Chapter 3

LINEAR MODELS FOR OUTLIER DETECTION

"My nature is to be linear, and when I'm not, I feel really proud of myself." – Cynthia Weil

1. Introduction

The different dimensions in real data sets are highly correlated with one another. This is because the different attributes are usually generated by the same underlying process in closely related ways. In the classical statistics literature, this is referred to as *regression modeling*, a parametric form of correlation analysis. Some forms of correlation analysis attempt to predict individual attribute values from others, whereas other forms summarize the entire data in the form of latent variables. An example of the latter is the method of *principal component analysis*. Both forms of modeling can be very useful in different scenarios of outlier analysis. This chapter will discuss the different methods for using linear correlation analysis for outlier detection.

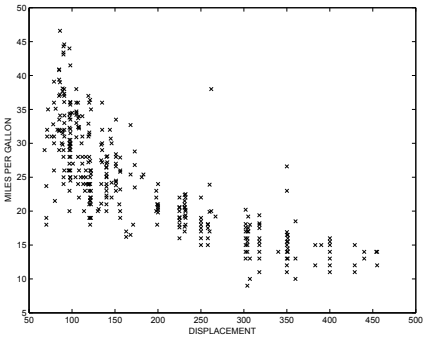
The main assumption of this model is that the data is embedded in a lower dimensional subspace. In the case of proximity-based methods, which will be discussed in the next chapter, the goal is to determine specific *regions of the space* in which outlier points behave very differently from other points. On the other hand, in linear methods, the goal is to find *lower dimensional subspaces*, in which the outlier points behave very differently from other points. This can be viewed as an orthogonal point of view to clustering- or nearest-neighbor based methods, which try to summarize the data *horizontally* (i.e. on the rows or data values), rather than *vertically* (i.e. on the columns or dimensions). As will be discussed in the chapter on high-dimensional outlier detection, it is in

C.C. Aggarwal, *Outlier Analysis*, DOI 10.1007/978-1-4614-6396-2_3,

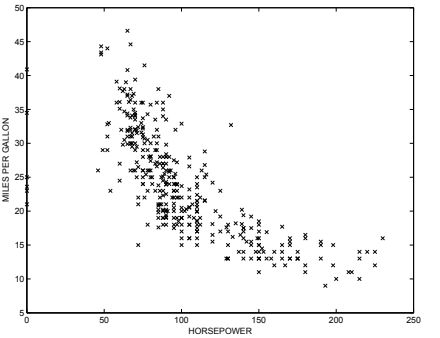
principle, possible to combine these methods for more general local subspace models, which can determine outliers on the basis of a combination of horizontal and vertical criteria.

The assumption of approximately linear correlations is a critical one for ensuring the effectiveness of the model. This may or may not be true for a given data set. For example, consider the behavior of two data sets from the *UCI Machine Learning Repository* [169]. In particular, consider the behavior of the *Autompg* and *Arrythmia* data sets from this repository. The first data set measures various characteristics of cars, and relates them to the mileage (mpg) of the cars. The second data set contains different kinds of features derived from ECG readings of human patients.

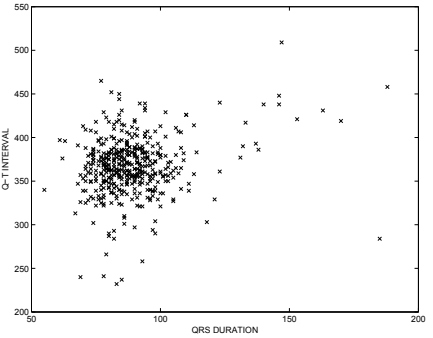
In the first set of Figures 3.1(a) and (b), the dependence of the *Miles per Gallon* attribute has been shown on each of the *displacement* and *horsepower* attributes respectively for the *Autompg* data set. It is evident that a significant level of correlation exists between these attributes. While a significant amount of noise exists in the data, the linear dependence between the attributes is apparent. In fact, it can be shown for this data set, that with increasing dimensionality (by picking more attributes from the data set), the data can be aligned along much lower dimensional planes. This is also evident in the 3-dimensional plot of Figure 3.1(e). On the other hand, when various views along three of the measured dimensions of the *Arrythmia* data set (Figures 3.1(c), (d) and (f)) are examined, it is evident that the data separates out into two clusters, one of which is slightly larger than the other. Furthermore, it is rather hard to embed this kind of data distribution into a lower dimensional subspace. This data set is much more suitable for proximity-based analysis, which will be presented in Chapter 4. The reason for introducing this example is to revisit the point made in the first chapter about the impact of the choices made during the crucial phase of picking the correct data model. In general, the most difficult case is when different views of the *same* data set may be suitable for different models. Such data sets are best addressed with the use of subspace methods discussed in Chapter 5, which can combine the power of row and column selection for outlier analysis. However, in many cases, simplified models such as linear models or proximity-based models are sufficient, without incurring the complexity of subspace methods. From a model-selection perspective, exploratory and visual analysis of the data is rather critical in the first phase of outlier detection in order to find out whether a particular data model is suitable for a particular data set. This is particularly true in the case of unsupervised data models.



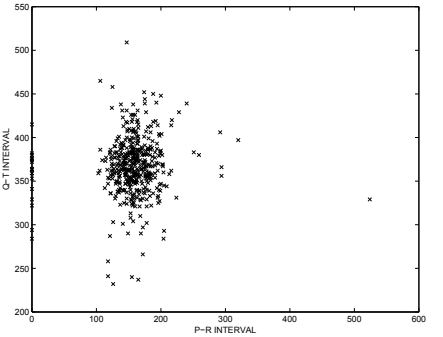
(a) View 1 (*Autompg*)



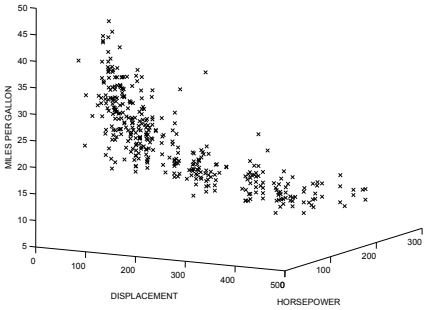
(b) View 2 (*Autompg*)



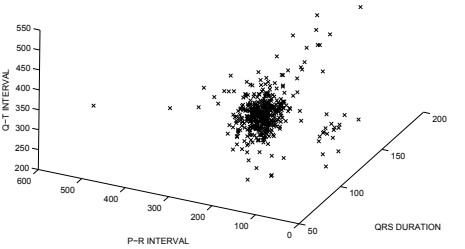
(c) View 1 (*Arrythmia*)



(d) View 2 (*Arrythmia*)



(e) 3-d View (*Autompg*)



(f) 3-d View (*Arrythmia*)

Figure 3.1. Effectiveness of linear assumption is data set dependent

In this chapter, two main classes of linear models will be studied. The first class of models uses statistical regression modeling between dependent and independent variables in order to determine *specific kinds of dependencies* in the data. Such forms of modeling are more useful when some of the attributes in an application should be monitored on a prioritized basis (eg. the *last* value of a time-series, where the previous history of values are the independent variables used for modeling). The second class of models uses principal component analysis in order to treat all attributes in a homogeneous way, and determine the lower dimensional subspaces of projection. At a technical and mathematical level, both forms of modeling are quite similar, and use very similar methods in order to derive the optimal lower dimensional representations. The main difference is in how the objective function of the two models is formulated.

