Minimum **R**egularized Covariance Determinant Estimator

Honey, we shrunk the data and the covariance matrix

Kris Boudt
(joint with: P. Rousseeuw, S. Vanduffel and T. Verdonck)

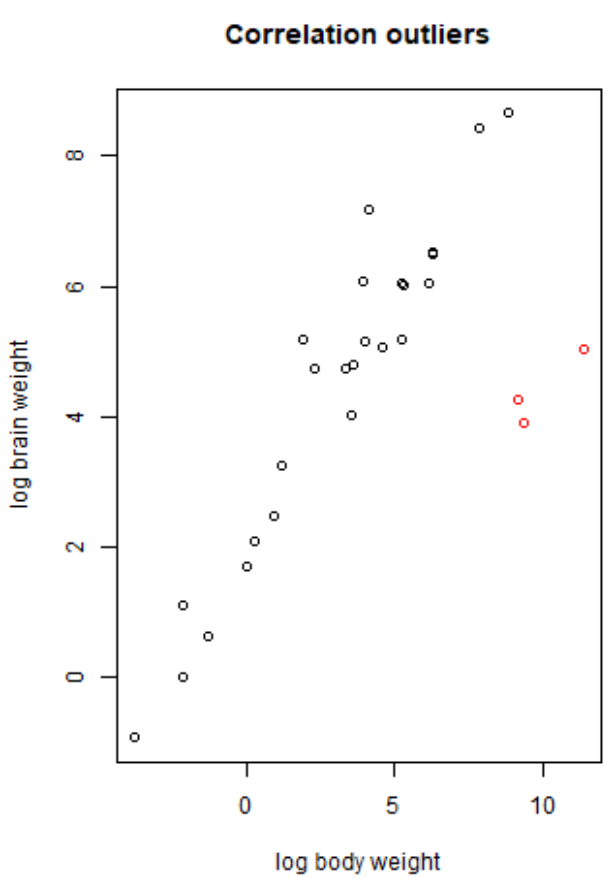
Vrije Universiteit Brussel/Amsterdam

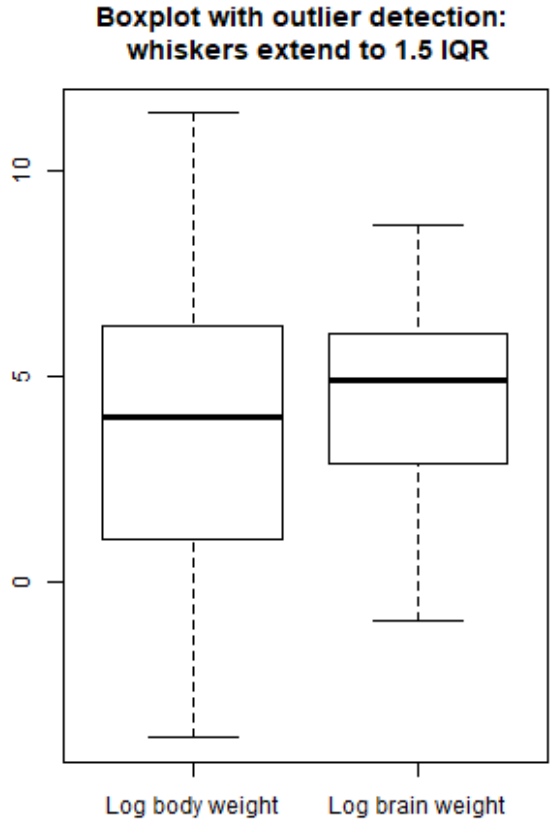
June 1, 2018

Peter Rousseeuw's MCD estimator

- Goal: Shrink the data to the subset of h "good" observations and estimate the covariance on that subset. Typically: $h = \lceil 0.5n \rceil$ or $h = \lceil 0.75n \rceil$.
- More formally: Given an $n \times p$ data matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ with $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$, the MCD finds the $h < n$ observations whose sample covariance matrix has the lowest possible determinant.
- The mean and covariance of that sample is the MCD mean and scatter matrix.

Why is the MCD useful? Example of animal body and brain weight.

