Proof of Concept
Identifying patches with different statistical properties
Bill Menke, Oct 31, 2023

The problem is to identify different patches in a dataset that have distinct statistical properties and to finding the boundary between them. I consider here a one-dimensional problem with two patches divided by a point.

Suppose a dataset $d(x)$ has left and right hand parts, $d_L(x)$ and $d_R(x)$, respectively, with each drawn from different pdfs, described by autocorrelation functions $C_L(r)$ and $C_R(r)$, respectively, with $r = |x_1 - x_2|$ and no correlation between left and right parts, and where the point dividing the parts is x_0. The problem is to estimate x_0.

Method. I use a grid search over x_0 to determine the x_0^{est} that minimizes the generalized error $E(x_0)$, where the both the prediction error and the error in prior information contribute to E. The autocorrelation functions $C_L(r)$ and $C_R(r)$ are assumed to be known. The error in prior information is computed using an overall variance, ½ $\left(C_L^2(0) + C_R^2(0)\right)$ that does not vary between regions.

Gaussian processes regression is used to estimate $d^{pre}(x, x_0)$ using an overall covariance matrix

$$\begin{bmatrix} \mathbf{C}_L & 0 \\ 0 & \mathbf{C}_R \end{bmatrix} \quad \text{where} \quad \begin{matrix} [\mathbf{C}_L]_{ij} = C_L(r_{ij}) \ \text{ for } i,j < x_0 \\ [\mathbf{C}_R]_{ij} = C_L(r_{ij}) \ \text{ for } i,j \geq x_0 \end{matrix} \ \text{and} \ \ r = |x_i - x_j|$$

Although I hold C_L and C_R fixed in this work, I imagine that it would be possible to view them as functions of hyper-parameters and then augment the grid-search to search over their possible values.

In multidimensional cases, I imagine that it would be possible to parameterize the boundary as a curve with a curve whose shape is controlled by a few parameters, and then grid search over the parameter.

1.

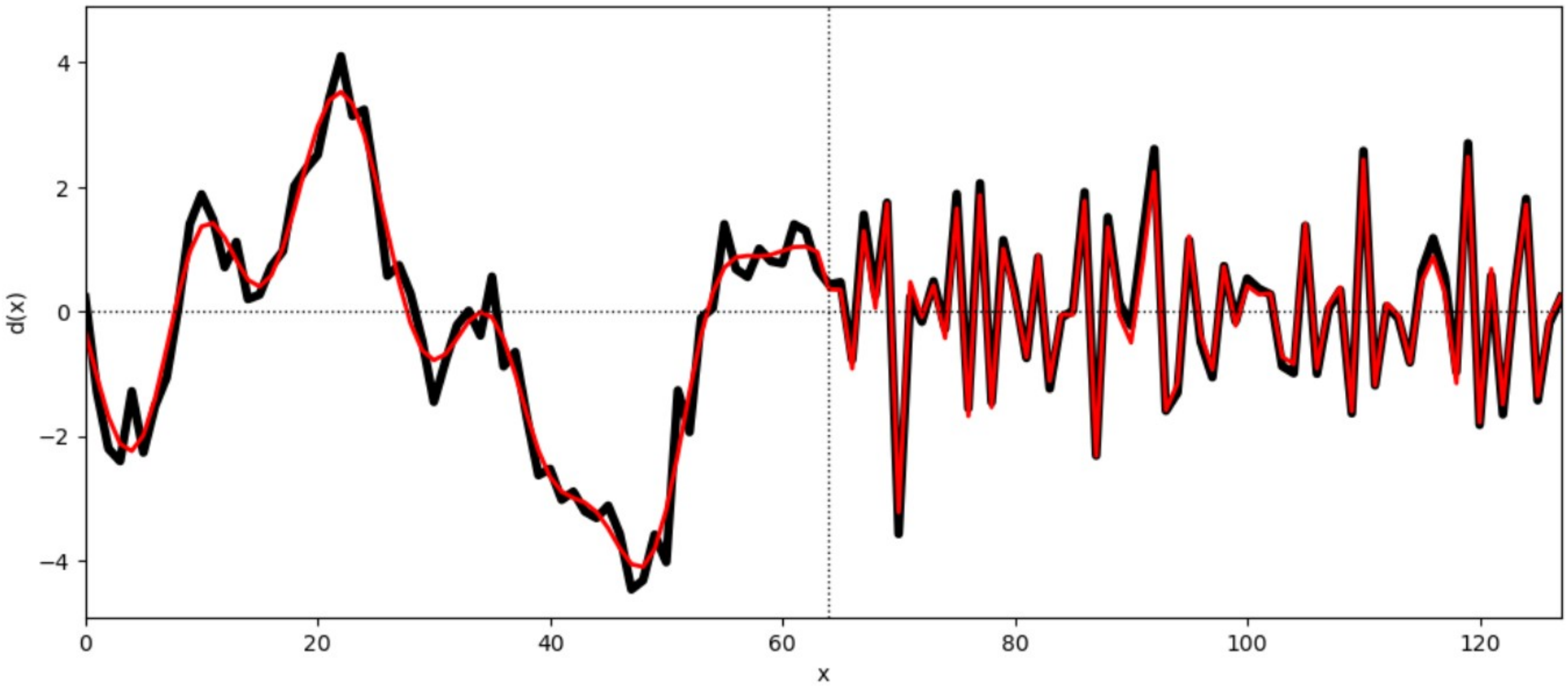
Experiment 1.

