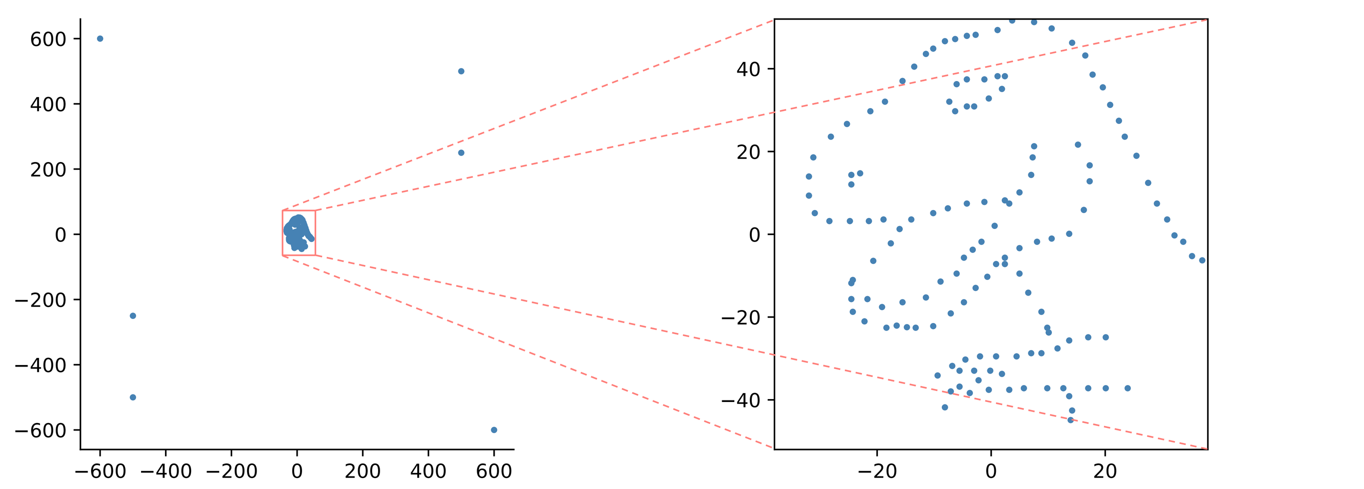
CLIPPR: Maximally Informative CLIPped Projection with Bounding Regions

Bo Kang[1], Dylan Cashman[2], Remco Chang[2], Jefrey Lijffijt[1], Tijl De Bie[1]

Ghent University[1], Tufts University[2]

Motivation

- Plot with large scale lacks small-scale details (limited resolution)
- Zooming-in for details loses further away points
- *Example: plotting the full 2D data (left) misses detailed structure*

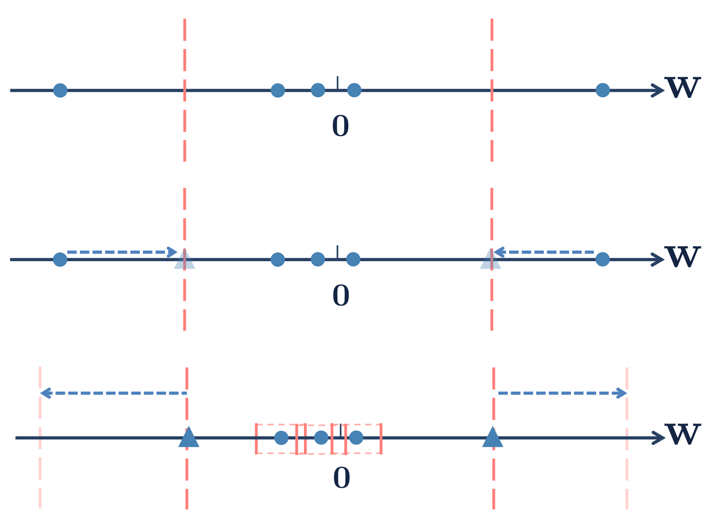


http://www.thefunctionalart.com/2016/08/download-datasaurus-never-trust-summary.html

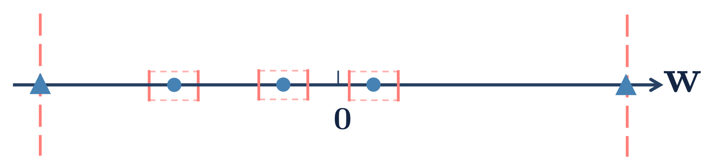
- Can we balance scale and detail automatically?

