Help on: OGK (Orthogonalized Gnanadesikan and Kettenring) method

This method, proposed by Maronna and Zamar (2002), is based on a simple robust bivariate covariance estimator s_{jk} suggested by Gnanadesikan and Kettenring (1972). For a pair of random variables Y_j and Y_k, and a standard deviation function $\sigma(.), s_{jk}$ is defined by Gnanadesikan and Kettenring as:

$$s_{jk} = \frac{1}{4}[\sigma(\frac{Y_j}{\sigma(Y_j)} + \frac{Y_k}{\sigma(Y_k)})^2 - \sigma(\frac{Y_j}{\sigma(Y_j)} - \frac{Y_k}{\sigma(Y_k)})^2]$$

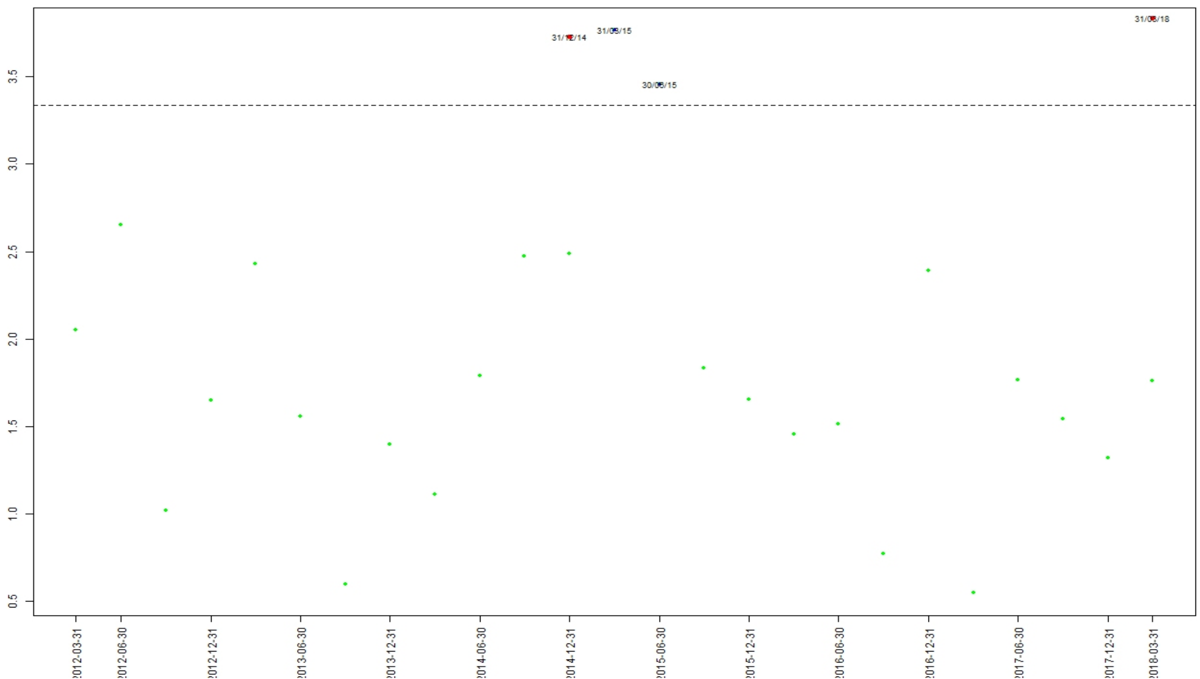
If a robust function is chosen for $\sigma(.)$ then s_{jk} is also robust and an estimate of the covariance matrix can be obtained by computing each of its elements s_{jk} for each j = 1,...,p and k = 1,...,p using the above equation. This estimator does not necessarily produce a positive definite matrix (although symmetric) and it is not affine equivariant. Maronna and Zamar (2002) overcome the lack of positive definiteness by the following steps:

- Define $y_i = D^{-1}x_i$, i = 1,...,n with $D = diag(\sigma(X_1), ..., \sigma(X_p))$ where $X_l, l = 1, ..., p$ are the columns of the data matrix $X = x_1, ..., x_n$. Thus, a normalized data matrix $Y = y_1, ..., y_n$ is computed.
- Compute the matrix $U = (u_{jk})$ in which $u_{jk} = s_{jk} = s(Y_j; Y_k)$ if $j \neq k$ and $u_{jk} = 1$ otherwise. Here Y_l, $l = 1, ..., p$ are the columns of the transformed data matrix Y and s(. , .) is a robust estimate of the covariance of two random variables like the one in the above equation.
- Obtain the principal component decomposition (PCA) of Y by decomposing $U = E\Lambda E^T$ where Λ is a diagonal matrix $\Lambda = diag(\lambda_1, ..., \lambda_p)$ with the eigenvalues λ_j of U and E is a matrix with columns of the eigenvalues e_j of U.
- Define $z_i = E^T y_i = E^T D^{-1} x_i$ and $A = DE$. Then, the estimator of the scatter, Σ is $C_{OGK} = A\Gamma A^T$ where $\Gamma = diag(\sigma(Z_j)^2)$, $j = 1, ..., p$ and the location estimator is $T_{OGK} = Am$ where, $m = m(z_i) = (m(Z_1), ..., m(Z_p))$ is a robust mean function.

This can be iterated by computing C_{OGK} and T_{OGK} for $Z = z_1, ..., z_n$ obtained in the last step of the procedure and then transforming back to the original coordinate system. Simulations (Maronna and Zamar, 2002) show that iterations beyond the second did not lead to improvement.

Screen example is depicted below: The upper figure presents outliers (values above a chosen cutoff level that is determined by the user). Green points represent normal observations, red points common (mahalanobis) outliers, and blue triangles OGK's robust outliers. The lower figure presents bilateral: histograms (including normal curves as bencmarks), correlation coefficients (and their respective t-statistics), and common (within red ellipses) and OGK's robust (within blue ellipses) distances from the center of bilateral distributions.

OGK Robust and Classic Distances - all variables



OGK Bilateral Robust and Classic Distances