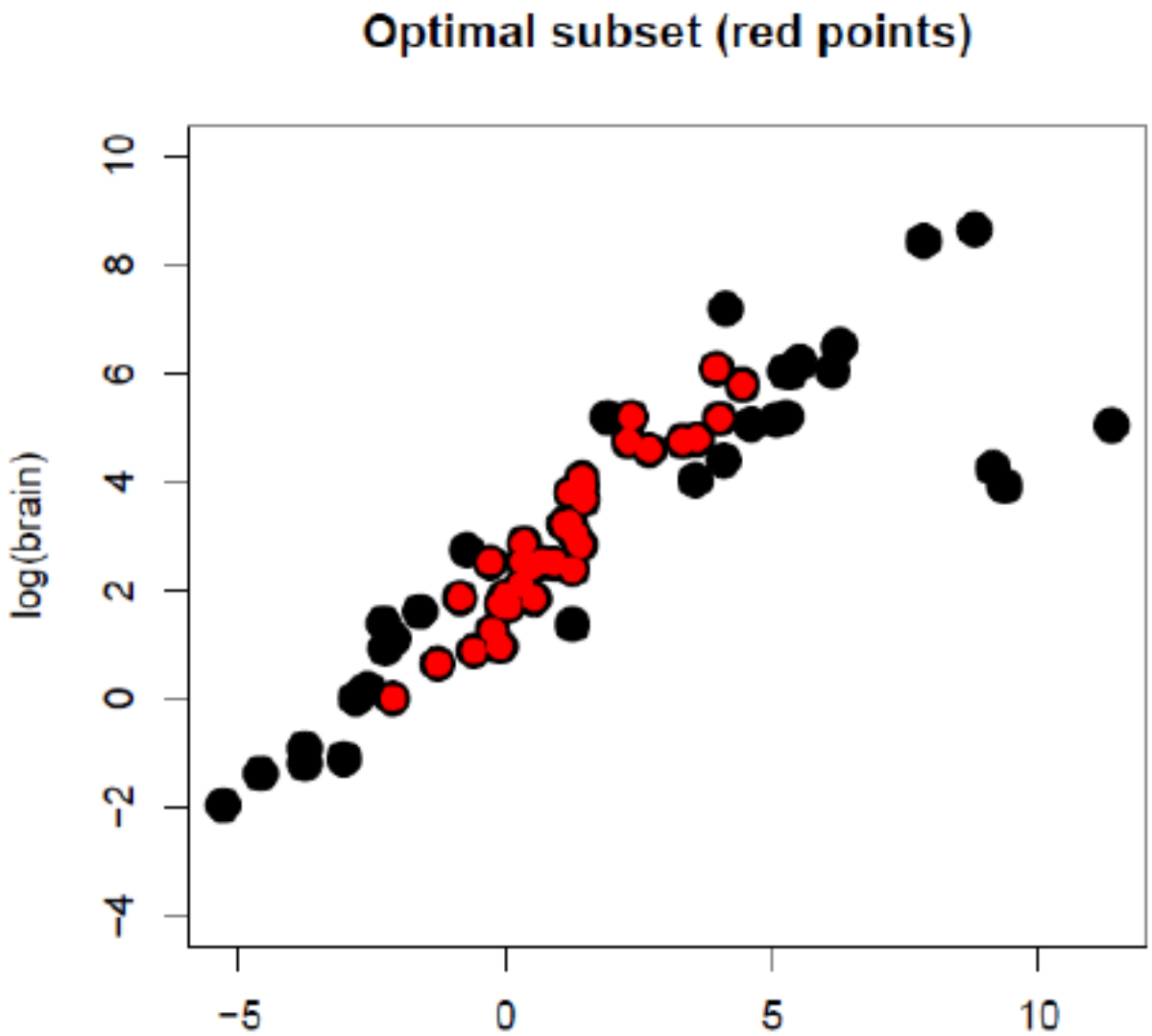




Index	Species	Body weight (in kilograms)	Log body weight	Brain weight (in grams)	Log brain weight
1	Mountain Beaver	1.350	0.300	8.1	2.092
2	Cow	465.000	6.142	423.0	6.0473
3	Gray wolf	36.330	3.593	119.5	4.783
4	Goat	27.660	3.320	115.0	4.745
5	Guinea pig	1.040	0.039	5.5	1.705
6	**Diplodocus**	11700.000	9.367	50.0	3.912
7	Asian elephant	2547.000	7.843	4603.0	8.434
8	Donkey	187.100	5.232	419.0	6.038
9	Horse	521.000	6.256	655.0	6.485
10	Potar monkey	10.000	2.303	115.0	4.745
11	Cat	3.300	1.194	25.6	3.243
12	Giraffe	529.000	6.271	680.0	6.522
13	Gorilla	207.000	5.333	406.0	6.006
14	Human	62.000	4.127	1320.0	7.185
15	African elephant	6654.000	8.803	5712.0	8.650
16	**Triceratops**	9400.000	9.148	70.0	4.248
17	Rhesus monkey	6.800	1.917	179.0	5.187
18	Kangaroo	35.000	3.555	56.0	4.025
19	Hamster	0.120	−2.120	1.0	0.00
20	Mouse	0.023	−3.772	0.4	−0.916
21	Rabbit	2.500	0.916	12.1	2.493
22	Sheep	55.500	4.016	175.0	5.165
23	Jaguar	100.000	4.605	157.0	5.056
24	Chimpanzee	52.160	3.954	440.0	6.087
25	**Brachiosaurus**	87000.000	11.374	154.5	5.040
26	Rat	0.280	−1.273	1.9	0.642
27	Mole	0.122	−2.104	3.0	1.099
28	Pig	192.000	5.257	180.0	5.193

Table: Body and brain weight for 28 Animals.