It should be emphasized that regression-analysis is used extensively to detect anomalies in time-series data, and many of the basic techniques discussed in this chapter are applicable to that scenario as well. However, since the time-series aspect of the problem is also based on dependencies of *temporally adjacent* data values, there are a number of subtle differences in how anomalies are detected in those cases. Therefore, in this chapter, the much simpler case of multidimensional outlier analysis will be addressed. At the same time, the discussion will be general enough, so that the fundamentals necessary for the discussion of applying regression analysis in the time-series scenario (Chapter 8) are introduced.

This chapter is organized as follows. In section 2, the basic linear regression models for outlier analysis will be introduced. In section 3, the principal component method for outlier analysis will be introduced. This can be considered an important special case of linear regression models, which is used frequently in outlier analysis. Therefore it is given a dedicated treatment in its own section. Section 4 will study the limitations of linear models for outlier analysis. Section 5 contains the conclusions and summary.

2. Linear Regression Models

In linear regression, the observed values in the data are modeled using a linear system of equations. Specifically, the different dimensions in the data are related to one another using a set of linear coefficients. Since the number of observed values are typically much larger than the dimensionality of the data, this system of equations is an *over-determined* one, and cannot be solved exactly. Therefore, these models optimize

the square error of the deviations of data points from values predicted by the linear model. The exact choice of the error function determines whether a particular variable is treated specially (i.e. error of predicted variable value), or whether variables are treated homogeneously (i.e. error distance from estimated lower dimensional plane). These different choices of the error function do *not* lead to the same model. In fact, as the following discussion will show, the models can be very different *especially in the presence of outliers*.

Regression analysis is generally considered an important application of its own in statistics. In classical instantiations of this application, it is desirable to learn a specific dependent variable from a set of independent variables. This is a common scenario in time-series analysis, which will be discussed in detail in Chapter 8. Thus, a specific variable is treated *specially* from the other variables. Most applications on outlier analysis do not treat any particular variable as special, and the definition of outliers is generally based on the *overall* distribution of the underlying data points. However, the special case of regression analysis with dependent variables is also important in many applications. This is because in many real-life domains such as temporal and spatial data, the attributes are partitioned into *contextual* and *behavioral* attributes. In such cases, a particular behavioral attribute value is predicted as a function of the behavioral attributes in its *contextual* neighborhood in order to determine deviations from expected values. Therefore, the importance of the dependent variable is paramount. In such cases, outliers are defined on the basis of how other independent variables impact the dependent variable, and anomalies within the relationships of independent variables with each other are considered less important. The identification of outliers in such cases is also very useful for *noise reduction* in regression modeling, which is an important problem in its own right. This problem is considered so important, that an entire book has been devoted to this subject [387]. Therefore, the special case of regression analysis with dependent variables will be studied first. Then, the general application of regression methods to outlier analysis will be discussed. The focus in this section is to discuss the impact of outliers on the linear modeling process of a *dependent variable* on a set of *explanatory* variables. The discussion of this case also sets the stage for a more detailed discussion for the cases of time-series data in Chapter 8, and spatial data in Chapter 10.

In a later subsection, the more general problem of utilizing regression modeling for generic outlier analysis will be discussed. In that case, no particular variable is considered special, and regression modeling is a *tool* (rather than an application in its own right). Such a tool may be used

either to remove noise for other applications, or to identify interesting anomalies. This latter form of the problem is the focus of most of this book, though dependent variable regression analysis is also important in many applications such as time-series data.

2.1 Modeling with Dependent Variables

A variable Y can be modeled as a linear function of d dependent variables as follows:

$$Y = \sum_{i=1}^{d} a_i \cdot X_i + a_{d+1}$$

The variable Y is the response variable or the dependent variable, and the variables $X_1 \ldots X_d$ are the independent or the explanatory variables. The coefficients $a_1 \ldots a_{d+1}$ need to be learned from the data. The data may contain N different instances, which provide examples of how Y may be related to the different values of X_i. The jth instances of the data are denoted by $(x_{j1} \ldots x_{jd})$ and y_j. The jth instance of the response variable is related to the explanatory variables as follows:

$$y_j = \sum_{i=1}^{d} a_i \cdot x_{ji} + a_{d+1} + \epsilon_j$$

Here ϵ_j represents the error in modeling the jth instance. In *least squares regression*, the goal is to determine the regression coefficients $a_1 \ldots a_{d+1}$, which minimize the error $\sum_{j=1}^{N} \epsilon_j^2$. The $N \times (d+1)$-matrix whose j-th row is $(x_{j1} \ldots x_{jd}, 1)$ is denoted by U, and the $N \times 1$ matrix of the different values of Y is denoted by V. Thus, the first d dimensions of U can be considered a d-dimensional data set containing the N instances of the independent variables, and V is corresponding vector of response variables. The $(d+1) \times 1$ column vector of coefficients $a_1 \ldots a_{d+1}$ is denoted by A. This creates an *over-determined* system of equations denoted by:

$$V \approx U \cdot A \tag{3.1}$$

The least-squares error *of predicting the response variable* is optimized by minimizing $||V - U \cdot A||$ over all values of the coefficient A. It will be seen later, that more general ways of formulating the error function may exist, rather than simply predicting the error of the response variable. Clearly, the choice of the error function has an impact on the optimal hyperplane found by the regression analysis process. It can be shown through simple optimization methods via differential calculus, that the optimal coefficients for this minimization problem is provided by the

following equation:

$$A = (U^T \cdot U)^{-1} \cdot (U^T \cdot V) \tag{3.2}$$

Note that $U^T \cdot U$ is a $(d+1) \times (d+1)$ matrix, which needs to be inverted in order to solve this system of equations. The system of equations above thus needs to be over-determined in order for the matrix $U^T \cdot U$ to have full rank, and be invertible. The closed form solution to this problem is particularly convenient, and is one of the cornerstones of regression analysis in classical statistics. It is useful to examine the special case of two dimensional data:

$$Y = a_1 \cdot X_1 + a_2 \tag{3.3}$$

In this case, the estimation of the coefficient a_1 has a particularly simple form, and it can be shown that the best estimate for a_1 is as follows:

$$a_1 = \frac{Cov(X_1, Y)}{Var(X_1)}$$

Here $Var(\cdot)$ and $Cov(\cdot)$ correspond to the variance and covariance of the underlying random variables. The value a_2 can further be easily estimated, by plugging in the means of X_1 and Y into the linear dependence, once a_1 has been estimated. In general, if X_1 is regressed on Y instead of the other way around, one would have obtained $a_1 = \frac{Cov(X_1,Y)}{Var(Y)}$. Note that the regression dependencies would have been different for these cases. This shows the impact of the error term on the final regression plane which is found by the method.

The set of coefficients $a_1 \ldots a_{d+1}$ define a lower dimensional hyperplane which fits the data as well as possible in order to optimize the error in the dependent variable. This hyperplane may be different for the same data set, depending upon which variable is chosen as the dependent variable. In order to explain this point, let us examine the behavior of two attributes from the *Auto-Mpg* data set of the UCI Machine Learning repository [169].

Specifically, the second and the third attributes of the *Auto-Mpg* data set correspond to the *Displacement* and *Horsepower* attributes in a set of records corresponding to cars. The scatter plot for this pair of attributes is illustrated in Figure 3.2. Three regression planes have been shown in this figure, which are as follows:

- One regression plane is drawn for the case, when the *Horsepower (Y-axis)* is dependent on the *Displacement (X-axis)*. The residual in this case is the error of prediction of the *Horsepower* attribute. The sum of squares of this residual is optimized.

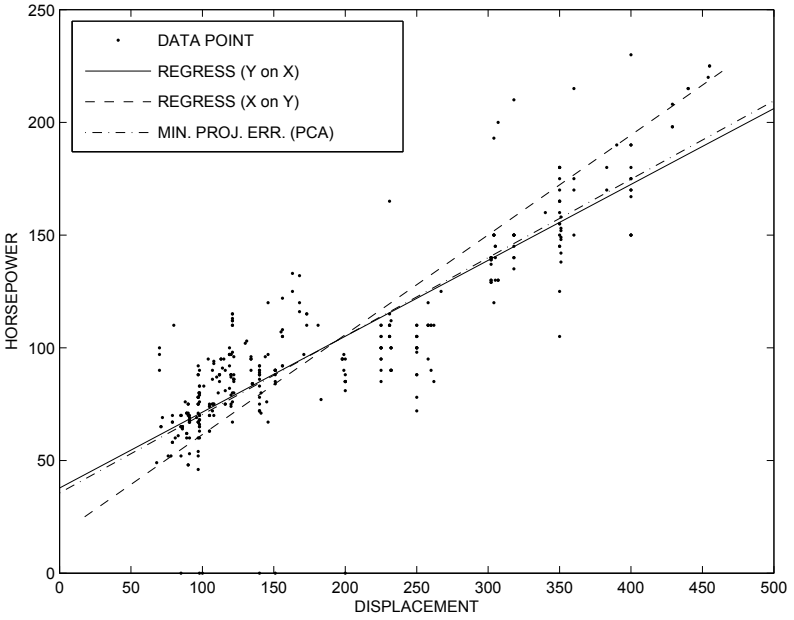
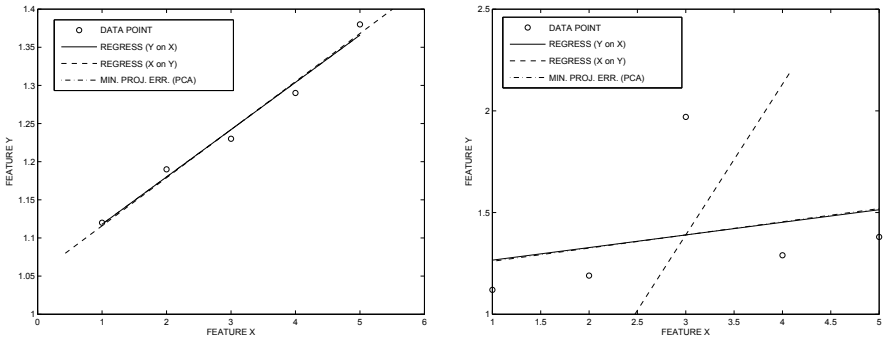


Figure 3.2. Optimal regression plane depends upon the choice of residual which is optimized

- The second regression plane is drawn for the case, when the *Displacement (X-axis)* is dependent on the *Horsepower (Y-axis)*. The residual in this case is the error in prediction of the *Displacement* attribute.
- In the last case, the goal is to optimize the mean square error of the data points in terms of their absolute distance to the best fitting hyperplane. Thus, the residual in this case is the distance of each point to the hyperplane, in a direction which is normal to the hyperplane. Thus, this hyperplane minimizes the mean square distances between the data points, and their projection into the hyperplane. So far, the determination of such a hyperplane has not been discussed. This will be done in a later section on Principal Component Analysis (PCA).

It is evident from Figure 3.2 that the optimal hyper-planes in these different cases are quite different. While the optimization of the mean square projection distance produces a hyperplane which is somewhat similar to the case of Y-on-X regression, the two are not the same. This is because these different cases correspond to different choices of errors on the residuals which are optimized, and therefore correspond to



(a) Regression with no outliers

(b) Regression with one perturbed point

Figure 3.3. Drastic effects of outliers on quality of regression analysis

different best fitting hyperplanes. It is also noteworthy that the three projection planes are collinear and pass through the mean of the data set.

When the data fits the linear assumption very well, all these hyperplanes are likely to be very similar and not very different from one another. However, the presence of noise and outliers can result in rather drastic negative effects on the modeling process, when some of the outliers show significant deviations. In order to illustrate this point, a variation of an example from [387] is used. In Figure 3.3, the different regression planes for two sets of five data points have been presented corresponding to different dependent variables. The two sets of five data points in Figures 3.3(a) and (b) are different by only one point, in which the Y-coordinate was assumed to be somehow perturbed during data collection. As a result, this point does not fit the remaining data very well.

