Quiz 3 (slides 2 and 3):

Assume that you have a sample, sample size n, from some p-variate distribution with finite expected value and finite covariance matrix. Assume that you have applied principal component transformation to the (n times p) data matrix X and obtained the (n times p) score matrix Y.

3 a) How do you perform the transformation in practice?

3 b) How do you measure the quality of representation of the individual r by the principal axis i?

3 c) Mention at least two applications of PCA.

3 d) Is PCA transformation affine invariant?

3 e) Is PCA sensitive to the effect of outlying observations?

3 f) PCA is sensitive to scaling of the variables. How can you address this problem?

3 g) In PCA, the signs of the eigenvectors used in the transformation are not uniquely defined. Why that is not a problem?

Quiz 4 (slides 4):

4 a) Explain, in your own words (informally) what, in statistics, is meant by robustness?

4 b) The influence function is said to be a measure of local robustness. Why?

4 c) The breakdown point is said to be a measure of global robustness. Why?

4 d) Explain how to calculate the influence function and the empirical influence function in practice.

4 e) What does it tell us, if the influence function is bounded?

4 f) Does the sample mean have a bounded empirical influence function? Based on that, is the sample mean a robust estimator?

4 g) Does the sample median have a bounded empirical influence function? Based on that, is the sample median a robust estimator?

4 h) Explain how to calculate the break down point in practice.

4 i) What is the break down point of the sample mean? Based on that, is the sample mean a robust estimator?

4 j) What is the break down point of the sample median? Based on that, is the sample median a robust estimator?

4 k) Assume that your observations come from an elliptical distribution. Explain how you can now robustify PCA.

Answers:

3 a) Calculate the sample mean vector and the eigenvalues and orthonormal eigenvectors of the sample covariance matrix. Now, for each p-variate data point x_r, let y_r=G^T(x_r-m), where m is the sample mean vector and the column vectors of G are the orthonormal eigenvectors of the sample covariance matrix ordered such that the corresponding eigenvalues are in decreasing order. The vectors y_r are the rows of the score matrix Y.

3 b) Consider the vector from 0 to point x_r-m, where m is the sample mean vector and the vector from zero to g_i, where g_i is the ith column vector of G. (The matrix G is defined in the previous answer.) Calculate now the squared cosine of the angle between these two vectors. Values close to 1 indicate that the quality of representation is good. Values close to 0 indicate that the quality of representation is not good. See also Lecture 2, Note 2.4.

3 c) Dimension reduction, outlier detection, heuristic clustering.

3 d) No, it is not.

3 e) Yes, it is. Traditional PCA is very non-robust method.

3 f) One standardize the variables first. The data can be standardized by subtracting the sample mean vector from each observation, and then dividing each variable by the square root of the corresponding sample variance.

3 g) If an eigenvector is multiplied by -1, that leads to the corresponding scores to be multiplied by -1 as well. Thus, the original observations can still be given in terms of the scores and the eigenvectors.

4 a) A robust method or statistics is a method that does not behave too badly in the presence of outlying points or small deviations from the assumed distribution.

4 b) It measures what happens under point mass (local) contamination or (if we consider the empirical version) in the presence of one contaminated point.

4 c) It measures how large proportion of the points we have to contaminate in order to make the statistics at hand to give completely unreasonable results (break down, to approach infinity).

4 d) See pages 8 and 13 of the lecture slides 4.

4 e) A statistic that has bounded influence function is considered (locally) robust.

4 f) No. Based on that, the sample mean is not a robust estimator.

4 g) Yes. Based on that, the sample median is a robust estimator.

4 h) See page 17 of the lecture slides 4.

4 i) The break down point of the sample mean is 0. Based on that, the sample mean is not a robust estimator.

4 j) The break down point of the sample median is 1/2. Based on that, the sample median is a robust estimator.

4 k) If one can assume ellipticity, PCA can be robustified by replacing the traditional (sample) mean vector and the traditional (sample) covariance matrix by robust affine equivariant location and scatter functionals/estimators.