemodynamic stability, which could be useful and interesting to practitioners.

Lastly, in Ch. 6, we aimed to apply our multi-view framework to supervised learning. We proposed a classification method using our clustering method as a foundation. Hypothesizing that multi-view relationships could be employed as discriminative factors between classes, our method learned CLS clusterings on each class. We classified new observations by checking which class fit best with a philosophy similar to the nearest-neighbor algorithm, except with multi-view relationships rather than spatial features. This method could be considered a nonlinear multiclass generalization of CCA detection. The method was evaluated on the medical dataset and showed fair quantitative performance. Additionally, it was deployed on two tasks in non-intrusive load monitoring and performed better than the state of the art in both on a commonly benchmarked dataset. These results illustrated that multi-view relationships could effectively characterize different classes compared to modern single-view methods that operated on spatial relationships. We included case analysis to show how our method leveraged multi-view structure to avoid mistakes made by alternatives because select features from a single view could appear as a different class while the multi-view relationship was more robust.

In sum, our work established the utility of multi-view relationships as units of analysis. Using relationships known through domain knowledge, we improved the state of the art in an applied domain by performing inference on explainable latent variables. Then we showed how to fit unknown relationships from data. We presented empirical evidence that such relationships can characterize useful structure in unsupervised learning and serve as discriminative factors in supervised learning, exploiting a severely underutilized property of multi-view data.

7.2 Future work

Our work exposes many additional intriguing questions for future consideration. First, in Part II although we considered a few specific application domains, our methodology can easily be applied to many fields. For instance, our multi-view approach to time series analysis can be used in virtually any temporal data. The method for event detection in NILM can be carried over to any type of change detection. To recap the method, views are represented by windows before and after each time step in order to characterize dynamics of temporal change. Features are computed for each window through various time series featurizations such as the Fourier transform. Then we can apply our framework to arbitrary classification tasks. Additionally, our clustering procedure from medical data can be generally utilized to segment multiple time series temporally and cross-sectionally, assuming temporal alignment. A potential domain to try both procedures would be in financial markets. Change detection could be applied to classify time steps as market crashes, while clustering could be used to find similar patterns of behavior across industry and time. For example, one could investigate parsimony and phenotypical structure among patterns of response or various stocks or industries to change-points in the market.

Also, we presented methods for clustering and classification; conspicuously missing is regression. One approach might be to modify CLS classification to predict continuous output; however, it is difficult to see a natural way to do so. Instead, it may be possible to extend our ideas to regression by a different strategy, though one that still leverages a mixture of linear multi-view relationships. We assume that each mixture component expresses a multi-view relationship in the inputs and a corresponding regression function. A natural choice of mechanism under these conditions is the Mixture of Experts (Jacobs et al., 1991), which learns a mixture of regression functions f_j on inputs z_i and a gating network g to assign mixture weights to each point. Many forms can be selected for the f_j such as linear regression. An important change would be that to assign mixture weights, rather than feed each point z_i directly to the gating network as $g(z_i)$, we would apply the gating network to the multi-view relationship expressed by multiple points. For example, if each z_i contained two views x_i and y_i, we would compute the cross-correlation matrix C between the views of a set of points $\{z_i\}_{i \in S_r}$ given by an index set S_r. Then all these points would be assigned the same mixture weights $g(C)$. This change corresponds to a representation of multi-view relationships using correlation matrices rather than latent projections, but the main idea of the method is still to find clusters based on multi-view relationships. The sets S_r would cover all input indices and be predefined in the data. For instance, in time series data they could be contiguous windows of points. One of the main challenges here would be to simplify this representation of relationships because of the relatively high number of data points demanded to compute the correlation matrix.

