CASE STUDY: Spline Approximation of Two-Dimension Data (spline_sum, st_pen, meanabs_pen, logexp_sum, logistic)

Background

This case study approximates two-dimension data by a spline using PSG function *spline_sum*. Input data for building spline:

$\vec{x}$= vector of independent values,
$\vec{y}$= vectors of dependent values,
D= degree of the spline,
K= number of polynomial pieces in the spline,
S= smoothing degree of the spline.

PSG risk functions, Standard Deviation (*st_pen)*, Mean Absolute Error (*meanabs_pen),* and Maximum Likelihood for Logistic Regression (*logexp_sum)* are minimized to find the best approximation.

A matrix with data (so called Matrix of Scenarios) contains vectors $\vec{x}$ and $\vec{y}$ and a matrix with parameters specifies *D*, *K*, *S*.

Notations

J=number of points (observations) of independent variable, *j* = index for points, *j*=1,...,*J*;

x_j = point of independent variable, *j*=1,...,*J*. Points x_j are ordered, i.e., if $j_1 < j_2$ then $x_{j_1} \leq x_{j_2}$;

y_j = point of dependent variable corresponding to the point x_j, *j*=1,...,*J*;

$\vec{x} = (x_1, x_2, \dots, x_J)$ = vector of points x_j;

$\vec{y} = (y_1, y_2, \dots, y_J)$ = vector of points y_j;

D = degree of spline, $D \geq 0$, integer;

K = number of polynomial pieces in the spline, $K > 0$, integer;

S = smoothing degree of a spline, $0 \leq S \leq D$, integer;

$I = K \cdot (D + 1)$ = number of unknown coefficients of polynomial pieces in a spline;

a_{dk}= decision variable = coefficient for degree$d$in polynomial piece$k$, $d = 0, \dots, D$, $k = 1, \dots K$;

$\vec{a} = (a_{01}, a_{11}, \dots, a_{D1},\ a_{02}, a_{12}, \dots, a_{D2}, \dots, a_{0K}, a_{1K}, \dots, a_{DK})$ = vector of decision variables (coefficients).

Important! These decision variables are not part of input data. They are generated by PSG automatically. Names of decision variables are based on names of independent factors.

$X = \{X_0, X_1, \dots, X_K\}$ = set of points (knots) partitioning segment $[x_1, x_J]$ insub-segments, $k = 1, \dots K$; every sub-segment $[X_{k-1}, X_k]$ contains at least one point x_j ($X_0 = x_1, X_K = x_J$);

J_k = sub-set of indexes *j*=1,...,*J* corresponding to sub-segment $[X_{k-1}, X_k]$, $J_k = \{j | x_j \in [X_{k-1}, X_k]\}$;

$L_j(\vec{a}) = y_j - G_j^0(\vec{a}) = y_j - \sum_{d=0}^{D} a_{dk} \cdot x_j^d$ = Loss Function valueatpoint x_j ,$j \in J_k$, $k = 1, \dots, K$;

$G_j^0(\vec{a}) = \sum_{d=0}^{D} a_{dk} \cdot x_j^d$ = Gain Functions with zero scenario benchmark at pointx_j ,$j \in J_k$, $k = 1, \dots, K$;

spline_sum$(D, K, S, \vec{x}, \vec{y}, \vec{a}) = \{L_1(\vec{a}), L_2(\vec{a}), \dots, L_J(\vec{a})\}$ = PSG function. Spline_sum generates a set of loss scenarios $L_j(\vec{a})$ using initial data and smoothing constraint;

$\mathbf{st\_pen}(\mathbf{spline\_sum}(D, K, S, \vec{x}, \vec{y}, \vec{a}) \quad) = \mathbf{st\_pen}(L_1(\vec{a}), L_2(\vec{a}), \dots, L_J(\vec{a})) = \left(\frac{1}{J}\sum_{j=1}^{J} L_j^2(\vec{a})\right)^{1/2}$= PSG functionStandard Penalty applied to Spline_sum;

$\mathbf{meanabs\_pen}(\mathbf{spline\_sum}(D, K, S, \vec{x}, \vec{y}, \vec{a}) \quad) = \mathbf{meanabs\_pen}(L_1(\vec{a}), L_2(\vec{a}), \dots, L_J(\vec{a})) = \frac{1}{J}\sum_{j=1}^{J}\left|L_j(\vec{a})\right|$= PSG functionMean Absolute Penalty applied to Spline_sum;

$\mathbf{logexp\_sum}(\mathbf{spline\_sum}(D, K, S, \vec{x}, \vec{y}, \vec{a})) = \mathbf{logexp\_sum}(\vec{y}, G_1^0(\vec{a}), G_2^0(\vec{a}), \dots, G_J^0(\vec{a})) = \frac{1}{J}\sum_{j=1}^{J}\left(y_j G_j^0(\vec{a}) - \ln\left(1 + \exp\left(G_j^0(\vec{a})\right)\right)\right)$= PSG functionLogarithms Exponents Sum applied to Spline_sum. In this case all y_j should be 0 or 1. Spline_sum is logistic regression log-likelihood function.

