Minimum Regularized 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
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 $X = (x_1, \ldots, x_n)'$ with $x_i = (x_{i1}, \ldots, x_{ip})'$, 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.

Correlation outliers

Boxplot with outlier detection: whiskers extend to 1.5 IQR

log brain weight

log body weight

Log body weight

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

**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.
Honey, I Shrunk the Sample Covariance Matrix

Olivier Ledoit and Michael Wolf

- Extremely popular in finance.
- 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 data ("Industry_10")
rets (in %): 200 rows, 100 columns
# on this data, the shrinkage intensity equals 0.21
# what happens if we replace the first observation by an outlier of
size k, with k ranging from -100 (%) to 200 (%)?

vrho <- rep(NA, 100)
vk <- seq(-100, 200, length.out = 100)
i <- 1
contrets <- rets
for (k in vk) {
    contrets[1,] <- k
    vrho[i] <- intensity(contrets)
    i <- i + 1
}
plot(vk, vrho, type = "l", xlab = "contamination value", ylim = c(0, 1), ylab = "Shrinkage intensity")
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.
Define regularized covariance estimator as convex combination of

1. predetermined, symmetric, positive definite target matrix $T$. For simplicity in notation, assume here $T = I_p$.
2. sample covariance estimate $S(H)$ based on a subset $H$

\[ K(H) = \rho \, I + (1 - \rho) c_\alpha S(H), \]

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

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

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

Starting from $h$-subset $H_1$, compute $m_1 = \frac{1}{h} \sum_{i \in H_1} x_i$ and $S_1 = \frac{1}{h} \sum_{i \in H_1} (x_i - m_1)(x_i - m_1)'$. The matrix $K_1 = \rho T + (1 - \rho) S_1$ is positive definite hence invertible, so we can compute

$$d_1(i) = (x_i - m_1)' K_1^{-1} (x_i - m_1) \quad i = 1, \ldots, n.$$  

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 $m_2 = \frac{1}{h} \sum_{i \in H_2} x_i$, $S_2 = \frac{1}{h} \sum_{i \in H_1} (x_i - m_2)(x_i - m_2)'$ and $K_2 = \rho T + (1 - \rho) S_2$. Then:

$$\det(K_2) \leq \det(K_1),$$

with equality iff $m_2 = m_1$ and $K_2 = K_1$. 
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 $m^i$ and scatter estimates $S^i$ ($i = 1, \ldots, 6$).

2. **C-step.** Determine the subsets $H^i_0$ containing the $h$ observations with lowest Mahalanobis distance in terms of $m^i$ and $S^i$.

3. **Calibrate intensity parameter.** For each subset $H^i_0$, determine the smallest value of $0 \leq \rho^i < 1$ for which $\rho^i I + (1 - \rho^i) c_\alpha S(H^i_0)$ is well-conditioned. Denote this value as $\rho^i_0$.

4. If $\max_i \rho^i_0 \leq 0.1$, set $\rho = \max_i \rho^i_0$, else set $\rho = \max\{0.1; \text{median}_i \rho^i_0\}$.

5. **Repeat C-steps till convergence.** For the initial subset $H^i_0$ for which $\rho^i_0 \leq \rho$, repeat the generalized C-steps using $\rho I + (1 - \rho) c_\alpha S(H^i_0)$ until convergence. Denote the resulting subsets as $H^i$.

6. Let $H_{MRCD}$ be the subset for which $\rho I + (1 - \rho) c_\alpha S(H^i)$ 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.
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 \( 800 \times 100, 200 \times 100, \) and \( 200 \times 400 \).

We let the fraction of contamination \( \varepsilon \) be either 0\% (clean data), 20\% or 40\% (medium sized \(( k = 50, \) along the eigenvector direction of \( \Sigma \) 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):

\[
MSE = \frac{1}{M} \frac{1}{p^2} \sum_{m=1}^{M} \sum_{k=1}^{p} \sum_{l=1}^{p} (S_m - \Sigma_m)_{k,l}^2.
\]
## Simulation Results: Clean data

<table>
<thead>
<tr>
<th></th>
<th>MSE</th>
<th>Average value of $\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$800 \times 100$</td>
<td>$200 \times 100$</td>
</tr>
<tr>
<td>$h = \lceil 0.5n \rceil$</td>
<td>0.0024</td>
<td>0.0087</td>
</tr>
<tr>
<td>$h = \lceil 0.75n \rceil$</td>
<td>0.0017</td>
<td>0.0064</td>
</tr>
<tr>
<td>$h = n$</td>
<td>0.0013</td>
<td>0.0050</td>
</tr>
<tr>
<td>OGK</td>
<td>0.0015</td>
<td>0.0060</td>
</tr>
</tbody>
</table>

- No outliers: best is $h = n$
- Regularize when needed
- Outperformance wrt OGK.
Simulation Results: 20% contamination

<table>
<thead>
<tr>
<th></th>
<th>MSE</th>
<th>Average value of $\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>800 $\times$ 100</td>
<td>200 $\times$ 100</td>
</tr>
<tr>
<td>$h = [0.5n]$</td>
<td>0.0024</td>
<td>0.0088</td>
</tr>
<tr>
<td>$h = [0.75n]$</td>
<td>0.0017</td>
<td>0.0066</td>
</tr>
<tr>
<td>$h = n$</td>
<td>17.4482</td>
<td>15.6942</td>
</tr>
<tr>
<td>OGK</td>
<td>0.0079</td>
<td>0.0187</td>
</tr>
</tbody>
</table>

- Breakdown for $h = n$.
- Otherwise robustness.
Simulation Results: 40% contamination

<table>
<thead>
<tr>
<th>h = [0.5n]</th>
<th>MSE</th>
<th>Average value of ρ</th>
</tr>
</thead>
<tbody>
<tr>
<td>800 × 100</td>
<td>0.0025</td>
<td>0.0011</td>
</tr>
<tr>
<td>200 × 100</td>
<td>0.0094</td>
<td>0.0227</td>
</tr>
<tr>
<td>200 × 400</td>
<td>0.0099</td>
<td>0.0367</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>h = [0.75n]</th>
<th>MSE</th>
<th>Average value of ρ</th>
</tr>
</thead>
<tbody>
<tr>
<td>800 × 100</td>
<td>2.8783</td>
<td>0.0367</td>
</tr>
<tr>
<td>200 × 100</td>
<td>3.462</td>
<td>0.1055</td>
</tr>
<tr>
<td>200 × 400</td>
<td>3.1857</td>
<td>0.1736</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>h = n</th>
<th>MSE</th>
<th>Average value of ρ</th>
</tr>
</thead>
<tbody>
<tr>
<td>800 × 100</td>
<td>66.8405</td>
<td>0.0367</td>
</tr>
<tr>
<td>200 × 100</td>
<td>60.5137</td>
<td>0.1055</td>
</tr>
<tr>
<td>200 × 400</td>
<td>16.4693</td>
<td>0.1736</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OGK</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.0398</td>
</tr>
</tbody>
</table>

- Breakdown for $h = n$ and $h = [0.75n]$.
- Robustness for $h = [0.5n]$. 
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:

$$d_i = \sqrt{(x_i - m_{MRCD})'K_{MRCD}^{-1}(x_i - m_{MRCD})}$$
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:

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