The MCD subset for $h = \lceil 0.5n \rceil$



Merits of the MCD estimator

- Intuitive: minimizes a clear objective function;
- Elliptical distributions: consistency, asymptotic normality;
- Resistance to outliers: High breakdown point, bounded influence function.
- Efficient algorithm exists to solve the problem, exploiting the C-step theorem. See covMcd in robustbase and rrcov of Valentin Todorov.

Drawback of the MCD estimator

- Use of subsets for estimation can be inefficient. Solution is to use a reweighted version of the MCD in which only the detected outliers receive a low weight.
- Practical implementation requires to invert the subset covariance, and is thus only applicable for $p < h$. For accuracy, recommendation is $n > 5p$.
  $\Rightarrow$ Problematic in case of fat data. **A big MCD is needed**.

Solution: Shrinking the covariance matrix

Honey, I Shrunk the Sample Covariance Matrix

Olivier Ledoit and Michael Wolf

The Journal of Portfolio Management Summer 2004, 30 (4) 110-119; DOI: https://doi.org/10.3905/jpm.2004.110

- Extremely popular in finance.
- Implemented in the covEstimation function of the RiskPortfolios package of Ardia, Boudt and Gagnon-Fleury (2017).
- Not robust. In fact, in case of the identity matrix as target, one can show that one big outlier is sufficient to put all the weight on the target.

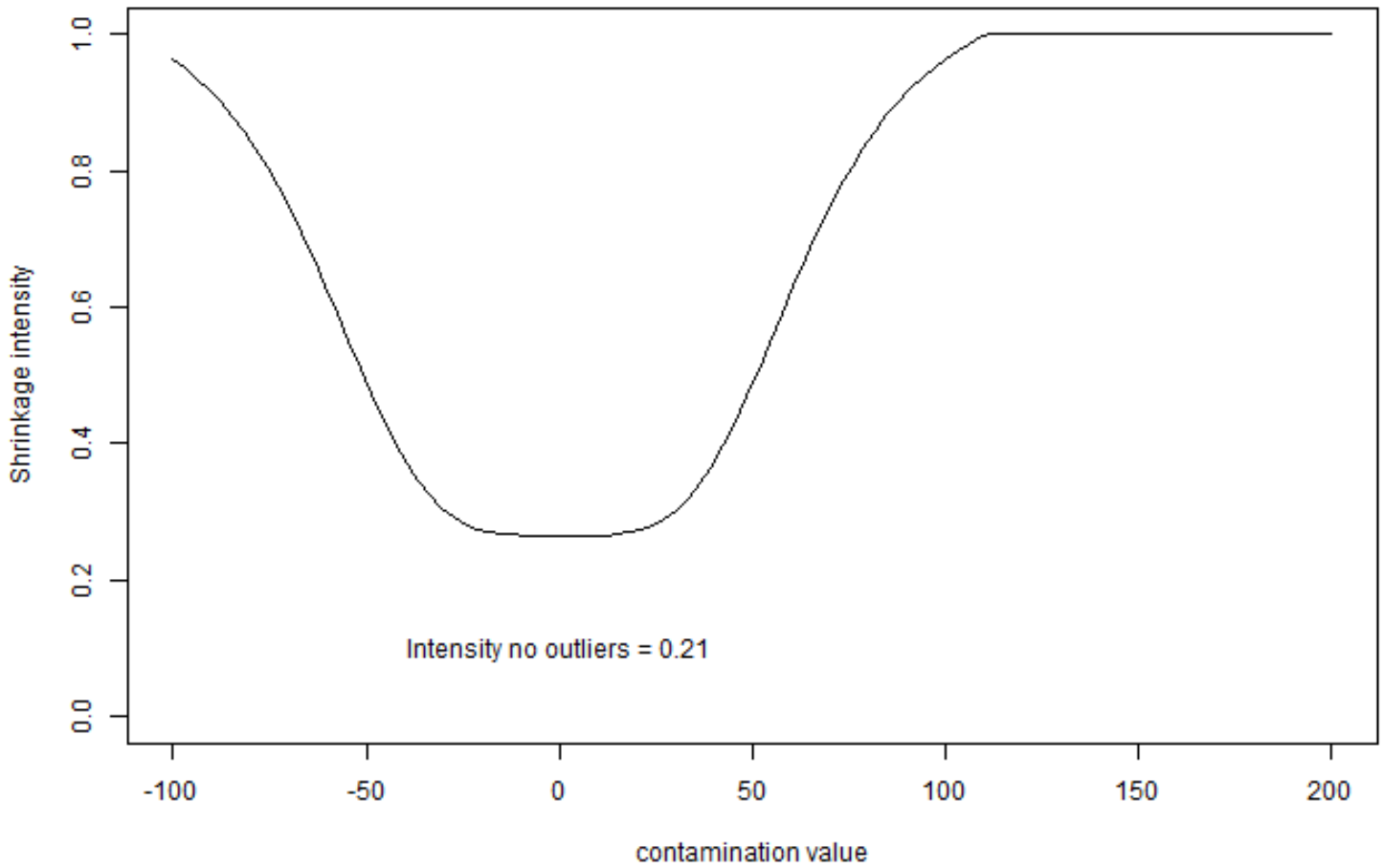
Non-robustness of shrinkage to the identity

```r
# Define function computing Shrinkage intensity when target = z* Identity
library(RiskPortfolios)
intensity <- function (rets)
{
  lwCovElement <- RiskPortfolios:::.lwCovElement(rets, type = "oneparm")
  t <- lwCovElement$t
  n <- lwCovElement$n
  mu <- lwCovElement$mu
  shiftRets <- lwCovElement$shiftRets
  smple <- lwCovElement$smple
  y <- lwCovElement$y
  meanvar <- mean(diag(smple))
  prior <- meanvar * diag(n)
  phiMat <- crossprod(y)/t - 2 * (crossprod(shiftRets)) * smple/t + smple^2
  phi <- sum(apply(phiMat, 2, sum))
  gamma <- norm(smple - prior, type = "F")^2
  kappa <- phi/gamma
  shrinkage <- pmax(0, pmin(1, kappa/t))
  return(shrinkage)
}
```