Next, there are avenues for theoretical contributions. For instance, although we know how our CLS clustering methodology works—by operating on cluster-wise linear multi-view relationships—we currently do not have any theoretical justification for why it could perform better or worse on different datasets. An interesting project would be to quantify the strength of multi-view relationships. For example, if we assume the data are generated by a mixture model similar to a mixture of CCA, we might consider stochastic noise added to the points that reduces the discriminative power of the multi-view relationships. Say one mixture component generates observed random variables X and Y from a latent variable Z as $X = f(Z) + U$ and $Y = g(Z) + V$ where U and V are stochastic noise. We could attempt to bound the performance of CLS clustering in terms of $\mathrm{Var}(U)$ and $\mathrm{Var}(V)$. A challenge in this problem would be to connect this probabilistic model to the distribution-free characteristics of our work.

On a related note, the applicability of our framework could be investigated. Udell and Townsend (2019) prove that observed data have a low-rank latent factorization under relatively general conditions. They consider a latent variable model in which random variables in two latent spaces are mapped by an arbitrary function to the observed space X. Under fairly relaxed conditions about boundedness and smoothness, the authors prove any $n \times d$ matrix of observations in X has approximate rank $O(p(\log(n + d)))$ where p is a finite degree polynomial. We propose to examine whether this relationship can be inverted to show that latent variables have multiple observed factorizations. In particular, we could consider a latent variable model in which random variables in one latent space are mapped to to multiple observed spaces. The aim would be to prove a matrix of these latent variables would have logarithmic rank in the dimensions of the observed space. If so, it would suggest that meaningful multi-view relationships can be commonly found in real data.

Also, our work in Part II is limited to two views. To truly claim generality, it would need to be extended to an arbitrary number. A possible approach to do so is highlighted by Horst (1961) with Generalized Canonical Correlation Analysis (GCCA), whose main idea is to learn canonical covariates shared by all views. More precisely, let X_j be the data matrix of the j-th view. Then GCCA solves

$$\max_{U_j,G} \sum_j \|G - U_j^\top X_j\|_{\mathcal{F}}^2.$$

This problem can be solved by an eigendecomposition in a similar manner to classical CCA. We could conceivably replace CCA in our methods with GCCA. Also, it might be feasible to generalize CLS in in the same fashion. Doing so would extend our methodology to more than two views.

An additional line of questioning would be to connect this work to mainstream machine learning. Our framework could be compared to well-known baselines by applying it to common multi-modal problems. Especially popular in this area is deep learning, which is well-suited for vision, text, and audio. Multi-modal data are trivially multi-view, and there exist many interesting problems such as audiovisual speech classification (Ngiam et al., 2011), object recognition from captions on images (Srivastava and Salakhutdinov, 2014; Eitel et al., 2015), and emotion recognition in video (Kahou et al., 2016). Furthermore, many current multi-view classification methods are tailored to these data types. These tasks represent the best opportunity to compare the more specific but well-known methods to our framework. It must be understood, however, that our work fundamentally differs in concept because virtually all these other classifiers leverage multiple views by finding agreement between them, not by explicitly analyzing relations between them, so they operate on separate properties of data that could often disagree.

Furthermore, it should be possible to generalize our framework for learning multi-view relationships to arbitrary nonlinear relationships rather than cluster-wise linear ones. A possibility would be to replace Pearson correlation in CCA, which only measures linear correlation, with a measure such as mutual information. Then we would consider an optimization problem such as

$$\max_{u,v} I(Xu; Yv)$$

where $I(A; B)$ is the mutual information between random variables A and B. Such a setting would also easily extend to more than two views. Alternatively, we could map features nonlinearly to higher dimensions by kernel methods such as Kernel CCA (Akaho, 2006) or deep learning methods such as Deep CCA (Andrew et al., 2013). Similar to our idea for GCCA, we could replace CCA with these nonlinear variants in our methodology. The result would have the ability to model general nonlinear multi-view structure, eliminating the need to cluster. Principal challenges would include overfitting and lack of interpretability.

