On Eigenfunction Based Spatial Analysis for Outlier Detection in High-Dimensional Datasets

Atulya K. NAGAR,
Intelligent and Distributed Systems Laboratory,
Deanery of Business and Computer Sciences,
Liverpool Hope University,
Liverpool L16 9JD, U.K.

ABSTRACT

This paper is concerned with two methods, one based on eigenvalue analysis, and the other, a modified version of singular value decomposition (SVD) called pseudo-SVD, for detecting outliers in high-dimensional data sets. The eigenvalue analysis approach examines the spatial relationship among the column vectors of object-attribute matrix to obtain an insight into the degree of inconsistency in a cluster of data. The pseudo-SVD method, in which the singular values are allowed to have a sign, looks at the direction of vectors in the object-attribute matrix and based on the degree of their orthogonality detects the outliers. The pseudo-SVD algorithm is formulated as an optimisation problem for clustering the data on the basis of their angular inclination. The methods have been applied to two case studies: one pertaining to a dermatological dataset and the other related to an engineering problem of state estimation. Further research directions are also discussed.

Keywords: Eigenfunction, Pseudo-SVD, Spatial, Orthogonal, Data Mining, Optimisation.

1. BACKGROUND

Outlier detection is an important data-mining aspect that aims to find exceptional behaviours of certain objects or data from the bulk of the source data. Outliers in the data arise from either recording errors, known as statistical bad data, or from noisy data of various kinds. Extracting these behaviours poses extraordinary attention when revealed. Therefore, detecting outliers may be as significant as discovering general patterns. Outlier detection [3] is used in various applications such as credit card fraud detection, customer and market segmentation, computer intrusion, discovering criminal behaviours, detection of bad-data and outliers from SCADA (supervisory control and data acquisition) database used in large distribution systems like Power and Water [8, 9].

The problem of mining outliers from large data sets lies in computational costs. For instance, the singular value decomposition (SVD) algorithm has been used in data mining for extracting clusters in high dimensional data sets [2]. Several outlier mechanisms are categorised as distance-based, depth-based, distribution-based, clustering-based and density-based [6]. Our method can be considered as clustering-based (in this case the metric is angular-based), as objects or vectors of similar angular (or spatial) inclination can be grouped together. Most algorithms, for example, those presented in [7, 10], which define outliers by using full dimensional distances between points, suffer from dimensionality and therefore present performance costs that ultimately has an impact on the quality of the clustered data.

Eigenvalue problems arise in number of applications of computational science for biological and engineering systems; they are useful mainly for two reasons: firstly because the matrices can be transformed in terms of a basis of eigenfunctions (thus speeding up the computational solution) and secondly because eigenvalues can provide an insight, by graphically visualising the values into the behaviour of an evolving matrix systems [11]. Brown [1] proposed eigenanalysis based method to solve the 'optimal meter placement' problem (also known as observability analysis), for large power distribution systems, by studying the spatial orientation of the observability matrix. The eigenvalue analysis approach makes use of the well established fact in linear algebra that if the coefficients of any pair of equations are approximately proportional, we will encounter difficulty solving the equations – if there are small measurement errors then this will be reflected as large errors in the unknowns. In other words, small errors in the measurements will render the object-attribute matrix as ill-conditioned. This can be seen in terms of analytical geometry as amounting to oblique intersection of lines, planes, or hypersurfaces as the case may be. Thus, it can be seen that if we have fuzzyness or outliers or error of uncertainty in our recording of data or measurements, the corresponding error in the output (or solution) is rendered to a minimum if the lines, or surfaces intersect are orthogonal; conversely, the outliers (or fuzzyness) gets more pronounced with the degree of obliqueness of the intersecting angle. In higher dimensional matrices, this amounts to spatial relationship among the vectors of the object-attribute matrix. Orthogonality amongst these vectors corresponds to right-angle intersections and leads to minimum fuzzyness/uncertainty or outliers in the solution or output. Conversely, if in our high-dimensional space there exists a vector which is approximately normal to all the vectors of the object-attribute matrix, we will have the situation of minimum fuzzyness in corresponding direction. Thus the analysis of the column vectors of the object-attribute matrix gives us the information of the degree of inconsistency in the high-dimensional datasets.

The pseudo-SVD method was first used by Featherstone et al. [3] for identification and control of large-scale chemical plant processes. They formulated the SVD method so that singular values are allowed to have a sign and thereby they introduced the term pseudo-SVD for such decompositions. A similar approach is used in this paper. Having formulated the SVD as a pseudo-SVD, we recast the problem as an optimisation problem to identify the outliers in the data sets and cluster the data sets based on their angular inclination. Our method, therefore, can be classified as cluster-based with the metric defined by angular inclination.