True data, $d^{true}(x)$ (black). Observed data, $d^{obs}(x)$ (red)

Left hand part drawn from data with autocorrelation $C_L(r)$ and right and from $C_R(r)$,

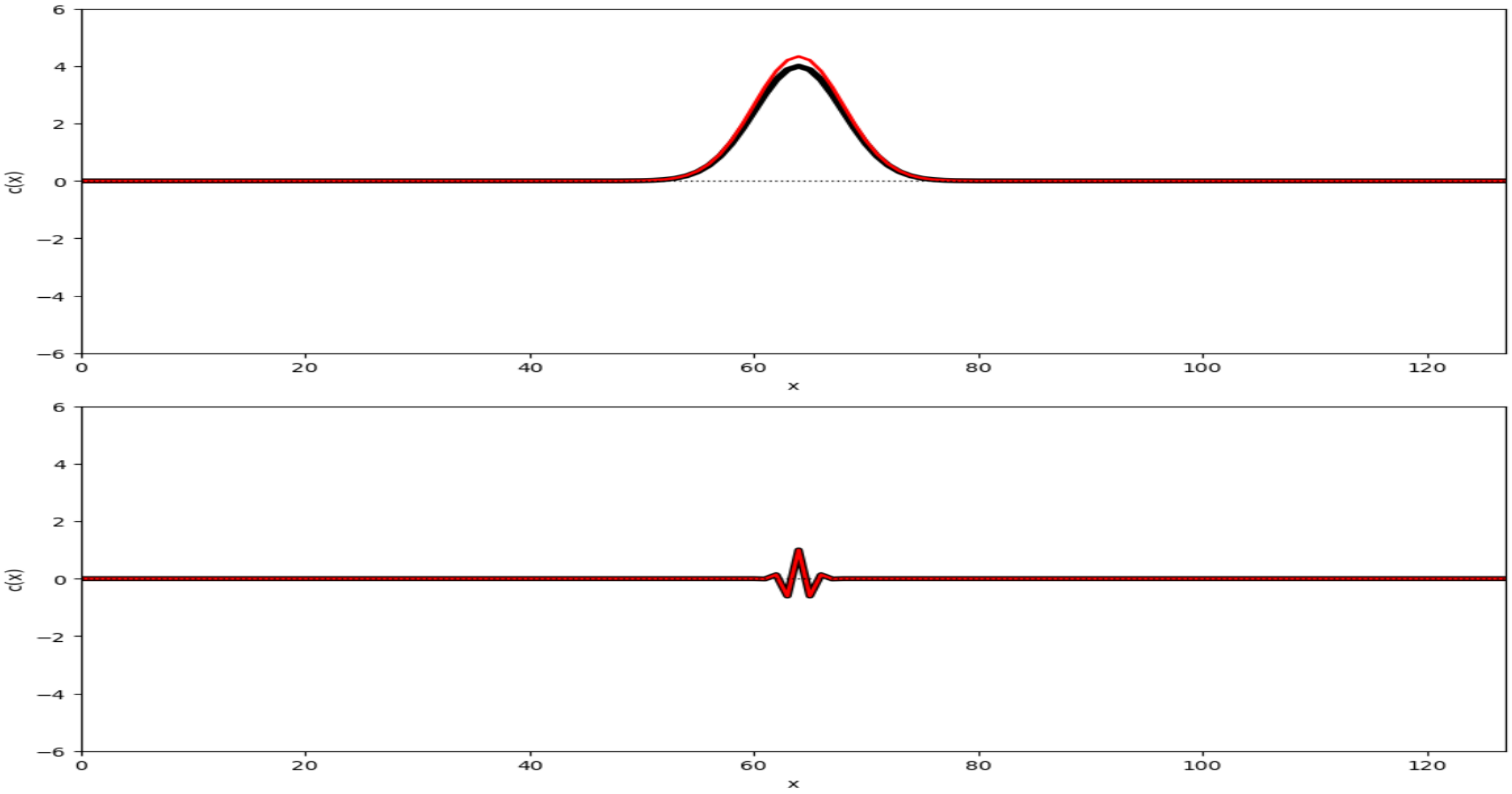
(with $r = |x_1 - x_2|$) and no correlation between left and right parts

