Anomaly Detection Using Dictionary Learning

Mark Eisen*, Mengjie Pan†, Zachary Siegel‡and Sara Staszak§

July 22, 2013

MAXIMA REU
Summer 2013
Institute for Mathematics and its Applications
University of Minnesota
Faculty advisor: Alicia Johnson (Macalester College)
Problem poser: Jarvis Haupt (University of Minnesota)

Abstract

This report applies dictionary learning and sparse coding algorithms to data in the interest of developing a better method of anomaly detection without *a priori* information about the anomalies themselves. These methods aim to find a sparse representation of data Y with respect to a learned basis, or dictionary D. Specifically, iterative learning algorithms are used to solve the minimization problem $\min_{X,D} \|Y - DX\|_2^2 + \lambda\|X\|_0$, where X is a set of coefficients and λ controls the sparsity of X. Sparsity helps assign semantic meaning to individual dictionary elements based upon their use in reconstructing data, which in turn highlights natural groupings and relationships among the data points. Thus, though traditional applications of dictionary learning include image denoising, novel methods for identification of *anomalous* or *salient* data points can also be derived from such structural features. To this end, we develop sparsity-informed metrics for defining and identifying anomalies with broad applications. Our results are promising and competitive with previous methods for flagging anomalous data in both images and propagating wavefield video.

*University of Pennsylvania
†Bryn Mawr College
‡Pomona College
§Macalester College

Contents

1 Introduction

Material diagnostics is, understandably, increasingly focusing on methods of non-destructive testing (NDT), which allow for a material's condition to be assessed without causing damage to the material itself. The Scanning Laser Dropler Vibrometer (SLDV) has proven an effective tool for NDT. The SLDV directs a laser at each point in the material and uses Doppler shift measurements to determine the amplitude and velocity of the vibration at each point in time and space [38]. This allows for a complete reconstruction of the propagating wavefield across the surface of the material. Wavefield videos have in the past been used for defect localization in materials [25] [33]. Gonella and Haupt [21] look specifically at wavefield data as would be obtained from a SLDV and use dimensionality reduction techniques to locate the position of defects in the material. This method is powerful in that it can be performed without any knowledge of material properties.

In this paper, we construct an anomaly detection scheme inspired from work in [21] that utilizes more recent and generalized dimensionality reduction concepts to localize structural defects in materials from the SLDV wavefield video. Specifically, we use the techniques of sparse coding and dictionary learning, which have led to advancements of image denoising and have since been shown to have a wide array of applications. Although material diagnositcs is our primary motivation, we develop dictionary learning-informed methods for detecting anomalies in any type of data.

1.1 Anomaly Detection

Anomaly detection is an area of study that has a number of applications beyond structural defect localization, ranging from finding visual saliency in natural images to flagging intruders in large-scale data networks [22] [44] [19] [10]. In each application, anomalies may take on one or multiple forms, such as single points, repeating patterns, or entire spatial or temporal regions. Methods for detecting anomalies are devised in both supervised [19] [23] and unsupervised form [15] [24], though our interest lies primarily with the latter, in which little to no *a priori* information is given about typical and anomalous data. In this case, anomalies are most often flagged for not following some greater trend the rest of the data follows. We seek a detection scheme that is robust against various types of both input data and anomalies.

1.2 Existing Methods

Existing approaches to anomaly detection often identify a simple model for the typical system behavior and identify the points that deviate from this model. Many of these are based on the idea of dimensionality reduction. For example, Gonella and Haupt [21] perform singular value decomposition to obtain a low rank model for wavelet data and flag the deviations as structural defects. Similarly Shyu, et al. [37] use robust principal component classifiers to identify data that does not fit with a learned model. In a slighlty different approach, Breunig et al. [7] assign to data points a degree of being an outlier, called the local outlier factor (LOF), through comparison of the local density at a data point and its surroundings.

1.3 Proposed Method

We aim to extend the dimensionality reduction concept of previous anomaly detection methods to include a recent development in signal processing called dictionary learning [41]. Dictionary learning generalizes the assumption that "typical" data points inhabit a low-dimensional subspace of the ambient space by proposing that points may all lie in the union of many very low-dimensional subspaces. Specifically, using sparse coding techniques, dictionary learning represents each data point as a linear combination of only a few basis elements, ie. dictionary atoms. Since each data point is closely related to its corresponding atoms, points using the same atoms are presumably semantically related, naturally grouping the data. This suggests that anomalous data will exhibit at least one of three properties:

1. Anomalous data will not be well represented by a learned dictionary as long as the dictionary is constrained to a sufficiently small number of atoms. Therefore, anomalies can be identified as having large residuals.

2. The learned dictionary will be more influenced by anomalies than by data points that follow the greater trend, since anomalies will be much farther from the union of the spans of small subsets of the dictionary than from the span of the entire dictionary. That is, anomalous data will have high leverage on the model relative to "typical" data.

3. If the dictionary contains a sufficiently large number of atoms or the anomalies occur frequently, these data points will fit the model well but will use "rare" basis atoms. These atoms are typically not used by regular data and are included in the learned model primarily to "accommodate" the anomalous data. Alternatively, typical basis elements may be used in atypical combinations to represent anomalous points.

The remanider of this paper is organized as follows. Section 2 discusses sparse coding, dictionary learning and tuning parameters. Section 3 states our detection methods in detail. Results of our detection methods on wavefield data and natural images are presented in Section 4 and 5, respectively. Section 6 discusses extended techniques used to improve detection or aid parameterization of dictionary learning. Section 7 provides a conclusion and brief discussion of future directions for this work.

2 Methodology

Dictionary learning aims to solve the following joint optimization problem:

$$\min_{X,D} \|Y - DX\|_2^2 \quad s.t.\ \|X_i\|_0 \leq T,\ \ \forall i \in [1, \ldots, n] \tag{1}$$

where

- $Y = (Y_1, \ldots, Y_n)$ consists of the vector data (length m) to be represented,
- $D = (D_1, \ldots, D_K)$ is an overcomplete basis of "dictionary atoms" (length m) for the column space of Y,
- $X = (X_1, \ldots, X_n)$ is the set of coefficient vectors (length K) that represent each signal with respect to D,
- $\|\cdot\|_0$ is the l_0 pseudonorm, which counts nonzero entries, and
- T is the imposed sparsity.

In other words, a solution is sought in which each vector Y_i can be best reconstructed as a linear combination of at most T columns of D; each reconstruction $DX_i \approx Y_i$ is confined to the union of the T-dimensional subspaces spanned by columns of D. In (1), the joint minimization on X and D is typically solved iteratively, optimizing for each term seperately while the other is kept fixed. This allows for easier, convex optimization algorithms to be used in the learning process.

2.1 Sparse Coding

For a fixed D, the process of finding an optimal set of sparse coefficients X is known as sparse coding:

$$\min_{X} \|Y - DX\|_2^2 \quad s.t.\ \|X_i\|_0 \leq T,\ \ \forall i \in [1, \ldots, n] \tag{2}$$

Greedy algorithms are often employed to find approximate solutions to an explicitly defined sparsity contraint, such as the most commonly used Orthogonal Matching Pursuit (OMP) [32]. However, due to the computational complexity of (2), a number of convex relaxations are often employed. In general, the minimization of a least-squares objective function with a regularization term of the form:

$$\|Y - DX\|_2^2 + \lambda\Omega(X), \tag{3}$$

can be substituted, where $\Omega(X)$ is a sparsity-inducing operator and λ is a control parameter that tunes the trade-off between goodness of fit and sparsity of the coefficients. Many different operators have been used

to produce sparse results, such as l_2 and l_i/l_q-norms [16] [46] [3] [2]. The most common of these includes a relaxation of the l_0-norm to the l_1-norm in what is referred to as the LASSO problem [40]:

$$\min_X \frac{1}{2}\|Y - DX\|_2^2 + \lambda\|X\|_1, \tag{4}$$

where the l_1 norm is defined as the sum of abolsute values of X. By minimizing the l_1-norm of X, we find a solution in which X contains a very small number of values, effectively approximating the l_0-norm. Algorithms used to solve this minimization include Least Angle Regression (LARS) and In-Crowd [13] [20].

Solutions to (4) have proven to be, with the right choice of dictionary, an effective tool in recovering signals from relatively few or noisy measurements, as is the goal of compressive sensing [9]. Many natural signals have shown to have sparse representations in well known domains, such as wavelets or Fourier bases [36] [39]. For this reason, sparse coding has become a popular tool in image processing tasks such as image recovery (demosaicing) and facial reconstruction [14] [45]. Another strong advantage of sparse representation relates to the concept of interpretability. By restricting each data point to comprise only a couple of dictionary elements, greater contextual meaning can be assigned to each atom by noting how it maps to certain data points. It is this semantic meaning that is believed to be a unique and important benefit of sparse representations, and the primary motivation behind the third proposed property for anomalous data introduced in Section 1.3.

2.2 Dictionary Learning

Different dictionaries may permit better or worse sparse fits than others, making the choice of dictionary a crucial step in the coding process. In the absence of a predefined dictionary that will provide a strong sparse representation, the best results can be obtained by finding the optimal dictionary specific to the data set.

In the dictionary learning step, we fix the set of coefficients X and solve for the dictionary D that will provide the best reconstruction of Y. Common dictionary learning methods include probabilistic approaches, such as maximum likelihood optimization [31], and clustering-based approaches, such as K-SVD [1]. We will briefly discuss two popular and state-of-the-art algorithms: K-SVD and Online Dictionary Learning.

The K-SVD algorithm is a generalization of k-means clustering, in which data points are clustered into k groups according to their distance from the group mean. After sparse coding is performed, each column of D is iteratively updated using singular value decomposition (SVD) to solve for (1). Assuming the sparse coding stage is performed well, K-SVD is shown to guarantee convergence to a local minimum of the objective function [1]. Online dictionary learning, on the other hand, is a more recently developed algorithm that iteratively refines each Y_i by, at each step, efficiently minimizing a quadratic surrogate function of (1) over the set of constraints. Specifically, it looks to solve the LASSO-based problem for D:

$$\min_D \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}\|Y_i - DX_i\|_2^2 + \lambda\|X_i\|_1\right), \tag{5}$$

where n is the number of columns in Y. Online dictionary learning is also shown to have faster convergence to a stationary point than previous learning algorithms [29].

Note that a dictionary learned in the interest of sparse coding will be different from one learned otherwise. For example, SVD matrix factorization also leads to an accurate representation of data by a small number of bases, in which sparsity can be induced afterwards by truncating small coefficients. However, while this can confer some benefits of sparse coding, as used in classical methods of data denoising, learning an overcomplete dictionary will ultimately permit a more accurate representation under such sparsity constraint. While the coefficients in the SVD can suggest ''hidden" variables, they may lack the semantic meaning offered by sparsity in an overcomplete dictionary. A learned dictionary may point toward more interpretable patterns by the occasion of atoms' individual or concurrent use.

2.3 Choosing and Parameterizing a Model

To detect anomalies via dictionary learning, the numerous tuning paramaeters must be understood. As the desired dictionary model is inherently different from one constructed in the interest of denoising or

demosaicing, existing heuristics cannot be relied upon. Although nonparametric dictionary learning methods have been explored using a variety of parameter learning techniques, including beta processes [47] and minimum description length (MDL) [34], our discussion relies more on an intuitive understanding of the parameters to be chosen in the context of anomaly detection.

2.3.1 The Equivalent Minimization Problems

Dictionary learning and sparse coding denote general *types* of strategies, not specific algorithms, and can take on different forms for their various applications. Each problem calls for the solution of one of several idealized formulations where each has numerous computable approximations. All such algorithms require parametric tuning to suit their applications.