This paper is organised as follows: in Section 2 we discuss the eigenvalue approach for outlier detection. Section 3 describes the pseudo-SVD approach which leads to the formulation of an optimisation problem. Test datasets pertaining to two different case studies have been used, to illustrate the approach, in Section 4. Further research directions, for the solution and visualisation of the optimisation problem (formulated in Section 3), have been outlined in Section 5 along with the concluding remarks.

2. EIGENVALUE APPROACH

For the high-dimensional data sets if the m objects (say, denoted by $Z \in \mathbb{R}^m$) are related to their corresponding n attributes (say $X \in \mathbb{R}^n$), where $m \geq n$, via the $m \times n$ object-attribute matrix $M \in \mathbb{R}^{m \times n}$, then we have the following linear relationship:

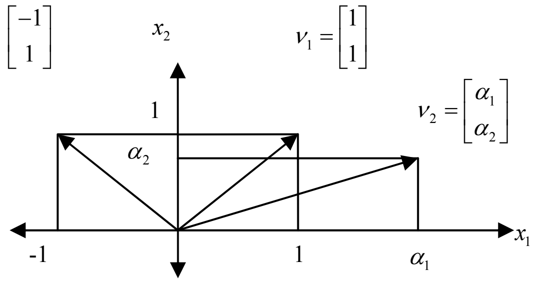
$$Z = MX \qquad (2.1)$$

In Linear Algebra it is a well known fact that if the coefficients of any pair of the above equations are approximately proportional, we will encounter difficulty in solving the equations because small measurement or data recording errors (in Z) will reflect or show as large errors in the unknowns (X). This becomes clear once we recognise the fact that the elements of the column vectors of the object-attribute matrix M are respective coefficients of a set of linear equations relating the corresponding object vector Z and attribute variable X. With this in mind the entire theory of Linear Algebra can be brought to bear on the problem under consideration. The rank of the matrix M is considered to be the necessary and sufficient condition for the solution of the equation (2.1). However, even if the matrix M has maximal (i.e. full) rank n, when there are uncertainties or outliers in the data matrix M, the object-attribute matrix may become ill conditioned and consequently the calculated solution to Eq. (2.1) may differ from the ideal or preferred solution to Eq. (2.1). In order to understand this further let us look at a simple example of the spatial-orientation concept of vectors.

An illustrative example

Consider a matrix $M = \begin{bmatrix} 1 & -\alpha_1 \\ 1 & -\alpha_2 \end{bmatrix}$ whose rank is seen to be 2 provided we have $\alpha_1 \neq \alpha_2$. It can be clearly seen that as $\alpha_1 \to \alpha_2$ the column vectors of M become coincident; the system thus becomes unstable and unsolvable. A vector $\begin{bmatrix} -1 \\ 1 \end{bmatrix}$, which is approximately normal to both the column vectors of the matrix M, is clearly the direction of maximum fuzzyness, uncertainty or missing information, in this case. It is in the direction of this "most orthogonal" vector that attention needs to be paid as it contains considerable 'information'. This illustration is shown in figure (2.1) where the sign of the second vector is reversed for convenience. Such a geometric interpretation becomes cumbersome when we consider higher dimensional data matrices. In order to gain understanding of the outliers and fuzzyness in the object-attribute matrix M, without actually solving the system of equations, we need a tool which will provide us with the insight into the intractable problem of detecting outliers from the unbridled growth of contaminated data. In doing, so we will exploit the fact that the inner-product of two orthogonal vectors is zero.

Figure 2.1: An illustration of noisy vectors: orthogonal vector gives direction of minimum disturbance.


$\begin{bmatrix} -1 \\ 1 \end{bmatrix}$
x_2
$v_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$
1
$v_2 = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}$
α_2
x_1
-1
1
α_1

As we are interested in the angles between the column vectors of the matrix M, we can normalise the column vectors of the object-attribute matrix without affecting our results. Let us denote this normalised matrix by M_n. Because the inner-product of two orthogonal vectors is zero, the most orthogonal vector can be defined as the vector that minimises the sum of the squares of the inner-products between it and each of the column vectors of the object-attribute matrix M_n. It can be shown [4], using simple calculus and the theory of simultaneous linear algebraic equations, that the most orthogonal vector of the column vectors of the object-attribute matrix M_n is in fact the eigenvector associated with the smallest eigenvalue of the matrix $M_n M_n^T$. If all the eigenvalues of this matrix are equal then we have an outlier free data set. Below we outline the steps involved in this algorithm.