Appendix A

Kalman filter updates

The updates for the Kalman filter in Ch 2 are as follows. Let $\hat{x}_{t|t}$ be the *a posteriori* state estimate given observations $\{y_1, \ldots, y_t\}$. Let $P_{t|t}$ be the *a posteriori* error covariance matrix of $\hat{x}_{t|t}$ given the same observations. Then

$$\hat{x}_{t|t-1} = \hat{x}_{t-1|t-1},$$

$$\begin{aligned} P_{t|t-1} &= P_{t-1|t-1} + Q_t, \\ K_t &= P_{t|t-1} C^\top (R_t + C P_{t|t-1} C^\top)^{-1}, \\ \hat{x}_{t|t} &= \hat{x}_{t|t-1} + K_t (y_t - C \hat{x}_{t|t-1}), \\ P_{t|t} &= P_{t|t-1} - K_t C P_{t|t-1}. \end{aligned}$$

Appendix B

Rebinning algorithm

Here we describe our method for shifting energy bins of counts to simulate gain drift in Ch. 2 and Ch. 4. The inputs are the counts data, the energy frequencies of the current bin boundaries, and the energy frequencies of the desired bin boundaries. The output is the counts data adjusted for the new bins. The method is simply to linearly interpolate counts where new boundaries intersect old ones. For example, if the old boundaries are [0, 10, 20] and the new boundaries are [0, 7, 14, 21], then 70% of counts in the old [0, 10] bin are moved to the new [0, 7] bin. The remaining 30%, plus 40% of the old [10, 20] bin, are moved to the new [7, 14] bin. The same routine is applied to the remaining bins. In the experiments here, the desired bin boundaries are set to the original boundaries multiplied by the gain drift coefficient.

Appendix C

Energy window algorithm

Here we give the algorithm to compute the energy window in Ch. 4.

Algorithm 2 Energy window computation

```
procedure OPTIMIZE(source template $s \in \mathbb{R}^d$, mean background spectrum $\mu \in \mathbb{R}^d$)
    for $k = 1, \ldots, d$ do
        $r_k \leftarrow s_k / \mu_k$
    for $k = 1, \ldots, d$ do
        $\mathcal{C} \leftarrow$ set of bins with $k$ highest $r_j$ values
        $S_k \leftarrow \sum_{j \in \mathcal{C}} s_j$
        $B_k \leftarrow \sum_{j \in \mathcal{C}} \mu_j$
    $k^* \leftarrow \operatorname{argmax} S_k / \sqrt{B_k}$
    return set of bins with $k^*$ highest $r_j$ values
```

Bibliography

Aage, H. K. and Korsbech, U. (2003). Search for lost or orphan radioactive sources based on NaI gamma spectrometry. *Applied Radiation and Isotopes*, 58(1):103–113.

Akaho, S. (2006). A kernel method for canonical correlation analysis. *arXiv preprint cs/0609071.*

Akesson, B. M., Jorgensen, J. B., Poulsen, N. K., and Jorgensen, S. B. (2008). A generalized autocovariance least-squares method for Kalman filter tuning. *Journal of Process Control*, 18(7-8):769–779.

Anderson, K. D., Berges, M. E., Ocneanu, A., Benitez, D., and Moura, J. M. (2012). Event detection for non intrusive load monitoring. *IECON Proceedings (Industrial Electronics Conference)*, pages 3312–3317.

Anderson, K. K., Jarman, K. D., Mann, M. L., Pfund, D. M., and Runkle, R. C. (2008). Discriminating nuclear threats from benign sources in gamma-ray spectra using a spectral comparison ratio method. *Journal of Radioanalytical and Nuclear Chemistry*, 276(3):713–718.

Andrew, G., Arora, R., Bilmes, J., and Livescu, K. (2013). Deep canonical correlation analysis. *International Conference on Machine Learning.*

Anscombe, F. J. (1948). The Transformation of Poisson, Binomial and Negative-Binomial Data. *Biometrika Trust*, 35:246–254.

Aucott, T. J., Bandstra, M. S., Negut, V., Curtis, J. C., Chivers, D. H., and Vetter, K. (2014). Effects of background on gamma-ray detection for mobile spectroscopy and imaging systems. *IEEE Transactions on Nuclear Science*, 61(2):985–991.