Idea of method

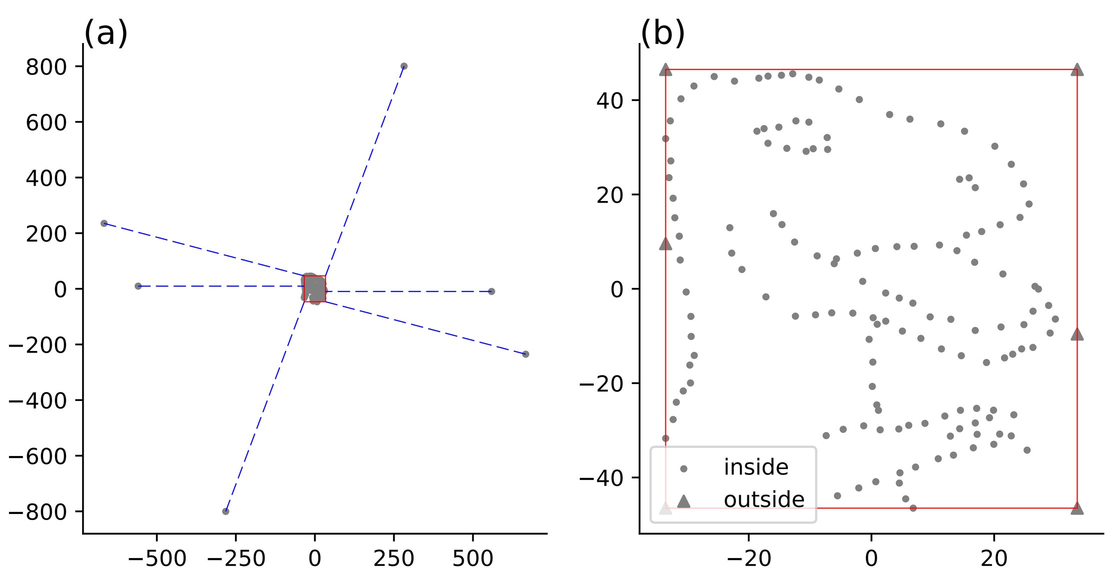
1) Overlay bounding box on scatter plot
2) Clip points outside to border and present them with a different marker
3) Zoom-in to fill plotting area



- For points inside, we learn their position up to the **resolution**
- For points on the border, we learn their direction

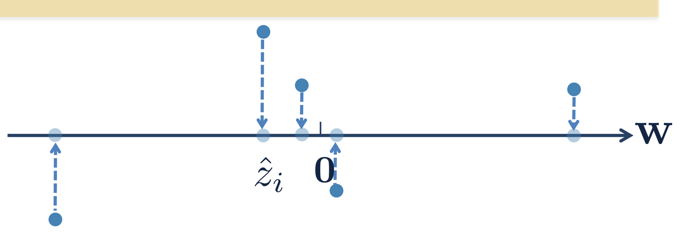


- We can quantify the **information content** of this visualization
- And hence, optimize the **trade-off** between scale and detail
- Example: actual result on synthetic data

