Robust scale estimation with extensions

Garth Tarr, Samuel Müller and Neville Weber

School of Mathematics and Statistics
THE UNIVERSITY OF SYDNEY

Outline

Outline

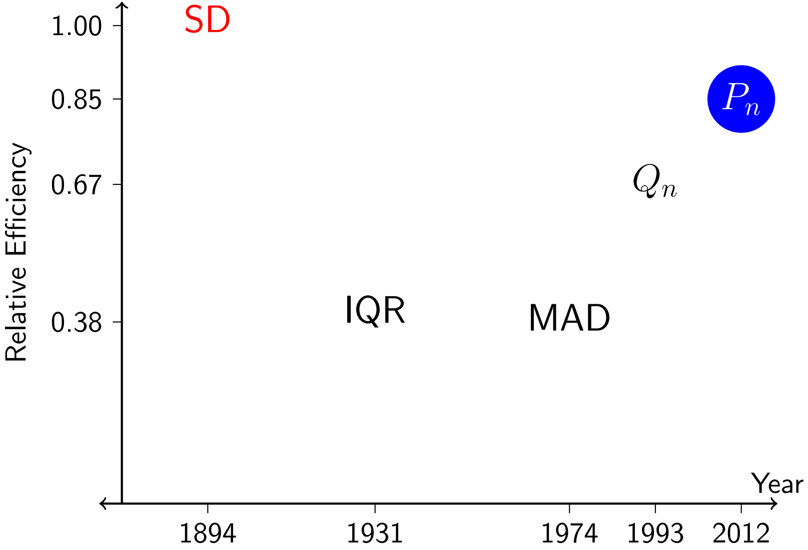
The robust scale estimator P_n

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,

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

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

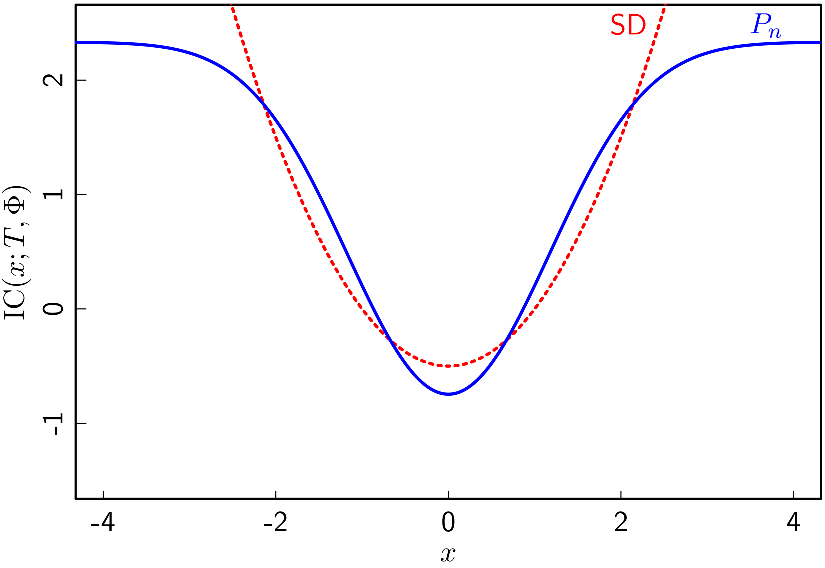
$$H_n(t) = \binom{n}{2}^{-1} \sum_{i<j} \mathbb{I}\left\{ \frac{X_i + X_j}{2} \leq t \right\}, \quad \text{for } t \in \mathbb{R}.$$

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



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,

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

- 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 $\boldsymbol{P_n^2}$ 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.
- γ_P inherits the **13.4%** breakdown value from P_n.
- Optionally, adaptive trimming $\implies$ 50% breakdown value.

