Curves and Modality

Arne Kovac

Given noisy bivariate observations $(x_i, y_i), i = 1, \ldots, n$ at n different time points $t_1, \ldots, t_n$ we consider the problem of specifying a smooth curve $f = (f^X, f^Y)$ such that f approximates the data and is simple in the sense that the number of local extreme values in the curvature function is as small as possible. In Figure 1 the top left panel shows a spiral with added bivariate Gaussian noise and the right panel a reconstruction obtained from a kernel estimator. The curve is smooth, but does not approximate the data very well as can be seen in the bottom panel where the residuals in x- and y-direction are plotted.

We adopt a bivariate version of the multiresolution criterion by Davies and Kovac (2001) and require the sums of the residuals in x- and y-direction over stretches of different sizes and locations all to be smaller than what would be expected from white noise. More specifically we require an approximation to satisfy simultaneously

$$\left|\sum_{i\in I}(y_i - f_i^y)\right| < w_I \cdot \sigma, \quad \left|\sum_{i\in I}(x_i - f_i^x)\right| < w_I \cdot \sigma$$

with $w_I = \sqrt{|I| \cdot 2\log(2n)}$ for all intervals I of some family $\mathcal{I}$ of subintervals of $\{1, \ldots, n\}$.

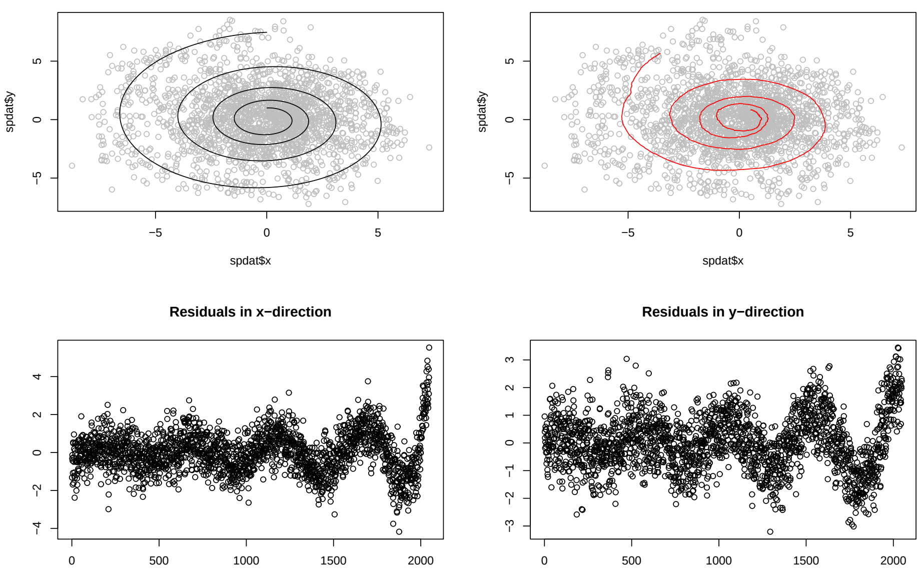

Figure 1. Noisy spiral and kernel estimator. Top left: Original spiral, Top right: Kernel estimator, Bottom left and right: Residuals in x- and y-direction

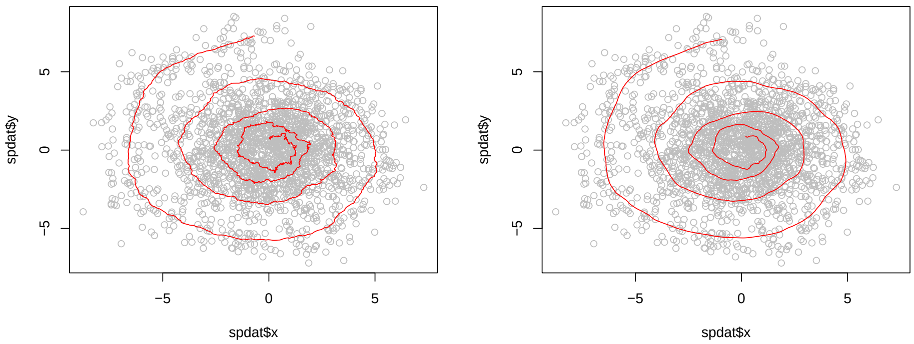

FIGURE 2. Noisy spiral and two approximations that just satisfy the multiresolution criterion. Left: Kernel estimator, Right: Total variation penalty

One choice for $\mathcal{I}$ is to take all possible subintervals

$$\mathcal{I}_1 = \{\ \{j, j+1, \ldots, k\} \text{ for all } 1 \leq j \leq k \leq n\}.$$

Computational complexity can be reduced by considering a smaller collection like all intervals with dyadic end points

$$\mathcal{I}_2 = \{\ \{2^j k + 1, \ldots, 2^j(k+1)\} \text{ for all } 0 \leq j \leq \lfloor \log_2(n) \rfloor, k = 0, 1, \ldots, \lceil \frac{n}{2^j} \rceil\}.$$

