Robust scale estimation with extensions
Garth Tarr, Samuel Müller and Neville Weber

School of Mathematics and Statistics
THE UNIVERSITY OF SYDNEY
Outline

The robust scale estimator $P_n$

Robust covariance estimation

Robust covariance matrices

Conclusion and key references
<table>
<thead>
<tr>
<th>Covariance</th>
<th>Multivariate</th>
<th>Conclusion</th>
</tr>
</thead>
</table>

## Outline

The robust scale estimator $P_n$

Robust covariance estimation

Robust covariance matrices

Conclusion and key references
History of scale efficiencies at the Gaussian \((n = 20)\)
Pairwise mean scale estimator: $P_n$

- Consider the $U$-statistic, based on the pairwise mean kernel,

$$U_n(X) = \left( \frac{n}{2} \right)^{-1} \sum_{i<j} \frac{X_i + X_j}{2}.$$ 

- Let $H(t) = P((X_i + X_j)/2 \leq t)$ be the cdf of the kernels with corresponding empirical distribution function,

$$H_n(t) = \left( \frac{n}{2} \right)^{-1} \sum_{i<j} \mathbb{I} \left\{ \frac{X_i + X_j}{2} \leq t \right\}, \quad \text{for } t \in \mathbb{R}.$$ 

**Definition (Interquartile range of pairwise means)**

$$P_n = c \left[ H_n^{-1} (0.75) - H_n^{-1} (0.25) \right],$$

where $c \approx 1.048$ is a correction factor to ensure $P_n$ is consistent for the standard deviation when the underlying observations are Gaussian.
Influence curve

\[ IC(x; T, \Phi) \]
Asymptotic variance and relative efficiency

- Tarr, Müller and Weber (2012) show that $P_n$ is asymptotically normal with variance, $V$, given by the expected square of the influence function.

- When the underlying data are Gaussian,

$$V(P_n, \Phi) = \int IC(x; P_n, \Phi)^2 d\Phi(x) = 0.579.$$  

- This equates to an asymptotic efficiency of 0.86.

We have a robust scale estimator that is still quite efficient at the Gaussian distribution but how can we extend this to the multivariate setting?
Outline

The robust scale estimator $P_n$

Robust covariance estimation

Robust covariance matrices

Conclusion and key references
From scale to covariance: the GK device

• Gnanadesikan and Kettenring (1972) relate scale and covariance using the following identity,

\[ \text{cov}(X, Y) = \frac{1}{4\alpha\beta} \left[ \text{var}(\alpha X + \beta Y) - \text{var}(\alpha X - \beta Y) \right], \]

where \( X \) and \( Y \) are random variables.

• In general, \( X \) and \( Y \) can have different units, so we set \( \alpha = 1/\sqrt{\text{var}(X)} \) and \( \beta = 1/\sqrt{\text{var}(Y)} \).

• Replacing variance with \( P_n^2 \) we can similarly construct,

\[ \gamma_P(X, Y) = \frac{1}{4\alpha\beta} \left[ P_n^2(\alpha X + \beta Y) - P_n^2(\alpha X - \beta Y) \right], \]

where \( \alpha = 1/P_n(X) \) and \( \beta = 1/P_n(Y) \).
Robustness properties

**Breakdown value**

- $P_n$ can breakdown if 25% of the pairwise means are contaminated.
- This can occur when 13.4% of observations are contaminated.
- $\gamma_P$ inherits the **13.4%** breakdown value from $P_n$.
- Optionally, adaptive trimming $\Rightarrow$ 50% breakdown value.

**Influence curve**

- Following Genton and Ma (1999), we have shown that **influence curve** and therefore the **gross error sensitivity** of $\gamma_P$ can be derived from the IC of $P_n$.
- Key point: the IC is **bounded** and hence the gross error sensitivity is **finite**.
Outline

The robust scale estimator $P_n$

Robust covariance estimation

Robust covariance matrices

Conclusion and key references
Covariance matrices

“Good” covariance matrices should be:

1. Positive-semidefinite.
2. Affine equivariant. That is, if $\hat{\mathbf{C}}$ is a covariance matrix estimator, $\mathbf{A}$ and $\mathbf{a}$ are constants and $\mathbf{X}$ is random then,

$$\hat{\mathbf{C}}(\mathbf{A}\mathbf{X} + \mathbf{a}) = \mathbf{A}\hat{\mathbf{C}}(\mathbf{X})\mathbf{A}' .$$

How can we ensure that a matrix full of pairwise covariances is a “good” covariance matrix?

