MS-E2112 Multivariate Statistical Analysis (5cr)
Lecture 4: Measures of Robustness, Robust Principal Component Analysis

Lecturer: Pauliina Ilmonen
Slides: Ilmonen/Kantala

Contents

Measures of Robustness

Influence Function

Empirical Influence Function

Breakdown Point

Robust PCA

References

Measures of Robustness

Robust Statistical Methods

In statistics, robust methods are methods that perform well – or do not perform too poorly – in the presence of outlying observations.

Robust Statistical Methods, Example

Mean vs median...

Measures of Robustness

Let x denote a random variable or a random vector with a cumulative distribution function F_x, and let $X = \{x_1, x_2, ..., x_n\}$, where $x_1, x_2, ..., x_n$ are n independent and identically distributed observations from the distribution F_x. Consider functional $Q(F_x)$ (or $Q(F_n)$). We wish to measure robustness of that functional.

Influence Function

Influence Function

Influence function measures the effect on functional Q when the underlying distribution deviates slightly from the assumed one.

$$IF(y, Q, F_x) = \lim_{0<\varepsilon\to 0} \frac{Q((1-\varepsilon)F_x + \varepsilon\delta_y) - Q(F_x)}{\varepsilon},$$

where δ_y is the cumulative distribution function having all its probability mass at y i.e.

$$\delta_y(t) = \begin{cases} 0, & t < y, \\ 1, & t \geq y. \end{cases}$$

Influence Function

Influence function measures the effect of point-mass contamination, and thus it is considered as a measure of local robustness.

A functional with bounded influence function (with respect to for example L_2 norm) is considered as robust and desirable.

Example, Population Mean (Expected Value)

$$
\begin{aligned}
IF(y, \mu, F_x) &= \lim_{0<\varepsilon\to 0} \frac{\mu((1-\varepsilon)F_x + \varepsilon\delta_y) - \mu(F_x)}{\varepsilon} \\
&= \lim_{0<\varepsilon\to 0} \frac{E[(1-\varepsilon)x + \varepsilon y] - E[x]}{\varepsilon} \\
&= \lim_{0<\varepsilon\to 0} \frac{E[x - \varepsilon x + \varepsilon y] - E[x]}{\varepsilon} \\
&= \lim_{0<\varepsilon\to 0} \frac{E[x] - \varepsilon E[x] + \varepsilon E[y] - E[x]}{\varepsilon} \\
&= \lim_{0<\varepsilon\to 0} \frac{-\varepsilon E[x] + \varepsilon y}{\varepsilon} \\
&= \lim_{0<\varepsilon\to 0} -E[x] + y = -E[x] + y = y - E[x] = y - \mu(F_x).
\end{aligned}
$$

This is not bounded with respect to y.

Empirical Influence Function

Empirical Influence Function

Empirical influence function (also called the sensitivity curve) is a measure of the dependence of the estimator on the value of one of the points in the sample.

The empirical influence function can be seen as an estimate of the theoretical influence function.

Empirical Influence Function

Let $X = \{x_1, x_2, ..., x_n\}$, and let $X_y = \{x_1, x_2, ..., x_n, y\}$. Now

$$IF_E(y, Q, F_n) = \frac{Q((1 - \frac{1}{n+1})F_n + \frac{1}{n+1}\delta_y) - Q(F_n)}{\frac{1}{n+1}}$$

$$= (n+1)(Q((1 - \frac{1}{n+1})F_n + \frac{1}{n+1}\delta_y) - Q(F_n))$$
$$= (n+1)(Q(X_y) - Q(X)).$$

Example, Sample Mean

$$IF_E(y, \hat{\mu}, F_n) = (n+1)(\hat{\mu}(X_y) - \hat{\mu}(X)$$

$$= (n+1)(\frac{1}{n+1}(\sum_{i=1}^{n} x_i + y) - \frac{1}{n}\sum_{i=1}^{n} x_i)$$

$$= \sum_{i=1}^{n} x_i + y - \frac{n+1}{n}\sum_{i=1}^{n} x_i$$

$$= y - (\frac{n+1}{n} - 1)\sum_{i=1}^{n} x_i$$

$$= y - (\frac{n+1-n}{n})\sum_{i=1}^{n} x_i$$

$$= y - \frac{1}{n}\sum_{i=1}^{n} x_i = y - \hat{\mu}(X).$$

This is not bounded with respect to y. Note that the empirical influence function estimates the theoretical influence function.

Breakdown Point

Breakdown Point

Another very often used measure of robustness is the breakdown point. Whereas influence function measures local robustness, the breakdown point can be seen as a measure of global robustness.

