University of Antwerp
Faculty of Science

Cellwise outliers

From identification to regression

Jakob Raymaekers
Dept. of Mathematics, University of Antwerp, Belgium
Peter J. Rousseeuw
Dept. of Mathematics, KU Leuven, Belgium

DSSV-26 Conference, June 2026

Types of outliers

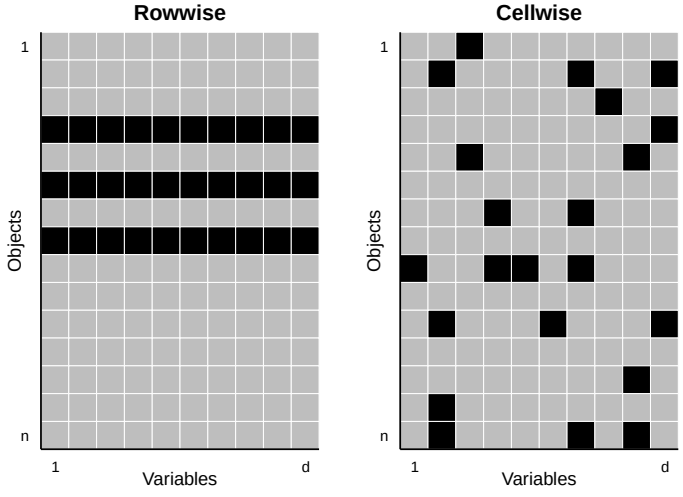
We want to analyze a data matrix $\boldsymbol{X}$ with n rows (cases) and $d > 1$ columns (variables). But the data may contain outliers.

The usual **rowwise contamination model** of Tukey (1960), Huber (1964),... assumes that some **rows** $\boldsymbol{x}_i$ have been replaced by arbitrary rows.

Rowwise robust methods downweight outlying rows and require at least 50% of clean rows.

The **cellwise contamination model** of Alqallaf, Van Aelst, Yohai and Zamar (2009 AOS) assumes that some **cells** x_{ij} have been replaced.

Downweighting/dropping an entire row loses a lot of information.


Rowwise
Cellwise
1
n
Objects
1
d
Variables

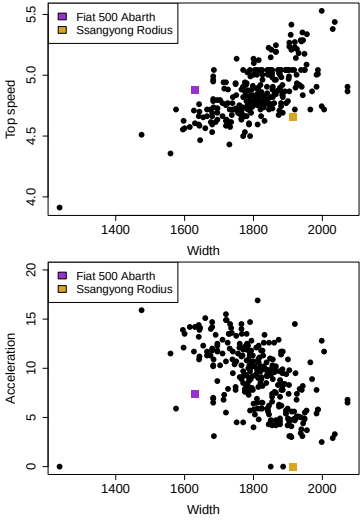
Detecting cellwise outliers

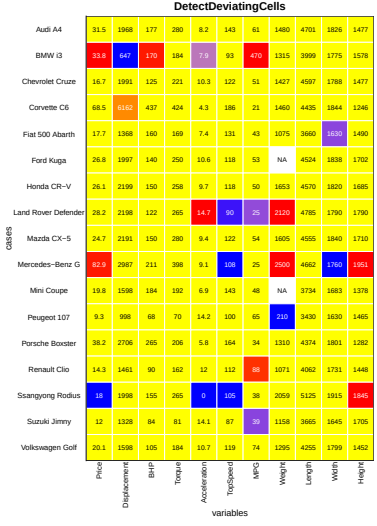
To **detect** outlying cells:

- By column only: based on the usual robust z-scores, or with the univariate Gervini-Yohai filter (2002 AOS)
- Modified Stahel-Donoho outlyingness (Van Aelst et al., 2012 CSDA)
- Bivariate filters, e.g. Leung et al (2017)
- Detecting Deviating Data Cells (Rousseeuw and Van den Bossche, Technometrics 2018).

The majority of approaches do not quantify cellwise outlyingness by the **residual from a robust fit**.

The challenge of identifying cellwise outliers





Cellwise deviation with respect to a fit

CellHandler (Raymaekers and Rousseeuw 2021, JDSSV) was our first answer to the challenge of detecting outlying cells based on estimates $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$.

The main idea is to find a sparse $\boldsymbol{\delta}$ which shifts a few cells of an observation $\boldsymbol{z}$ to obtain $\boldsymbol{z} - \boldsymbol{\delta}$ which fits better with the model suggested by $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$.