In the context of sparse coding, this work primarily deals with the idealized problem (2) in the computable form of the LASSO problem (4). Both the LASSO and the idealized formulation have three equivalent formulations:

$$\min_X \frac{1}{2}\|Y - DX\|_2^2 + \lambda\|X\|_* \tag{4}$$

$$\min_X \|Y - DX\|_2^2 \quad s.t.\ \|X\|_* \leq T \tag{2}$$

$$\min_X \|X\|_* \quad s.t.\ \|Y - DX\|_2^2 < \epsilon \tag{6}$$

where $*$ denotes the sparsity-inducing norm being considered. These problems' *equivalence* means that for any λ, T, or ϵ, the other two can be chosen so the same X solves all three of (4), (2), and (6) [41].

2.3.2 Sparsity Constraints λ, T, and ϵ

The accuracy of the dictionary reconstruction of Y is quantified by the residual, $||Y - DX||_p^p$. The sparsity of a representation is described by the sparsity-inducing l_0 pseudonorm, which counts the number of nonzero coefficients in the representation X. In general, sparsity can be measured by any appropriate regularization operator, as in (3), as discussed in Section 2.1.

The sparsity constraint λ in (4) controls penalization of non-sparse coefficients in the dictionary representation X; λ represents a tradeoff between accuracy and sparsity in the minimization solution. As $\lambda \to 0$, the sparsity-regularization operator has zero weight in the objective function (4), and thus no sparsity is enforced. In this case, the solution X is any solution to $Y = DX$, an underdetermined system as D is overcomplete. On the other hand, as $\lambda \to \infty$, full sparsity is enforced and $X \to 0$, an all-zero matrix.

In the formulation (2), the case $T = 1$ ensures each Y_i is represented entirely by one dictionary atom, which amounts to a clustering problem solvable by any K-means algorithm, where K would be the size of the dictionary. For any other integer T, (2) is solved by any algorithm designed for the LASSO problem, such as K-SVD or online dictionary learning as discussed in Section 2.2. Due to the equivalence of the above formulations, these demands on the number of nonzero coefficients can be enforced explicitly through T, or implicitly through ϵ and λ; e.g., $T = 1$ would correspond to a very high λ and a high ϵ.

For anomaly detection, rather than choosing a single value for λ (or any parameter), overall results can improve when anomalous patches are flagged at various parameter values, and those patches found most frequently are determined to be true anomalies.

2.3.3 Partitioning and K

The size K of the dictionary should be proportional to the quantity of data. For example, for 64 data points, a dictionary size of 100 would be, in a sense, "too" overcomplete and would allow each patch to be represented by one atom that equals itself, i.e. K-means clustering on K data points (over-fitting). Identifying anomalies based on both residuals and the use of "rare" atoms (as described in Section 1.3) thus only makes sense when using an appropriately-sized dictionary. Experimental performance at various dictionary sizes is described in Section 4.

Certain data types come with a natural orientation, such as spatially-oriented pixels in an image, temporally-oriented audio data, and spatio-temporally-oriented video data. Such data sets are *partitioned*

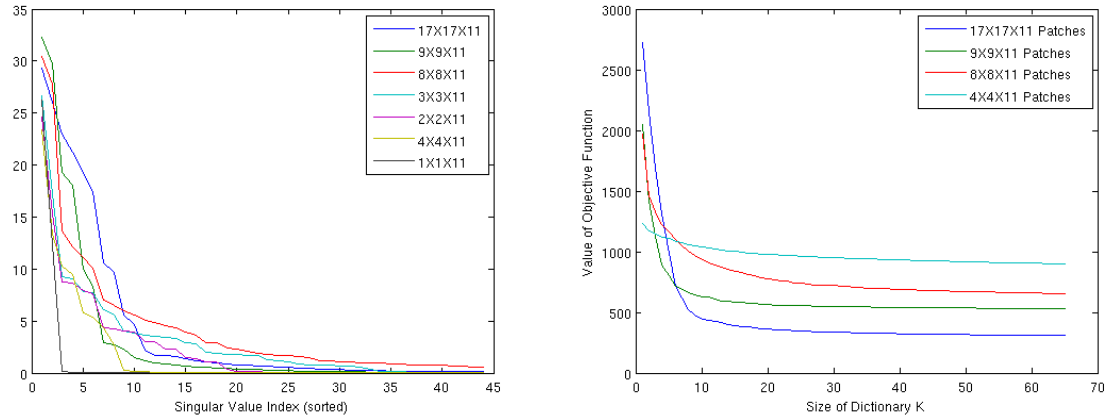
by placing nearby pieces of information and vectorizing them. Images, for example, are broken into $m \times m$ squares or pixels, whose values are then vectorized. These vectors become the columns of the $m^2 \times n$ matrix Y.

Different partitionings call for differently sized dictionaries. As discussed above, the size of the dictionary intuitively should scale down with the number of data points n, which increases with a finer partitioning. However, the ambient dimension of the data will then decrease, presumably reducing the number of atoms that would define an "overcomplete" dictionary. Thus, it is unclear whether the size of the dictionary should a priori increase with the number of data points if that increase is the result of subdivision.

In the case of wavefield video data, performance and mathematical justification suggest that K *should* increase with finer partitioning. For one, in finer partitionings, although the ambient dimension decreases, the effective dimension of the data remains fairly consistent through a range of partitionings. Figure 1a shows that the number of significant singular values, and indeed those singular values themselves, do not greatly fluctuate at the different subdivisions of video data. To ground this concept, if data were partitioned into single-pixel patches, each patch would be 1-dimensional and there would be exactly one singular value; if the entire image were vectorized as one patch, there would only be one data point, and hence also one singular value. This phenomenon therefore only occurs at some mid-level range of patch size.

The behavior of the dictionary learning and sparse coding is of course affected by dictionary size as a larger dictionary allows for a sparser and more accurate reconstruction of data. The minimized dictionary-learning objective function (1) decays with a larger K as presented in Figure 1b. The "knee" in this plot lies approximately at the knee in the decay of singular values, which is considered the 'numerical rank' of the data, and therefore should scale with the size of the dictionary that can effectively span that data. As is seen with the numerical rank of differently partitioned data, the knee in this plot remains fairly consistent at the four presented partitionings, though does increase slightly with finer partitioning, as we've seen our dictionary size should. These knees may not always be an optimal dictionary size, but suggest a reasonable range for unfamiliar data. Figure 2 shows that images of various types exhibit this phenomenon of constant dimensionality at different partitions.

Figure 1: Numerical dimensionality of data and dimensionality of a dictionary reconstructing that data



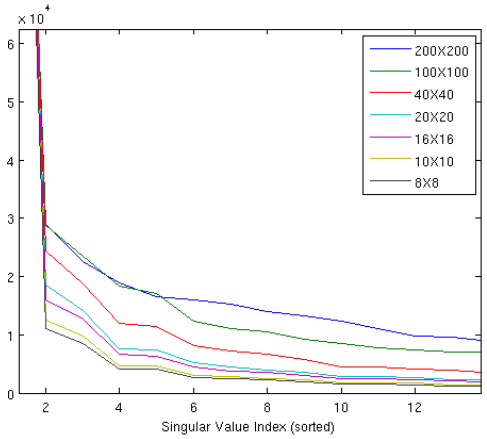
(a) The singular values of wavefield video data partitioned at various resolutions into a matrix

(b) The value of the minimized objective function (1) at various dictionary sizes, presented at four different partitioning resolutions.

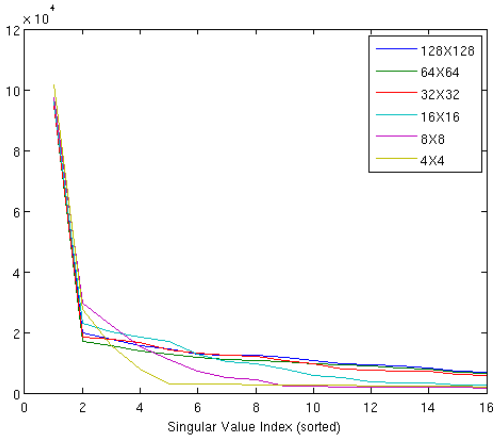
3 Detection Methods

With the dictionary learning framework in mind, we now detail our proposed anomaly detection methods from Section 1.3. Specifically, we aim to flag an anomalous data point based on its residual in the model reconstruction, its influence on the learned dictionary, and its use of rare dictionary atoms.

Figure 2: Numerical dimensionality of partitioned image data



(a) The singular values of animated image data partitioned at various resolutions



(b) The singular values of geometric test image data partitioned at various resolutions

3.1 Residual Thresholding

The first and most basic method employed to identify anomalies consists of learning a dictionary to model all of the data elements and locating features with reconstructions that have large residuals. For data Y, learned dictionary D, and sparse coefficient matrix X, the elements Y_i that are considered anomalous are those such that

$$\|R_i\|_p^p := \|Y_i - \hat{Y}_i\|_p^p > \epsilon, \tag{7}$$

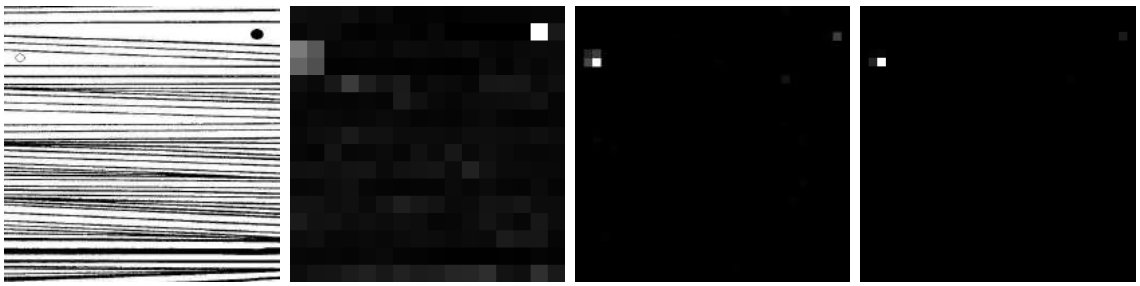
where $\hat{Y}_i = DX_i$ is the reconstruction of Y_i and ϵ is an error threshold. Both p and ϵ can be tuned for specific applications. In general, p will reflect the expected size of anomaly relative to patch resolution and ϵ will control how conservative the definition of anomaly is. Setting $p = 2$ is a natural choice in selecting the norm, as the objective function in the learning step is minimized in this way, and is used most commonly in implementation. However, it is worth noting that different values of p can highlight different sizes of anomalies relative to the dimension of the data. Take, for example, the image in Figure 3a, which contains two salient circles of different sizes and intensities, and otherwise consists entirely of straight lines. By extracting patches of 25×25 pixels from the image and learning a dictionary and sparse coefficients to fit the data, we can look at the residuals taken with respect to various p-norms.

The images shown in Figures 3b-3d show the magnitude of the residuals at each patch of the image, taken with respect to different norms. For low p, as shown in Figure 3b, the largest residual appears in the top right corner, corresponding to the location of the larger, dark circle. With larger norms, the magnitude of the residual in the top right corner begins to decrease while the magnitude of the rediduals in patches containing the smaller, clear circle become larger (seen in Figures 3c and 3d). This most likely reflects the fact that, relative to patch size, smaller anomalies will contain smaller average residuals taken over an entire patch. However, taken to higher powers, small but intense areas of saliency will create larger residuals.

Determining a proper range of threshold ϵ can be difficult when little is known about the range of values the data and its residuals take on. Simple, nonparametric thresholds based on interquartile ranges can be used as a quick, naive approach, along with more complex nonparametric outlier detection methods [6] [4]. More statistically meaningful thresholds can be determined when more is known about the underlying probability distributions of the residuals. Because the residual is minimized as a two-norm in (4) and (5), it can be reasonably assumed that the residuals in each row of the residual matrix $R = Y - \hat{Y}$ will have a normal distribution. It may be tempting to immediately conclude that, from (7), each column of $\|R_i\|_2^2$ is the sum of square normal random variables, and thus

$$\|R_i\|_2^2 \sim \chi^2(m), \tag{8}$$

Figure 3: Test image and magnitude of residuals of each image patch with respect to various norms p



(a) Test image (b) Residuals, p = 5 (c) Residuals, p = 20 (d) Residuals,p = 50

where m is the row dimension of Y. However, the rows of R are likely not independent from one another, especially in the case of images or other types of data with clear spatial relationships between data points. In any case, if instead each row of $\|R_i\|_2^2$ is treated as a seperate chi-squared random variable with degrees of freedom $k = 1$, then the summation of sufficiently large number of rows will always converge to a normal distribution from the Central Limit Theorem (or, in other words, $\chi^2(k)$ is approximatley normal for large k).