Outlier detection algorithm

Step 1. Determine the $M_n M_n^T$ matrix

Step 2. Find eigevalues and eigenvectors

Step 3. Identify the eigenvector corresponding to the smallest eigenvalue

Step 4. Use this eigenvector to identify the direction of fuzzyness in data.

The most desirable data to be added to the initial measurements or data set should be the one which is as close as possible to the most orthogonal vector (identified in step 3 above) as it is this direction where there is deficiency of data. This amounts to at least matching one or more of the highest components of this most orthogonal vector. Also, if all the smallest eigenvectors are equal then any linear combination of the associated eigenvectors is also an eigenvector [4]. Thus, if this situation arises, the direction of the most orthogonal vector can be computed as the linear combination of the eigenvectors corresponding to equal smallest eigenvalues. The steps can be performed iteratively unless there is a situation of minimum fuzzyness or uncertainty.

3. PSEUDO-SINGULAR VALUE DECOMPOSITION

Eigenanalysis approach outlined in the previous section is useful if the matrix M has associated eigenvalue decomposition. It is known that even if the eigenvalues do exist, an infinitesimal perturbation may in general remove them [11]. Singular value decomposition (SVD) makes use of two different bases and all matrices (even rectangular ones) have singular values. Spectral analysis using SVD has been used for visualization of clusters in data-mining [2]. Here we present an optimisation problem formulation based on the pseudo-SVD [3]; the pseudo-singular values itself will help identify outliers and the solution of the resulting optimisation problem will yield clustering of data as meaningful granules or confidence bounds. The derivation and formulation of the optimisation problem presented here is based on the treatment given in [3]. For the object-attribute matrix M in equation (2.1), performing the SVD of M gives:

$$M = \hat{U}\Lambda\hat{V}^T \tag{3.1}$$

Let us define (as in derived in [3]) the diagonal matrix D_U which has each of its diagonal elements either +1 or -1, with the (i,i) th element equal to -1 if the dot product the two bases matrices, $(\hat{V}^i)^T\hat{U}^i$, is negative (where $\hat{U}$ and $\hat{V}$ are as defined in (3.1) and A^i is used to denote the i th column vector of matrix A).

Substituting the diagonal matrix D_U in Eq. (3.1), we get:

$$\begin{aligned} M &= \hat{U}\Lambda\hat{V}^T = (\hat{U}D_U)(D_U\Lambda)\hat{V}^T \\ &= U\Omega V^T \end{aligned} \tag{3.2}$$

where $\Omega = D_U\Lambda$ is a constant real diagonal matrix whose diagonal elements are known as 'pseudo-singular values' (a term coined by Featherstone and Braatz in [3]). The pseudo-singular values can thus be of any sign (and could also be zero), and are defined in such a way that the angle between the corresponding column vectors U^i and V^i is not greater than a right angle. The RHS of Eq. (3.2) is referred to as the pseudo-singular value decomposition (pseudo-SVD or p-SVD) [3].

Now expressing the RHS of equation (3.2) as a rank one decomposition we obtain:

$$M = \sum_{i=1}^{n}\Omega_{ii}U^i(V^i)^T \tag{3.3}$$

Since, as a result of SVD, the columns of V form an orthonormal basis, the attribute vector, X can be expressed in terms of the basis vector as follows:

$$X = \sum_{j=1}^{n}\alpha_j V^j \tag{3.4}$$

here the real scalar $\alpha_j = (V^j)^T X$ quantifies the amount of movement or inclination of X in the direction of the orthonormal basis vector V^j.

Similarly if there is noise (deterministic norm bounded or Gaussian white noise), then the effect of noise ε on the output can be expressed as:

$$\varepsilon = \sum_{j=1}^{n}\beta_j U^j \tag{3.5}$$

so now the real scalar $\beta_j = (U^j)^T\varepsilon$ represents the amount of noise in the direction of the orthonormal basis vector U^j.