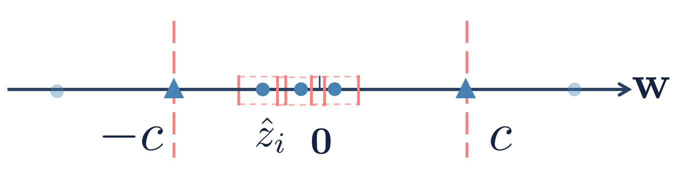


Clipped projection

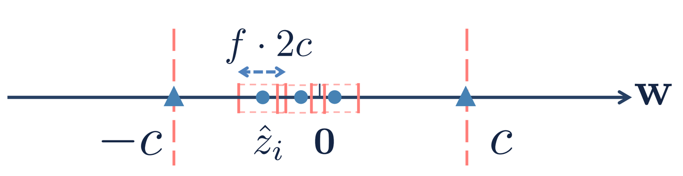
- Denote the 1D projection of data $\hat{\mathbf{x}}_i \in \mathbb{R}^d$ onto $\mathbf{w} \in \mathbb{R}^d\,(\mathbf{w}'\mathbf{w}=1)$ point as $\hat{z}_i = \hat{\mathbf{x}}_i'\mathbf{w}$



- A **bounding box** is a (centered) window $(-c, c)$, with $c \in \mathbb{R}_+$

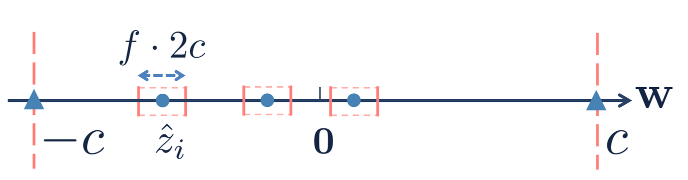


- Idea: For a **resolution parameter** f, projection $\hat{z}_i$ is specified up to a **pixel** of size $f \cdot 2c$



- A **clipped projection** is defined as

$$\hat{\mathbf{x}}_i'\mathbf{w} \in [l(\hat{z}_i, c), u(\hat{z}_i, c)]$$

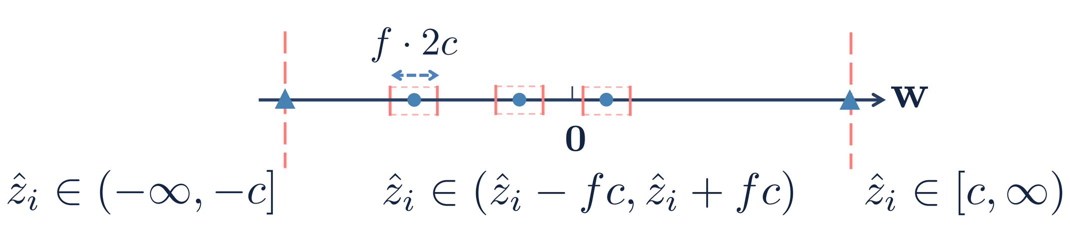


- Clipped point:

$$\hat{z}_i \in (-\infty, -c] \text{ or } \hat{z}_i \in [c, \infty)$$

- Unclipped point:

$$\hat{z}_i \in (\hat{z}_i - fc, \hat{z}_i + fc)$$

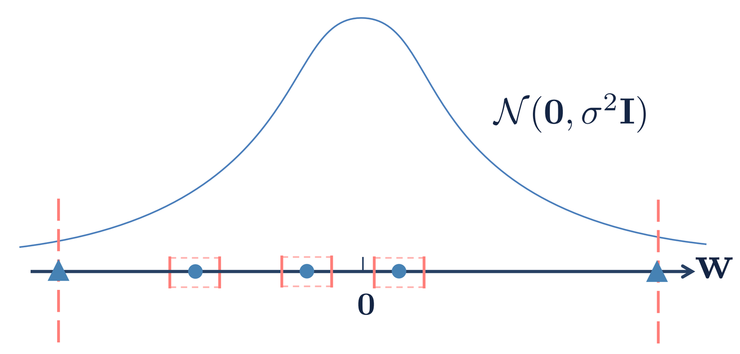


- For **projection** of data $\hat{\mathbf{X}} \in \mathbb{R}^{n\times d}$ onto $\mathbf{W} \in \mathbb{R}^{d\times k}\,(\mathbf{W}'\mathbf{W}=\mathbf{I})$ as $\hat{\mathbf{\Pi}}_{\mathbf{W}} \triangleq \hat{\mathbf{X}}\mathbf{W}$, we have multi-dimensional clipped projection:

$$\hat{\mathbf{X}}\mathbf{W} \in \left[\mathbf{L}\left(\hat{\mathbf{\Pi}}_{\mathbf{W}}, \mathbf{c}\right), \mathbf{U}\left(\hat{\mathbf{\Pi}}_{\mathbf{W}}, \mathbf{c}\right)\right]$$

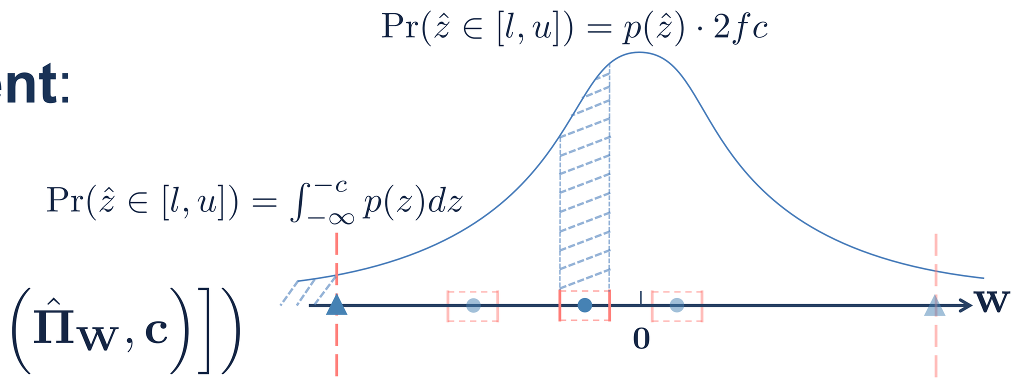
Find informative visualization

- Specify a background model to be $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, with $\sigma^2 = \text{Tr}(\hat{\mathbf{X}}'\hat{\mathbf{X}})/nd$



- Quantify **information content**:

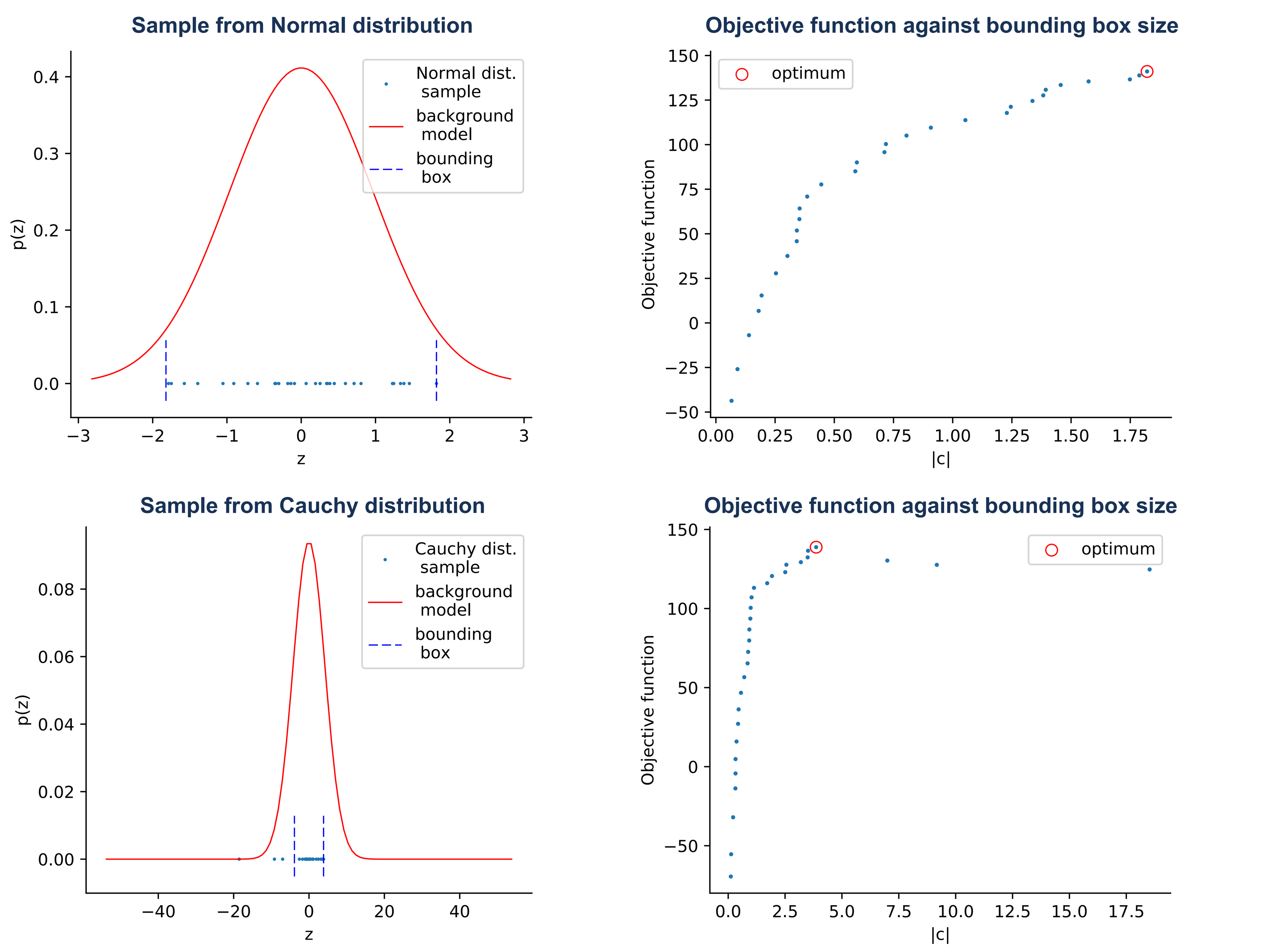
$$\text{IC}(\mathbf{W}, \hat{\mathbf{\Pi}}_{\mathbf{W}}, \mathbf{c}) = -\log \Pr\left(\hat{\mathbf{\Pi}}_{\mathbf{W}} \in \left[\mathbf{L}\left(\hat{\mathbf{\Pi}}_{\mathbf{W}}, \mathbf{c}\right), \mathbf{U}\left(\hat{\mathbf{\Pi}}_{\mathbf{W}}, \mathbf{c}\right)\right]\right)$$



- Maximize the information content over $\mathbf{W}$ and $\mathbf{c}$:

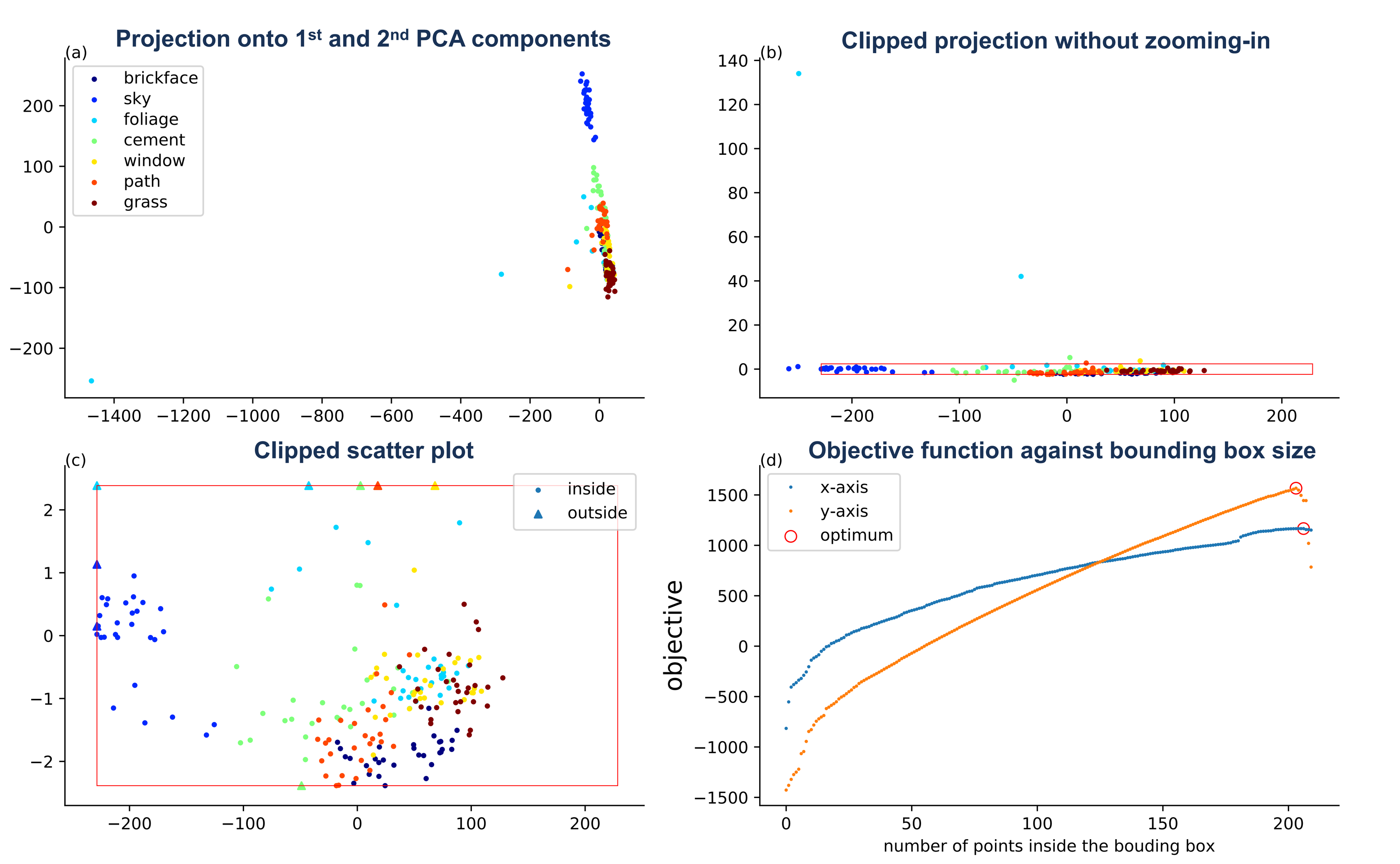
$$\underset{\mathbf{W},\mathbf{c}}{\text{argmax}} \quad \text{IC}(\mathbf{W}, \hat{\mathbf{\Pi}}_{\mathbf{W}}, \mathbf{c})$$
$$\text{s.t.} \quad \mathbf{W}'\mathbf{W} = \mathbf{I}$$
$$\mathbf{c} > \mathbf{0}.$$

- Example: optimize $\mathbf{c}$ for 1 dimensional data sampled from Normal distribution $\mathcal{N}(0,1)$ and Cauchy distribution $f(0,1)$



Case study: UCI segmentation dataset

- **Dataset:** $\hat{\mathbf{X}} \in \mathbb{R}^{210\times 19}$, 210 image patches ($3\times 3$ pixels) drawn randomly from a database of 7 outdoor images. Data points are described by 19 image features and are categorized into seven classes.
- **Results:** the principal components are dominated by a single outlier, while the **clipped scatter plot** shows variation in the center of the data.



This work was supported by the ERC under FP7/2007-2013 (Grant Agreement no. 615517), the EU's H2020 R & I programme and the FWO (MSC Grant Agreement no. 665501), the National Science Foundation (NSF CAREER IIS-1452977) and the NSF-ERC program.