Non-robustness of shrinkage to the identity

```
1  data("Industry_10") #rets (in %): 200 rows, 100 columns
2  # on this data, the shrinkage intensity equals 0.21
3  # what happens if we replace the first observation by an outlier of
       size k, with k ranging from -100 (%) to 200 (%)?
4  vrho <- rep(NA,100)
5  vk <- seq(-100,200,length.out=100)
6  i <- 1
7  contrets <- rets
8  for(k in vk ){
9    contrets[1,] <- k
10   vrho[i] <- intensity(contrets)
11   i <- i+1
12 }
13 plot(vk,vrho,type="l", xlab="contamination value",ylim=c(0,1),ylab=
       "Shrinkage intensity")
14 text( 0,0.1, paste("Intensity no outliers =", round(intensity(rets)
       ,2) ))
```

Explosion of shrinkage factor with outliers and identity as target



We thus need to shrink BOTH the data and the covariance matrix.

MRCD: Minimum Regularized Covariance Determinant estimator

- Define regularized covariance estimator as convex combination of
  (1) predetermined, symmetric, positive definite target matrix $\mathbf{T}$.
  For simplicity in notation, assume here $\mathbf{T} = I_p$.
  (2) sample covariance estimate $\mathbf{S}(H)$ based on a subset H

$$\mathbf{K}(H) = \rho\, \mathbf{I} + (1 - \rho) c_\alpha \mathbf{S}(H),$$

where c_α is a consistency factor and $\rho \in (0, 1]$ is regularization intensity parameter.

MRCD: Minimum Regularized Covariance Deterimanant estimator

- Find subset H that minimizes the determinant of $\mathbf{K}(H)$:

$$H_{MRCD} = \arg \min_{H \in \mathcal{H}_h} (\det \mathbf{K}(H))^{1/p}.$$

- Once optimal subset is determined, MRCD scatter is computed as $\mathbf{K}_{MRCD} = \mathbf{K}(H_{MRCD})$.

Calibration of regularization parameter

$$\mathbf{K}(H) = \rho\, \mathbf{I} + (1-\rho) c_\alpha \mathbf{S}(H)$$

- Set ρ such that $\mathbf{K}(H)$ is well-conditioned ($\lambda_{\text{max}}/\lambda_{\text{min}} \leq 1000$).
- Easy to implement, since the eigenvalues of $\mathbf{K}(H)$ equal $\rho + (1-\rho)\lambda$.
- We only use regularization when needed!

Key property: C-step theorem

Starting from h-subset H_1, compute $\mathbf{m}_1 = \frac{1}{h}\sum_{i \in H_1} \mathbf{x}_i$ and $\mathbf{S}_1 = \frac{1}{h}\sum_{i \in H_1}(\mathbf{x}_i - m_1)(\mathbf{x}_i - \mathbf{m}_1)'$. The matrix $\mathbf{K}_1 = \rho\mathbf{T} + (1-\rho)\mathbf{S}_1$ is positive definite hence invertible, so we can compute

$$d_1(i) = (\mathbf{x}_i - \mathbf{m}_1)'\mathbf{K}_1^{-1}(\mathbf{x}_i - \mathbf{m}_1) \qquad i = 1, \ldots, n.$$

Let H_2 be an h-subset for which

$$\sum_{i \in H_2} d_1(i) \leq \sum_{i \in H_1} d_1(i)$$

and compute $\mathbf{m}_2 = \frac{1}{h}\sum_{i \in H_2} \mathbf{x}_i$, $\mathbf{S}_2 = \frac{1}{h}\sum_{i \in H_1}(\mathbf{x}_i - m_2)(\mathbf{x}_i - \mathbf{m}_2)'$ and $\mathbf{K}_2 = \rho\mathbf{T} + (1-\rho)\mathbf{S}_2$. Then:

$$\det(\mathbf{K}_2) \leq \det(\mathbf{K}_1),$$

with equality iff $\mathbf{m}_2 = \mathbf{m}_1$ and $\mathbf{K}_2 = \mathbf{K}_1$.