Dividing point x_0 shown by dotted vertical line.

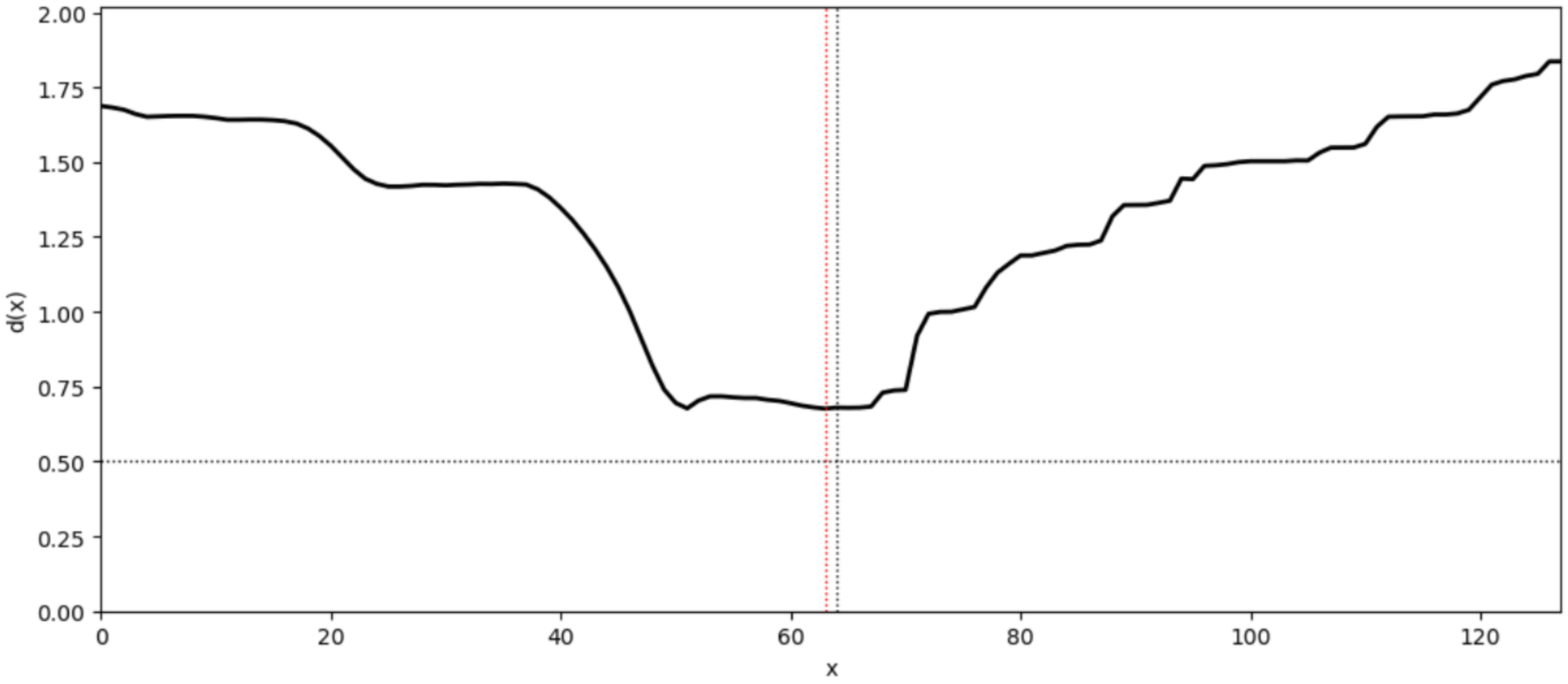


2.

True (black) and estimated (red) $C_L(r)$ (top) and true (black) and estimated (red) $C_R(r)$ (bottom). Estimates are from time series drawn from pdfs with these autocorrelation functions.