This collection has been used for the examples below. The multiresolution criterion requires the true value of σ. This may be estimated from the data by putting

$$\sigma = \frac{1.4826}{\sqrt{2}} \text{median} \left(|y_2 - y_1|, |x_2 - x_1|, \ldots, |x_n - x_{n-1}| \right)$$

(Davies and Kovac, 2001; Donoho et al, 1995)

We aim to find a curve f that satisfies this multiresolution criterion and is at the same time as simple as possible. Figure 2 shows in its left panel another approximation from an kernel estimator, but this time choosing the largest bandwidth such that the kernel estimate satisfies the multiresolution criterion above. Although this estimate approximates the data much better it contains a large number of spurious local extreme values.

In the univariate setting of non-parametric regression regularisation techniques based on total variation like the taut string method (Mammen and van de Geer, 1997; Davies and Kovac, 2001) and its generalisations (Dümbgen and Kovac, 2005) or quantile regression using total variation penalties (Koenker et al, 1994) have been shown to produce simple estimates.

We consider a two-dimensional total variation penalty and consider minimising the functional

$$T(f) = \sum_{i=1}^{n} (x_i - f_i^X)^2 + \sum_{i=1}^{n} (y_i - f_i^Y)^2 + \sum_{i=1}^{n-1} \lambda_i \sqrt{(f_{i+1}^X - f_i^X)^2 + (f_{i+1}^Y - f_i^Y)^2}$$

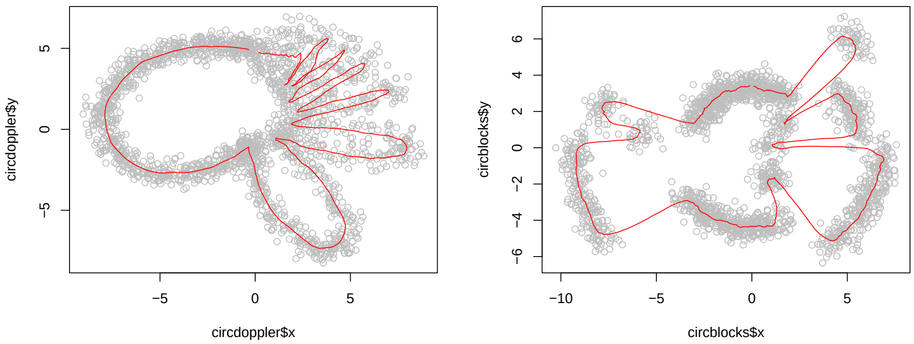

FIGURE 3. Circular noisy versions of Donoho and Johnstone's Doppler and Bumps signals and approximations using total variation penalties and the multiresolution criterion.

The smooth taut string functional by Kovac (2006) can be regarded as a special case of this functional in one dimension.

In order to obtain approximations that are as smooth and simple as possible we start with a large global penalty $\lambda_1 = \cdots = \lambda_{n-1}$ and successively reduce λ on intervals where the multiresolution criterion is not yet satisfied. This local squeezing technique has been described by Davies and Kovac (2001) and Dümbgen and Kovac (2005) in more detail. The application of this technique to the spiral data can be seen in the right panel of Figure 2. The approximation is much smoother than the kernel estimate.

Finally, Figure 3 shows approximations obtained from circular versions of the well known Doppler and Blocks functions by Donoho and Johnstone (1994). These were generated as $x_j = \cos(2\pi j/n)r_j$ and $y_j = \sin(2\pi j/n)r_j$ where $r_j = f(j/n) - \min_i(f(i/n), i = 1, \ldots, n) + 1$ and where f was successively the Doppler and the Blocks signal. The bivariate total variation penalties generate sharp discontinuities for the Blocks signal while the functions look smooth and simple elsewhere and approximate the data very well.

REFERENCES

[1] P. L. Davies and A. Kovac, *Local extremes, runs, strings and multiresolution (with discussion)*, Annals of Statistics 29 (2001), 1–65.

[2] D. L. Donoho, I. .M. Johnstone, G. Kerkyacharian and D. Picard, *Wavelet shrinkage: asymptopia?*, Journal of the Royal Statistical Society, Ser. B **57** (1995), 371–394.

[3] D. L. Donoho and I. .M. Johnstone, *Ideal spatial adaption by wavelet shrinkage*, Biometrika **81** (1994), 425–455.

[4] L. Dümbgen and A. Kovac, *Extensions of Smoothing via Taut Strings*, Technical report.

[5] R. Koenker, P. Ng and S. Portnoy, *Quantile Smoothing Splines*, Biometrika 81 (1994), 673–680.

[6] A. Kovac, *Smooth functions and local extreme values*, Computational Statistics and Data Analysis (to appear).

[7] E. Mammen and S. van de Geer, *Locally adaptive regression splines*, Annals of Statistics 25 (1997), 387–413.