Influence curve

- Following Genton and Ma (1999), we have shown that **influence curve** and therefore the **gross error sensitivity** of γ_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

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,

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

! 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

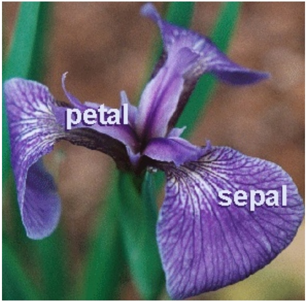
Definition (Principal Component Analysis)

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

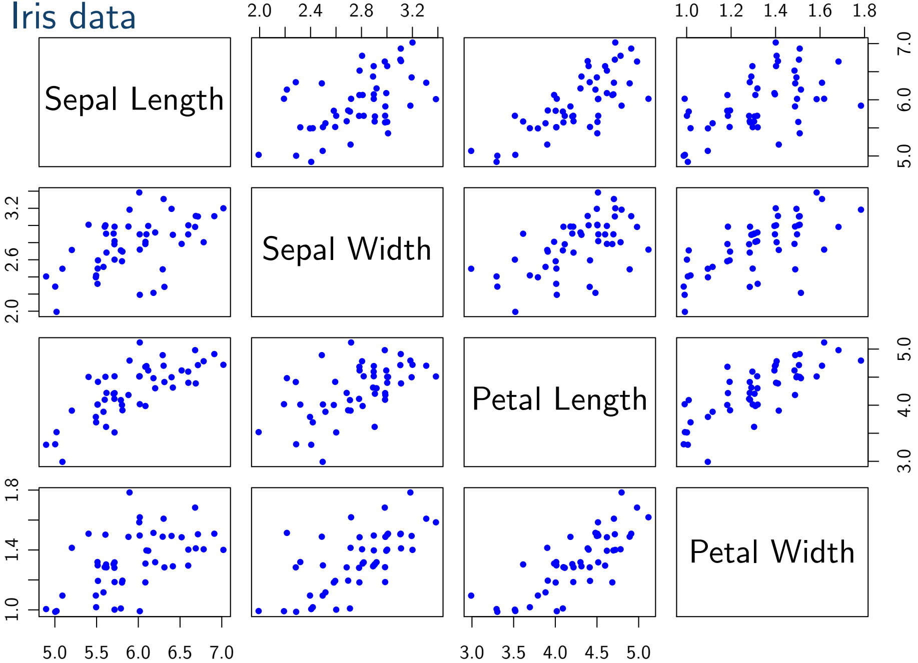
- PCA can be formulated as an eigenvalue decomposition of a covariance matrix and so is inherently susceptible to outliers.

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
2.0 2.4 2.8 3.2
1.0 1.2 1.4 1.6 1.8
5.0 6.0 7.0
2.0 2.6 3.2
3.0 4.0 5.0
1.0 1.4 1.8
5.0 5.5 6.0 6.5 7.0
3.0 3.5 4.0 4.5 5.0

Covariance matrices for the iris data

Classical covariance	Sepal Length	Sepal Width	Petal Length	Petal Width
Sepal.Length	0.27			
Sepal.Width	0.09	0.10		
Petal.Length	0.18	0.08	0.22	
Petal.Width	0.06	0.04	0.07	0.04

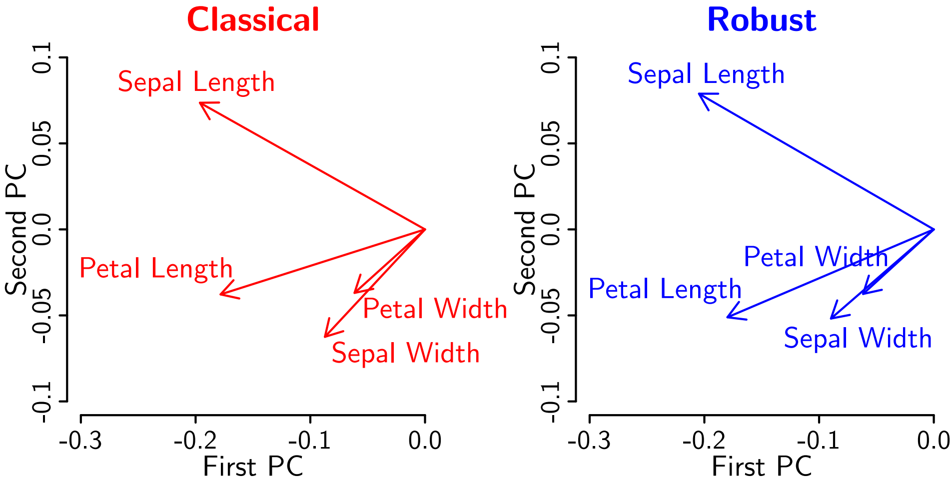
Robust before OGK	Sepal Length	Sepal Width	Petal Length	Petal Width
Sepal Length	0.28			
Sepal Width	0.09	0.14		
Petal Length	0.20	0.10	0.23	
Petal Width	0.06	0.05	0.08	0.03