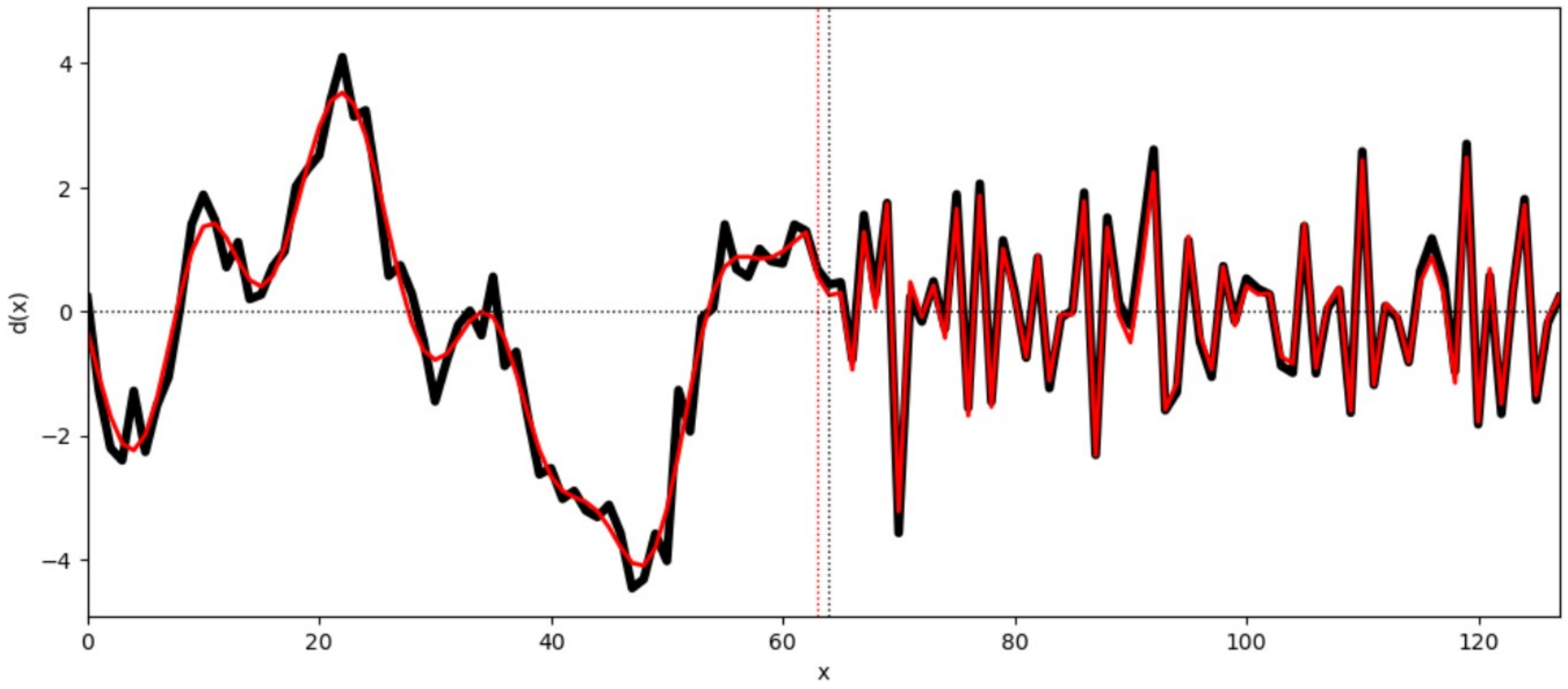


3. Generalized error for all possible x_0s, with minimum (red dotted vertical) line and true x_0 (black dotted vertical line)



4.