The original data set in Figure 3.3(a) fits the linear assumption very well. Therefore, all the three regression planes tend to be very similar to one another. However, after the perturbation of a single data point, the resulting projection planes are drastically perturbed. In particular, the X on Y-regression plane is significantly perturbed so as to no longer represent the real trends in the underlying data set. It is also noteworthy that the optimal projection plane is closer to the more stable of the two regression models. This is a general property of optimal projection planes, since they optimize their orientation in a stable way so as to globally fit the data well. The determination of such planes will be discussed in the next section.

Clearly, the removal of outliers is crucial in such applications, in order to improve the quality of the regression analysis. Therefore, a useful approach would be to examine the residuals ϵ_j, and remove those data points which are detrimental for outlier analysis. The mean of these residuals is expected to be 0, and the variance of these residuals can be estimated directly from the data.

The most common assumption for outlier analysis is to assume that the error term ϵ_i is a normal distribution, which is centered at zero. Then, the t-value test discussed in Chapter 2 can be used directly on the different residuals, and the outlying observations can be subsequently removed. The normal assumption on the residuals implies that the vector of coefficients is also normally distributed with mean and variances, as discussed earlier. When the outliers have drastic effects on the regression, such as in the case of the X-on-Y regression in Figure 3.3(b), the removal of outliers is likely to result in the removal of the wrong observations, since the regression parameters are drastically incorrect. On the other hand, in all cases, the projection based minimization seems to provide more robust results (as opposed to picking a particular dependent variable) to the presence of outliers. Therefore, even for dependent variable analysis, it may sometimes be helpful to use such projection-based error minimization. This is the method of *Principal Component Analysis (PCA)*. The formulation for this case will be discussed in the next subsection, and a more detailed discussion of the solution and different aspects of principal component analysis will be discussed in a dedicated section of its own.

2.2 Regression Modeling for Mean Square Projection Error

The previous section discussed the case, where a particular variable is considered special, and the optimal plane is determined in order to minimize the mean-square error of the residuals for this variable. In the most general form of regression-modeling, all variables are treated in a similar way, and the optimal regression plane is determined the minimize the *projection error* of the data to the plane. This can be considered an unsupervised form of outlier analysis, because the outliers are determined without treating any particular variable specially.

The projection error of the data to the plane is the sum of the squares of the distances of the points to their projection into the plane. The projection of a point to the plane is performed by using the normal direction to the plane which passes through the data point and the plane. The point at which this normal intersects the plane is the projection

point. Thus, in this case, let us assume that we have a set of variables $X_1 \ldots X_d$, and the corresponding regression plane is as follows:

$$a_1 \cdot X_1 + \ldots + a_d \cdot X_d + a_{d+1} = 0 \tag{3.4}$$

Each variable is associated with a coefficient, and the "special" (dependent) variable (without a coefficient) is missing in this case. For simplification of the subsequent discussion of computing distances of different observations to this plane, a normalization constraint will be assumed.

$$\sum_{i=1}^{d} a_i^2 = 1 \tag{3.5}$$

Note that the $(d+1)$th term (constant coefficient) is not used in the normalization. As before, let U be a $N \times (d+1)$ matrix containing the set of N observations corresponding to the variables $X_1 \ldots X_d, 1$. The last column in the matrix U corresponds to the constant term, and therefore only contains unit values. Let A be a column vector containing $a_1 \ldots a_{d+1}$. It can be shown that the N-dimensional column vector of distances for the different data points to this regression plane is given by $U \cdot A$. The L_2-norm $||U \cdot A||_2$ of the column vector of distances is the objective function, which needs to be minimized over the different possible values of the coefficients $a_1 \ldots a_{d+1}$, under the normalization constraint. It can be shown that a effective (and much more general) solution to the problem can be obtained with Principal Component Analysis (PCA). Because of its importance to outlier analysis, this method will be discussed in a dedicated section of its own, along with corresponding applications.

3. Principal Component Analysis

The least-squares formulation of the previous section simply tries to find a *single* $(d-1)$-dimensional hyperplane which has an optimum fit to the data values. The principal component analysis method can be used to solve a *generalized* version of this problem. Specifically, it can find optimal representation hyperplanes of *any* dimensionality. Specifically, the PCA method can determine the k-dimensional hyperplane (for any value of $k < d$), which minimizes the squared projection error. In principal component analysis, the $d \times d$ covariance matrix over d-dimensional data is computed, where the (i, j)th entry is equal to the covariance between the dimensions i and j for the set of N observations of the variables $X_1 \ldots X_d$.

It is easier to think in terms of a multidimensional data set of dimensionality d and size N, rather than a set of d variables with N

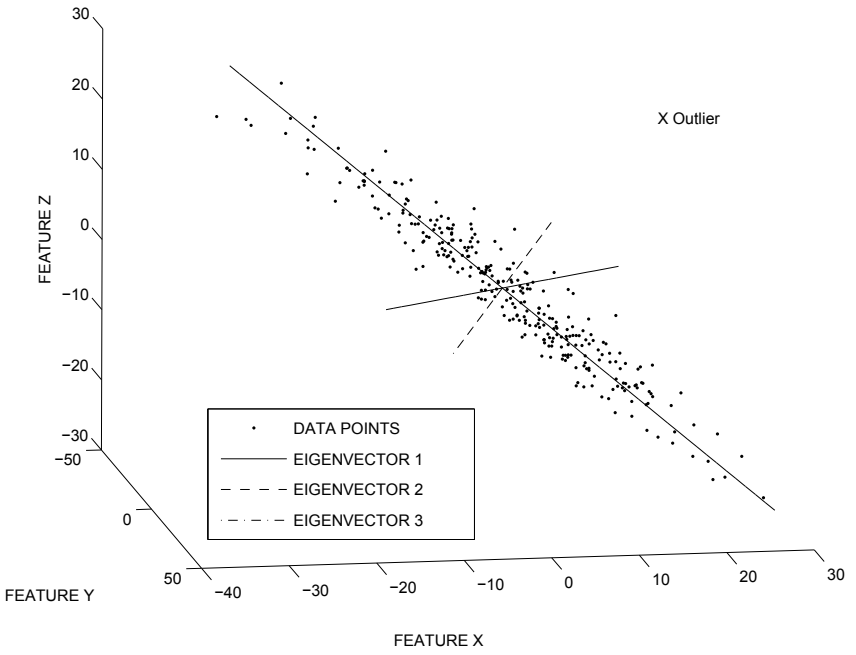


Figure 3.4. Eigenvectors correspond to directions of correlations in the data. A small number of eigenvectors can capture most of the variance in the data.