If we consider a system of equations (2.1) perturbed or affected by noise (deterministic norm bounded or Gaussian white noise), then equation (2.1) can be written as:

$$Z = MX + \varepsilon \tag{3.6}$$

Substituting equations (3.3), (3.4) and (3.5) into equation (3.6) yields:

$$\begin{aligned} Z &= \sum_{i=1}^{n}\Omega_{ii}U^i(V^i)^T\sum_{j=1}^{n}\alpha_j V^j + \sum_{j=1}^{n}\beta_j U^j \\ &= \sum_{j=1}^{n}(\Omega_{jj}\alpha_j + \beta_j)U^j \end{aligned} \tag{3.7}$$

(This has been achieved keeping in view that $(V^i)^T V^j = \delta_{ij}$, V being orthonormal matrix; and $\delta_{ij} = \begin{cases} 0, \text{ if } i \neq j \\ 1, \text{ if } i = j \end{cases}$ is the Kronecker delta).

Let us now consider the projection of Z vector in the direction of orthonormal basis vector U^j:

$$(U^j)^T Z = \Omega_{jj}\alpha_j + \beta_j \tag{3.8}$$

From here can interpret that if the vector Z is in the direction of U^j such that $\Omega_{jj} = 0$ then we have the situation of noisy data with the parameter $\beta_j = (U^j)^T Z$ quantifying the extent or amount of noise in that direction. Similarly, from Eqs. (3.4) and (3.8) we infer that for minimum error in the vector Z in the direction of U^j, X in Eq. (3.4) must be such that the parameter $\alpha_j = -\beta_j/\Omega_{jj}$ exists. In other words if Ω_{jj} does not have appropriate sign (or direction), then the RHS of Eq. (3.8) will be greater than β_j, i.e. the amount of noise. For small or minimum 'signal-to-noise' ratio, which can be measured by the ratio of $(U^j)^T Z$ to β_j, we must have small Ω_{jj}. In the next section we apply the ideas developed and discussed here to two application areas: one related to a dermatological dataset and the other to an engineering problem of state estimation.

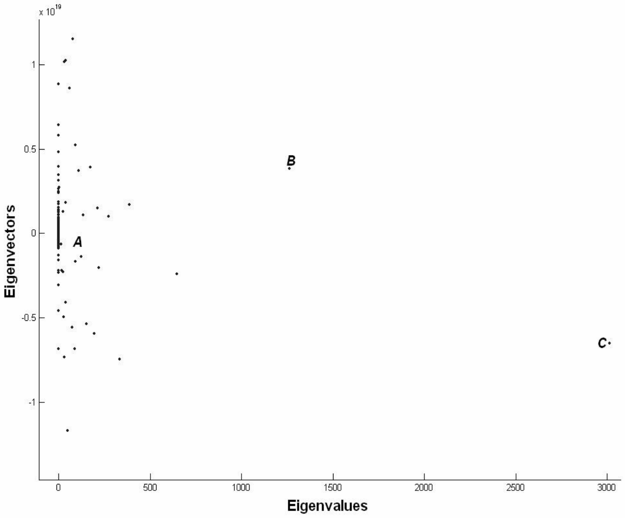
4. APPLICATIONS, RESULTS, AND DISCUSSIONS

Application 1: Dermatological dataset

The algorithm for eigenfunction analysis has been applied to a public domain (real world) dermatology dataset [2] with 358 instances and 34 attributes. The data records information about differential diagnosis of erythemato-squamous disease, a real problem in dermatology. The data presents two types of investigation, clinical and histopathological, with these diseases. All such diseases share clinical features such as erythema and scaling with very little differences. It is known that patients at first observation (clinically) may have different features of another disease and show characteristic features of erythemato-squamous at the following stages. These feature variations of the disease are identified as clusters and outliers detected.

The object-attribute matrix thus consists of 33 columns and 358 rows represented as equation (2.1); where Z is the object and M is the coefficient matrix with 33 attributes (column-wise) and 358 equations (row-wise). Our problem is to determine an eigenvector corresponding to the smallest eigenvalue. This most orthogonal eigenvector suggests identification of clusters as shown in Figure 4.1; the eigenvector corresponding to the smallest eigenvalue points to cluster **A**; in fact that is the direction of the most orthogonal vector to the column vector of the object-attribute matrix. Any additional data to be classified as belonging to the cluster **A** should be as much as possible in the direction of this most orthogonal eigenvector. The outliers in the figure are shown with **B** and **C**. The pseudo-SVD method validates these results as pseudo-SVD found that Ω_{jj} does not have appropriate sign, and thus the RHS of Eq. (3.8) is greater than β_j, the amount of noise. In this case the signal-to-noise ratio, which is measured by the ratio of $(U^j)^T Z$ to β_j, had large Ω_{jj} values.