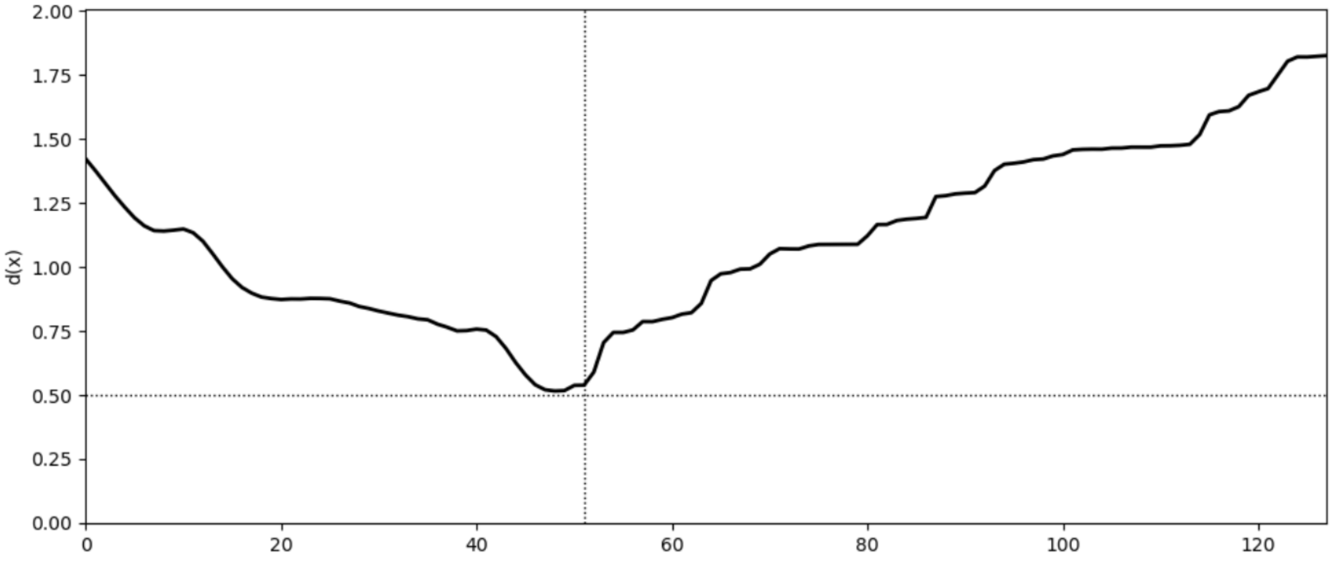
Observed data, $d^{obs}(x)$ (black) and true x_0 (black dotted vertical line). Predicted data, $d^{pre}(x)$ (red) and estimated x_0 (black dotted vertical line).