$\mathbf{logistic}(\mathbf{spline\_sum}(D, K, \vec{f}, \vec{y}, \vec{a}) \quad)$= vector with components:

$\exp\left(G_j^0(\vec{a})\right) / \left(1 + \exp\left(G_j^0(\vec{a})\right)\right)$, j=1,...,J.

Optimization Problem 1

minimizingStandard Penaltyfor building spline

$$\mathbf{min}_{\vec{a}}\mathbf{st\_pen}(\mathbf{spline\_sum}(D, K, S, \vec{x}, \vec{y}, \vec{a}))$$

calculation of Mean Absolute Penalty and Spline Sum for built spline

calculate

$\mathbf{meanabs\_pen}(\mathbf{spline\_sum}(D, K, S, \vec{x}, \vec{y}, \vec{a}))$

$\mathbf{spline\_sum}(D, K, S, \vec{x}, \vec{y}, \vec{a})$

Optimization Problem 2

minimizing Mean AbsolutePenalty for building spline

$$\mathbf{min}_{\vec{a}}\mathbf{meanabs\_pen}(\mathbf{spline\_sum}(D, K, S, \vec{x}, \vec{y}, \vec{a}))$$

calculation of Standard Penalty and Spline Sum for built spline

calculate

$\mathbf{st\_pen}(\mathbf{spline\_sum}(D, K, S, \vec{x}, \vec{y}, \vec{a}))$

$\mathbf{spline\_sum}(D, K, S, \vec{x}, \vec{y}, \vec{a})$

Optimization Problem 3

maximizing Logarithms Exponents Sum for building spline

$$\mathbf{max}_{\vec{a}}\mathbf{logexp\_sum}(\mathbf{spline\_sum}(D, K, S, \vec{x}, \vec{y}, \vec{a}))$$

calculation of Logarithms Exponents Sum and Logistic for built spline

calculate

$$\mathbf{logexp\_sum}(\mathbf{spline\_sum}(D, K, S, \vec{x}, \vec{y}, \vec{a}))$$

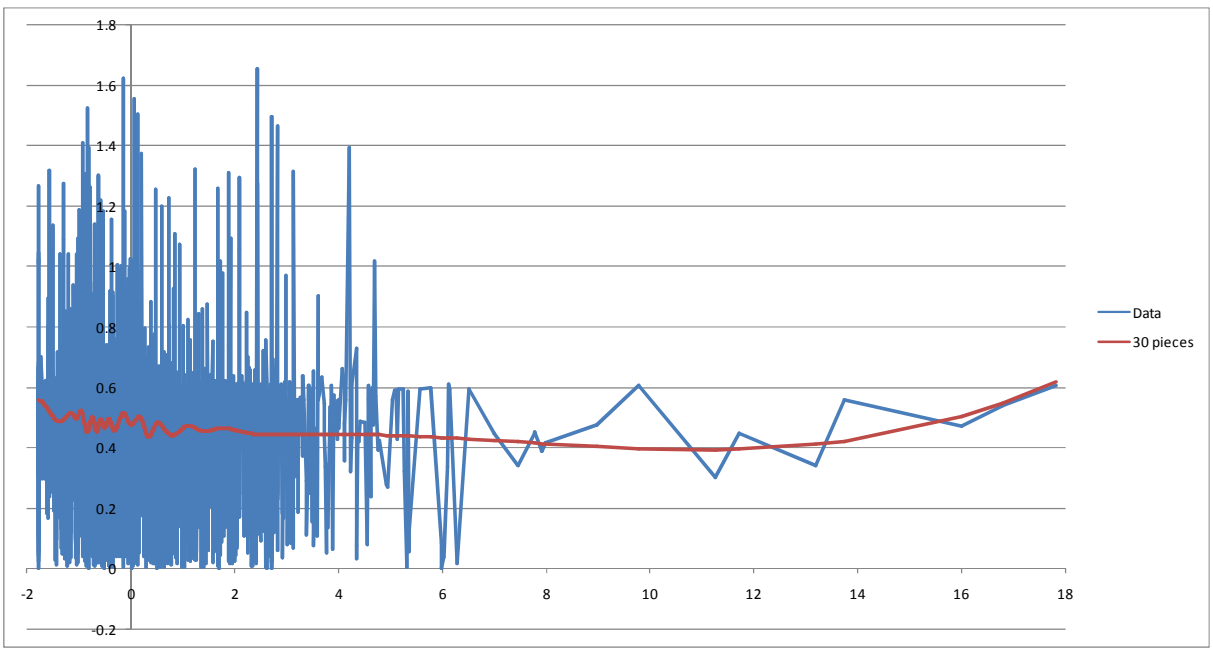
$$\mathbf{logistic}(\mathbf{spline\_sum}(D, K, S, \vec{x}, \vec{y}, \vec{a}))$$