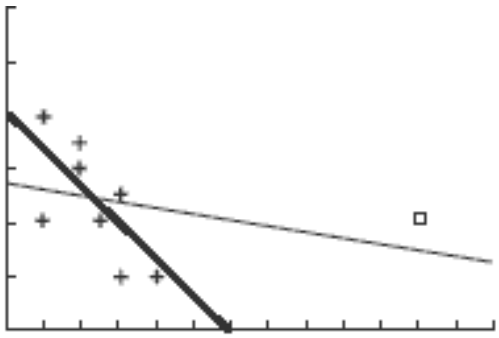
For correlated chi-squared random variables, this convergence is not reached quickly, so depending on the size of m, one cannot always assume that simply summing the squared residuals in each column will have converged to a normal random variable. There are, however, some well-known transformations to approximate a normal random variable from a chi-squared [18] [42], with varying degrees of accuracy [8]. A better approach is to implement transforms that do not approximate the normal directly but greatly speed up the convergence of the sum of many random variables. In [5], the natural logirithm of the chi-squared $\ln(\chi^2)$ is shown to converge to normality much faster due to the reduction of the skewness of the distribution.

If a sufficient approximation to the normal distribution can be obtained, we can view the total residual for each reconstructed data vector as an instance of the normal random variable. From there, common stastical measures of outlier detection can be used, such as flagging all points three standard deviations from the mean. This approach benefits from setting the threshold based on natural percentiles of a normal curve, though depending on the dimension of the data in question, may not always be applicable. In our experimental results discussed in Sections 4.1 and 5, examples of both non-parametric and distribution-based thresholds are shown.

3.2 Influence and RANSAC

Due to a potentially strong influence in the creation of the model itself, not every anomaly will necessarily have a poor model fit. Take, for example, the simple case of linear regression shown in Figure 4. In this plot, the square anomalous point does not have a large residual from the thinner line that was fit to the entire

Figure 4: Effect of influential data point (square) in simple linear regression



data set. The darker line is fit to the data when the anomaly is ignored and fits the rest of the data more closely. The difference between these two lines is therefore a more appropriate measure for flagging this type of outlier. To measure a datapoints' influence, or leverage, we first look at the classical measure of *Cook's Distance*. Cook's Distance, δ_j, for Y_j is given by

$$\delta_j \propto \frac{\sum_{i=1}^{n} ||\hat{Y}_i - \hat{Y}_{i(j)}||_2^2}{\sum_{i=1}^{n} ||\hat{Y}_i - Y_i||_2^2} \tag{9}$$

where $\hat{Y}_i$ is the full model's reconstruction of data point i and $\hat{Y}_{i(j)}$ is the reconstruction of data point i using a model learned without access to data point j [11]. The distances are normalized by dividing by the total square error in the full model reconstruction. It is suggested in [12] that $\delta_j > 1$ indicates that Y_j is highly influential. This measure can easily be applied to dictionary learning. While the effect of removing a data point can be analytically described with PCA, a sparsely-coded model might behave differently and confer more information in such a process. As with traditional Cook's Distance methods, higher-order information regarding the concurrent removal of any combination of data points can reveal different patterns, potentially identifying similar anomalous behavior in one or several data points. The semantic associations gained by sparse coding may make an influence-based measure particularly fruitful in predicting anomalous behavior.

However, using Cook's Distance in its complete form can be quite computationally expensive for large data sets and thus difficult to fine tune the parameters for any particular application due to the time required to run a single test. While this tool can prove effective when used for second-order detection on a set of previously flagged data points (see Section 4.4), for a more complete influence-based detection scheme we look at a more efficient algorithm called *Random Sample Consensus*.

Random Sample Consensus (RANSAC), detailed in [17], is an iterative algorithm that looks to find the set of data points whose inclusion in the model fitting creates the best model for fitting those data points. Intuitively, if there exists a set of "inliers" and "outliers" in the data set, then there is also a model that will provide a very strong fit of the inliers. Therefore, the model that provides the strongest fit for the inliers will be the model that is created when all of the influential points, or outliers, are ignored. RANSAC iteratively takes random partitionings of the data to find this set of inliers in which the error of the inlier reconstruction is minimized. This method is beneficial in that it is robust against a large ratio of outlier to inlier data points, as well as being considerably faster than computing Cook's Distance for each data point seperately.

3.3 Use of Rare Dictionary Atoms

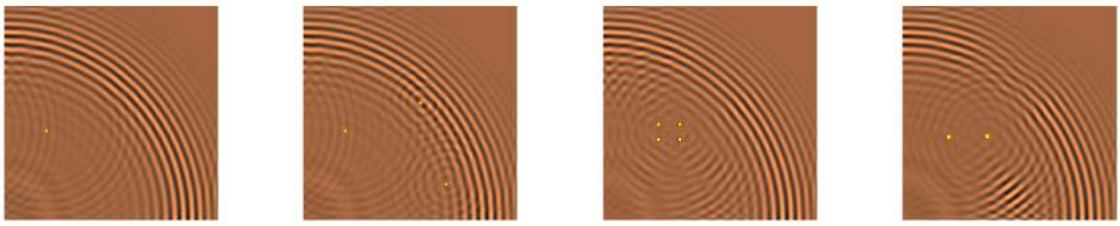
Our final definition of an anomaly is a data point that uses "rare" dictionary atoms. We define and quantify rarity in the following two perspectives:

1. A rare atom is infrequently used to reconstruct the data. Since each row in X corresponds to a column, or an atom, in Y, we sum up absolute values of entries in each row in X as a measure of the frequency of atoms being used. We rank $\sum_{i=1}^{n} |X_i^j|$, where X^j is the jth row of X, for each dictionary atom D_j, and we determine that a significantly small sum corresponds to a rare atom.

2. A rare atom is independently used or infrequently used in combination with other atoms to reconstruct the data. We assess an atom's correlation with others based on its covariance. We obtain a covariance matrix $\mathrm{C} = X \cdot X^T$, and we rank $\sum_{i=1}^{K} |C_i^j|$ for each dictionary atom D_j. A small sum means that an atom is infrequently grouped with other atoms to reconstruct data, and thus an atom with a significantly small sum is rare.

After we find the rare dictionary elements, we look for patches that have a significant use of the rare dictionary atoms. We consider a patch an anomaly if its coefficent corresponding to the rare atom is considered an outlier based on a common nonparametric flagging method. Quantitatively, given a rare atom D_j, we examine a set of coefficients $X_i^j \ \forall i$. Y_i is an anomaly if $X_i^j > Q_3 + 1.5 \times IQR$, where Q_3 and IQR denote the third quartile and the interquartile range of the set of coefficients, respectively. We believe that this detection method is not only the most unique to dictionary learning, but also provides an interesting and more generalized criteria for anomaly detection. An example of this added layer of robustness is seen later in Section 5.

Figure 5: Wave propagation across material with one, three, square and line defects.



(a) Single defect (b) Three defects (c) Square defect (d) Line defect

4 Wavefield Propagation Results

As our primary motivation for this work was to detect structural defects in materials based on the propagating wavefield video obtained from a SLDV, we ran extensive tests on synthetic data of this type. Similar to the test data in [21], still images of the waves propagating across a material with a single defect, three defects, a square, and a line defect are shown in Figures 5a-5d. A diagonal slice of the video data of propagating wavefield was taken so that, at each point, we looked at the t time steps after the propagation reaches that point. The wavefield image at each patch at all t time steps was then combined into a $mt \times n$ matrix. Here, m is the number of pixels in a single square patch and n is the number of total patches. In all cases we used $t = 11$ time steps in order to avoid boundary reflections from the edge of the material.

As previously discussed, the resolution of a given data set has an impact on the ability of these techniques to detect any anomalies. Therefore, we also tested our methods on different partitionings of the video data. For our tests we used three different resolutions: $m = 1089$ and $n = 37$, $m = 289$ and $n = 152$, and $m = 81$ and $n = 587$. Note that these correspond to patch sizes of 32x32, 17x17, and 9x9 pixels, respectively.

We used the open-source Sparse Modeling Software (SPAMS) package to implement the algorithms used for sparse coding and dictionary learning. In particular, the LARS algorithm was used in the sparse coding step and the optimized Online dictionary learning was used to learn the dictionaries [29] [28]. In addition, all dictionaries were learned from (4) with $\lambda = 1.5\sigma_Y$, where σ_Y is the standard deviation of the wavefield data.

4.1 Identifying patches with large residuals

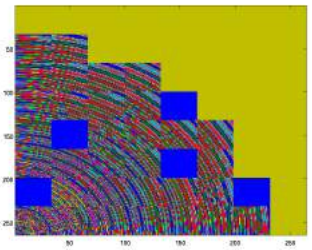
From (7), we used the common $p = 2$-norm when defining the residuals for all resolutions, and used both nonparametric and normal percentile based error thresholds ϵ.

Our nonparametric threshold ϵ_n was based on a simple interquartile-range based measure:

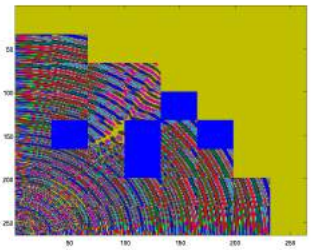
$$\epsilon_n = Q3(\|R_i\|_2^2) + BIQR(\|R_i\|_2^2), \tag{10}$$

where $Q3(\|R_i\|_2^2)$ and $IQR(\|R_i\|_2^2)$ is the third quartile and interquartile range of residuals $\|R_i\|_2^2$, respectively, and B is a constant to control sensitivity. With this nonparametric threshold, it is difficult to get sense for B without first inspecting the residual data. Shown in Figure 8 is an example of potential ϵ_n for different values of B in the case of three defects at the resolution of $m = 289$ and dictionary size $K = 24$. It is clear that a smaller B will flag more anomalies giving a conservative count of defects. A visual inspection of this plot shows that, for this data, $B = 4$ captures all of the substantial anomalies. A similar visual inspection was done for the resolution $m = 1089$ and $m = 81$, yielding roughly "optimal" $B = 2$ and $B = 8$, respectively. Shown in Figure 6 are the results for all sets of defects for each resolution using nonparametric thresholding. Note than for the 33x33 pixel case in Figures 6a-6c, the defects were too small relative to patch size to be detected by the l_2-norm of the residuals, so in (7) the larger $p = 5$ was used, though it still performed quite poorly relative to other patch sizes. As discussed in Section 2.3.3, the "optimal" dictionary size K varied between resolutions. For example, the 1089 pixel patches required a dictionary of size $K = 16$, while the 289 and 81 pixel patches performed best with sizes $K = 24$ and $K = 32$, respectively. A more

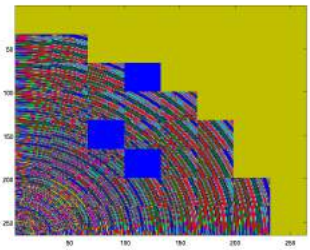
Figure 6: Non-parametric residual threshold results, "optimal" K



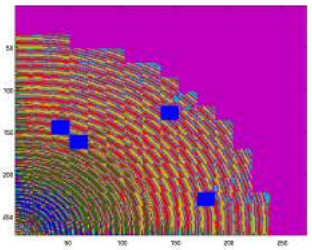
(a) Three defects, 33x33 pixel patches, $K = 16$, $B = 2$



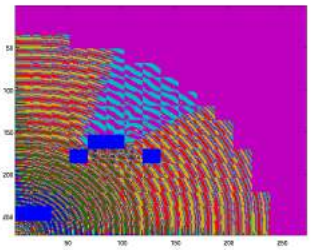
(b) Crack defect, 33x33 pixel patches, $K = 16$, $B = 2$