Bach, F. R. and Jordan, M. I. (2006). A probabilistic interpretation of canonical correlation analysis. *Dept Statist Univ California Berkeley CA Tech Rep*, 688:1–11.

Bai, E. W., Chan, K. S., Eichinger, W., and Kump, P. (2011). Detection of radionuclides from weak and poorly resolved spectra using Lasso and subsampling techniques. *Radiation Measurements*, 46(10):1138–1146.

Bandstra, M. S., Aucott, T. J., Brubaker, E., Chivers, D. H., Cooper, R. J., Curtis, J. C., Davis, J. R., Joshi, T. H., Kua, J., Meyer, R., Negut, V., Quinlan, M., Quiter, B. J., Srinivasan, S., Zakhor, A., Zhang, R., and Vetter, K. (2016). RadMAP: The Radiological Multi-sensor Analysis Platform. *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 840(May):59–68.

Bavdekar, V. A., Deshpande, A. P., and Patwardhan, S. C. (2011). Identification of process and measurement noise covariance for state and parameter estimation using extended Kalman filter. *Journal of Process Control*, 21(4):585–601.

Bickel, S. and Scheffer, T. (2004). Multi-view clustering. In *IEEE International Conference on Data Mining*, number December 2004, pages 19–26.

Bilton, K. J., Joshi, T. H., Bandstra, M. S., Curtis, J. C., Quiter, B. J., Cooper, R. J., and Vetter, K. (2019). Non-negative matrix factorization of gamma-ray spectra for background modeling, detection, and source identification. *IEEE Transactions on Nuclear Science*, 66(5):1–1.

Blaschko, M. B. and Lampert, C. H. (2008). Correlational spectral clustering. *Computer Vision and Pattern Recognition*.

Borgwardt, K. M., Gretton, A., Rasch, M. J., Kriegel, H. P., Schölkopf, B., and Smola, A. J. (2006). Integrating structured biological data by Kernel Maximum Mean Discrepancy. *Bioinformatics*, 22(14):49–57.

Boyd, J. H., Forbes, J., Nakada, T.-a., Walley, K. R., and Russell, J. A. (2011). Fluid resuscitation in septic shock: a positive fluid balance and elevated central venous pressure are associated with increased mortality. *Critical care medicine*, 39(2):259–265.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Chang, H. H., Lian, K. L., Su, Y. C., and Lee, W. J. (2014). Power-spectrum-based wavelet transform for nonintrusive demand monitoring and load identification. *IEEE Transactions on Industry Applications*, 50(3):2081–2089.

Chaudhuri, K., Kakade, S., Livescu, K., and Sridharan, K. (2009). Multi-view clustering via canonical correlation analysis. In *International Conference on Machine Learning*, pages 1–8.

Damman, K., van Deursen, V. M., Navis, G., Voors, A. A., van Veldhuisen, D. J., and Hillege, H. L. (2009). Increased central venous pressure is associated with impaired renal function and mortality in a broad spectrum of patients with cardiovascular disease. *Journal of the American College of Cardiology*, 53(7):582–588.

De-Arteaga, M., Dubrawski, A., and Huggins, P. (2015). Canonical Autocorrelation Analysis. *arXiv preprint*.

De Baets, L., Ruyssinck, J., Develder, C., Dhaene, T., and Deschrijver, D. (2017). On the Bayesian optimization and robustness of event detection methods in NILM. *Energy and Buildings*, 145:57–66.

De Baets, L., Ruyssinck, J., Develder, C., Dhaene, T., and Deschrijver, D. (2018). Appliance classification using VI trajectories and convolutional neural networks. *Energy and Buildings*, 158:32–36.

de Paiva Penha, D. and Garcez Castro, A. R. (2018). Home appliance identification for NILM systems based on deep neural networks. *International Journal of Artificial Intelligence & Applications*, 9(2):69–80.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 39(1):1–38.

Du, Q., Wei, W., May, D., and Younan, N. H. (2010). Noise-adjusted principal component analysis for buried radioactive target detection and classification. *IEEE Transactions on Nuclear Science*, 57(6 PART 2):3760–3767.