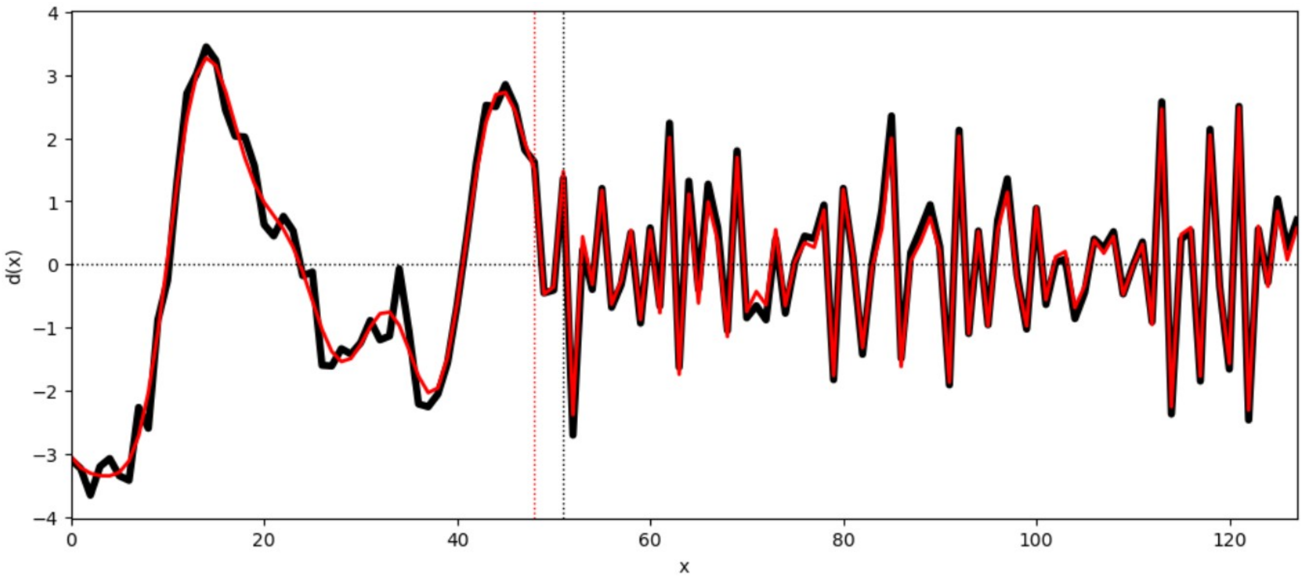
Experiments with a different x_0

5.
Experiment 2

(Top) Generalized error for all possible choices of x_0.



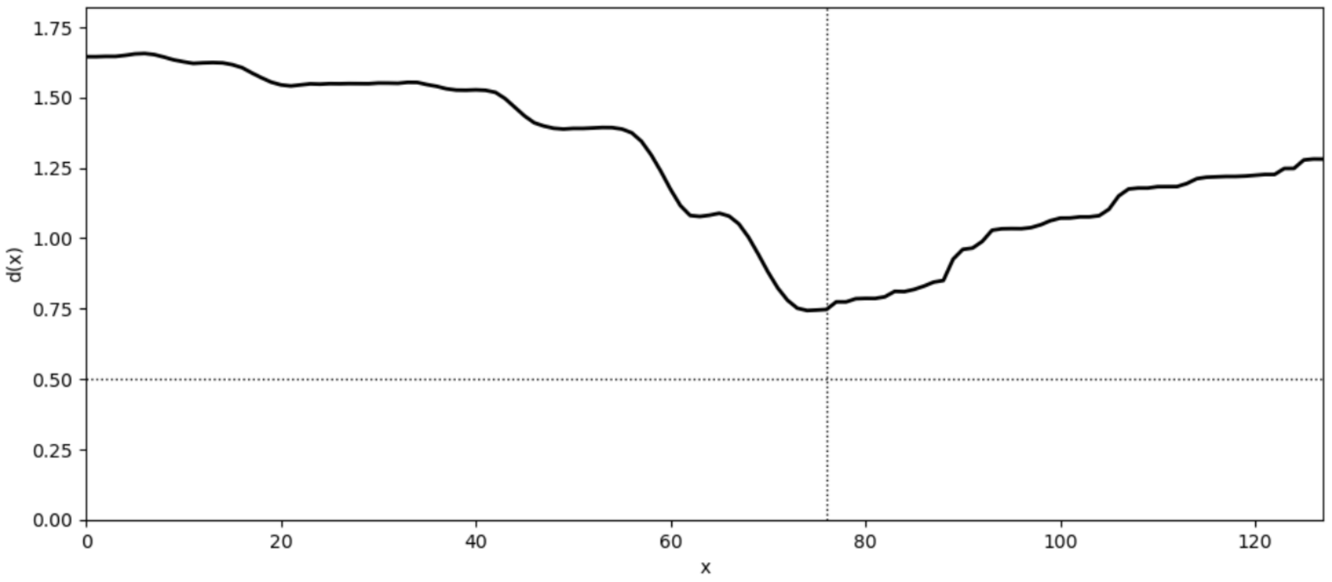
(Bottom)True data, $d^{true}(x)$ and true x_0 (black). Estimated data, $d^{obs}(x)$ and estimated x_0 (red)