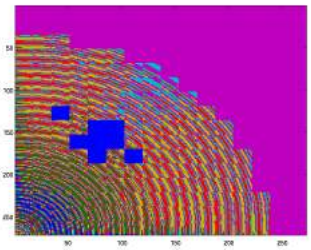
(c) Square defect, 33x33 pixel patches, $K = 16$, $B = 2$



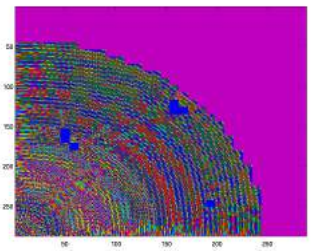
(d) Three defects, 17x17 pixel patches, $K = 24$, $B = 4$



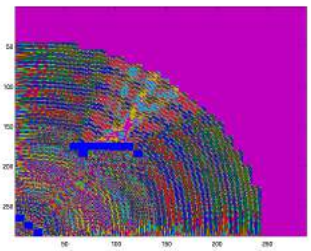
(e) Crack defect, 17x17 pixel patches, $K = 24$, $B = 4$



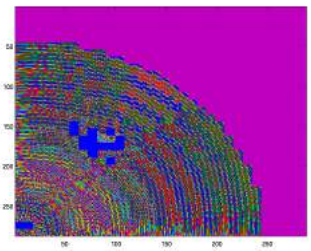
(f) Square defect, 17x17 pixel patches, $K = 24$, $B = 4$



(g) Three defects, 9x9 pixel patches, $K = 36$, $B = 8$

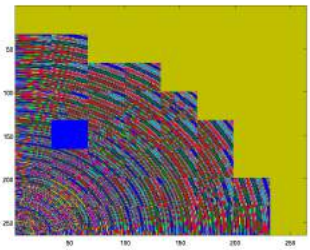


(h) Crack defect, 9x9 pixel patches, $K = 36$, $B = 8$

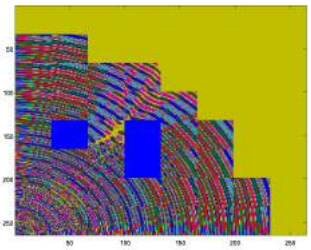


(i) Square defect, 9x9 pixel patches, $K = 36$, $B = 8$

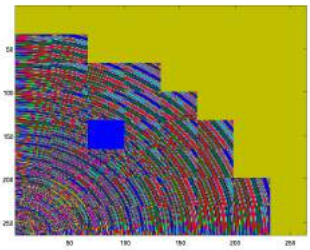
Figure 7: Non-parametric residual threshold results, $15 \leq K \leq 45$



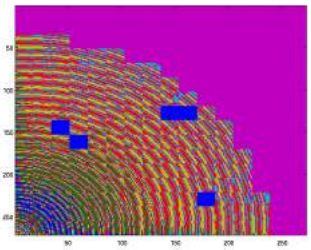
(a) Three defects, 33x33 pixel patches, $B = 2$



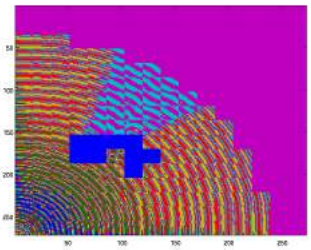
(b) Crack defect, 33x33 pixel patches, $B = 2$



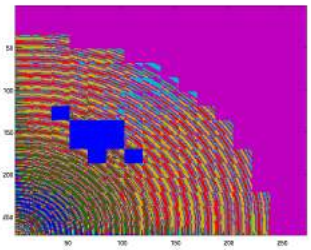
(c) Square defect, 33x33 pixel patches, $B = 2$



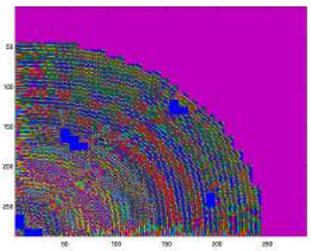
(d) Three defects, 17x17 pixel patches, $B = 4$



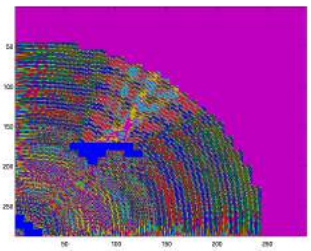
(e) Crack defect, 17x17 pixel patches, $B = 4$



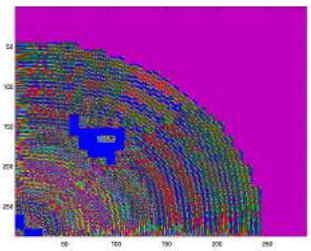
(f) Square defect, 17x17 pixel patches, $B = 4$



(g) Three defects, 9x9 pixel patches, $B = 8$

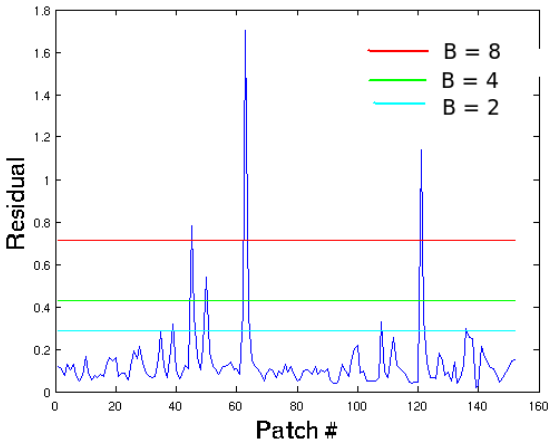


(h) Crack defect, 9x9 pixel patches, $B = 8$



(i) Square defect, 9x9 pixel patches, $B = 8$

Figure 8: Nonparametric thresholds for various B



robust approach we used was to simply perform the test for a range of values of K and add together the norm of the residual for all cases. Results for this approach ($15 \leq K \leq 45$) are shown in Figure 7.

In addition to the nonparametric threshold, we also ran tests that flagged anomalies based on a standard normal distribution. Recall from Section 3.1 that $\chi^2(m)$ will approximate the sum of square residuals and, for large m, this will closely resemble a normal distribution. To ensure convergence to normal, we took the natural logarithm of $\|R_i\|_2^2$, as discussed earlier, and then simply flagged points that were above 3 standard deviations from the mean as anomalous. The results for this approach over the range of K are shown in Figure 9. In the 33x33 pixel patch case (Figures 9a-9c), the ratio of non-anomalous to anomalous patches is too small due to small n. For this reason, the outlier data contribute highly to the properties of the normal curve and instead are less than 2 standard deviations from the mean. Again, the large patches perform the worst, while the other two resolutions find most anomalies with false positives around the region of excitation.

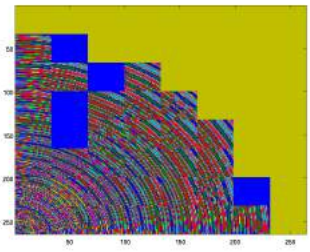
4.2 Identifying patches using RANSAC

We employed RANSAC to find the inliers of wavefield data highlighted patches not included in the consensus set as anomalies. Shown in Figure 10 is our results of detecting three point anomalies, crack anomalies and rectangular anomalies at the resolution of $m = 289$ and $m = 81$. While the optimal dictionary sizes for both resolutions were generally between 20-25, the specific value of K used to generate each figure is noted. Our tests showed that RANSAC detected regions of three point anomalies at both resolutions with occasional false positives around the excitation point, though it missed the exact locations a few times (shown in Figure 10a and 10d). RANSAC also detected crack and rectangular anomalies correctly, with some false positives close to the anomalies(shown in Figure 10b, 10c, 10e and 10f). False positives commonly occured around the excitation point (shown in Figure 10c), since the excitation point does not fit in the dictionary model. Note that RANSAC was not able to perform at the resolution of $m = 1089$ because the corresponding n is very small and thus the sample size is too small to get a model for other data to fit in. We also want to point out that, since RANSAC takes random samples on every try, it is possible that results for detecting anomalies change every time, even with the same parameters. However, the regions detected are usually the same, which shows that RANSAC is stable though the samples are randomly chosen.

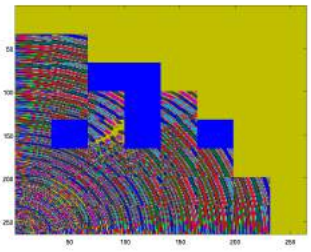
4.3 Identifying rare dictionary atoms

We next applied usage of the rare dictionary atoms method to anomaly detection on wavefield data. At the resolution $m = 81$ and with $K = 20$, this method detected some of the point anomalies and parts of the crack and square anomalies (Figure 11a-11c). However, several false positives were found and there were sometimes more false positives than true anomalies detected. For example, in Figure 11b, the method only

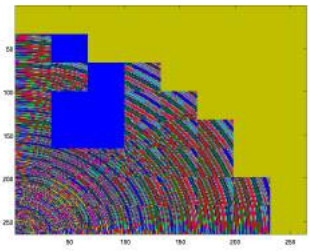
Figure 9: Normal quantile-based residual threshold results, $15 \leq K \leq 45$



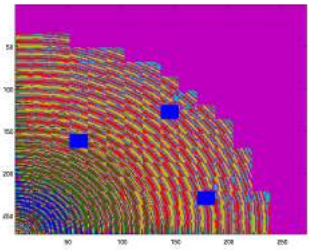
(a) Three defects, 33x33 pixel patches (1 std)



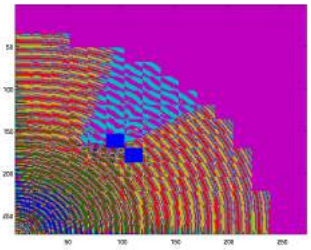
(b) Crack defect, 33x33 pixel patches (1 std)



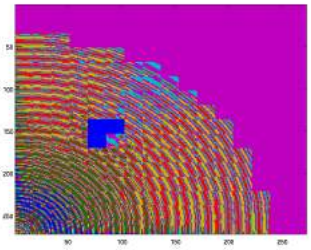
(c) Square defect, 33x33 pixel patches (1 std)



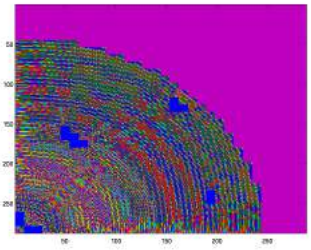
(d) Three defects, 17x17 pixel patches



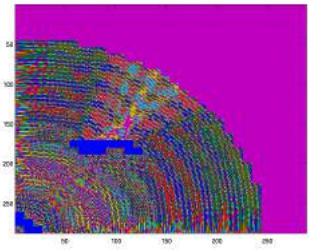
(e) Crack defect, 17x17 pixel patches



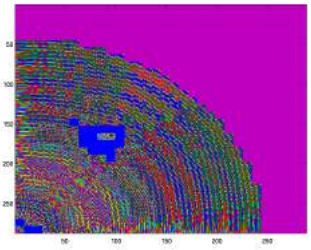
(f) Square defect, 17x17 pixel patches



(g) Three defects, 9x9 pixel patches

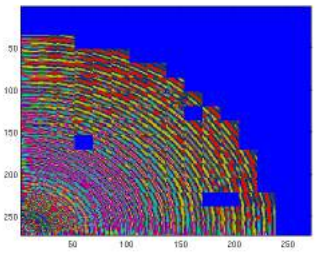


(h) Crack defect, 9x9 pixel patches

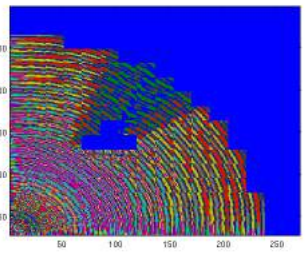


(i) Square defect, 9x9 pixel patches

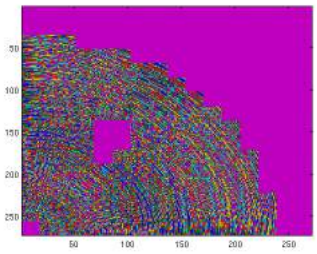
Figure 10: Detecting anomalies using RANSAC