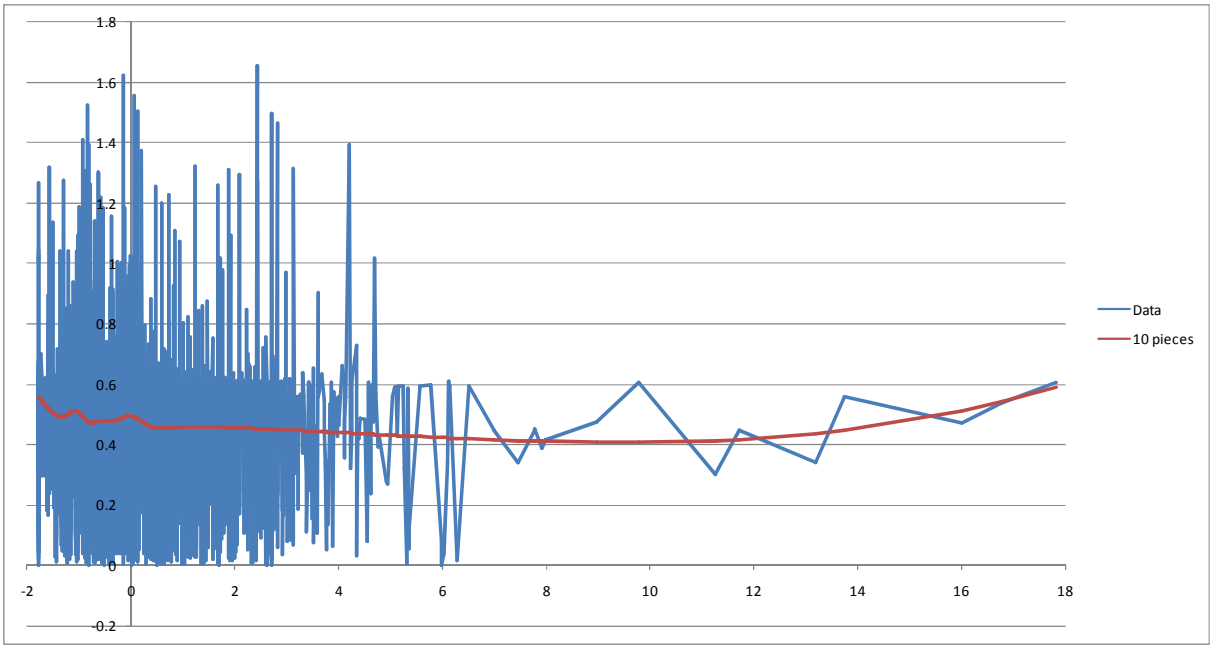
Solution Output

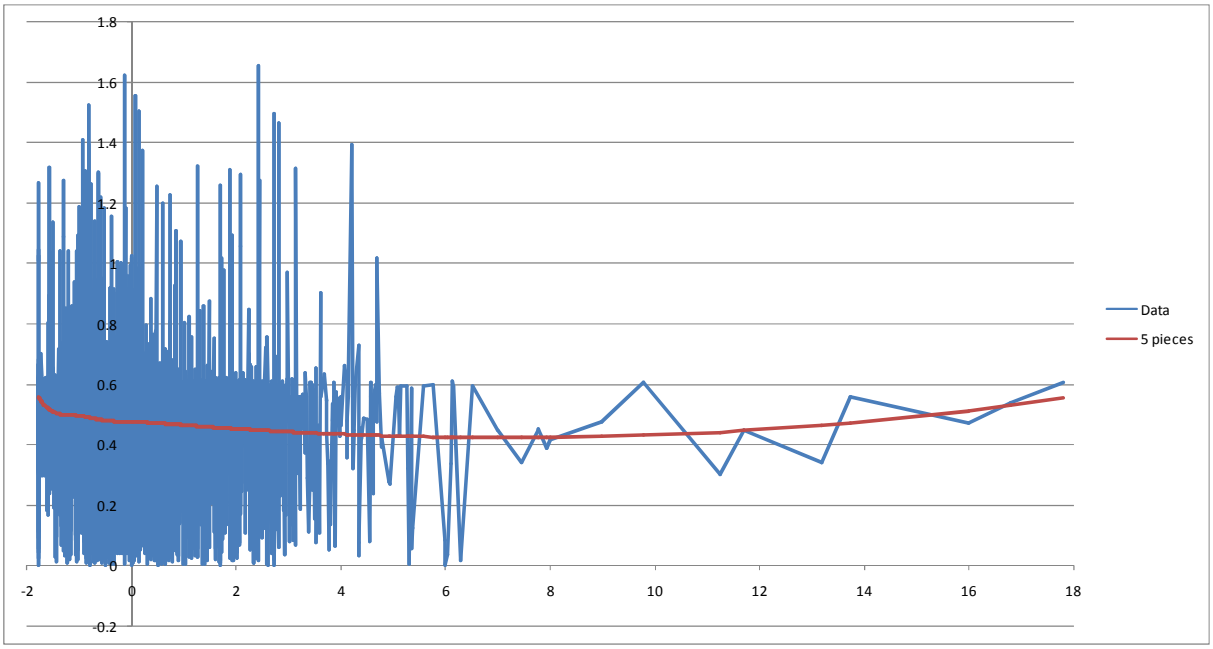
For the considered optimization problems, solvers provide a standard output containing values of objective, constraints, and functions, and a point with solution vector containing generated decision variables a_{dk}. In addition to the standard output, the solver gives *Spline_sum* function report containing two additional matrices: matrix with *knots* and matrix with *quant*. The first matrix