Eitel, A., Springenberg, J. T., Spinello, L., Riedmiller, M., and Burgard, W. (2015). Multimodal deep learning for robust RGB-D object recognition. *IEEE International Conference on Intelligent Robots and Systems*, 2015-Decem:681–687.

Fagan, D. K., Robinson, S. M., and Runkle, R. C. (2012). Statistical methods applied to gamma-ray spectroscopy algorithms in nuclear security missions. *Applied Radiation and Isotopes*, 70(10):2428–2439.

Fern, X. and Friedl, M. (2005). Correlation clustering for learning mixtures of canonical correlation models. In *SIAM International Conference on Data Mining*, pages 439–446.

Figueiredo, M. B., Almeida, A. D., and Ribeiro, B. (2011). An experimental study on electrical signature identification of non-intrusive load monitoring (NILM) systems. *International Conference on Adaptive and Natural Computing Algorithms*.

Filip, A. (2011). BLUED : A fully labeled public dataset for event-based non-intrusive load monitoring research. *2nd Workshop on Data Mining Applications in Sustainability*.

Giri, S. and Bergés, M. (2017). An error correction framework for sequences resulting from known state-transition models in Non-Intrusive Load Monitoring. *Advanced Engineering Informatics*, 32:152–162.

Hardoon, D. R., Szedmak, S., and Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664.

Hart, G. W. (1992). Nonintrusive appliance load monitoring system. *Proceedings of the IEEE*, 80(12):1870–1891.

Hathaway, R. J. and Bezdek, J. C. (1993). Switching regression models and fuzzy clustering. *IEEE Transactions on Fuzzy Systems*, 1(3):195–204.

Horst, P. (1961). Generalized canonical correlations and their applications to experimental data. *Journal of Clinical Psychology*, 17(4):331–347.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3/4):321–377.

Huggins, P., Jin, J., Dubrawski, A., Labov, S., and Nelson, K. (2014). Using Gaussian rate priors with Poisson data likelihoods for improved detection of sources of known types in cluttered background scenes.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87.

Jin, Y., Tebekaemi, E., Berges, M., and Soibelman, L. (2011a). A time-frequency approach for event detection in non-intrusive load monitoring. *Signal Processing, Sensor Fusion, and Target Recognition XX*, 8050(May 2011):80501U.

Jin, Y., Tebekaemi, E., Berges, M., and Soibelman, L. (2011b). Robust adaptive event detection in non-intrusive load monitoring for energy aware smart facilities. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, (May 2016):4340–4343.

Kahou, S. E., Bouthillier, X., Lamblin, P., Gulcehre, C., Michalski, V., Konda, K., Jean, S., Froumenty, P., Dauphin, Y., Boulanger-Lewandowski, N., Chandias Ferrari, R., Mirza, M., Warde-Farley, D., Courville, A., Vincent, P., Memisevic, R., Pal, C., and Bengio, Y. (2016). EmoNets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, 10(2):99–111.

Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35.

Kelly, J. and Knottenbelt, W. (2015). Neural NILM. pages 55–64.

Klami, A. and Kaski, S. (2008). Probabilistic approach to detecting dependencies between data sets. *Neurocomputing*, 72(1):39–46.

Kumar, A., Anel, R., Bunnell, E., Habet, K., Zanotti, S., Marshall, S., Neumann, A., Ali, A., Cheang, M., Kavinsky, C., et al. (2004). Pulmonary artery occlusion pressure and central venous pressure fail to predict ventricular filling volume, cardiac performance, or the response to volume infusion in normal subjects. *Critical care medicine*, 32(3):691–699.

Kumar, A., Rai, P., and Daumé, H. (2011). Co-regularized multi-view spectral clustering. *Neural Information Processing Systems*, pages 1413–1421.

Kump, P., Bai, E. W., Chan, K. S., and Eichinger, W. (2013). Detection of shielded radionuclides from weak and poorly resolved spectra using group positive RIVAL. *Radiation Measurements*, 48(1):18–28.