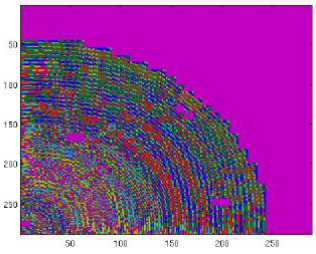
(a) Three defects, 17x17 pixel patches, $K = 24$



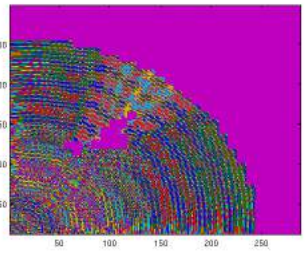
(b) Crack defect, 17x17 pixel patches, $K = 24$



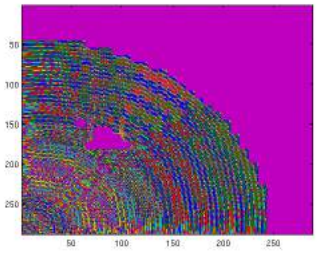
(c) Square defect, 17x17 pixel patches, $K = 20$



(d) Three defects, 9x9 pixel patches, $K = 24$



(e) Crack defect, 9x9 pixel patches, $K = 24$



(f) Square defect, 9x9 pixel patches, $K = 24$

detected the anomaly in the middle of the crack but missed the rest and had two false positives. The method detected anomalies correctly at the resolution $m = 289$ at $35 \leq K \leq 40$ (Figure 11d-11f). It detected the region of anomalies at the resolution $m = 1089$ at $30 \leq K \leq 35$, but missed a few patches (Figure 11g-11i). Few false positives were found at the later two resolutions. Using rare dictionary atoms for detection is sensitive to K at resolution $m = 81$ and $m = 289$, but it is stable at resolution $m = 1089$. We also tested this method on wavefield data with no anomalies. At small dictionary size ($K < 20$), it did not detect any anomalies; while at large dictionary size ($K > 30$), it flagged the excitation point as anomalies.

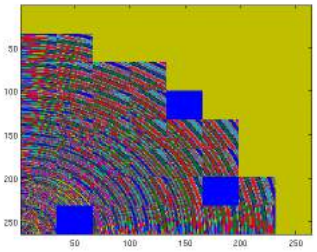
4.4 Use of influence as a second-order detection

The methods we presented above often detect false positives around the excitation point and near anomalies. In order to decrease false positives, we apply the influence method to patches considered as potential anomalies as a second-order detection. This second-order detection eliminates false positives around anomalies, but it does not always eliminate false positives around the excitation point since the excitation point sometimes has greater influence than the anomalies, especially for large patches. Figure 12 shows performance of detecting three anomalies before and after applying influence method. The influence method eliminates three false positives and keeps all three anomalies.

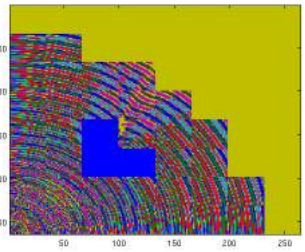
4.5 Comparing method performance on wavefield data

While each of our anomaly detection methods can pick out the correct patches for several types of anomalies in the wavefield data, the methods do interact differently with the parameters. Given the same set of data, the residual method will perform better with smaller values of K while the rare atom method will have more success with larger values. This makes sense intuitively because with a larger dictionary size it is more likely that the anomalous regions will get one or more atoms that are then not used by the rest of the reconstruction. Similarly, with fewer dictionary atoms, since it is less likely that anomalies will have their own atoms to use, they will be poorly reconstructed and thus have larger residuals by which they can be detected.

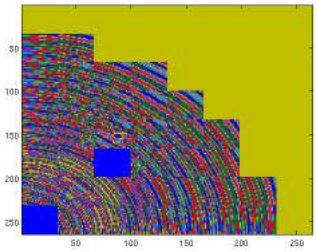
Figure 11: Detecting anomalies using rare dictionary method



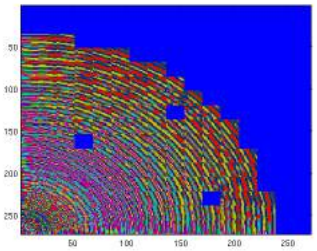
(a) Three defects, 33x33 pixel patches, $K = 20$



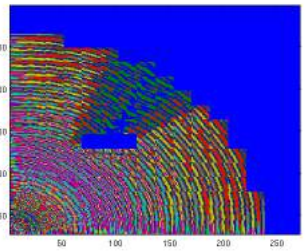
(b) Crack defect, 33x33 pixel patches, $K = 20$



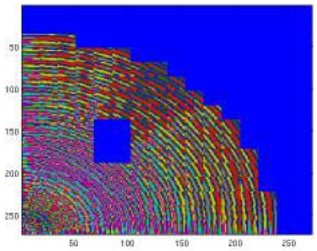
(c) Square defect, 33x33 pixel patches, $K = 21$



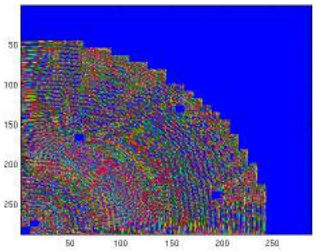
(d) Three defects, 17x17 pixel patches, $K = 39$



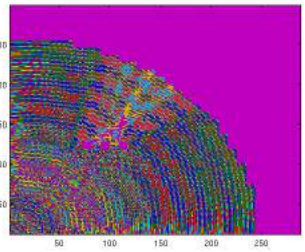
(e) Crack defect, 17x17 pixel patches, $K = 34$



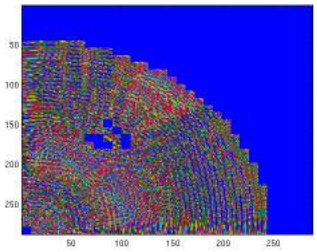
(f) Square defect, 17x17 pixel patches, $K = 37$



(g) Three defects, 9x9 pixel patches, $K = 30$

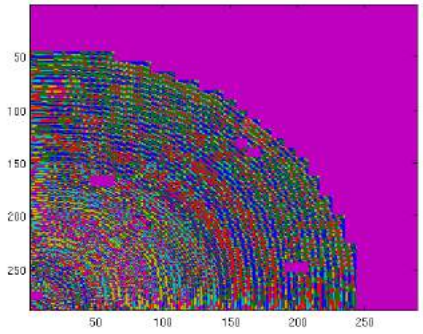


(h) Crack defect, 9x9 pixel patches, $K = 30$

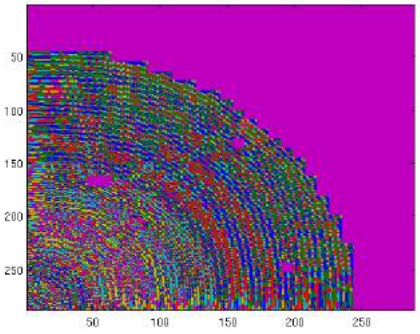


(i) Square defect, 9x9 pixel patches, $K = 35$

Figure 12: Comparing methods for detecting saliency based on orientation

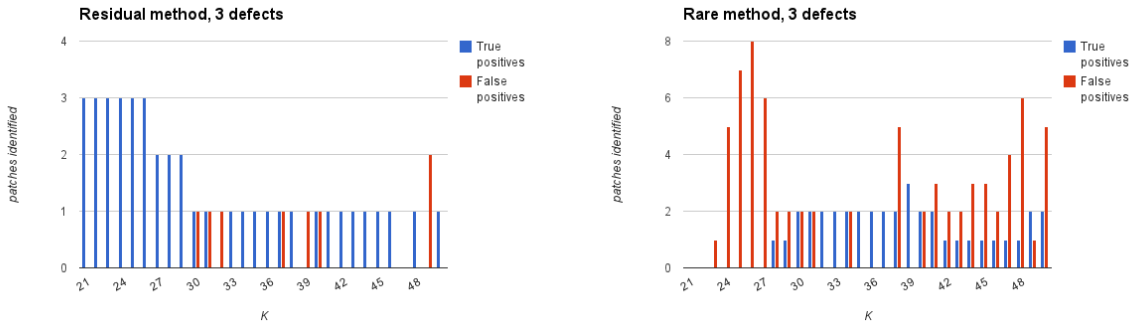


(a) Detecting three anomalies only using RANSAC, $K = 24$



(b) Detecting three anomalies using influence as a second-order detection, $K = 24$

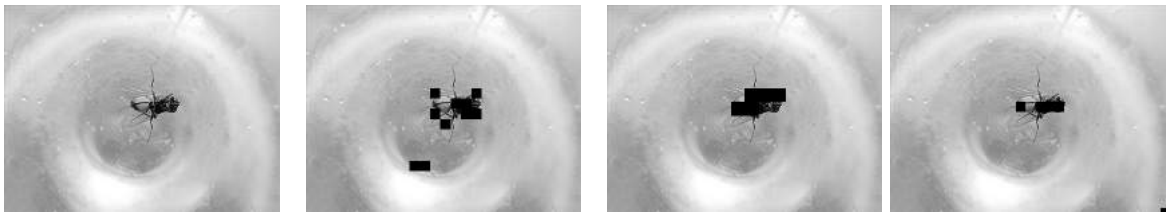
Figure 13: Patches identified by each method for different values of K



(a) Residual method

(b) Rare atom method

Figure 14: Method performance for detecting saliency in "bee" image



(a) Original image

(b) Residual normal quantile-thresholding, $15 \leq K \leq 45$

(c) RANSAC, $K = 12$

(d) Anomaly detection using rare atom method

The charts in Figure 13 show the relationship between both true and false positives for different dictionary sizes using the wavefield data with three anomalies. At some of the higher values of K shown in 13a, the residual method does not identify any anomalies, which supports our intuition that the residual method would perform better with smaller dictionaries. For the lowest values of K shown in figure 13b, the rare atom method only identifies false positives. While there is generally a trade off between more true positives and fewer false positives, this method is less successful with smaller dictionaries.

5 Natural Image Results

We tested each of our three detection methods on detecting salient features in a small set of natural images taken from the MSRA Salient Object Database [26]. Note that each of these images are entirely in grayscale, meaning saliency is determined only through an object's shape or texture, and not color. The detection results for images using residusal thresholding, RANSAC, and use of rare dictionary elements are shown in Figures 14-16.

In all cases, image patches correspdonding to the salient object were detected, though with varying degrees of completeness. For example, in Figure 14, the RANSAC method in 14c captured almost the entire bee while residual thresholding in 14b only flagged the edges. All methods had trouble capturing the entire region of the image containing the bottle in Figure 15. To capture larger anomalous structures, however, we develop a spatially-adjusted scoring framework in 6.2.

Natural image data is also informative in how it highlights the various benefits and applications of each particular method. If the feature of interest is relativley large in comparison to other image features, it may use dictionary atoms that are unique to that feature, thus making it easy for the rare atom method to identify but difficult for the residual thresholding method. As with the general case, the choice of parameters can have an impact as well when determining the efficacy of each method with different data.