Algorithm

0. **Standardize the data** using the median and the Qn estimator for univariate location and scale.
1. **Initialization of subset selection.** Follow Subsection 3.1 in Hubert et al. (2012) to obtain six robust, well-conditioned initial location estimates $\mathbf{m}^i$ and scatter estimates $\mathbf{S}^i$ $(i = 1, \ldots, 6)$.
2. **C-step.** Determine the subsets H_0^i containing the h observations with lowest Mahalanobis distance in terms of $\mathbf{m}^i$ and $\mathbf{S}^i$.
3. **Calibrate intensity parameter.** For each subset H_0^i, determine the smallest value of $0 \leq \rho^i < 1$ for which $\rho^i\ \mathbf{I} + (1 - \rho^i) c_\alpha \mathbf{S}(H_0^i)$ is well-conditioned. Denote this value as ρ_0^i .
4. If $\max_i \rho_0^i \leq 0.1$, set $\rho = \max_i \rho_0^i$, else set $\rho = \max\{0.1; \text{median}_i \rho_0^i\}$.
5. **Repeat C-steps till convergence.** For the initial subset H_0^i for which $\rho_0^i \leq \rho$, repeat the generalized C-steps using $\rho\ \mathbf{I} + (1 - \rho) c_\alpha \mathbf{S}(H_0^i)$ until convergence. Denote the resulting subsets as H^i .
6. Let H_{MRCD} be the subset for which $\rho\ \mathbf{I} + (1 - \rho) c_\alpha \mathbf{S}(H^i)$ has the lowest determinant among the candidate subsets.

Conclusion: Merits of the MRCD estimator

- *Honey, we shrunk the data and the covariance matrix*.
- Intuitive: minimizes a clear objective function.
- For fixed p, same asymptotic properties as MCD since no regularization asymptotically.
- Resistant to outliers: High breakdown point, bounded influence function.
- Practical: algorithm based on C-steps
- Next: Simulation study confirms accuracy and that it outperforms the only other multivariate robust covariance estimator available for fat data, namely the OGK estimator of Maronna and Zamar (2002), implemented as covOGK in robustbase.

Simulation study

- We focus on MRCD with the identity matrix as target.
- We follow Agostinelli et al. (2015) by simulating from a p-variate normal distribution with a correlation matrix that is randomly generated in each replication
- We take $n \times p$ as either 800×100, 200×100, and 200×400.
- We let the fraction of contamination ε be either 0% (clean data), 20% or 40% (medium sized ($k = 50$, along the eigenvector direction of $\mathbf{\Sigma}$ with smallest eigenvalue, since this is the direction where the contamination is hardest to detect).
- As performance measure we show the Mean Squared Error (MSE):

$$MSE = \frac{1}{M}\frac{1}{p^2}\sum_{m=1}^{M}\sum_{k=1}^{p}\sum_{l=1}^{p}(\mathbf{S}_m - \mathbf{\Sigma}_m)^2_{k,l} \quad .$$

Simulation Results: Clean data

	MSE			*Average value of* ρ		
	800×100	200×100	200×400	800×100	200×100	200×400
$h = \lceil 0.5n \rceil$	0.0024	0.0087	0.0105	0	0.0047	0.0108
$h = \lceil 0.75n \rceil$	0.0017	0.0064	0.0066	0	0.0001	0.0080
$h = n$	0.0013	0.0050	0.0049	0	0	0.0064
OGK	0.0015	0.0060	0.0058			

- No outliers: best is $h = n$
- Regularize when needed
- Outperformance wrt OGK.

Simulation Results: 20% contamination

	MSE			*Average value of* ρ		
	800×100	200×100	200×400	800×100	200×100	200×400
$h = \lceil 0.5n \rceil$	0.0024	0.0088	0.0102	0	0.0023	0.0053
$h = \lceil 0.75n \rceil$	0.0017	0.0066	0.0066	0	0	0.0039
$h = n$	17.4482	15.6942	4.3830	0.0220	0.1234	0.2251
OGK	0.0079	0.0187	0.0146			

- Breakdown for $h = n$.
- otherwise robustness.