Caption: dobs (black) and dest (red). R.m.s. error 0.2076

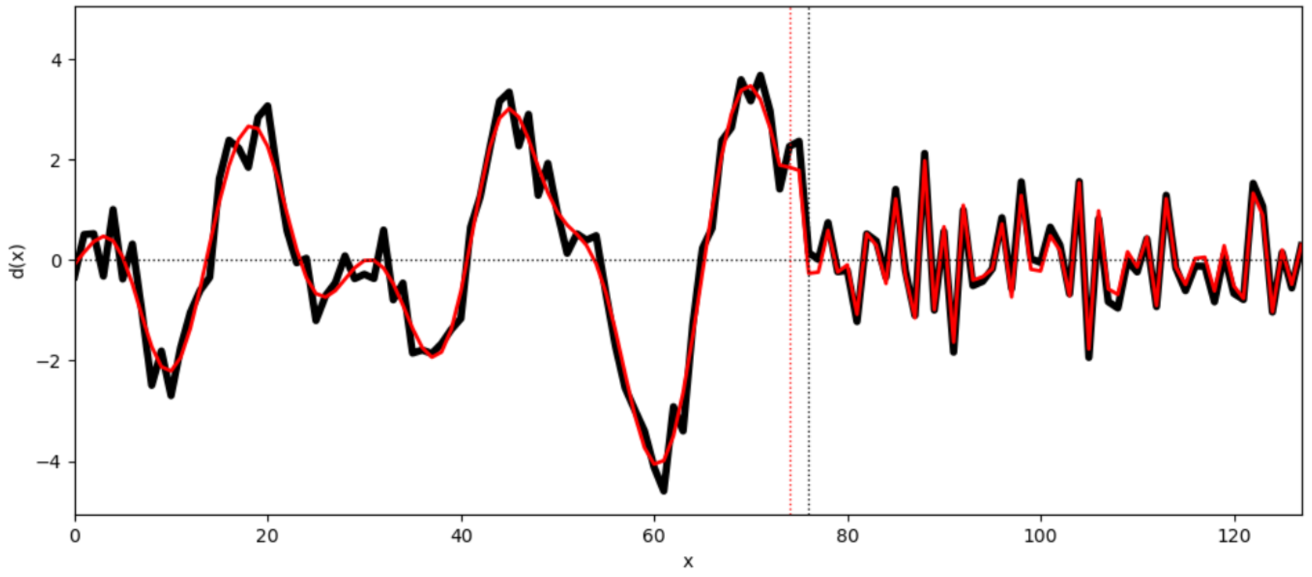
6.
Experiment 3

(Top) Generalized error for all possible choices of x_0.



Caption: R.m.s. generlized error 0.3404

(Bottom)True data, $d^{true}(x)$ and true t_0 (black). Estimated data, $d^{obs}(x)$ and estimated x_0 (red)



Caption: dobs (black) and dest (red). R.m.s. error 0.3387

Results

The estimated x_0 is found to be close to the true value, as long as the $C_L(r)$ and $C_R(r)$ are sufficiently different that one leads to a poorer fit, when applied to data drawn from the other.