observations (as presented in the earlier portions of this chapter). Thus, in the context of a multidimensional data set, the value of d represents the dimensionality, and the value of N represents the number of records (or rows). The i-th record is a row of the multidimensional data set, and is denoted by $R_i = [x_{i1} \ldots x_{id}]$, where x_{ij} is the ith observation for the jth variable X_j. Let us denote the $d \times d$ covariance matrix of the data set by Σ, in which the (i, j)th entry is the covariance between the ith and jth dimensions. This matrix can be shown to be symmetric and positive semi-definite. It can therefore be diagonalized as follows:

$$\Sigma = P \cdot D \cdot P^T$$

Here D is a diagonal matrix, and P is an orthonormal matrix, whose columns correspond to the (orthonormal) eigenvectors of Σ. The corresponding entries in the diagonal matrix D provide the eigenvalues. These orthonormal vectors provides the axes directions along which the data should be projected. The key properties of principal component analysis, which are relevant to outlier analysis, are as follows:

PROPERTY 3.1 (PCA PROPERTIES) *Principal component analysis provides a set of eigenvectors satisfying the following properties:*

- *If the top-k eigenvectors are picked (by largest eigenvalue), then the k-dimensional hyperplane defined by these eigenvectors, and passing through the mean of the data, is a plane for which the mean square distance of all data points to it is as small as possible among all hyperplanes of dimensionality k.*
- *If the data is transformed to the axis-system corresponding to the orthogonal eigenvectors, the variance of the transformed data along each eigenvector dimension is equal to the corresponding eigenvalue. The covariances of the transformed data in this new representation are 0.*
- *Since the variances of the transformed data along the eigenvectors with small eigenvalues are low, significant deviations of the transformed data from the mean values along these directions may represent outliers.*

A formal proof of these properties may be found in [244]. Note that this provides a *much* more general solution than the determination of the optimal coefficients of Equation 3.4. Specifically, the optimal solution for the coefficients of Equation 3.4 may be simply derived as the coefficients of the top *one* eigenvector representing $a_1 \ldots a_d$, and the constant term a_{d+1} may be inferred by substituting the mean of the data in Equation

3.4. On the other hand, the PCA-solution provides a recursive solution of *any* dimensionality by picking the top k eigenvectors.

The data can be transformed to this new axis system, with transformed d-dimensional records denoted by $Y_1 \ldots Y_N$. This can be achieved by using the product between the original vector representation R_i and the orthonormal eigenvector matrix P containing the new axis-system:

$$Y_i = [y_{i1} \ldots y_{id}] = R_i \cdot P$$

In this new representation, the inter-attribute covariances of Y_i are zero, and most of the variances along the individual attributes correspond to the coordinates along the eigenvectors with the largest eigenvalues. In fact, the eigenvalues represent the variances of the transformed vectors Y_i along these directions in the new coordinate system. For example, if the jth eigenvalue is very small, then the value of y_{ij} in this new transformed representation does not vary much over the different values of i. The beautiful part about PCA is that, in a single shot, it provides all the key directions of *global* correlation, which retain most of the information in the underlying data. These directions are also referred to as the *principal components* in the data, since their second-order correlations are zero, and most of the variance of the data is retained along these directions. In many real scenarios involving very high-dimensional data sets, a very large fraction of the eigenvalues often turn out to be very close to zero. This essentially means that most of the data aligns along a *much lower dimensional subspace*. This is very convenient from the perspective of outlier analysis, because the observations which lie very far away from these directions of projection can be assumed to be outliers. For example, for an eigenvector j which has a small eigenvalue, a large deviation of y_{ij} for the ith record from other values of y_{kj} is indicative of outlier behavior. This is because the values of y_{kj} do not vary much, when j is fixed and k is varied. Therefore, the value y_{ij} is unusual.

The effectiveness of principal component analysis in exposing outliers from the underlying data set can be illustrated with an example. Consider the scatterplot of the 3-dimensional data illustrated in Figure 3.4. In this case, the corresponding eigenvectors have been ordered by decreasing eigenvalues (variances), though this is not immediately obvious from the figure in this 2-d perspective. In this case, the standard deviation along the first eigenvector is three times that along the second eigenvector and nine times that along the third eigenvector. Thus, most of the variance would be captured in the lower-dimensional subspace formed by the top two eigenvectors, though a significant amount of variance would also be captured by picking only the first eigenvector. If the normal distances of the original data points to the 1-dimensional line

corresponding to the first eigenvector (and passing through the mean of the data) are computed, the data point 'X' in the figure would be immediately exposed as an outlier. In the case of high-dimensional data, most of the variance of the data can be captured along a much lower k-dimensional subspace. The residuals for the data points can then be computed by examining the projection distances to this k-dimensional hyperplane passing through the mean of the data points. Data points which have very large distances from this hyperplane can be discarded as outliers. As before, it is possible to model these residuals as a normal distribution, and perform a Z-value test for the corresponding statistical significance.

A more accurate way of modeling the abnormality level without picking any particular set of k dimensions, would be to use the eigenvalue to compute the normalized distance of the data point to the centroid along the direction of *each principal component*. Let $\overline{e_j}$ be the jth eigenvector with a variance (eigenvalue) of λ_j along that direction. The overall normalized outlier score of a data point $\overline{X}$, to the centroid $\overline{\mu}$ of the data is given by the sum of squares of these values:

$$Score(\overline{X}) = \sum_{j=1}^{d} \frac{|(\overline{X} - \overline{\mu}) \cdot \overline{e_j}|^2}{\lambda_j} \tag{3.6}$$

It is important to note that most of the contribution to the outlier score is provided by deviations along the principal component with small values of λ_j, when a data point deviates significantly along such directions. The sum of the squares of these values over all dimensions is a χ^2-distribution with d degrees of freedom. The value of the aggregate residual is compared to the cumulative distribution for the χ^2-distribution in order to determine a probability value for the level of anomalousness. The aforementioned approach was first used in [406].

While it may not be immediately apparent, the score computed above is closely related to the multivariate extreme value analysis method discussed in section 3.4 of Chapter 2. Specifically, the Mahalanobis distance value between $\overline{X}$ and $\overline{\mu}$ computed in that section is *exactly the same*[1] as the score above, except that the eigenvector analysis above provides a better understanding of how this score is decomposed along the different directions of correlation. This decomposition also allows the ability to use only the dimensions with the small eigenvalues in order to obtain an outlier score, which ignores the long eigenvalues. It is possible to use a score which is constructed with only the *smallest* $\delta < d$ eigenvalues.

[1]See Exercise 11 of this chapter for the systematic steps.