Labov, S. and Nelson, K. (2019). Private communication.

Lei, E., Miller, K., Labov, S., Nelson, K., and Dubrawski, A. (2016). Radiological threat detection for an unknown energy window by canonical correlation analysis. *Nuclear Science Symposium*.

Lei, E., Miller, K., Labov, S., Nelson, K., and Dubrawski, A. (2017a). Robust detection of radiation threat by simultaneous estimation of source intensity and background. *Nuclear Science Symposium*.

Lei, E., Miller, K., Pinsky, M., and Dubrawski, A. (2017b). Bleeding detection by multi-view correlation clustering of central venous pressure. *Workshop for Machine Learning for Healthcare at NeurIPS*.

Lei, E., Miller, K., Pinsky, M., and Dubrawski, A. (2019). Characterization of multi-view hemodynamic data by learning mixtures of multi-output regressors. *International Symposium on Intensive Care and Emergency Medicine*.

Lin, Y. H., Tsai, M. S., and Chen, C. S. (2011). Applications of fuzzy classification with fuzzy c-means clustering and optimization strategies for load identification in NILM systems. *IEEE International Conference on Fuzzy Systems*, (May):859–866.

Liu, J., Wang, C., Gao, J., and Han, J. (2013). Multi-view clustering via joint nonnegative matrix factorization. In *SIAM International Conference on Data Mining*, pages 252–260.

Marik, P. E. and Cavallazzi, R. (2013). Does the central venous pressure predict fluid responsiveness? an updated meta-analysis and a plea for some common sense. *Critical care medicine*, 41(7):1774–1781.

Meehan, P., McArdle, C., and Daniels, S. (2014). An efficient, scalable time-frequency method for tracking energy usage of domestic appliances using a two-step classification algorithm. *Energies*, 7(11):7041–7066.

Mehra, R. (1970). On the identification of variances and adaptive Kalman filtering. *IEEE Transactions on Automatic Control*, AC-15(2):175–184.

Michard, F. and Teboul, J.-L. (2000). Using heart-lung interactions to assess fluid responsiveness during mechanical ventilation. *Critical Care*, 4(5):282.

Miller, K., Huggins, P., Labov, S., Nelson, K., and Dubrawski, A. (2016). Evaluation of coded aperture radiation detectors using a Bayesian approach. *Nuclear Instruments and Methods in Physics Research, Section A*, 839:29–38.

Nelson, K. and Labov, S. (2009). Detection and alarming with sords unimaged data: Background data analysis. *Lawrence Livermore National Lab Technical Report*.

Nelson, K. and Labov, S. (2010). Aggregation of mobile radiation data. *Lawrence Livermore National Lab Technical Report*.

Nelson, K. and Labov, S. (2012). Aggregation of mobile data. *Lawrence Livermore National Lab Technical Report*, 2.2(1):2–3.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. (2011). Multimodal deep learning. *International Conference on Machine Learning*, page 1.

Nie, F., Zeng, Z., Tsang, I., Xu, D., and Zhang, C. (2011). Spectral embedded clustering: A framework for in-sample and out-of-sample spectral clustering. *IEEE Transactions on Neural Networks*, 22(11):1796–1808.

Odelson, B. J., Rajamani, M. R., and Rawlings, J. B. (2016). A new autocovariance least-squares method for estimating noise covariances. In *Texas-Wisconsin Modeling and Control Consortium*, volume 28, pages 391–397.

Ottersten, B., Stoica, P., and Roy, R. (1998). Covariance matching estimation techniques for array signal processing applications. *Digital Signal Processing*, 8(3):185–210.

Pereira, L. and Larsys, M.-i. (2017). Developing and evaluating a probabilistic event detector for non-intrusive load monitoring.

Pfund, D., Anderson, K., Detwiler, R., Jarman, K., McDonald, B., Milbrath, B., Myjak, M., Paradis, N., Robinson, S., and Woodring, M. (2016). Improvements in the method of radiation anomaly detection by spectral comparison ratios. *Applied Radiation and Isotopes*, 110:174–182.