Breakdown Point

Let $X_n = \{x_1, x_2, ..., x_n\}$, where $x_1, x_2, ..., x_n$ are n independent and identically distributed observations from the distribution F_x. Assume that $m < n$ and replace $x_1, x_2, ..., x_m$ with $x_1^*, x_2^*, ..., x_m^*$. Let $X_n^* = \{x_1^*, x_2^*, ..., x_m^*, x_{m+1}, ..., x_n\}$.

Now, the maximum bias

$$maxBias(m, X_n, Q) = \sup_{x_1^*, x_2^*, \ldots, x_m^*} d(Q(X_n), Q(X_n^*)),$$

where $d(\cdot, \cdot)$ denotes some distance function (for example the Euclidean distance).

The finite sample breakdown point is now given by

$$BP(Q, n) = \min_m \{\frac{m}{n} \mid maxBias(m, X_n, Q) = \infty\},$$

and the (asymptotic) breakdown point

$$BP(Q) = \lim_{n \to \infty} BP(Q, n).$$

Breakdown Point

A functional with large breakdown point is considered as robust. If $BP(Q) = \frac{1}{2}$, then Q is very robust (according to its breakdown point), and if $BP(Q) = 0$, then Q is very nonrobust. When the value is in between $\frac{1}{2}$ and 0, then it is a matter of taste ;-).

Example, Sample Mean

Let $X_n = \{x_1, x_2, ..., x_n\}$ a sample of be independent and identically distributed observations from some distribution F_x. Let $\hat{\mu}(X_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$ and let $\hat{\mu}(X_n^*) = \frac{1}{n}(\sum_{i=2}^{n} x_i + x_1^*)$. Let $d(\hat{\mu}(X_n), \hat{\mu}(X_n^*))$ be the Euclidean distance between $\hat{\mu}(X_n)$ and $\hat{\mu}(X_n^*)$. If now $x_1^* \to \infty$, then also $\hat{\mu}(X_n^*) \to \infty$ and consequently

$$maxBias(1, X_n, \hat{\mu}) = \sup_{x_1^*} d(\hat{\mu}(X_n), \hat{\mu}(X_n^*)) = \infty.$$

Contaminating just one data point is enough to make the Euclidean distance arbitrarily large. Thus the finite sample breakdown point

$$BP(\hat{\mu}, n) = \frac{1}{n}$$

and the (asymptotic) breakdown point of the sample mean is

$$BP(\hat{\mu}) = \lim_{n\to\infty} BP(\hat{\mu}, n) = \lim_{n\to\infty} \frac{1}{n} = 0.$$

Example, Sample Median (1/4)

Let $X_n = \{x_1, x_2, ..., x_n\}$ be a sample of independent and identically distributed observations from some distribution F_x. Let $Med(X_n)$ be the sample median calculated from the original sample and let $Med(X_n^*)$ be the sample median calculated from the contaminated sample $X_n^* = \{x_1^*, x_2^*, ..., x_m^*, x_{m+1}, ..., x_n\}$. Let $d(Med(X_n), Med(X_n^*))$ be the Euclidean distance between $Med(X_n)$ and $Med(X_n^*)$.

Example, Sample Median (2/4)

Assume first that n is even. When n is even, the sample median is the average of the two middle values of the ordered observations. Now, one has to contaminate at least half of the observations in order to make the sample median and consequently the Euclidean distance arbitrarily large. Thus, for even n, the number of contaminated observations m has to be at least $n/2$ for

$$maxBias(m, X_n, Med) = \sup_{x_1^*, x_2^*, \ldots, x_m^*} d(Med(X_n), Med(X_n^*)) = \infty$$

to hold, and the finite sample breakdown point is then

$$BP(Med, n) = \min_m \{ \frac{m}{n} \mid maxBias(m, X_n, Med) = \infty \} = \frac{n/2}{n} = \frac{1}{2}.$$

Example, Sample Median (3/4)

Assume now that n is odd. When n is odd, the sample median is the middle value of the ordered observations. Now, one has to contaminate at least $(n+1)/2$ observations in order to make the sample median and consequently the Euclidean distance arbitrarily large. Thus, for odd n, the number of contaminated observations m has to be at least $(n+1)/2$ for

$$maxBias(m, X_n, Med) = \sup_{x_1^*, x_2^*, \ldots, x_m^*} d(Med(X_n), Med(X_n^*)) = \infty$$

to hold, and the finite sample breakdown point is then

$$BP(Med, n) = \min_m \{\frac{m}{n} \mid maxBias(m, X_n, Med) = \infty\} = \frac{(n+1)/2}{n} = \frac{n+1}{2n}.$$