It is based on the identity

$$\text{MD}^2(\boldsymbol{z} - \boldsymbol{\delta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = ||\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{\delta}||_2^2,$$

with $\tilde{\boldsymbol{Y}} := \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{z} - \boldsymbol{\mu})$ and $\tilde{\boldsymbol{X}} := \boldsymbol{\Sigma}^{-1/2}$ with coefficient vector $\boldsymbol{\delta}$.

Finding such a sparse $\boldsymbol{\delta}$ can now be done through solving a penalized regression!

Unfortunately, CellHandler does not integrate naturally into a covariance estimator with attractive properties, but it did give us direction.

Likelihood for missing data

Rather than shifting cells, we instead **select** cells. For incomplete data, denote the missingness pattern of the $n \times d$ data matrix $\boldsymbol{X}$ by the $n \times d$ matrix $\boldsymbol{W}$ with entries w_{ij} that are 0 for missing x_{ij} and 1 otherwise.

The **observed likelihood** of row i is defined as:

$$f(\boldsymbol{x}_i^{(\boldsymbol{w}_i)}, \boldsymbol{\mu}^{(\boldsymbol{w}_i)}, \boldsymbol{\Sigma}^{(\boldsymbol{w}_i)}) = \frac{1}{(2\pi)^{d(\boldsymbol{w}_i)/2} \left|\boldsymbol{\Sigma}^{(\boldsymbol{w}_i)}\right|^{1/2}} e^{-\mathrm{MD}^2(\boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma})/2}$$

in which $\mathrm{MD}(\boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the **partial Mahalanobis distance**. Here

- $\boldsymbol{x}_i^{(\boldsymbol{w}_i)}$ is the vector with only the entries for which $w_{ij} = 1$
- $\boldsymbol{\Sigma}^{(\boldsymbol{w}_i)}$ is the submatrix of $\boldsymbol{\Sigma}$ for the variables j with $w_{ij} = 1$.
- $d(\boldsymbol{w}_i)$ is the dimension of $\boldsymbol{x}_i^{(\boldsymbol{w}_i)}$, i.e. the number of non-missing entries of $\boldsymbol{x}_i$.

Likelihood for missing data

The MLE of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for the observed likelihood thus minimizes

$$\sum_{i=1}^{n} L(\boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^{n} \left(\ln |\boldsymbol{\Sigma}^{(\boldsymbol{w}_i)}| + d(\boldsymbol{w}_i) \ln(2\pi) + \mathrm{MD}^2(\boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \right) \quad .$$

This estimate is typically computed by the EM algorithm (Dempster et al., 1977).

When constructing a cellwise MCD, the matrix $\boldsymbol{W}$ is unknown and must be estimated!

Cellwise MCD

The **cellwise MCD** (**cellMCD**) finds $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{W})$ that minimize the objective function

$$\sum_{i=1}^{n} L(\boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \sum_{j=1}^{d} q_j ||\mathbf{1}_d - \boldsymbol{W}_{\cdot j}||_0$$

under the constraints $||\boldsymbol{W}_{\cdot j}||_0 \geqslant h$ for all $j = 1, \ldots, d$ and $\lambda_d(\boldsymbol{\Sigma}) \geqslant a > 0$.

Here $||\mathbf{1}_d - \boldsymbol{W}_{\cdot j}||_0$ is the number of zero weights in column j of $\boldsymbol{W}$, i.e. the number of cells in column j of $\boldsymbol{X}$ flagged as outlying.

From covariance to linear regression

Consider a standard linear regression model

$$\boldsymbol{y} = \alpha + \boldsymbol{X}\boldsymbol{\beta} + \text{noise}$$

Now both $\boldsymbol{y}$ and $\boldsymbol{X}$ can contain cellwise outliers.
We aim to:

- rely as little as possible on assumptions of ellipticity
- handle *test* data with potential outliers
- treat in-sample and out-of-sample in a uniform way

Simply cellMCD?

It may be tempting to simply treat $\boldsymbol{Z} = [\boldsymbol{y}\ ; \boldsymbol{X}]$ as a $(d+1)$-dimensional data matrix and apply cellMCD, i.e. find $(\boldsymbol{\mu}_Z, \boldsymbol{\Sigma}_Z, \boldsymbol{W})$ that minimize:

$$\sum_{i=1}^{n} L(\boldsymbol{z}_i, \boldsymbol{w}_i, \boldsymbol{\mu}_Z, \boldsymbol{\Sigma}_Z) + \sum_{j=1}^{d+1} q_j ||\mathbf{1}_d - \boldsymbol{W}_{\cdot j}||_0,$$

and then recover $\hat{\alpha}, \hat{\beta}$ from $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}$. This is similar in spirit to the 3SGS estimator. Unfortunately:

- this strongly relies on ellipticity of $\boldsymbol{Z}$
- $\boldsymbol{w}_{ij}$ now depends on all $\boldsymbol{w}_{ik}$, $k \neq j$ hence the information in the response is used to flag outliers in the predictors. This cannot be replicated out-of-sample.