Figure 4.1: Eigenvalue vs. eigenvectors: identification of clusters (**A**) and detection of outliers (**B**, **C**)



Application 2: State estimation of distribution systems: outlier detection

In the case of water distribution systems the purpose of a state-estimator is to provide the best possible information about the flows and pressures in a system given all the available data [8]. It is capable of producing a validated set of on-line information about a particular network. These data can then be used for a wide range of purposes, e.g.: security analysis, poor quality water tracing, decision support and pump scheduling.

The observability problem in water-system state estimation consists essentially in determining whether the measurements currently available to the state estimator provide sufficient information to allow the computation of the estimates. Such tests are important both as a design tool in meter placement studies performed offline and in the online real-time implementation of the state estimator. The uncertainty in measurement data is transferred through the state estimation process where it is compounded by uncertainties in the model. The traditional 'observability test' based on rank of the technology matrix provides a yes/no answer to the question of the adequacy of the measurement set [1, 8]. The eigenfunction approach reported in this research seeks to add a more realistic test that predicts the performance of an estimator before it is installed through an analysis of the uncertainties arising from the available data. This is achieved through analysis of the spatial direction of the M matrix in order to provide an informed answer to the observability question 'if observable, how observable?'. Measurements can then be added to the direction of measurement deficient direction revealed by the eigenvector corresponding to the smallest eigenvalue. The p-SVD method also provides useful insights into the datasets containing missing data, statistically bad data, and noisy data, and uncertainties (stochastic or deterministic). The p-SVD problem above can be formulated as an optimisation problem to provide confidence bounds on the estimated. In the next subsection we briefly discuss this optimisation problem formulation and for the state estimation problem we report results showing confidence bounds.

Pseudo-SVD optimisation problem

The formulation above in section 3 can be recast as the following constraint optimisation problem (also discussed by Nagar et al. in [8, 9] - but using a different approach and, therefore, different formulation):

$$\begin{aligned}&\min\ (Z - MX)^T R^{-1}(Z - MX)\\ &\text{subject to: } \theta_j \le \alpha_j \le \omega_j\end{aligned} \tag{3.9}$$

where R is a diagonal $m \times m$ matrix whose elements are the measurement weighting factors. The measurement weighting factor is taken as the reciprocal of the error variance, $\alpha_j = (V^j)^T X$, which, as seen above, is a metric that quantifies the orientation or inclination of the attribute vector X such that X lies in a particular cluster or granule (here bounds) of data set. The lower θ_j bound and the upper bound ω_j could be determined either statistically (confidence ellipsoids) or these could be probabilistic bounds or alternatively, these could, in fact, be defined as fuzzy bounds. These bounds can then be tightened iteratively to obtain clusters. In this work we obtain ellipsoid-of-confidence bounds which are obtained by projecting the ellipsoid along the co-ordinate axes.

The method has been applied to an engineering problem of state estimation for a test water distribution network system. A water system simulator and a state estimator were designed using MATLAB environment; solution of Eq. (3.9) requires interior-point methods. Before applying the state estimation, the eigenfunction based spatial analysis was performed to identify observable islands for the test network under consideration [9]. Figure 4.2 shows the results of state estimation for the test case. As a result of solving the optimisation problem in Eq. (3.9), point estimates are found to coincide with the centre of the confidence bounds (similar results were observed in [9] using an Leaner Fractional Transformation (LFT) based approach leading to a Semi-definite programming (SDP) formulation). The actual confidence bounds are shown in Figure 4.3.

By observing the bounds on the estimates one can infer the quality of the metering configuration for a water distribution network (or an observable sub-network known as an island) and determine whether the installation of new meters would be desirable. Confidence bounds thus help in answering the question of observability (if observable, how observable?) and in detecting anomalies and outliers in the multi-dimensional datasets from the distribution systems. The centre of the confidence ellipsoid gives the maximum-likely point state estimates. State estimation can also be used to detect anomalies or potential problems in the system (or part of the system).