However, it should also be noted that the approach already performs a kind of soft pruning because of the inverse weighting by the eigenvalues in the score. By explicitly pruning the score, the danger is that if a long eigenvector is relevant to the outlier, then that outlier will be missed. It is not uncommon for a rare value to also align along a long eigenvector. An unusual deviation of a similarly correlated nature in two correlated attributes will cause such a situation. In the event that a pruned score is used, the score may be modeled as a χ^2 distribution with δ degrees of freedom. Therefore, the score may be converted into a probability. This is quite desirable, because it provides a clear idea of the outlierness of the underlying object.

Principal component analysis is much more stable to the presence of a few outliers, than the dependent variable analysis methods. This is because principal component analysis computes the errors with respect to the *optimal hyperplane*, rather than a *particular variable*. When more outliers are added to the data, the optimal hyperplane usually does not change drastically enough to impact the choice of data points which should be considered outliers. Therefore, such an approach is more likely to pick the correct outliers, because the regression model is more accurate to begin with. If desired, this approach can be combined with a sequential ensemble methodology of Chapter 1 in order to determine the outliers robustly. In each iteration, the obvious outliers are removed, and a more refined PCA model is constructed. The final outliers are deviation levels in the last iteration of the sequential ensemble.

3.1 Normalization Issues

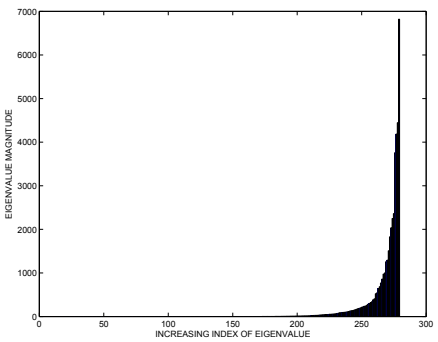
The use of PCA can sometimes provide results which are not very informative, when the scales of the different dimensions are very different. For example, consider a demographic data set containing attributes such as *Age* and *Salary*. The *Salary* attribute may range in the tens of thousands, whereas the *Age* attribute is almost always less than a hundred. The use of PCA would result in the principal components being dominated by the high-variance attributes. For example, for a 2-dimensional data set containing only *Age* and *Salary*, the largest eigenvector will be almost parallel to the *Salary* axis, irrespective of very high correlations between the *Age* and *Salary* attributes. This can reduce the effectiveness of the outlier detection process. Therefore, a natural solution is to normalize the data, so that the variance along each dimension is one unit. This is achieved by dividing each dimension with its standard deviation. This implicitly results in the use of a *correlation matrix* rather than the *covariance matrix* during principal component analysis. Of course, this

issue is not unique to linear modeling, and it is often advisable to use such pre-processing for most outlier detection algorithms.

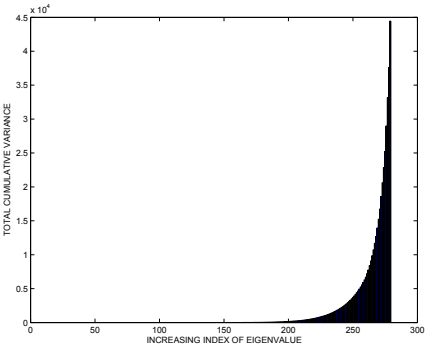
3.2 Applications to Noise Correction

Most of this book is devoted to either removal of outliers as noise, or identification of outliers as anomalies. However, in many applications, it is possible that even though parts of a data record may be erroneous, and may show up as outliers, it may be useful to correct that data record, under the assumption that it should show similarity to the broad patterns in the data. Principal Component Analysis (PCA) provides an approach for achieving this goal. In this case, the core idea of the approach is that *projection of the data point onto the k-dimensional hyperplane corresponding to the largest eigenvalues (and passing through the data mean) provides the optimal correction to the data values.* Obviously such an approach is likely to correct the outlier points significantly more than most of the other normal data points. Some theoretical results (along with experimental evidence) of why such an approach is likely to reduce noise and improve data quality for a variety of applications is provided in [18]. A similar approach to PCA (called *Latent Semantic Indexing*) has also been used in the context of text data, in order to reduce the noise, and significantly improve retrieval quality [133, 355]. In particular, it has been observed in [355] that the use of such dimensionality reduction methods in text data significantly improves the effectiveness of similarity computations, because of the reduction in the noise effects of *synonymy* and *polysemy*. Text representations are inherently noisy because the same word may mean multiple things (synonymy) or the same concept can be represented with multiple words (polysemy). This leads to numerous challenges in virtually all similarity-based applications. The technique of LSI [133] is essentially a variant of PCA, which was originally developed for efficient indexing and retrieval. However, it was eventually observed that the quality of similarity computations, in terms of the underlying precision and recall, actually improves with the use of LSI [355]. This observation was taken to its logical conclusion in [18], where it was theoretically and experimentally shown that significant noise reduction is likely to occur, with the proper use of PCA-based techniques.

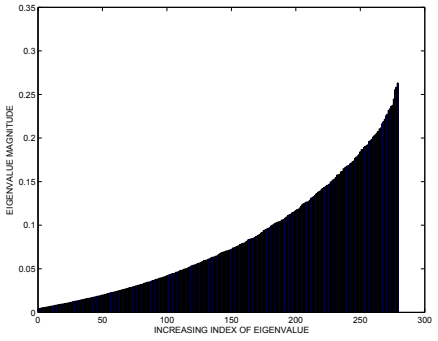
An even more effective approach for noise correction is to combine outlier removal and re-insertion with the correction process. The first step is to perform PCA, and remove the top outliers on the basis of a t-test with respect to the optimal plane of representation. Subsequently, PCA is performed again on this cleaner data set in order to generate



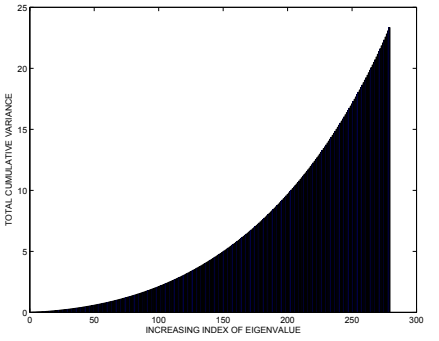
(a) Magnitude of Eigenvalues (Increasing Index): *Arrythmia*



(b) Variance in smallest k Eigenvalues: *Arrythmia*



(c) Magnitude of Eigenvalues (Increasing Index) *Uniform: 452 records only*



(d) Variance in smallest k Eigenvalues *Uniform: 452 records only*

Figure 3.5. Most of the Energy is Retained in a Small Number of Eigenvalues for the *Arrythmia* data set

the projection subspaces more accurately. The projections can then be performed on this corrected subspace. This process can actually be repeated iteratively, if desired in order to provide further refinement. A number of other approaches to perform regression analysis and outlier removal in a robust way are presented in [387].

