A GRAPHICAL TOOL FOR THE DETECTION OF MODES IN CONTINUOUS DATA

Thomas Burger & Thierry Dhorne (Lab-STICC)
OUTLINES

1. Visual representations/mode estimation of small size continuous-valued datasets

2. Density estimation and time-frequency analysis

3. A graphical tool for continuous data representation

4. Conclusion
OUTLINES

1. Visual representations/mode estimation of small size continuous-valued datasets

2. Density estimation and time-frequency analysis

3. A graphical tool for continuous data representation

4. Conclusion
MODE ESTIMATION

- The mode is one of the most explicit information about a dataset.

- In [Bi03], a method is proposed to find the mode of mono-modal continuous datasets.

- No extension to this work to our knowledge.

- How to determine the number of modes?

Here, we propose a graphical tool that helps in the visualization of the distribution of a continuous dataset.

VISUAL ANALYSIS OF CONTINUOUS DATASETS

Visualization provides a good mean to determine the number of modes. Moreover, it helps in the crucial steps of understanding the dataset.

Figure 1: There is no problem to visualize the distribution when the population is important enough (constant width/surface histograms, density estimation, etc.), but when the samples are not numerous enough, it is more complicated...
OUTLINES

1. Visual representations/mode estimation of small size continuous-valued datasets

2. **Density estimation and time-frequency analysis**

3. A graphical tool for continuous data representation

4. Conclusion
DENSITY ESTIMATION BY KERNEL METHOD

- Convolution of the dataset and a dedicated kernel

- Implemented in the \texttt{R} function \texttt{density()}

- Choice of the “shape” of the kernel? (gaussian, epanechnikov, triangular, cosine, etc.)

- Choice of the kernel size, depending on the density of the dataset (interval between items).

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{smoothing_effect.png}
\caption{The smoothing property of convolution is used to estimate the density.}
\end{figure}
CONVOLUTION IN SIGNAL PROCESSING

Convolutions are widely used in signal processing:

- To identify a pattern (kernel = pattern to find)
- To smooth/filter a signal
- etc.

In general, it is the basis for time-frequency analysis:

- Convolution in the time domain corresponds to product in Fourier domain
- Fourier analysis applied to sliding windows leads to temporal analysis
- Wavelet theory is based on convolution (sliding windows) analysis at various scales (various kernel sizes)

Figure 3: Sliding window Fourier representation.
Pattern Recognition and Shape Description

- Similar problem in Computer Vision: time-frequency analysis to describe the parametric curve of shape.

- CSS (Curvature Scale Space) descriptors [Mok92] are amongst the most efficient shape descriptors (MPEG7).

- CSS descriptors are based on the multi-scale convolution of a parametric curve with a gaussian kernel.

Figure 4: [Mok92] The CSS captures the global distribution of a shape at various scales.

APPLICATION TO STATISTICS

- Performing a multi-scale description of the dataset.
- The dataset is considered as a shape to describe (i.e. as a histogram).
- Kernel: Gaussian (as with the CSS descriptors).
- This idea has already been presented [Gri**] in 2005 in PAMI (the same journal as for [Mok92]).
- The point was to apply the mean shift algorithm at various scales to find the mode of the distribution.
- Practically, it corresponds to traverse the plots of the multiscale representation to find a maximum value.
- It remains unpublished...

OUTLINES

1. Visual representations/mode estimation of small size continuous-valued datasets

2. Density estimation and time-frequency analysis

3. A graphical tool for continuous data representation

4. Conclusion
APPLICATION TO VISUALIZATION
DETAILS OF THE CODE

Basically, the algorithm loops on the `density()` function with various sizes of kernel:

...  

# MatConv = matrix of the graphical representation
# It is constructed line by line

    for (ibw in (1):(length(axeOrd))) {
        mode <- density(data , bw=axeOrd[ibw],
                        kernel = "gaussian",
                        n=length(axeAbs),
                        from=newMinData, to=newMaxData);
        valueLine <- mode$y/max(mode$y);           # the values are normalized
        maxLine <- localMode(valueLine );          # Local max
        MatConv[ibw,] <- valueLine + maxLine ;     # artifact for representation
    }

    # display

    ...
PARAMETERS

**data:** Vector of the mono-valued dataset.

**percentmargin:** Size of the margin, so that the extremal value are not stuck to the border of the image.

**sizeKerMin:** Minimal value for the size of the kernel.

**sizeKerMax:** Maximal value for the size of the kernel.

**bwLen:** Number of convolutions with a different kernel. It corresponds to the number of lines in the display.

**ImWidth:** Width of the display.

**jitterOrHist:** Flag indicating the representation of the data in the lower part of the graphical representation. - 0 : automatic 1 : jittered density diagram 2 : histogram.
PERFORMANCE

- Execution time: between 5 and 10 seconds for a reasonable number of iterations of the `density()` function.

- The code is rather light.

- Most of the resources are necessary for the display.

- It is possible to run it even on large datasets (several hundreds of items) and on which classical visualization tools are efficient.

- The limits come from the size of the screen which limits the resolution of the display rather than the size of the dataset.
OUTLINES

1. Visual representations/mode estimation of small size continuous-valued datasets

2. Density estimation and time-frequency analysis

3. A graphical tool for continuous data representation

4. Conclusion
IN A NUTSHELL...

Efficient visualization tool:

- for small sample continuous datasets
- adaptable thanks to several parameters
- computationally acceptable

Based on:

- Multiscale gaussian convolutions
- Classical shape description methods
- Previous work has attempted to adapt this computer science background to statistics
OUTLOOK

- Dendrogram-like plot

- Interests for classification

- Future work will be focused on extracting knowledge from this “dendrogram”
QUESTION SESSION

- Thank you for your attention.
- Do you have any question?