To address these shortcomings, we take two measures: **symmetrization** and a **two-step procedure**.

Symmetrization

For the **population** model $Y = \boldsymbol{\beta}^\top \boldsymbol{X} + E$, and an independent copy $(\boldsymbol{X}', E', Y')$ of $(\boldsymbol{X}, E, Y)$, we have $Y - Y' = \boldsymbol{\beta}^\top (\boldsymbol{X} - \boldsymbol{X}') + E - E'$. We thus obtain

$$\tilde{Y} = \boldsymbol{\beta}^\top \tilde{\boldsymbol{X}} + \tilde{E},$$

with $\tilde{Y} := Y - Y'$, $\tilde{\boldsymbol{X}} := \boldsymbol{X} - \boldsymbol{X}'$ and $\tilde{E} := E - E'$.
By construction, we have $\mathbb{E}[\tilde{Y}] = \mathbb{E}[\tilde{E}] = 0$, $\mathbb{E}[\tilde{\boldsymbol{X}}] = \boldsymbol{0}$ and $\mathbb{E}[\tilde{\boldsymbol{X}}\,\tilde{\boldsymbol{X}}^\top] = 2\,\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top]$.

For **finite samples**:
We symmetrize $\boldsymbol{y} = \{y_1, \ldots, y_n\}$ by $\mathrm{Sym}(\boldsymbol{y}) = \{y_\ell - y_i \,;\, i \neq \ell\}$.

We also symmetrize $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ to $\{\boldsymbol{x}_\ell - \boldsymbol{x}_i \,;\, i \neq \ell\}$.

To speed up computations, we can work with a subset of the pairwise differences.

cellLTS: A two-step procedure

In *cellLTS*, we execute 2 steps:

Step 1: Obtain a cleaned data matrix of regressors $\widehat{X}$.

Step 2: Obtain robust estimates of α and β.

cellLTS: A two-step procedure

Step 1: Obtain a cleaned data matrix of regressors $\widehat{X}$.

(a) Obtain $\widehat{\Sigma}_X$ through cellMCD with fixed center on Sym($\boldsymbol{X}$) and dividing by 2.

(b) Obtain $\widehat{\mu}_X$ by a robust univariate location estimate on each column of $\boldsymbol{X}$.

(c) Obtain $\boldsymbol{W_X}$ by iterations that optimize the cellMCD objective with respect to $\boldsymbol{W_X}$, while keeping $\widehat{\mu}_{\boldsymbol{X}}$ and $\widehat{\Sigma}_{\boldsymbol{X}}$ fixed. The starting value is based on a marginal detection rule.

(d) Obtain $\widehat{\boldsymbol{X}}$ by

$$\widehat{x}_{ij} = \begin{cases} x_{ij} & \text{if} \quad \boldsymbol{W}_{ij} = 1 \\ \widehat{\mu}_j + \widehat{\Sigma}_{j,(\boldsymbol{w}_{i\cdot})} \left(\widehat{\Sigma}^{(\boldsymbol{w}_{i\cdot})} \right)^{-1} (\boldsymbol{x}_i^{(\boldsymbol{w}_{i\cdot})} - \widehat{\boldsymbol{\mu}}^{(\boldsymbol{w}_{i\cdot})}) & \text{if} \quad \boldsymbol{W}_{ij} = 0 \, . \end{cases}$$

Step 2: Obtain robust estimates of α and β.

cellLTS: A two-step procedure

Step 1: Obtain a cleaned data matrix of regressors $\widehat{X}$.

Step 2: Obtain robust estimates of α and β.

(e) Symmetrize $Z = [y\ ; \widehat{X}]$ and standardize the result with the scales $(\hat{\sigma}_y, \sqrt{\text{diag}(\widehat{\Sigma}_X)})$ to obtain $\tilde{Z} = [\tilde{y}\ ; \tilde{X}]$.