Covariance matrices for the iris data

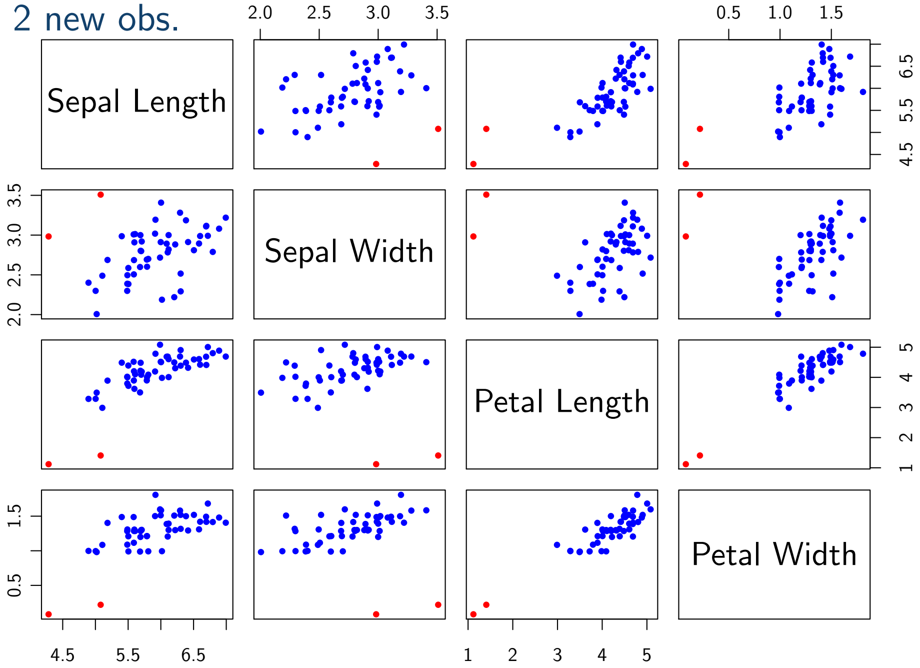
Robust before OGK	Sepal		Petal	
	Length	Width	Length	Width
Sepal Length	0.28			
Sepal Width	0.09	0.14		
Petal Length	0.20	0.10	0.23	
Petal Width	0.06	0.05	0.08	0.03

Robust after OGK	Sepal		Petal	
	Length	Width	Length	Width
Sepal Length	0.29			
Sepal Width	0.09	0.10		
Petal Length	0.19	0.09	0.23	
Petal Width	0.06	0.04	0.08	0.04

Iris data loadings

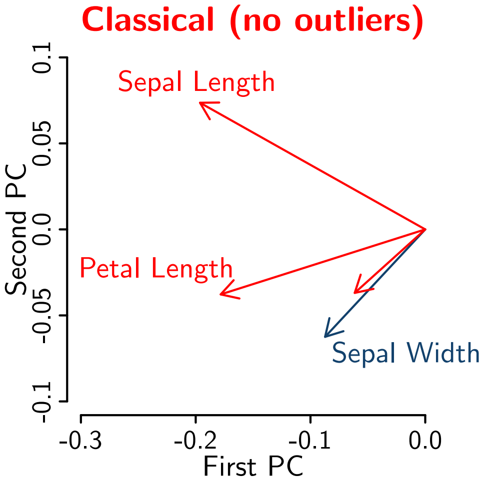


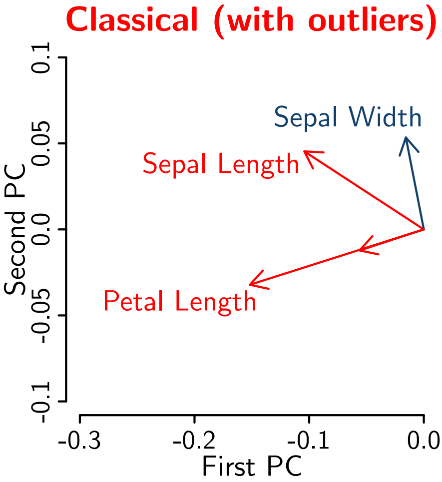
2 new obs.