3.3 How Many Eigenvectors?

As discussed earlier, the eigenvectors with the largest variance provide the most informative subspaces for data representation, and outlier analysis. In many applications such as noise correction, the data needs to be projected into a subspace of lower dimensionality by picking a specific

number of eigenvectors. Therefore, a natural question arises, as to how the dimensionality k of the projection subspace should be determined.

One observation in most real data sets is that the vast number of eigenvalues are relatively small, and most of the variance is concentrated in a few eigenvectors. An example illustrated in Figure 3.5 shows the behavior of the 279 eigenvectors of the *Arrythmia* data set of the UCI Machine Learning Repository [169]. Figure 3.5(a) shows the absolute magnitude of the eigenvalues in increasing order, whereas Figure 3.5(b) shows the total amount of variance retained in the top-k eigenvalues. In essence, Figure 3.5(b) is derived by using the cumulative sum over the eigenvalues in Figure 3.5(a). While it was argued at the beginning of the chapter that the *Arrythmia* data set is weakly correlated along many of the dimensions, on a pairwise basis, it is interesting to note that that it is still possible[2] to find a small number of directions of global correlation along which most of the variance is retained. In fact, it can be shown that the first 215 eigenvalues (out of 279) *cumulatively* contain less than 1% of the variance in the data set.

In other words, most eigenvalues are very small. Therefore, it pays to retain the eigenvectors corresponding to extremely large values, with respect to the average behavior of the eigenvalues. How to determine, what is "extremely large"? This is a classical case of extreme value analysis methods, which were introduced in Chapter 2. Therefore, each eigenvalue is treated as a data sample, and the statistical modeling is used to determine the large values with the use of hypothesis testing. A challenge in this case is that the sample sizes are small. Even for relatively high dimensional data sets (eg. 50-dimensional data sets), the number of samples (50 different eigenvalues) available for hypothesis testing is relatively small. Therefore, this is a good candidate for the t-value test. The t-value test can be used in conjunction with a particular level of significance and appropriate degrees of freedom in order to determine the number of eigenvectors which should be picked for analysis.

[2]Part of the reason for this is that the data set is relatively small with only 452 records. In such cases, it is much easier to find a small number of directions of correlation. As an example, the results of Figure 3.5(c) and (d) show that even for a uniformly distributed data set of the same size, it is possible to find some skews in the eigenvalues. This is one of the limitations of regression analysis, which will be discussed in a later section. Furthermore, the cumulative effects of even weak correlations become magnified with increasing dimensionality, when it is desired to find a much lower dimensional subspace contain the informative projections. This is of course a strength of Principal Component Analysis.

4. Limitations of Regression Analysis

Regression analysis has a few limitations as a tool for outlier detection. The most significant of these shortcomings was discussed at the very beginning of this chapter, in which the data-specific nature of regression analysis was explored. In particular, the data needs to be highly correlated, and aligned along lower dimensional subspaces, in order for regression analysis techniques to be effective. When the data is uncorrelated, but highly clustered in certain regions, such methods may not work effectively. On the other hand, even when the data is weakly correlated on a pairwise basis between different dimensions, it is often the case that subspaces of much lower dimensionality contain most of the variance in the data, because of the cumulative effect of inter-attribute correlations.

Another related issue is that the correlations in the data may not be global in nature. A number of recent analytical observations [7] have suggested that the subspace correlations are specific to particular localities of the data. In such cases, the global subspaces found by PCA are sub-optimal for outlier analysis. Therefore, it can sometimes be useful to combine linear models with proximity-models (discussed in the next chapter), in order to create more general local subspace models. This will be the topic of high-dimensional and subspace outlier detection, which is discussed in detail in Chapter 5.

As with any model-based approach, overfitting continues to be an issue, when used with a small set of data records. In this context, the relationship of the number of records to the data dimensionality is important. For example, if the number of data points are less than the dimensionality, it is possible to find one or more directions along which the variance is zero. Even for cases, where the data size is of greater (but similar) magnitude as the data dimensionality, considerable skew in the variances may be observed. This is evident from the results of Figure 3.5(c) and (d), where there is considerable skew in the eigenvalues for a small set of uniformly distributed data. This skew reduces, as the data size is increased. This is a classic case of overfitting, and it is important to interpret the results of linear modeling carefully, when the data set sizes are small.

The interpretability of regression-based methods is rather low. These methods project the data into much lower dimensional subspaces, which are expressed as a linear (positive or negative) combination of the original feature space. This cannot be easily interpreted in terms of physical significance in many real application. This also has the detrimental effect of reducing the intensional knowledge of the user for a particular

application. This is undesirable, because it is usually interesting to be able to explain *why* a data point is an outlier in terms of the features of the original data space.

Finally, the computational complexity of the approach may be an issue when the dimensionality of the data is large. When the data has dimensionality of d, this results in an $d \times d$ covariance matrix, which may be rather large. Furthermore, the diagonalization of this matrix will slow down at least quadratically with increasing dimensionality. A number of techniques have recently been proposed, which can perform PCA in faster time than quadratic dimensionality [191]. With advances in methods for matrix computation and the increasing power of computer hardware, this issue has ceased to be as much of a problem in recent years. Such dimensionality reduction techniques are now easily applied to large text collections with a dimensionality of several hundreds of thousands of words.

5. Conclusions and Summary

This chapter presents linear models outlier detection. Many data sets show significant correlations among the different attributes. In such cases, linear modeling may provide an effective tool for removing the outliers from the underlying data. Since linear modeling is a tool in of itself for other regression-based applications, the removal of outliers can be very useful for improving the effectiveness of such applications. In most cases, principal component analysis provides the most effective methods for outlier removal, because it is more robust to the presence of a few outliers in the data. A major limitation of linear modeling is that it does not try to recognize that the correlation behavior of the data in different localities may be different, and tries to fit the data into a single global model. However, it provides a general framework, which can be used for generalized local linear models, which are discussed in Chapter 5.