(f) Both $\tilde{X}$ and $\tilde{y}$ have $\tilde{n} := n(n-1)$ rows, and we will consider $\tilde{h}$-subsets $\tilde{H}$ of $\{1, \ldots, \tilde{n}\}$ with $\tilde{h} = h(h-1)$ elements. The cellLTS estimator of β is now

$$\widehat{\beta} = \underset{\beta(\tilde{H})}{\text{argmin}} \Big(\sum_{\ell \in \tilde{H}} r^2(\tilde{y}_\ell, \tilde{x}_\ell, \beta) + \lambda ||\beta||_2^2 \Big)$$

with residuals $r(\tilde{y}_\ell, \tilde{x}_\ell, \beta) := \tilde{y}_\ell - \tilde{x}_\ell^\top \beta$.

(g) Estimate α by the univariate MCD of the pseudo residuals $r^* = y - \widehat{X}\,\widehat{\beta}$.

(h) Transform the regression coefficients to undo the standardization in step (e).

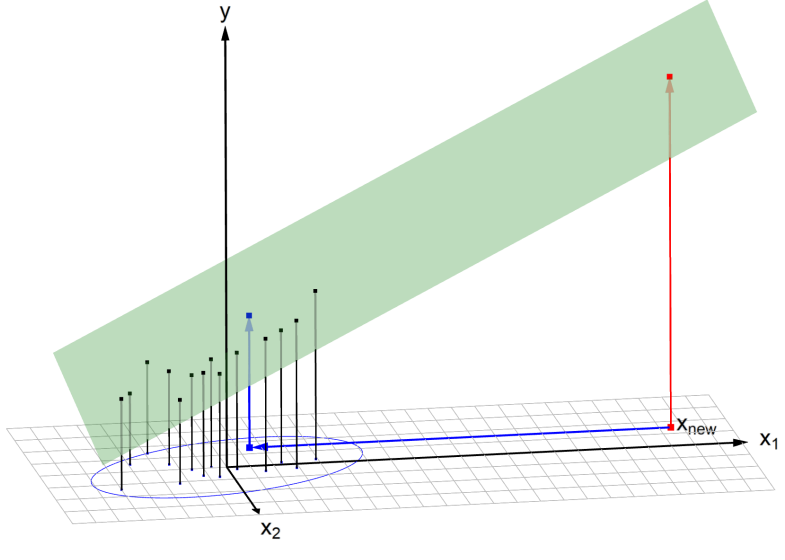
Out-of-sample predictions

We want to obtain the prediction at a new $\boldsymbol{x}$, which may itself contain missing values and/or outlying cells! Therefore we cannot just use $\hat{\alpha} + \boldsymbol{x}^\top \widehat{\boldsymbol{\beta}}$. Instead:

1. Obtain an initial $\boldsymbol{w}^{(0)}$ through the same marginal detection rule as in (c).
2. Obtain $\boldsymbol{w}$ by cellmcd iterations with fixed $\widehat{\boldsymbol{\mu}}_{\boldsymbol{X}}$ and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X}}$ as in (c)
3. Obtain $\hat{\boldsymbol{x}}$ as in (d)
4. Calculate $\widehat{y} = \hat{\alpha} + \hat{\boldsymbol{x}}^\top \widehat{\boldsymbol{\beta}}$

Note: the imputation in a new point $\boldsymbol{x}$ starts from the same marginal detection rule that we used for $\boldsymbol{W}_{\boldsymbol{X}}$, we treat in-sample prediction and out-of-sample prediction consistently.

Out-of-sample predictions



Breakdown value

Let $\boldsymbol{Z} = [\boldsymbol{y}\,; \boldsymbol{X}]$ as before and $\boldsymbol{\gamma} = (\alpha, \beta_1, \ldots, \beta_d)^\top$ the combined parameters. Denote with $\boldsymbol{Z}^{\text{con}}$ a contaminated matrix obtained by replacing at most m cells in each column of $\boldsymbol{Z}$ by arbitrary values.

The *cellwise finite-sample breakdown value* of the regression estimator $\widehat{\boldsymbol{\gamma}}$ at the dataset $\boldsymbol{Z}$ as

$$\varepsilon_n^*(\widehat{\boldsymbol{\gamma}}, \boldsymbol{Z}) := \min\left\{ \frac{m}{n} : \sup_{\boldsymbol{Z}^{\text{con}}} ||\widehat{\boldsymbol{\gamma}}(\boldsymbol{Z}^{\text{con}}) - \widehat{\boldsymbol{\gamma}}(\boldsymbol{Z})|| = \infty \right\}$$

If the dataset $\boldsymbol{X}$ is in general position and and $h \geqslant [(n + d + 1)/2]$, then the breakdown value of cellLTS with $\tilde{h} = h(h - 1)$ is

$$\varepsilon_n^*(\widehat{\boldsymbol{\gamma}}, \boldsymbol{Z}) = (n - h + 1)/n.$$