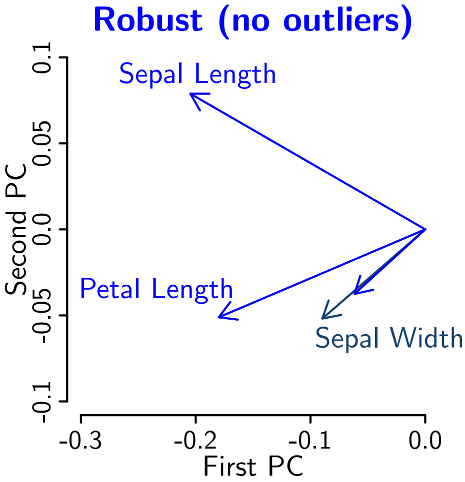
Sepal Length
Sepal Width
Petal Length
Petal Width
2.0
2.5
3.0
3.5
0.5
1.0
1.5
4.5
5.5
6.5
1
2
3
4
5

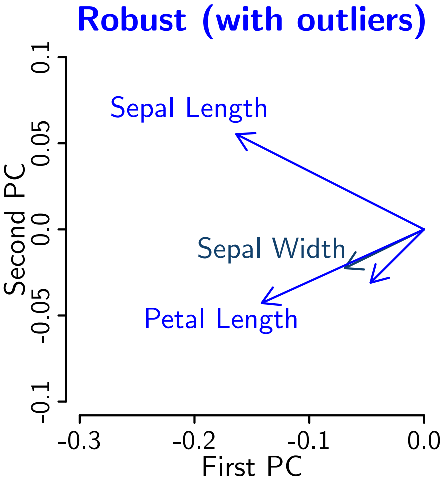
Iris data loadings (2 obs. from different species)



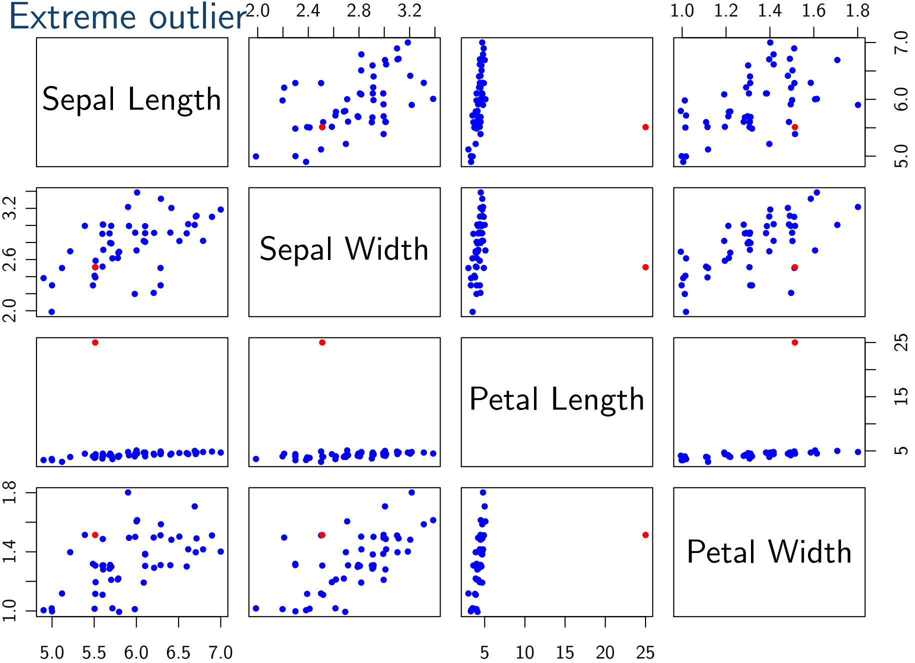


Iris data loadings (2 obs. from different species)