6. Bibliographic Survey

The relationships between the problems of regression and outlier detection has been explored extensively in the literature [387]. Outlier analysis is generally seen as an enormous challenge to robust regression in terms of the *noise* effects, and this has motivated an entire book on the subject. In many cases, the presence of outliers may lead to unstable behavior of regression analysis methods. An example of this was illustrated in in Figure 3.3(b) of this chapter, where a single outlier completely changes the regression slope to one which does not reflect the

true behavior of the data. It can be shown that under certain circumstances, a certain number of outliers can have an arbitrarily large effect on the estimation of the regression coefficients. This is also referred to as the *breakdown point* [202, 219] of regression analysis. Such circumstances are very undesirable in outlier analysis, because of the likelihood of very misleading results. Subsequently, numerous estimators have been proposed with higher breakdown points [387]. In such cases, a higher level of contamination would need to be present in the data in order for breakdown to occur.

The method of *Principal Component Analysis* is also used frequently in the classical literature [244] for regression analysis and dimensionality reduction. Its application for noise correction in the text domain was first observed in [355], and then modeled theoretically in [18]. It was shown that the projection of the data points onto the hyper-planes with the greatest variance provides a data representation, with higher quality of similarity computations, because of the effects of removing noise from the data. In the context of text data [355], a variant of PCA, known as Latent Semantic Indexing [133]. Initially, the approach was proposed as a dimensionality reduction technique for retrieval, and was not designed for noise reduction. However, over many years of experience with LSI, it was observed that the quality of retrieval actually improved with LSI, a point which was explicitly pointed out in [355], and later theoretically modeled in [18] for relational data. It should be noted that PCA and LSI are dimensionality reduction techniques which can summarize the data by finding linear correlations among the dimensions. In principle, any dimensionality reduction technique can be used for outlier analysis. An example of an outlier analysis method which uses a different dimensionality reduction technique such as matrix-factorization is discussed in [476]. The core principle is that dimensionality reduction methods provide an approximate representation of the data along with a corresponding set of residuals. These residuals can be used as the outlier scores.

PCA-based techniques have been used in order to detect outliers in a wide variety of domains such as statistics [93], astronomy [147], ecological data [231], network intrusion detection [280, 406, 448], and many kinds of time-series data. Some of the aforementioned applications are temporal, whereas others are not. Because of the relationship between PCA and time series correlation analysis, much of the application of such regression methods has been to the temporal domain. However, it should be emphasized that regression-based methods can also be applied to many non-temporal scenarios. In particular, the use of PCA for non-temporal and unsupervised outlier analysis seems to be relatively

unexplored, and is worthy of further study. Regression based methods will be re-visited in Chapter 8, where a number of methods for temporal outlier analysis will be discussed. In the context of temporal data, the outlier analysis problem is closely related to the problem of *time series forecasting*, where deviations from forecasted values in a time series are flagged as outliers. A variety of regression-based methods for noise reduction and anomaly detection in time-series sensor data streams are also discussed in [19]. In addition, a number of methods which resemble structural and temporal versions of PCA have been used for anomaly detection in graphs [229, 429]. In such methods, an augmented form of the adjacency matrix, or the similarity matrix may be used for eigenvector analysis. Such methods are commonly referred to as *spectral methods*, and are discussed in Chapter 11.

A more general model than global PCA is one in which the data is modeled as a probabilistic mixture of PCAs [451]. This is referred to as *Probabilistic PCA (PPCA)*. Such methods are quite prone to noise in the underlying data during the process of mixture modeling. A method proposed in [132] increases the robustness of PCA by modeling the underlying noise in the form of a student t-distribution. The effect of outliers on PCA-based clustering algorithms are significant. The work in [7] provides a methods for providing the outliers as a side product of the output of the clustering algorithm. Furthermore, methods for using local PCA in outlier analysis will be discussed in detail in Chapter 5 on outlier analysis in high dimensional data.

7. Exercises

1. Consider the data set of the following observations: { (1, 1), (2, 0.99), (3, 2), (4, 0,98), (5, 0,97) }. Perform a regression with Y as the dependent variable. Then perform a regression with X as the dependent variable. Why are the regression lines so different? Which point should be removed to make the regression lines more similar to one another?

2. Perform Principal Component Analysis on the data set of Exercise 1. Determine the optimal 1-dimensional hyperplane to represent the data. Which data point is furthest from this 1-dimensional plane?

3. Remove the outlier point found in Exercise 2, and perform regression analysis on the remaining four points. Now project the outlier point onto the optimal regression plane. What is the value of the corrected point?

4. Provide a formal derivation for the closed form of the estimates of the regression coefficients in least squares regression. [Hint: Use partial derivatives with respect to regression coefficients.]

5. Provide a formal derivation for the closed form of the optimal k-dimensional subspace in Principal Component Analysis.

6. Download the *KDD CUP 1999 data set* from the UCI Machine Learning Repository [169], and perform PCA on the quantitative attributes. What is the dimensionality of the subspace required to represent (i) 80% of the variance, (ii) 95% of the variance, and (iii) 99% of the variance.

7. Repeat Exercise 6 with the use of the *Arrythmia* data set from the *UCI Machine Learning Repository* [169].

8. Generate 1000 data points randomly in 100-dimensional space, where each dimension is generated from the uniform distribution in $(0, 1)$. Repeat Exercise 6 with this data set. What happens, when you use 1,000,000 data points instead of 1000?

9. Consider a 2-dimensional data set with variables X and Y. Suppose that $Var(X) << Var(Y)$. How does this impact the slope of the X-on-Y regression line, as compared to the slope of the Y-on-X regression lines. Does this provide you with any insights about why one of the regression lines in Figure 3.3(b) shifts significantly compared to that in Figure 3.3(a), because of the addition of an outlier?

10. Scale each dimension of the *Arrythmia* data set, such that the variance of each dimension is 1. Repeat Exercise 7 with the scaled data set. Does the scaling process increase the number of required dimensions, or reduce them? Why? Is there any general inference that you can make about an arbitrary data set from this?

11. Let Σ be the covariance matrix of a data set. Let the Σ be diagonalized as follows:

$$\Sigma = P \cdot D \cdot P^T$$

Here D is a diagonal matrix containing the eigenvalues λ_i, and D^{-1} is also a diagonal matrix containing the inverse of the eigenvalues (i.e. $1/\lambda_i$)

- Show that $\Sigma^{-1} = P \cdot D^{-1} \cdot P^T$
- For a given data point $\overline{X}$ from a data set with mean $\overline{\mu}$, show that the value of the Mahalanobis distance $(\overline{X} - \overline{\mu}) \cdot \Sigma^{-1} \cdot$

$(\overline{X} - \overline{\mu})^T$ between $\overline{X}$ and the mean $\overline{\mu}$ reduces to the same expression as the score in Equation 3.6.