Example: US Cancer dataset

Data on cancer rates in the period 2010-2016 in US counties, compiled from the U.S. Census Bureau and the National Cancer Institute.
The dataset contains 3047 rows, corresponding to the counties, and 33 columns.
The response variable is the death rate due to cancer (in cases per 100 000).
For illustrative purposes, we build a regression model using five covariates:

feature	explanation
`incidencerate`	Incidence rate of cancer (per 100 000).
`medincome`	Median income in the county (in thousands of dollars).
`medianage`	Median age in the county.
`pcths18_24`	Percentage of residents aged 18-24 whose highest attained education is a high school diploma.
`pctemployed16_over`	Percentage of people aged 16 and over who are employed.

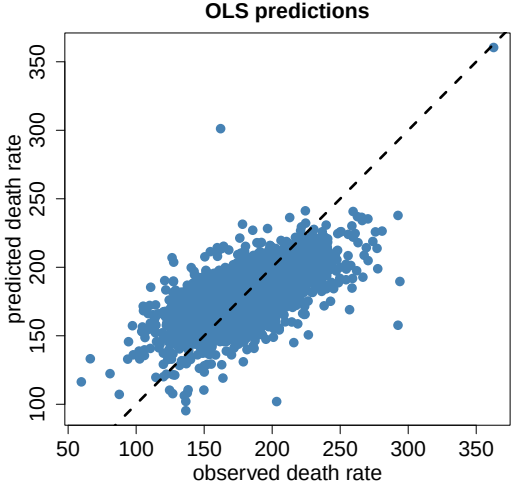
OLS vs. cellLTS on US Cancer data

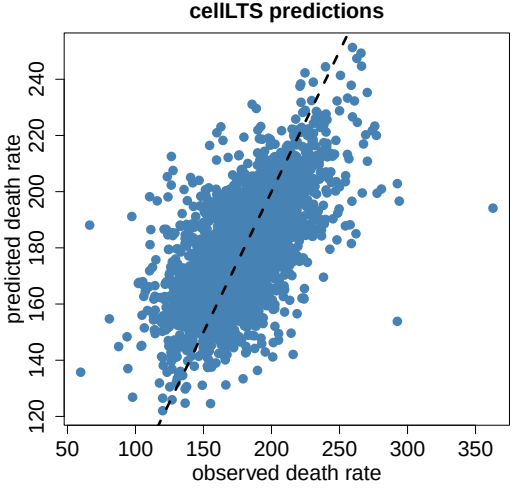
	OLS	cellLTS	OLS of cleaned
intercept	122.50	157.16	147.27
incidencerate	0.23	0.24	0.23
medincome	-0.66	-0.75	-0.65
medianage	-0.01	-0.73	-0.61
pcths18_24	0.47	0.62	0.55
pctemployed16_over	-0.56	-0.83	-0.62

The cleaned OLS fit is obtained by removing the gross errors in the medianage variable (values larger than 400).

OLS vs. cellLTS on US Cancer data

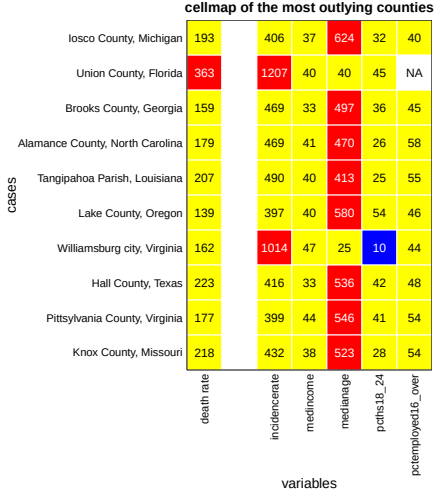
Predicted vs. observed death rate for OLS (left) and cellLTS (right).





Cellmap of US Cancer data

Cellmap of the 10 counties with the highest sum of absolute standardized cellwise residuals:



Conclusions

Implementation:
The software is in the R package *cellWise* on CRAN.

References:
R&R (2026). Least trimmed squares regression with missing values and cellwise outliers. *arXiv preprint* arXiv:2603.04632.

R&R (2026). Challenges of cellwise outliers. *Econometrics and Statistics*, 38, 6-25.

R&R (2024). The cellwise minimum covariance determinant estimator. *Journal of the American Statistical Association* 119.548, 2610-2621.

R&R (2021). Handling Cellwise Outliers by Sparse Regression and Robust Covariance. *Journal of Data Science, Statistics, and Visualisation*, 1.3.

Thank you!