Figure 15: Method performance for detecting saliency in "beer" image

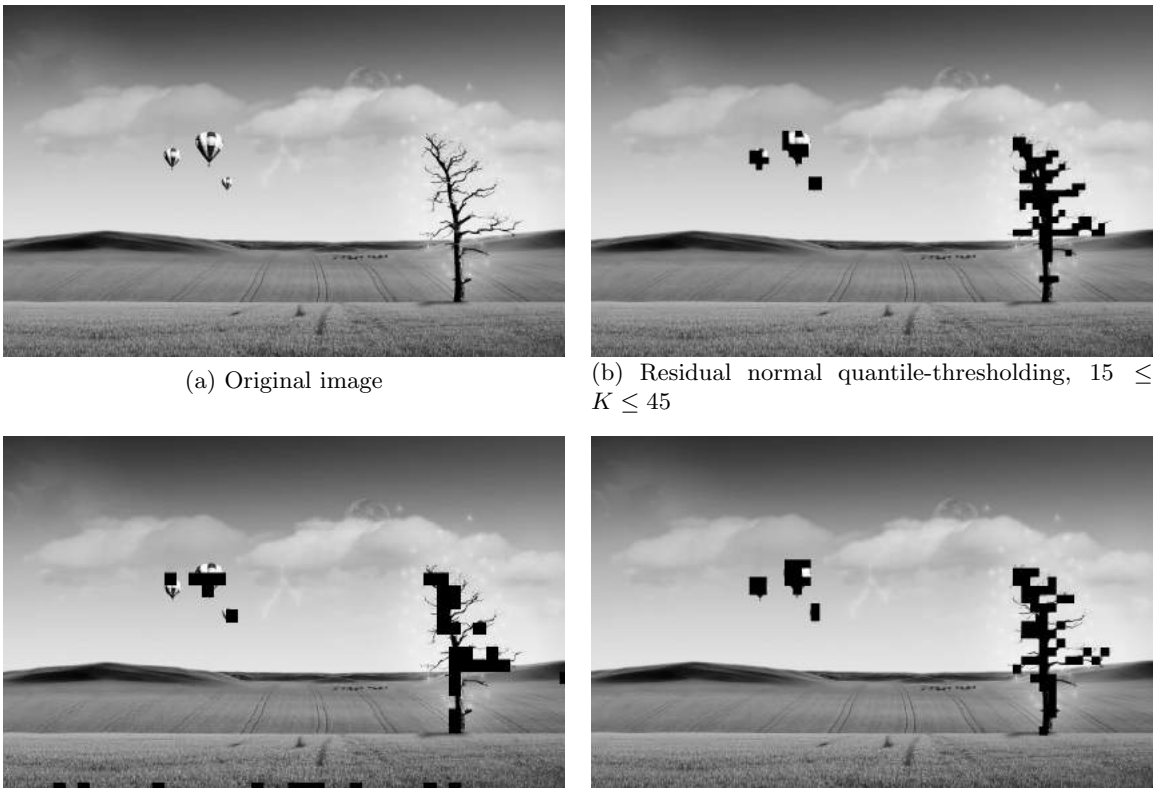
(a) Original image

(b) Residual normal quantile-thresholding, $15 \leq K \leq 45$

(c) RANSAC, $K = 13$

(d) Anomaly detection using rare atom method

Figure 16: Method performance for detecting saliency in "balloon" image [a]



(a) Original image

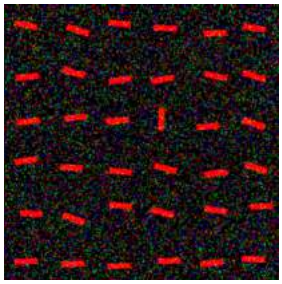
(b) Residual normal quantile-thresholding, $15 \leq K \leq 45$

(c) RANSAC, $K = 25$

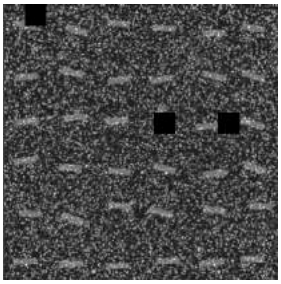
(d) Anomaly detection using rare atom method

[a] Image taken from http://nice-cool-pics.com

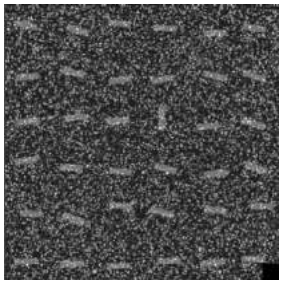
Figure 17: Comparing methods for detecting saliency based on orientation



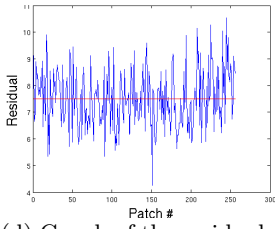
(a) Image with salient region based on orientation



(b) Anomaly detection using rare atom method



(c) Anomaly detection using residual method



(d) Graph of the residuals for figure 17a with line at the mean

To illustrate this, we peformed the residual and rare atom method on the test image shown in Figure 17a. In this image there is one clear saliency in the single bar that is oriented vertically. Using the rare atom method, three patches were flagged and the first patch identified contained the true "anomaly", shown in Figure 17b, whereas the residual method failed to identify the correct patch and in some cases did not identify any patches at all (shown in Figure 17c). An examination of the plot of residuals in Figure 17d shows that there isn't patch in which the residual is signifigantly larger than the others. The residuals are primarily close to the mean (the straight line) and patch 151, corresponding to the salient bar in fact has the smallest residual. It is likely that, for this image, dictionary learning is able to well represent the anomaly but only by learning a dictionary atom specific to that patch. As a result, this type of anomaly is impossible to detect with residual thresholding but quite simple to detect with the rare atom detection.

6 Extended Techniques

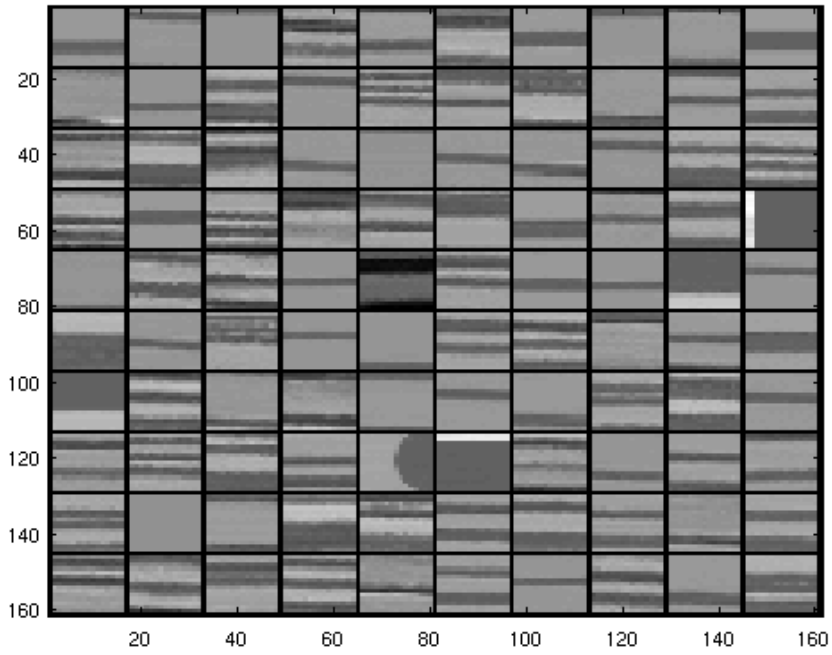
In this section we present a few additional, extended techniques that in certain contexts can improve detection or aid in the parameterization of the dictionary learning process.

6.1 Optimal dictionary size

As seen in Figure 13, finding the optimal dictionary size K is an important step in the paramaterization of the dictionary model. We've seen that, in general, the optimal size will scale more with the number of data points n than the dimension of the data m (see Section 2.3.3 for discussion). However, for an unsupervised approach such as ours, it is impractical to heuristically search for this optimal size by hand. Using a range of values for K is a potential solution to this problem; some results using a range are shown in Section 4.1, although finding a proper range will again require some trial-and-error. We instead wish to look into techniques that can automatically find an appropriate K or range of K through the clustering of nearby dictionary atoms.

Take, for example, the dictionary learned for a simple image shown in Figure 18. The dictionary is learned with $K = 100$, which, due to the number of redundant atoms, is clearly more than is necessary to represent the image. We can therefore cluster atoms that are close by a measure such as Euclidean distance or vector inner product, and the number of resulting clusters would represent a reasonable K. Figure 19a shows a portion of a dendrogram that indicates the hierarchical clusters of dictionary atoms that form at different resolutions based on the Euclidean distance between them. At a resolution of 0.22, the atoms underlined with red and blue will group into separate clusters. As seen in the dictionary in Figure 19b, these do in fact correspond to atoms that are almost visually indistinguishable. A conceptually related approach of clustering similar dictionary atoms is used in object categorization in images in [43]. Rao and Porikli in [35] also propose a method called clustering online dictionary learning (COLD) in which the "optimal" dictionary size is found by clustering nearby dictionary atoms using mean shift clustering at each step of the learning process.

Figure 18: Example dictionary learning for simple image, $K = 100$

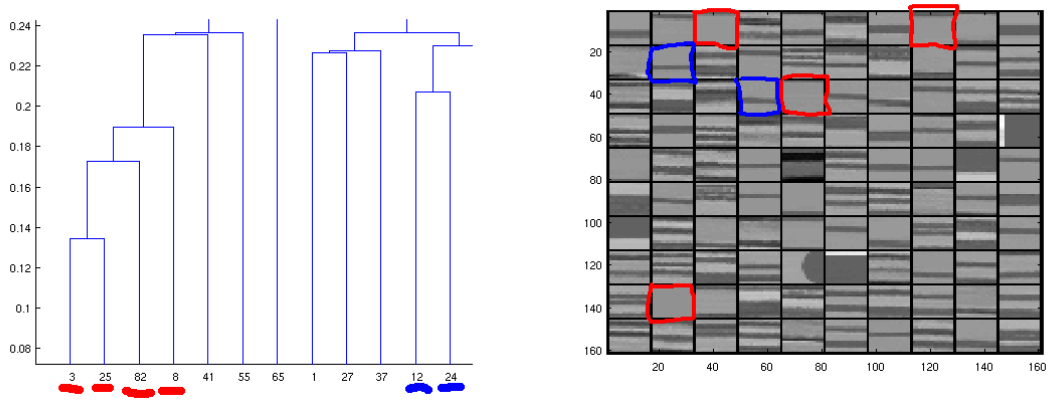


However, for the purposes of anomaly detection, we must be careful about clustering atoms together because there are some atoms that may be geometrically close to other atoms but carry important semantic signifigance in our detection schemes (e.g. rare or influential atom). Rather than performing clustering explicitly, we instead wish to use hierarchical clustering merely as a means to *identify* the "optimal" dictionary, after which we perform dictionary learning for that value of K rather than using the clustered dictionary. That is, we identify a resolution r and initially large dictionary size K_0 and learn a dictionary and perform hierarchical clustering. The number of clusters formed at resolution r becomes the new dictionary size K_1 and this process continues until $K_{i-1} = K_i$. Effectively, we look for a dictionary in which no atom is less than r from another atom but never manipulate the atoms themselves either during learning or in post-processing. Using this step will increase the running time, although in practice, finding the "optimal" size is faster than performing the tests over a range of K. It is also important to note that although this makes for an easier initial selection of K_0, the final size K_f is not completely independent of the intial value. This phenomenon, along with some initial results on propagating wavefield video, is presented below.

We performed preliminary tests of this approach on the wavefield video across materials with three structural defects (Figure 5b). These tests are in no way comprehensive and are provided only to demonstrate potential in this technique and some insight into selection of r and K_0. The tests were performed on the residual thresholding method and the rare dictionry atom usage method at two partitionings (17x17 pixel patches and 9x9 pixel patches). In the case of residual thresholding, we found that with $r = 6 * \sigma(Y)$ we were able to almost always find an "optimal size" within 8 iterations (in most cases only 3 or 4). The results for different initial dictionary sizes are shown in Figures 20 and 21. In almost every case K converges to the same value except in 20d and 21d when a larger K_0 actually causes it to converge to smaller final size. Interestingly, this convergence is not affected even when the dimensionality of the data changes for the two different partionings, which may be an effect similar to what was discussed in Section 2.3.3 regarding effective dimensionality of the data.

We also attempted to find optimal sizes for the use of the rare atoms method. This of course requires a different clustering threshold than a residual-based method, as we will need a larger K_f so that the anomalies will be well represented in the learned model. We experimentally found the strongest results at $r = 5 * \sigma(Y)$.