Figure 4.2: Point estimates coinciding with the centre of the confidence bounds

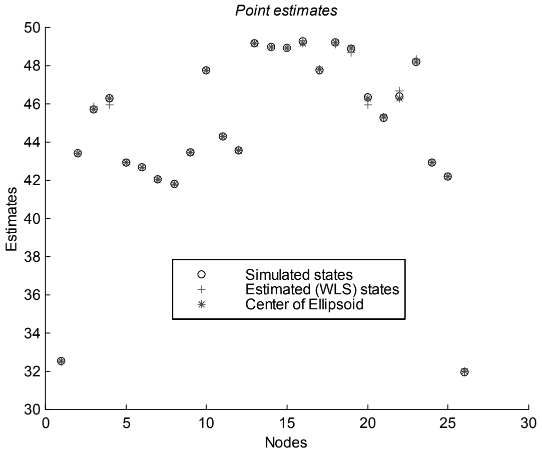
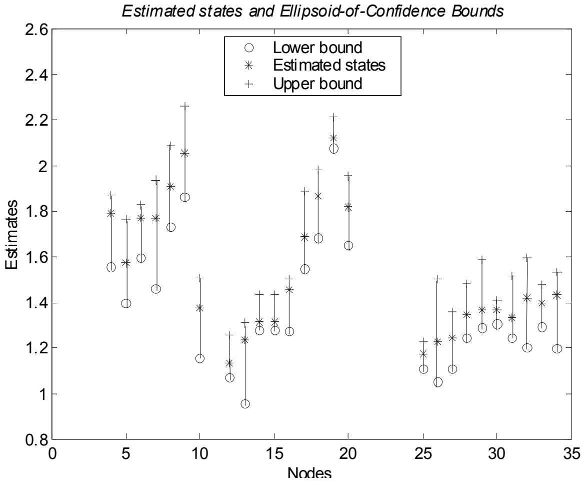


Figure 4.3: Confidence bounds for a nodal based state estimator



5. CONCLUDING REMARKS AND FURTHER WORK

This paper has introduced a novel technique to the problem of mining outliers from large datasets by looking at angular inclination of vectors, therefore known as cluster-based outlier detection using angular metric. Further work will investigate the optimisation problem of fuzzy or probabilistic bounds, as formulated in equation (3.9). Further implementation of this algorithm will consider visualisation of clusters and how variations of fuzzy bounds and consequently the angle affects the clusters. Implementation of the methodology to datasets pertaining to various other application areas, e.g. financial datasets, will also be reported elsewhere. Performance measures related work will also be investigated and reported later.

Acknowledgements: The work reported in this paper is an enhanced and evolved version of an initial paper presented in the SCI 2004 conference. Maybin Muyeba's useful discussions for the initial draft of the paper are acknowledged with thanks.

6. REFERENCES

[1] R.G. Brown, "Not Just Observable, but How Observable", **Proceedings of National Electronics Conference,** Vol. 22, 1966, pp. 409-714.

[2] V. Castelli, A. Thomasian, C-S. Li, "CSVD: Clustering and Singular Value Decomposition for Approximate Similarity Searches in High-dimensional Spaces"**, Research Report RC21755 (98001), IBM**, 30 May 2000.

[3] A.P. Featherstone, R.D. Braatz, "Integrated Robust Identification and Control of Large-Scale Processes", **Ind. Eng. Chem. Res.**, 1998, pp. 97-106.

[4] G.H. Golub, C.F. Van Loan, **Matrix Computations**, The Johns Hopkins University Press Books, Baltimore, third ed., 1997

[5] D. Hawkins, **Identification of Outliers**, Chapman and Hall, London, 1980.

[6] W. Jin, A.K Tung, and J. Han, "Mining Top-n Local Outliers in Large Databases*",* **Proceedings of the International Conference on Knowledge Discovery in Databases (KDD'2001)**, California, 2001, pp. 293-298.

[7] E.M. Knorr, R. T. Ng, "Algorithms for Mining Distance-based Outliers in Large Data Sets", **VLDB,** 1998, pp. 392-403.

[8] A.K. Nagar, R.S. Powell, "LFT/SDP Based Approach to the Uncertainty Analysis for State Estimation of Water Distribution Systems", **IEE Proc. of Control Theory and Applications**, Vol. 149, Issue 2, March 2002, pp. 137-142.

[9] A.K. Nagar, J.H. Andersen, and R.S. Powell, Chapter 2: "Mixed Uncertainty Analysis for State Estimation of Water Distribution Systems", **Water Software Systems: Theory and Applications**, Research Studies Press, Ed.: B. Ulanicki, B. Coulbeck and, J. Rance, ISBN 0-86380-273-7, September 2001.

[10] S. Ramaswamy, R. Rastogi, S. Kyuseok, "Efficient Algorithms for Mining Outliers from Large Data Sets", **SIGMOD 2000**, Dallas, 2000, Texas, pp. 93-104.

[11] L.N. Trefethen, "Pseudospectra of Linear Operators"*,* **SIAM Rev.,** Vol. 39, No. 3, September 1997, pp. 383-406.