Simulation Results: 40% contamination

	MSE			*Average value of* ρ		
	800×100	200×100	200×400	800×100	200×100	200×400
$h = \lceil 0.5n \rceil$	0.0025	0.0094	0.0099	0	0.0011	0.0022
$h = \lceil 0.75n \rceil$	2.8783	3.462	3.1857	0	0.0227	0.0842
$h = n$	66.8405	60.5137	16.4693	0.0367	0.1055	0.1736
OGK	0.0398	0.0744	0.0477			

- Breakdown for $h = n$ and $h = \lceil 0.75n \rceil$.
- Robustness for $h = \lceil 0.5n \rceil$.

Illustration on real data

- Outlier detection (chemistry);
- Regression analysis (criminology);
- Minimum variance portfolios.

Application to outlier detection: Octane data

- $n = 39$ gasoline samples with certain octane levels whose spectra have $p = 226$ wavelengths (variables).
- It is known that six samples (25, 26, 36-39) contain ethanol.
- For **outlier detection**, compute robust distance of each observation:

$$d_i = \sqrt{(\mathbf{x}_i - \mathbf{m}_{MRCD})' \mathbf{K}_{MRCD}^{-1} (\mathbf{x}_i - \mathbf{m}_{MRCD})}$$

Outlier detection

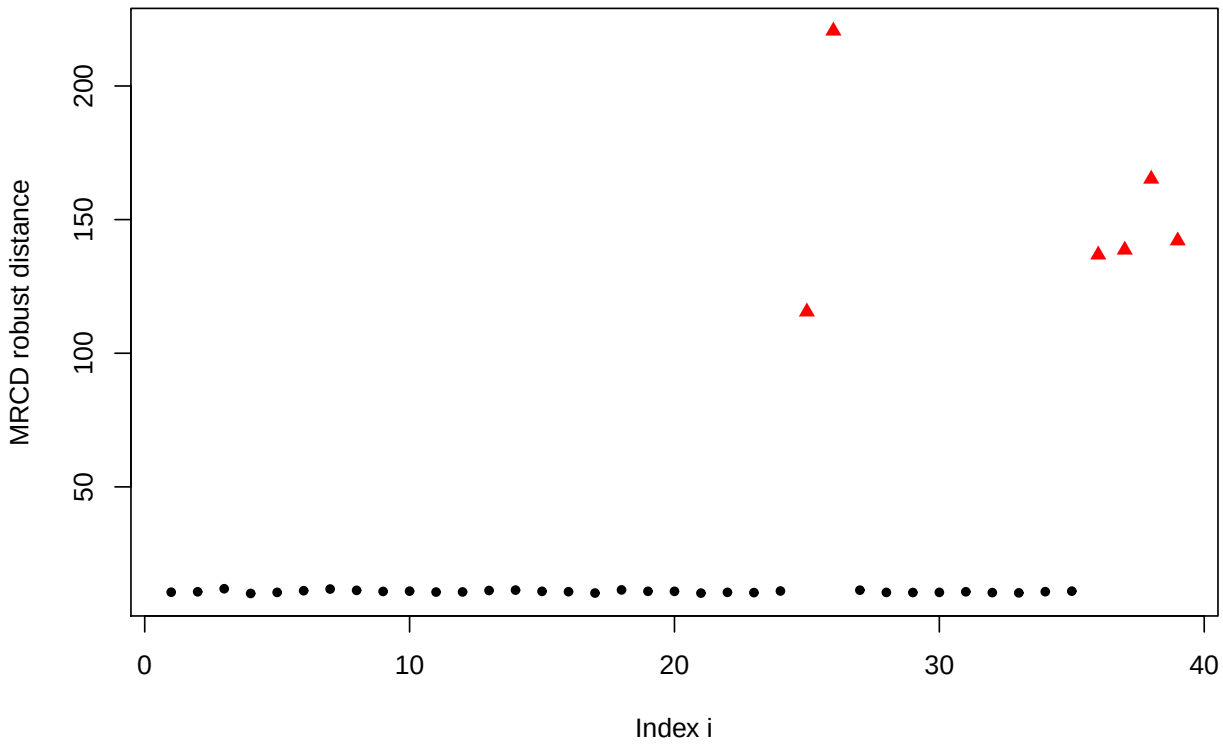


Figure: Robust distances of the octane data, based on the MRCD with $h = 33$.

Application to regression analysis: What explains the murder rate in US states?