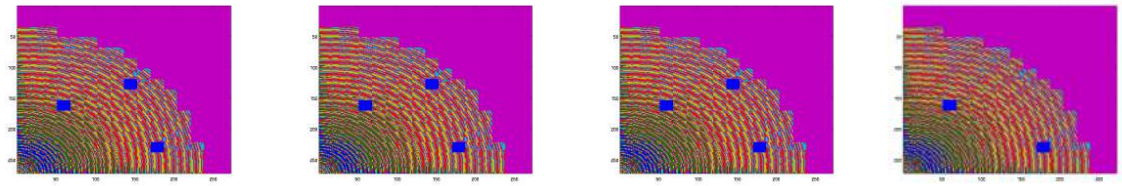
Figure 19: Clusters formed at resolution 0.22



(a) Strictly limiting false positives

(b) Allowing approximately one false positive per defect

Figure 20: Finding "optimal" size for residual thresholding, 17x17 pixel patches



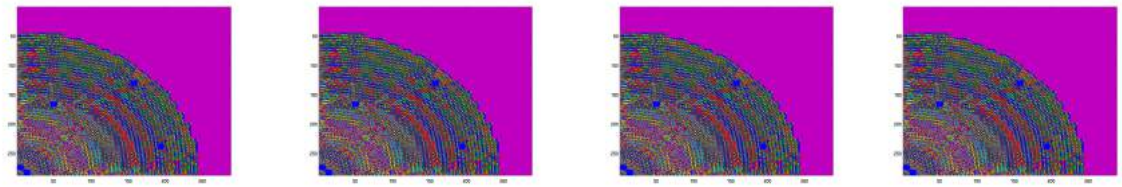
(a) $K_0 = 40$, $K_f = 27$ (b) $K_0 = 60$, $K_f = 27$ (c) $K_0 = 80$, $K_f = 27$ (d) $K_0 = 100$, $K_f = 20$

The results are shown in Figures 22 and 23. For both partitions, the same "optimal" size is found for all initial sizes, including the additional $K_0 = 150$. The value of K_f is not exactly the "optimal" size found experimentally in Section 4.3 and therefore performs slightly worse due to increased false positives. In general, this relatively simple procedure for determining "optimal" size shows promising results that with further study could prove a very effective tool in parameterization.

6.2 Spatial Scoring

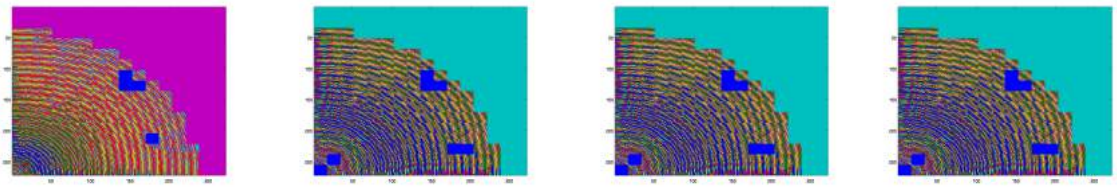
In datasets in which each data point's location relative to other data points carries important meaning, such as pixels in natural images, it may prove beneficial to include some of this spatial information when "scoring" each data point. For instance, in the case of an image, if all of image patches surrounding a small region are flagged as anomalous, then there is good chance that the inside of the region is anomalous as well. We let S_i

Figure 21: Finding "optimal" size for residual thresholding, 9x9 pixel patches



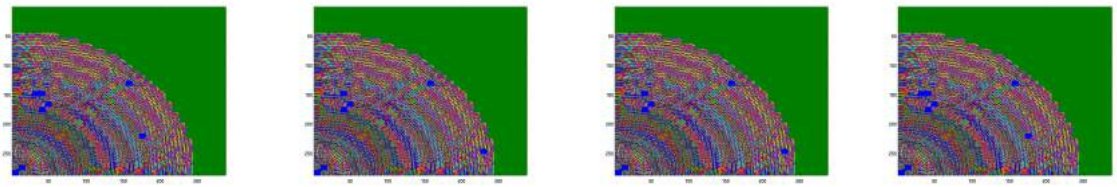
(a) $K_0 = 40$, $K_f = 27$ (b) $K_0 = 60$, $K_f = 27$ (c) $K_0 = 80$, $K_f = 27$ (d) $K_0 = 100$, $K_f = 22$

Figure 22: Finding "optimal" size for rare atom usage, 17x17 pixel patches



(a) $K_0 = 60$, $K_f = 42$ (b) $K_0 = 80$, $K_f = 42$ (c) $K_0 = 100$, $K_f = 42$ (d) $K_0 = 150$, $K_f = 42$

Figure 23: Finding "optimal" size for rare atom usage, 9x9 pixel patches



(a) $K_0 = 60$, $K_f = 53$ (b) $K_0 = 80$, $K_f = 52$ (c) $K_0 = 100$, $K_f = 52$ (d) $K_0 = 150$, $K_f = 53$

denote the "score" of Y_i and d_{ij} the distance between Y_i and Y_j. By "score" we mean to generalize any of the previously discussed anomaly indicators; e.g. sum of square residual. We then define a spatially-adjusted score, $\hat{S}_i$, as:

$$\hat{S}_i = \sum_{j=1}^{n} \alpha^{d_{ij}} S_j \mathbb{1}(d_{ij} < W), \tag{11}$$

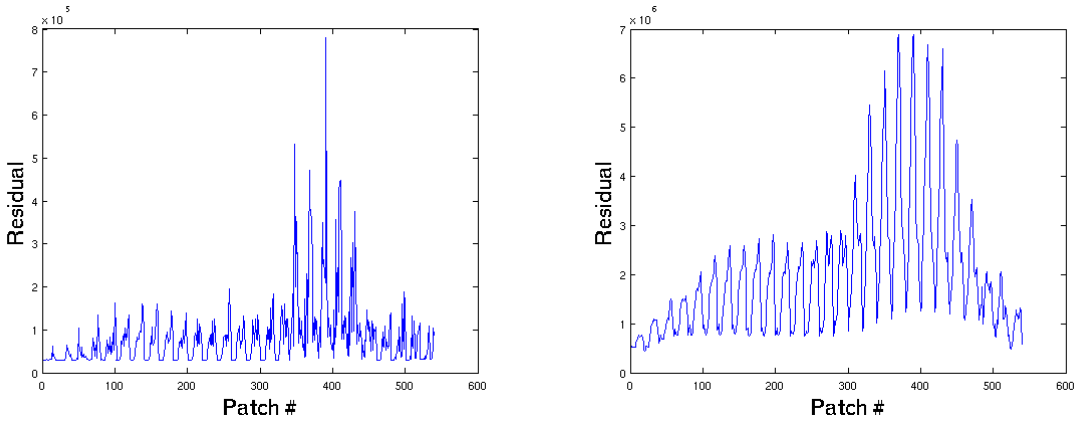
where $\alpha < 1$ is a decay parameter, W is the window in which we consider spatial proximity relevant, and $\mathbb{1}(\cdot)$ is the Boolean indicator function. By considering the scores of neighboring patches, we aim to improve the ability of our methods to detect larger anomalous structures relative to the size of the data partitioning. In images, for example, spatial scoring may help to detect the entirety of a salient object rather than just its outline. This is a partcularly important feature for saliency detection schemes, as otherwise the size of the patch partition can have a strong effect on detection. Larger patches may capture a wider physical range of saliency but, as briefly discussed in Section 3.1, may fail to detect comparitivley small anomalous points. Therefore, a method that uses small partitionings can still detect larger saliencies.

Another natural benefit of a spatially-adjusted scoring scheme is its ability to act as a filter of isolated and extraneous false positives. While legitimate anomalies can certainly be isolated from others, as patches get smaller it becomes less likely that a single anomalous point will not be in the neighborhood of another anomaly and is therefore a potential false positive. In this way, (11) can also be viewed as an averaging or low-pass filter, as seen in the residual plot in Figure 24. Examples of both of these applications of spatial scoring are shown in the following section.

To demonstrate the use of our spatial scoring system, we also compared a normal residual thresholding (nonparametric, $B = 1$, $K = 18$) result on the image in Figure 15a to a result that employed a spatial scoring. From (11), we set the window $W = 3$ and decay parameter $\alpha = 0.8$. These results are shown in Figure 25. Here we can see that more of the bottles' bases are detected as part of the salient object, though the necks are still being missed.

Another interesting thing to note is that the spatial scoring also helped to remove some false positives at the bottom of the image in Fig. 25a. This supports the proposed benefit of spatial scoring in removing isolated false positives. Another good example of this feature is shown in Figure 26, in which spatial scoring is used on non-parametric residual thresholding ($B = 4$, $K = 25$).

Figure 24: Example of residual, before and after spatial scoring



(a) Original residual scores

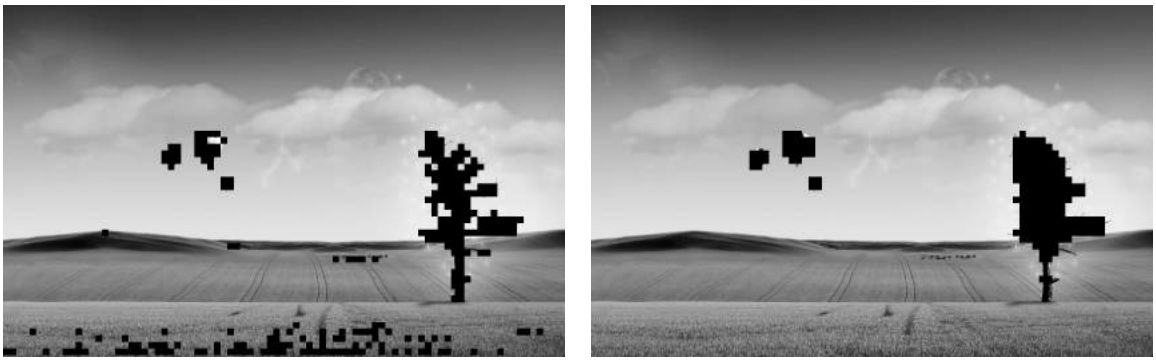
(b) Spatially-adjusted (averaged) residual scores

Figure 25: Spatial scoring on non-parametric residual thresholding ($B = 1$, $K = 18$) to detect larger anomalous structure

(a) Normal scoring

(b) Spatial scoring $W = 3$, $\alpha = 0.8$

Figure 26: Spatial scoring on non-parametric residual thresholding ($B = 4$, $K = 25$) to remove false positives



(a) Normal scoring

(b) Spatial scoring $W = 2$, $\alpha = 0.5$

6.3 Supervised Dictionary Learning

Dictionary learning, sparse coding, and all related anomaly-detection methods have clear extensions to *supervised* formulations. A dictionary can be learned on "clean" data with no anomalies, and then used to identify data patches that don't have an accurate sparsely-coded representation. Presumably, other "typical" data could be represented accurately and sparsely with such a dictionary, while anomalous data could not.

The advantages of *unsupervised* dictionary learning techniques should be clear from their descriptions, but certain, highly regular types of data lend themselves to the use of supervision. For example wavefield data follows a highly regular pattern in the absence of defects, and an off-the-shelf or tailor-made "wavefield dictionary" could be employed.

An overcomplete dictionary learned on defect-free data can represent anomalies less accurately (under a sparsity constraint) than a dictionary learned in the presence of those anomalies, because they might influence the dictionary to better represent them. While it has been shown above that an ad-hoc dictionary can successfully help identify anomalies on a clean background, a specialized dictionary learned on clean data might perform better for this reason.

Obtaining a useful premade dictionary might not be difficult. For image denoising and missing-data recovery, sparse coding performs well even on premade dictionaries consisting of *generic* "natural image patches" or simply wavelets [30]. A premade dictionary often comes along automatically with the data collection apparatus, such as in the case of security cameras, which accumulate more than enough training data to pick up on salient activity within a frame. This background detection has been performed in conjunction with dictionary learning for video data in [27].

Training data is particularly useful when the experimental data collection is undersampled. Intuitively, understanding the behavior of "typical" data is the only way to reconstruct missing points without making assumptions about those points; such assumptions, such as a minimal numerical rank, can lead to meaningful reconstructions and that body of research is termed "matrix completion problems." However, prior knowledge about typical data, such as an overcomplete dictionary learned from such data, allows very accurate reconstructions thereof and thus straightforward anomaly detection via large residuals.

