

INTISARI

ANALISIS KOMPONEN UTAMA ROBUST SPARSE (ROSPCA) PADA DATA BERDIMENSI TINGGI YANG MEMUAT OUTLIER

Oleh

Erika Eki Ovitasari

16/395559/PA/17306

Analisis Komponen Utama (AKU) merupakan salah satu analisis multivariat yang berguna untuk mereduksi data berdimensi tinggi yang memiliki korelasi sehingga membentuk data dengan variabel yang lebih sedikit tetapi tetap mampu menjelaskan keragaman dari data. Reduksi dimensi pada PCA klasik didasarkan pada matriks kovarian yang tidak *robust* terhadap *outlier*. Oleh karena itu diperlukan *Robust PCA* pada data yang mengandung *outlier*. Tujuan dari metode *Robust PCA* adalah untuk memperoleh komponen utama yang tidak terlalu banyak terpengaruh dengan keberadaan *outlier*. Telah diketahui juga bahwa *Classical PCA* dan *Robust PCA* merupakan kombinasi linear dari seluruh variabel. Hal ini akan mempersulit proses interpretasi komponen utama, karena peneliti tidak dapat mengetahui variabel mana saja yang berperan penting dalam pembentukan komponen utama. Untuk itu, diperlukan *Sparse PCA* guna mempermudah proses interpretasi komponen utama karena dengan *Sparse PCA* akan menghasilkan banyak *zero loadings* pada komponen utama yang terbentuk.

Pada skripsi ini akan dibahas metode pembentukan komponen utama *Robust Sparse PCA* (ROSPCA) pada data berdimensi tinggi yang memuat outlier, yaitu dengan mengintegrasikan *Sparse PCA* ke dalam ROBPCA yang merupakan gabungan konsep *projection-pursuit* dan estimator kovarian yang *robust* (FAST-MCD). Pada skripsi ini akan dilihat performa dari metode *Robust PCA* (ROBPCA), *Sparse Robust PCA* (SRPCA), dan *Robust Sparse PCA* (ROSPCA) dari sisi kemudahan dalam proses interpretasi komponen utama dan ke-*robust*-an terhadap *outlier*. Berdasar analisis yang telah dilakukan diperoleh kesimpulan bahwa metode *Robust Sparse PCA* (ROSPCA) merupakan metode yang terbaik.

Kata kunci : reduksi dimensi, *outlier*, data berdimensi tinggi, FAST-MCD, *projection-pursuit*, *robust*, *sparse*

ABSTRACT

ROBUST SPARSE PRINCIPAL COMPONENT ANALYSIS (ROSPCA) ON HIGH DIMENSIONAL DATA WITH OUTLIERS

By

Erika Eki Ovitasari

16/395559/PA/17306

Principal Component Analysis (PCA) is a multivariate analysis that is useful for reducing the high dimensional data of which have a correlation to form the data with less variables but still able to explain the diversity of the data. The dimension reduction of PCA based on covariance matrices that are not robust to outliers. Therefore Robust PCA is required on data containing outliers. The purpose of a Robust PCA method is to obtain the principal components that do not too much influenced by the outlier. It is known that Classical PCA and Robust PCA are linear combinations of all variables. This will inhibit the process of interpretation of the PCs, because researchers can't know which variable plays an important role in the formation of the PCs. For that, Sparse PCA required to simplify the process of interpretation. For this reason, Sparse PCA is needed to simplify the process of interpreting the main components because Sparse PCA will produce a lot of zero loadings on the main components that are formed.

This thesis will discuss about the Robust Sparse PCA on high dimensional data with outlier, by integrating Sparse PCA into ROBPCA which consists of the combination of a projection-pursuit concept and robust covariance estimator (FAST-MCD). In this paper we will see the performance of Robust PCA (ROBPCA), Sparse Robust PCA (SRPCA) and Robust Sparse PCA (ROSPCA) methods in terms of convenience in the process of interpretation of the PCs and the goodness of robustness towards outlier. Based on the result of the analysis process, we can conclude that Robust Sparse PCA (ROSPCA) is the best method among the others.

Keywords: dimensional reduction, outlier, high dimensional data, FAST-MCD, projection-pursuit, robust, sparse