- Maronna and Zamar (2002) outline the OGK routine which ensures that the resulting covariance matrix is positive-definite and approximately affine-equivariant matrix.
Application: Principal Component Analysis

Definition (Principal Component Analysis)

PCA is a dimension reduction technique that transforms a set of variables into a set of principal components (uncorrelated variables).

- PCA can be formulated as an eigenvalue decomposition of a covariance matrix and so is inherently susceptible to outliers.

Example (Versicolor iris data)

50 observations on 4 variables: length and width of petals and sepals.
Iris data

Sepal Length

Sepal Width

Petal Length

Petal Width

Conclusion
## Covariance matrices for the iris data

<table>
<thead>
<tr>
<th>Classical covariance</th>
<th>Sepal Length</th>
<th>Sepal Width</th>
<th>Petal Length</th>
<th>Petal Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sepal Length</td>
<td>0.27</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sepal Width</td>
<td>0.09</td>
<td><strong>0.10</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Petal Length</td>
<td>0.18</td>
<td>0.08</td>
<td>0.22</td>
<td></td>
</tr>
<tr>
<td>Petal Width</td>
<td>0.06</td>
<td>0.04</td>
<td>0.07</td>
<td>0.04</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Robust before OGK</th>
<th>Sepal Length</th>
<th>Sepal Width</th>
<th>Petal Length</th>
<th>Petal Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sepal Length</td>
<td>0.28</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sepal Width</td>
<td>0.09</td>
<td><strong>0.14</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Petal Length</td>
<td>0.20</td>
<td>0.10</td>
<td>0.23</td>
<td></td>
</tr>
<tr>
<td>Petal Width</td>
<td>0.06</td>
<td>0.05</td>
<td>0.08</td>
<td>0.03</td>
</tr>
</tbody>
</table>
Covariance matrices for the iris data

<table>
<thead>
<tr>
<th>Robust before OGK</th>
<th>Sepal</th>
<th>Petal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Length</td>
<td>Width</td>
</tr>
<tr>
<td>Sepal Length</td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td>Sepal Width</td>
<td>0.09</td>
<td>0.14</td>
</tr>
<tr>
<td>Petal Length</td>
<td>0.20</td>
<td>0.10</td>
</tr>
<tr>
<td>Petal Width</td>
<td>0.06</td>
<td>0.05</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Robust after OGK</th>
<th>Sepal</th>
<th>Petal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Length</td>
<td>Width</td>
</tr>
<tr>
<td>Sepal Length</td>
<td>0.29</td>
<td></td>
</tr>
<tr>
<td>Sepal Width</td>
<td>0.09</td>
<td>0.10</td>
</tr>
<tr>
<td>Petal Length</td>
<td>0.19</td>
<td>0.09</td>
</tr>
<tr>
<td>Petal Width</td>
<td>0.06</td>
<td>0.04</td>
</tr>
</tbody>
</table>
Iris data loadings
2 new obs.

Sepal Length

Sepal Width

Petal Length

Petal Width

Covariance

Multivariate

Conclusion
Iris data loadings (2 obs. from different species)

Classical (no outliers)

Classical (with outliers)
Iris data loadings (2 obs. from different species)

Robust (no outliers)

Robust (with outliers)

Sepal Length
Petal Length
Sepal Width
Petal Length
Sepal Width
Extreme outlier

Sepal Length

Sepal Width

Petal Length

Petal Width

Covariance

Multivariate

Conclusion
Iris data loadings (extreme outlier – petal length)
Iris data loadings (extreme outlier – petal length)

Robust (no outlier)

Robust (with outlier)

Sepal Length

Petal Length

Sepal Width

Petal Length

P Covariance Multivariate Conclusion

$Iris$ data loadings (extreme outlier – petal length)
Outline

The robust scale estimator $P_n$

Robust covariance estimation

Robust covariance matrices

Conclusion and key references
Summary

1. Aim

- Efficient, robust and widely applicable scale and covariance estimators.

2. Method

- $P_n$ scale estimator with the GK identity and an orthogonalisation procedure for pairwise covariance matrices.

3. Results

- 86% asymptotic efficiency at the Gaussian distribution.
- Robustness properties flow through to covariance estimates and beyond.
Robustness properties of dispersion estimators.

Robust estimates, residuals and outlier detection with multiresponse data.

Robust estimates of location and dispersion for high-dimensional datasets.

Alternatives to the median absolute deviation.

A robust scale estimator based on pairwise means.