Wavefield data in particular benefits from undersampled usability. Wavefield data is gathered through laser doppler vibrometers that scan over a mechanically excited solid and thus, to capture any phenomenon not observable in standing waves (such as interaction with a wavefront), excitation must be achieved and allowed to die down each time the scanning proceeds. This process takes time, and so the level of undersampling to which anomaly detection remains viable could be an important consideration in data collection.

The ability to reconstruct undersampled wavefield data is explored in depth in [21], which notes that data spatiotemporally sampled at twice its highest frequency (with respect to the Fourier transform) can be reconstructed *exactly*, as can data sampled at the laxer Landau rate. In theory, intelligent directional spectral decomposition could reconstruct a wavefield whose propagation's frequency is known from this minimal sampling, and thus only sampling rates below this threshold pose a *theoretical* challenge to any anomaly detection technique.

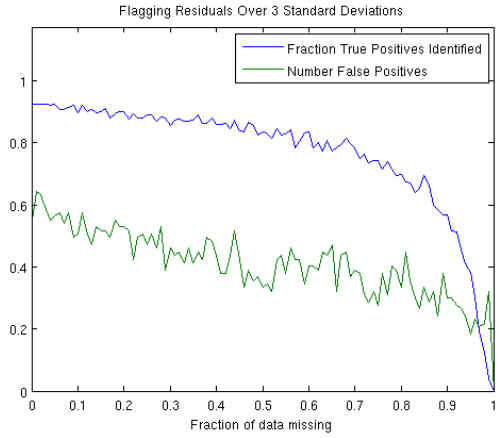
In [21], a sampling rate of 2.5% was found to correspond to this theoretical limit for a particular wavefield data set, and their anomaly-detection (based on singular value decomposition and sparse coding) worked well to around 6-7% sampling.

This report uses similar, if not identical data, and Figure 27 expresses the performance of anomaly detection using dictionary learning on training data, while Figure 28 shows the undersampled data being used, for which numerical techniques are clearly necessary, as the heavily undersampled anomalies are invisible to the naked eye. As can be seen, reasonable performance holds up until approximately 10% sampling and then drops off sharply. If false positives are not heavily penalized (allowing around one per detected anomaly), performance persists until 5% sampling. With ideal tuning these methods would perform better, but using flexible methods, such as averaging results over multiple parameter values, is more general and realistic.

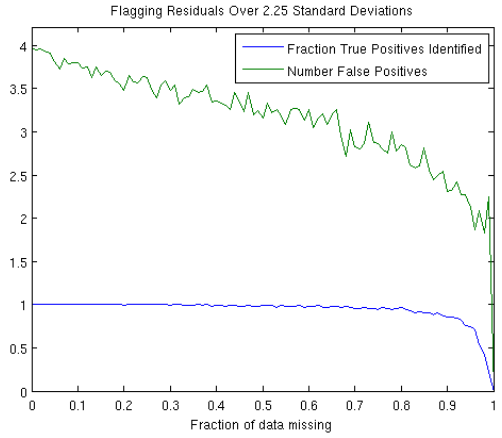
7 Conclusion

Using dictionary learning to detect structual anomalies in propagating wavefield data and saliency in natural images has been shown to be successful. Our three main approaches of finding patches with large residuals,

Figure 27: Performance of anomaly detection in undersampled data using supervised dictionary learning

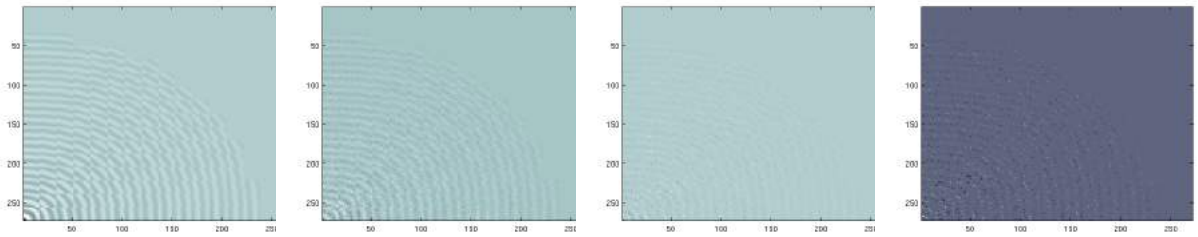


(a) Strictly limiting false positives



(b) Allowing approximately one false positive per defect

Figure 28: Undersampled wavefield data



(a) 100% sampling (b) 50% sampling (c) 20% sampling (d) 10% sampling

with large influence or with a significant use of a rarely used dictionary atom are able to detect anomalies correctly with tolerable false positives. While the three methods work best at specific dictionary sizes, we looked into clustering-based techniques to choose an optimal dictionary size. Additionally, we use spatial scoring to help detect the entirety of a salient object and we demonstrate supervised dictionary learning's ability to detect anomalies in undersampled data. Future work can begin to explore more advanced spare coding schemes, such as tree-sparse or nonlinear manifold representations, as well as perform more comprehensive testing on larger and more diverse datasets.

References

[1] Michal Aharon, Michael Elad, and Alfred Bruckstein. K-svd: Design of dictionaries for sparse representation. *Proceedings of SPARS*, 5:9–12, 2005.

[2] Andreas Argyriou, Rina Foygel, and Nathan Srebro. Sparse prediction with the k-support norm. *arXiv preprint arXiv:1204.5043*, 2012.

[3] Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Convex optimization with sparsity-inducing norms. *Optimization for Machine Learning*, pages 19–53, 2011.

[4] Vic Barnett and Toby Lewis. *Outliers in statistical data*, volume 3. Wiley New York, 1994.

[5] MAURICE S Bartlett and DG Kendall. The statistical analysis of variance-heterogeneity and the logarithmic transformation. *Supplement to the Journal of the Royal Statistical Society*, 8(1):128–138, 1946.

[6] Irad Ben-Gal. Outlier detection. In *Data Mining and Knowledge Discovery Handbook*, pages 131–146. Springer, 2005.

[7] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *ACM Sigmod Record*, volume 29, pages 93–104. ACM, 2000.

[8] J Randall Brown. Error analysis of some normal approximations to the chi-square distribution. *Journal of the Academy of Marketing Science*, 2(3):447–454, 1974.

[9] Emmanuel J Candès and Michael B Wakin. An introduction to compressive sampling. *Signal Processing Magazine, IEEE*, 25(2):21–30, 2008.

[10] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):15, 2009.

[11] R Dennis Cook. Detection of influential observation in linear regression. *Technometrics*, pages 15–18, 1977.

[12] R Dennis Cook and Sanford Weisberg. *Residuals and influence in regression*, volume 5. Chapman and Hall New York, 1982.

[13] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[14] Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *Image Processing, IEEE Transactions on*, 15(12):3736–3745, 2006.

[15] Eleazar Eskin, Andrew Arnold, Michael Prerau, Leonid Portnoy, and Sal Stolfo. A geometric framework for unsupervised anomaly detection. In *Applications of data mining in computer security*, pages 77–101. Springer, 2002.

[16] A Evgeniou and Massimiliano Pontil. Multi-task feature learning. In *Advances in neural information processing systems: Proceedings of the 2006 conference*, volume 19, page 41. The MIT Press, 2007.

[17] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, June 1981.

[18] Sir Ronald Aylmer Fisher, Statistiker Genetiker, Ronald Aylmer Fisher, Statistician Genetician, Great Britain, Ronald Aylmer Fisher, and Statisticien Généticien. *Statistical methods for research workers*, volume 14. Oliver and Boyd Edinburgh, 1970.

[19] Anup K Ghosh, James Wanken, and Frank Charron. Detecting anomalous and unknown intrusions against programs. In *Computer Security Applications Conference, 1998. Proceedings. 14th Annual*, pages 259–267. IEEE, 1998.

[20] Patrick R Gill, Albert Wang, and Alyosha Molnar. The in-crowd algorithm for fast basis pursuit denoising. *Signal Processing, IEEE Transactions on*, 59(10):4595–4605, 2011.

[21] Stefano Gonella and Jarvis Haupt. Automated defect localization via low rank plus outlier modeling of propagating wavefield data. Revised draft for Transactions on Ultrasonics, ferroelectrics and frequency control, IEEE, 2012.

[22] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[23] Wenjie Hu, Yihua Liao, and V Rao Vemuri. Robust anomaly detection using support vector machines. In *Proceedings of the international conference on machine learning*, pages 282–289, 2003.

[24] Kingsly Leung and Christopher Leckie. Unsupervised anomaly detection in network intrusion detection using clusters. In *Proceedings of the Twenty-eighth Australasian conference on Computer Science-Volume 38*, pages 333–342. Australian Computer Society, Inc., 2005.

[25] Tobias Leutenegger and Jürg Dual. Detection of defects in cylindrical structures using a time reverse method and a finite-difference approach. *Ultrasonics*, 40(1):721–725, 2002.

[26] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(2):353–367, 2011.

[27] Cewu Lu, Jianping Shi, and Jiaya Jia. Online robust dictionary learning.

[28] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 689–696. ACM, 2009.

[29] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research (JMLR)*, 11:19–60, 2010.

[30] Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman. Non-local sparse models for image restoration. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2272–2279. IEEE, 2009.

[31] Bruno A. Olshausen and David J. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 37(23):3311 – 3325, 1997.

[32] Yagyensh Chandra Pati, Ramin Rezaiifar, and PS Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*, pages 40–44. IEEE, 1993.

[33] Niels Portzgen, Dries Gisolf, and Gerrit Blacquiere. Inverse wave field extrapolation: A different ndi approach to imaging defects. *Ultrasonics, Ferroelectrics and Frequency Control, IEEE Transactions on*, 54(1):118–127, 2007.

[34] Ignacio Ramírez and Guillermo Sapiro. An mdl framework for sparse coding and dictionary learning. *Signal Processing, IEEE Transactions on*, 60(6):2913–2927, 2012.

[35] Nikhil Rao and Fatih Porikli. A clustering approach to optimize online dictionary learning. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 1293–1296. IEEE, 2012.

[36] Mark Rudelson and Roman Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics*, 61(8):1025–1045, 2008.

[37] Mei-Ling Shyu, Shu-Ching Chen, Kanoksri Sarinnapakorn, and LiWu Chang. A novel anomaly detection scheme based on principal component classifier. Technical report, DTIC Document, 2003.

[38] P Sriram, JI Craig, and S Hanagud. A scanning laser doppler vibrometer for modal testing. *International Journal of Analytical and Experimental Modal Analysis*, 5:155–167, 1990.

[39] Jean-Luc Starck, Fionn Murtagh, and Jalal M Fadili. *Sparse image and signal processing: wavelets, curvelets, morphological diversity*. Cambridge University Press, 2010.

[40] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[41] Ivana Tosic and Pascal Frossard. Dictionary learning. *Signal Processing Magazine, IEEE*, 28(2):27–38, 2011.

[42] Edwin B Wilson and Margaret M Hilferty. The distribution of chi-square. *Proceedings of the National Academy of Sciences of the United States of America*, 17(12):684, 1931.

[43] John Winn, Antonio Criminisi, and Thomas Minka. Object categorization by learned universal visual dictionary. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1800–1807. IEEE, 2005.

[44] Junchi Yan, Mengyuan Zhu, Huanxi Liu, and Yuncai Liu. Visual saliency detection via sparsity pursuit. *Signal Processing Letters, IEEE*, 17(8):739–742, 2010.

[45] Meng Yang, Lei Zhang, Jian Yang, and David Zhang. Robust sparse coding for face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 625–632. IEEE, 2011.

[46] Peng Zhao, Guilherme Rocha, and Bin Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37(6A):3468–3497, 2009.

[47] Mingyuan Zhou, Haojun Chen, Lu Ren, Guillermo Sapiro, Lawrence Carin, and John W Paisley. Non-parametric bayesian dictionary learning for sparse image representations. In *Advances in Neural Information Processing Systems*, pages 2295–2303, 2009.