- Murder rate per 100,000 residents in the $n = 50$ states of the US in 1980 on 25 demographic predictors.
- Robust regression:

$$\hat{\beta}_{MRCD} = \mathbf{K}_{xx}^{-1}\mathbf{K}_{xy} \ .$$

MRCD vs OLS coefficients

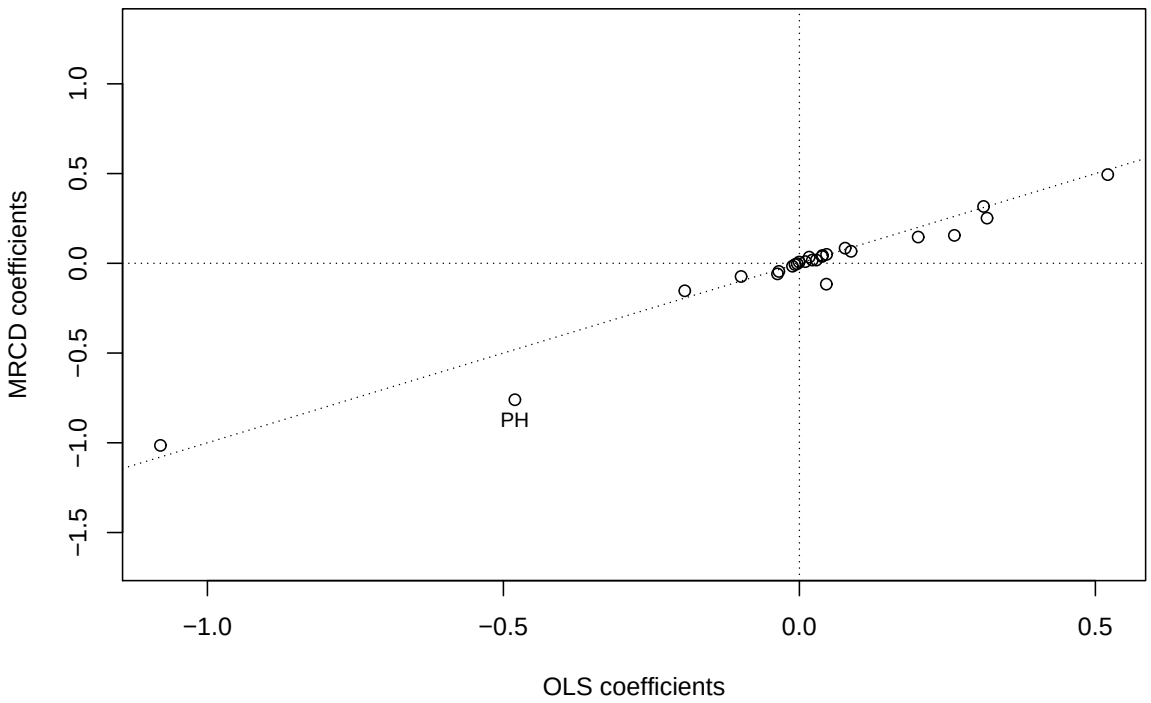


Figure: MRCD vs OLS coefficients of the multivariate regression of murder rate in 980 on 25 demographic predictors.

- Negative coefficient for telephone density in 1980.
- PH is proxy for the technological level of the state: on average the more technologically advanced the state, the lower the murder rate, other things being equal.

MRCD vs OLS coefficients

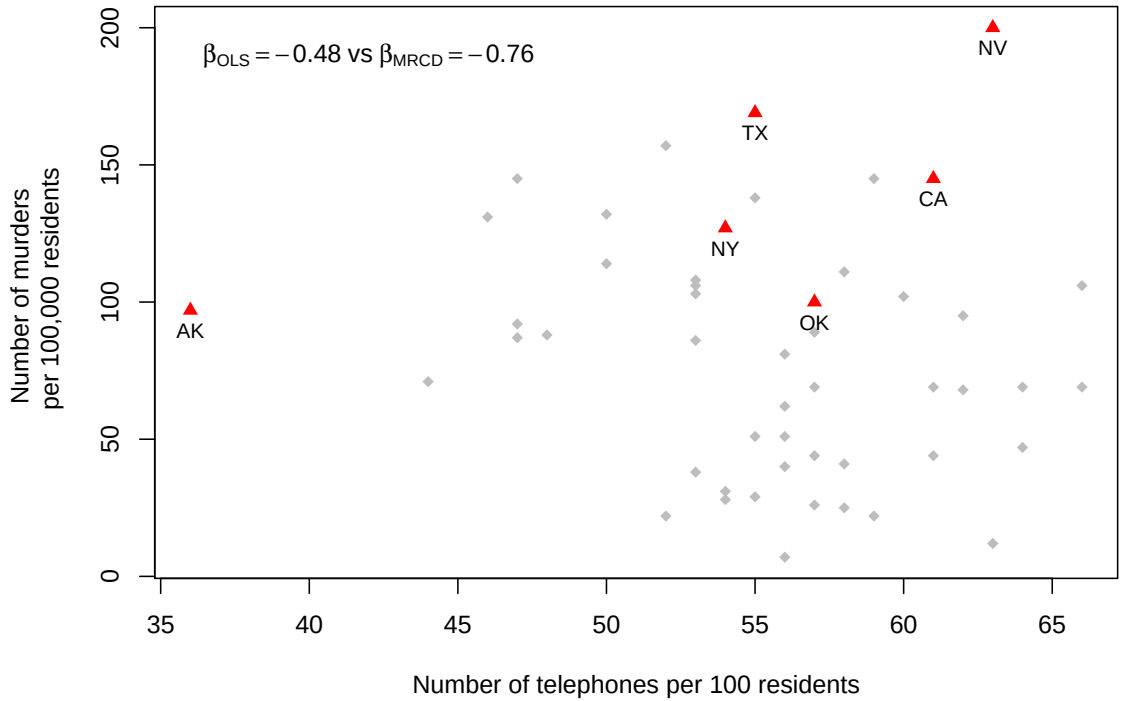


Figure: Scatter plot of murder rate per state against phone density. The red triangles indicate the observations that are not included in the final MRCD subset with $h = 44$.