Example, Sample Median (4/4)

If n is even, the (asymptotic) breakdown point

$$BP(Med) = \lim_{n\to\infty} BP(Med, n) = \lim_{n\to\infty} \frac{1}{2} = \frac{1}{2}.$$

If n is odd, the (asymptotic) breakdown point

$$BP(Med) = \lim_{n\to\infty} BP(Med, n) = \lim_{n\to\infty} \frac{n+1}{2n}$$

$$= \lim_{n\to\infty} (\frac{n}{2n} + \frac{1}{2n}) = \lim_{n\to\infty} (\frac{1}{2} + \frac{1}{2n}) = \frac{1}{2}.$$

Thus, the (asymptotic) breakdown point of sample median is $1/2$.

Breakdown Point, Some Remarks

- The applied distance does not have to be Euclidean.
- Sometimes $\mathit{maxBias}(m, X_n, Q) = \infty$ is not seen as the only "breaking down" case. For example Scatter $= 0$ can be seen as breaking down too.
- For matrices, breaking down is sometimes considered to be equal to the largest eigenvalue approaching ∞.
- It does not make much sense to try construct estimators that have breakdown point larger than $1/2$.

Robust PCA

Robust PCA

If the data can be assumed to arise from elliptical distribution, then principal component analysis can be robustified by replacing the sample covariance matrix with some robust scatter estimate. The reason for that is that, under elliptical distribution, all scatter estimates do estimate the same population quantity (up to the scale). Note that in general (without ellipticity assumption) this does not hold!

Minimum Covariance Determinant (MCD) Method

The determinant (volume) of a covariance matrix, can be seen as a measure of total variation of the data, and it is then called the generalized variance. Data points that are far away from the data cloud increase the volume of the covariance matrix.

Minimum Covariance Determinant (MCD) Method

Minimum Covariance Determinant (MCD) method is a well-known method for robustifying the estimation of the covariance matrix, and the mean vector, under the assumption of multivariate ellipticity.

MCD method is based on considering all subsets containing $p\%$ (usually 50%) of the original observations, and estimating the covariance matrix, and the mean vector, on the data of the subset associated with the smallest covariance matrix determinant. This is equivalent to finding the sub-sample with the smallest multivariate spread. The MCD sample covariance matrix, and the MCD sample mean vector, are then defined as the sample covariance matrix (up to the scale), and the sample mean vector, computed over this sub-sample.

Minimum Covariance Determinant (MCD) Method

Note that, as $\det(AB) = \det(A)\det(B)$ for all square matrices and as the point mass probability of continuous distributions is 0, MCD should be affine equivariant under continuous distributions. However, the fast versions of the algorithm are not necessarily affine equivariant. Some error might occur due to "smart" sub-sampling.

Robust PCA

Under the ellipticity assumption, PCA can be performed using the MCD scatter estimate instead of the traditional sample covariance matrix. MCD estimates are very robust, and thus as a consequence, robust PCA is obtained.

Robust PCA

Note that MCD is not the only possible robust scatter estimate - there exists several robust scatter estimates that all estimate the same population quantity (up to the scale) under the assumption of multivariate ellipticity.

Words of Warning

Some Words of Warning

- It is possible that a functional Q has bounded influence function, but its breakdown point is 0!
- Robust PCA, based on some robust scatter matrix, can be performed under the assumption of multivariate ellipticity. If the ellipticity assumption does not hold, instead of estimating the PCA transformation matrix Γ, one may be estimating some other population quantity.

Next Week

Next week we will talk about bivariate correspondence analysis (CA).

References

References I

K. V. Mardia, J. T. Kent, J. M. Bibby, Multivariate Analysis, Academic Press, London, 2003 (reprint of 1979).

References II

- R. V. Hogg, J. W. McKean, A. T. Craig, Introduction to Mathematical Statistics, Pearson Education, Upper Sadle River, 2005.
- R. A. Horn, C. R. Johnson, Matrix Analysis, Cambridge University Press, New York, 1985.
- R. A. Horn, C. R. Johnson, Topics in Matrix Analysis, Cambridge University Press, New York, 1991.

References III

P. J. Rousseeuw, Multivariate estimation with high breakdown point, *Mathematical Statistics and Applications* **8** (W. Grossmann, G. Pug, I. Vincze, W. Wertz, eds.), p. 283–297, 1985.

P. J. Rousseeuw, K. Van Driessen, A fast algorithm for the minimum covariance determinant estimator, *Technometrics* **41**, p. 212–223, 1999.