Chapter 5

High-Dimensional Outlier Detection: The Subspace Method

“In view of all that we have said in the foregoing sections, the many obstacles we appear to have surmounted, what casts the pall over our victory celebration? It is the curse of dimensionality, a malediction that has plagued the scientist from the earliest days.”– Richard Bellman

5.1 Introduction

Many real data sets are very high dimensional. In some scenarios, real data sets may contain hundreds or thousands of dimensions. With increasing dimensionality, many of the conventional outlier detection methods do not work very effectively. This is an artifact of the well-known curse of dimensionality. In high-dimensional space, the data becomes sparse, and the true outliers become masked by the noise effects of multiple irrelevant dimensions, when analyzed in full dimensionality.

A main cause of the dimensionality curse is the difficulty in defining the relevant locality of a point in the high-dimensional case. For example, proximity-based methods define locality with the use of distance functions on all the dimensions. On the other hand, all the dimensions may not be relevant for a specific test point, which also affects the quality of the underlying distance functions [263]. For example, all pairs of points are almost equidistant in high-dimensional space. This phenomenon is referred to as data sparsity or distance concentration. Since outliers are defined as data points in sparse regions, this results in a poorly discriminative situation where all data points are situated in almost equally sparse regions in full dimensionality. The challenges arising from the dimensionality curse are not specific to outlier detection. It is well known that many problems such as clustering and similarity search experience qualitative challenges with increasing dimensionality [5, 7, 121, 263]. In fact, it has been suggested that almost any algorithm that is based on the notion of proximity would degrade qualitatively in higher-dimensional space, and would therefore need to
be re-defined in a more meaningful way [8]. The impact of the dimensionality curse on the outlier detection problem was first noted in [4].

In order to further explain the causes of the ineffectiveness of full-dimensional outlier analysis algorithms, a motivating example will be presented. In Figure 5.1, four different 2-dimensional views of a hypothetical data set have been illustrated. Each of these views corresponds to a disjoint set of dimensions. It is evident that point ‘A’ is exposed as an outlier in the first view of the data set, whereas point ‘B’ is exposed as an outlier in the fourth view of the data set. However, neither of the data points ‘A’ and ‘B’ are exposed as outliers in the second and third views of the data set. These views are therefore noisy from the perspective of measuring the outlierness of ‘A’ and ‘B.’ In this case, three of the four views are quite non-informative and noisy for exposing any particular outlier ‘A’ or ‘B.’ In such cases, the outliers are lost in the random distributions within these views, when the distance measurements are performed in full dimensionality. This situation is often naturally magnified with increasing dimensionality. For data sets of very high dimensionality, it is possible that only a very small fraction of the views may be informative for the outlier analysis process.

What does the aforementioned pictorial illustration tell us about the issue of locally relevant dimensions? The physical interpretation of this situation is quite intuitive in practical scenarios. An object may have several measured quantities, and significantly abnormal behavior of this object may be reflected only in a small subset of these quantities. For ex-