- Arkansas is a bad leverage point. Nevada is a vertical outlier.
- Omitting them has led to a more negative value of this slope.

Application to minimum variance investing in the S&P 500

- Rolling sample of 3 years of monthly returns from 1987-2017
- Minimum variance portfolios with $0 \leq w_i \leq 5\%$

Minimum variance investing in the S&P 500

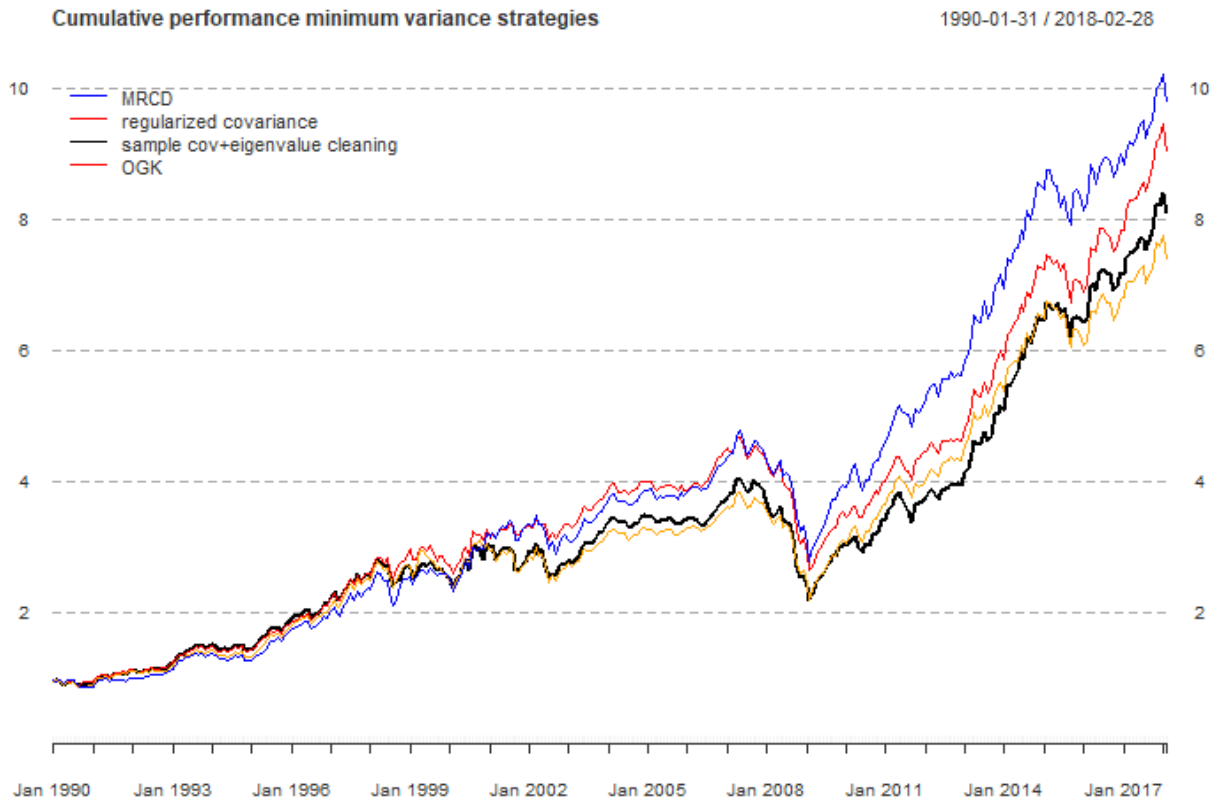


Figure: Cumulative performance charts of minimum variance portfolios.

Conclusions

- *Honey, we shrunk the data and the covariance matrix*
- We generalized the MCD approach by estimating the covariance matrix using a convex combination of a target matrix and the sample covariance matrix on the subset, chosen in order to minimize the determinant of this regularized covariance estimate.
- The resulting MRCD estimator preserves high breakdown properties of MCD and is well-conditioned, even when $p > n$.
- It only regularizes when needed.
- Broad range of applications, including finance.
- R code available at `https://wis.kuleuven.be/stat/robust/Programs/MRCD`.