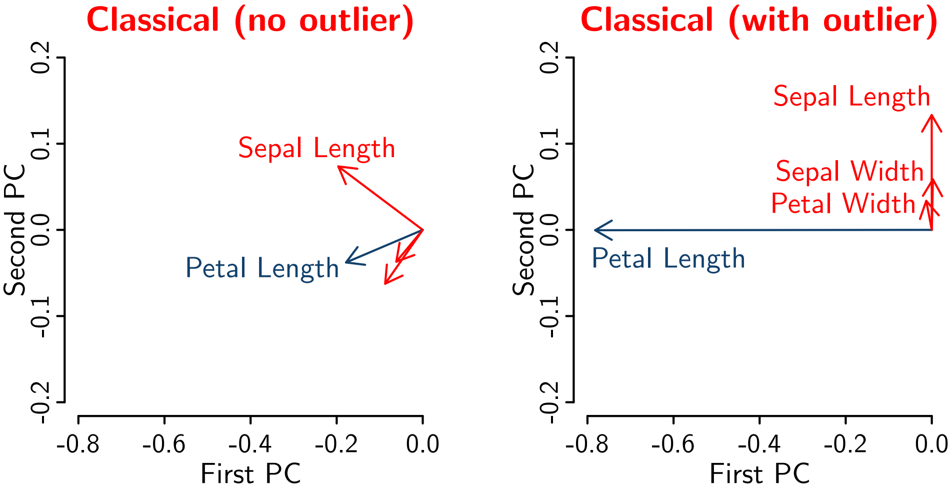




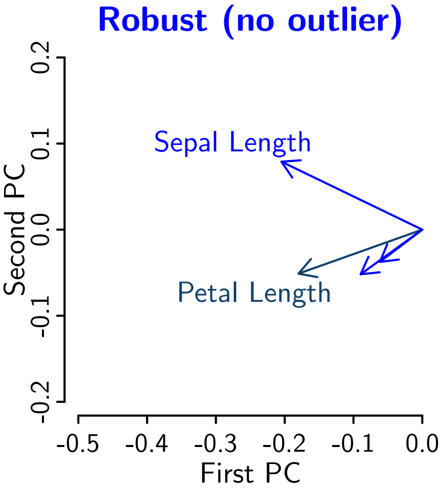
Extreme outlier

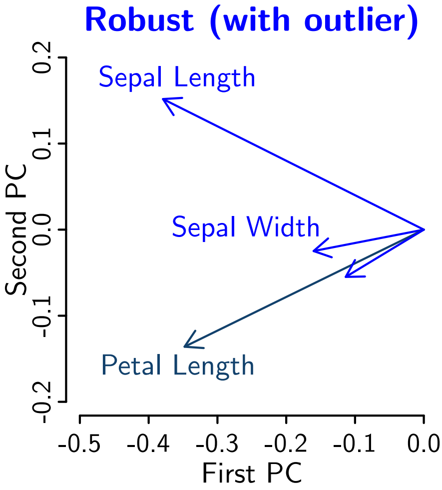


Iris data loadings (extreme outlier – petal length)



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.

References

Genton, M. and Ma, Y. (1999).
Robustness properties of dispersion estimators.
Statistics & Probability Letters, 44(4):343–350.

Gnanadesikan, R. and Kettenring J. R. (1972).
Robust estimates, residuals and outlier detection with multiresponse data.
Biometrics, 28(1):81–124.

Maronna, R. and Zamar, R., (2002).
Robust estimates of location and dispersion for high-dimensional datasets.
Technometrics, 44(4):307–317.

Rousseeuw, P. and Croux, C. (1993).
Alternatives to the median absolute deviation.
Journal of the American Statistical Association, 88(424):1273–1283.

Tarr, G., Müller, S. and Weber, N.C., (2012).
A robust scale estimator based on pairwise means.
Journal of Nonparametric Statistics, 24(1):187–199.