Pinsky, M. R. (1984). Instantaneous venous return curves in an intact canine preparation. *Journal of Applied Physiology*, 56(3):765–771.

Pinsky, M. R. and Payen, D. (2005). Functional hemodynamic monitoring. *Critical Care*, 9(6):566.

Rey, M. and Roth, V. (2012). Copula mixture model for dependency-seeking clustering. *International Conference on Machine Learning*, pages 927–934.

Ruder, S., Vulić, I., and Søgaard, A. (2017). A survey of cross-lingual word embedding models. 1:1–15.

Runkle, R. C., Tardiff, M. F., Anderson, K. K., Carlson, D. K., and Smith, L. E. (2006). Analysis of spectroscopic radiation portal monitor data using principal components analysis. *IEEE Transactions on Nuclear Science*, 53(3):1418–1423.

Sadeghianpourhamami, N., Ruyssinck, J., Deschrijver, D., Dhaene, T., and Develder, C. (2017). Comprehensive feature selection for appliance classification in NILM. *Energy and Buildings*, 151:98–106.

Shaw, S. R., Norford, L. K., Luo, D., and Leeb, S. B. (2002). Detection and diagnosis of HVAC faults via electrical load monitoring. *ASHRAE Transactions*, 108 PART 1(July):468.

Shimkin, N. (2009). Derivations of the discrete-time Kalman filter: The stochastic state-space model. Technical report.

Späth, H. (1982). A fast algorithm for clusterwise linear regression. *Computing*, 29(2):175–181.

Srivastava, N. and Salakhutdinov, R. (2014). Multimodal learning with Deep Boltzmann Machines. *Journal of Machine Learning Research*, 15:2949–2980.

Tandon, P. (2016). Bayesian aggregation of evidence for detection and characterization of patterns in multiple noisy observations. *AI Matters*, 2(3):25–26.

Tandon, P., Huggins, P., Maclachlan, R., Dubrawski, A., Nelson, K., and Labov, S. (2016). Detection of radioactive sources in urban scenes using Bayesian aggregation of data from mobile spectrometers. *Information Systems*, 57:195–206.

Tibshirani, R. (1996). Regression selection and shrinkage via the Lasso.

Turin, G. L. (1960). An introduction to matched filters. *IRE Transactions on Information Theory*, 6(3):311–329.

Udell, M. and Townsend, A. (2019). Why Are Big Data Matrices Approximately Low Rank? *SIAM Journal on Mathematics of Data Science*, 1(1):144–160.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416.

Wang, H., Nie, F., and Huang, H. (2013). Multi-view clustering and feature learning via structured sparsity. In *International Conference on Machine Learning*, volume 28, pages 352–360.

Witten, D. and Tibshirani, R. (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, 8(1):1–27.

Yeung, K. and Ruzzo, W. (2001). Details of the adjusted Rand index and clustering algorithms. *Bioinformatics*, (January):1–6.

Zhao, H., Ding, Z., and Fu, Y. (2017). Multi-view clustering via deep matrix factorization. *AAAI Conference on Artificial Intelligence*, pages 2921–2927.

Zimek, A. (2009). Correlation clustering. *ACM SIGKDD Explorations*, 11(1):53–54.

Ziock, K. and Goldstein, W. (2002). The lost source, varying backgrounds and why bigger may not be better. In *Unattended Radiation Sensor Systems for Remote Applications*, volume 632, pages 60–70. AIP Publishing.

Ziock, K.-P. and Nelson, K. E. (2007). Maximum detector sizes required for orphan source detection. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 579(1):357–362.

Zoha, A., Gluhak, A., Imran, M. A., and Rajasegarar, S. (2012). Non-intrusive Load Monitoring approaches for disaggregated energy sensing: A survey. *Sensors (Switzerland)*, 12(12):16838–16866.

Zong, L., Zhang, X., Zhao, L., Yu, H., and Zhao, Q. (2017). Multi-view clustering via multi-manifold regularized non-negative matrix factorization. *Neural Networks*, 88:74–89.