contains knots $\{X_1, \dots, X_{K-1}\}$, the second one contains numbers of points x_j in every sub-segment $[X_{k-1}, X_k]$. Names of these two matrices are based on the name of matrix of input data.

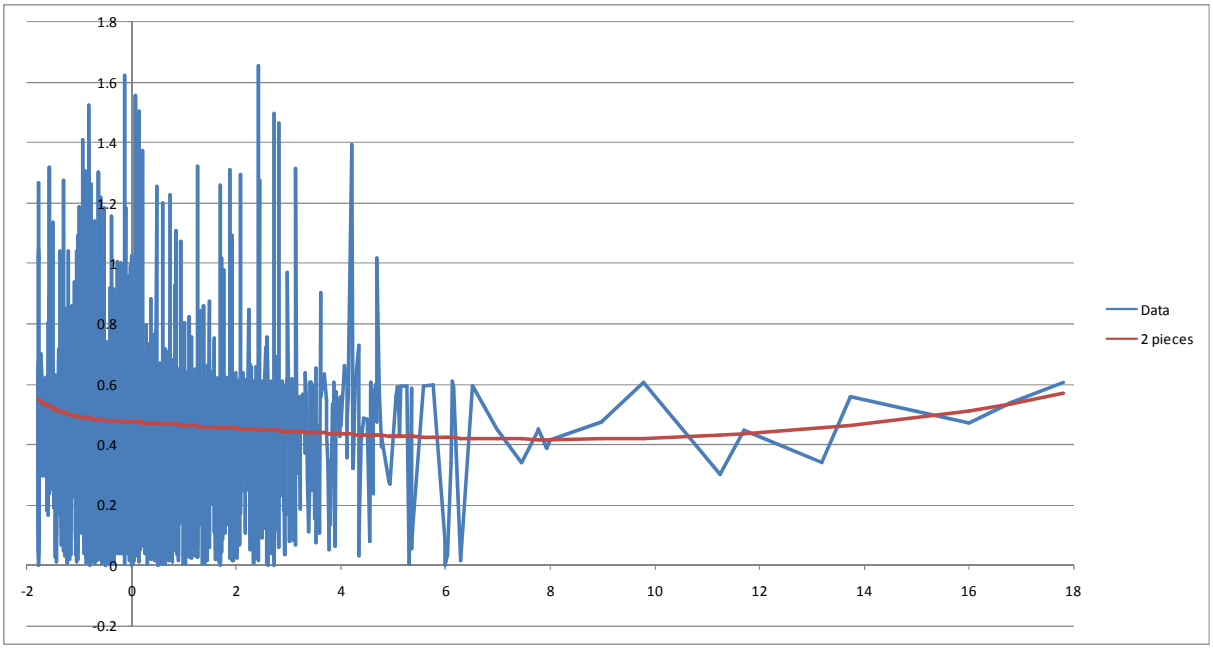
Solved Problem 1

Here are fourgraphspresenting data and spline approximationobtained with *Optimization Problem 1*. Data include 4,371 points connected by blue lines. The graphs show splines with different number of spline pieces: 30, 10, 5, 2.



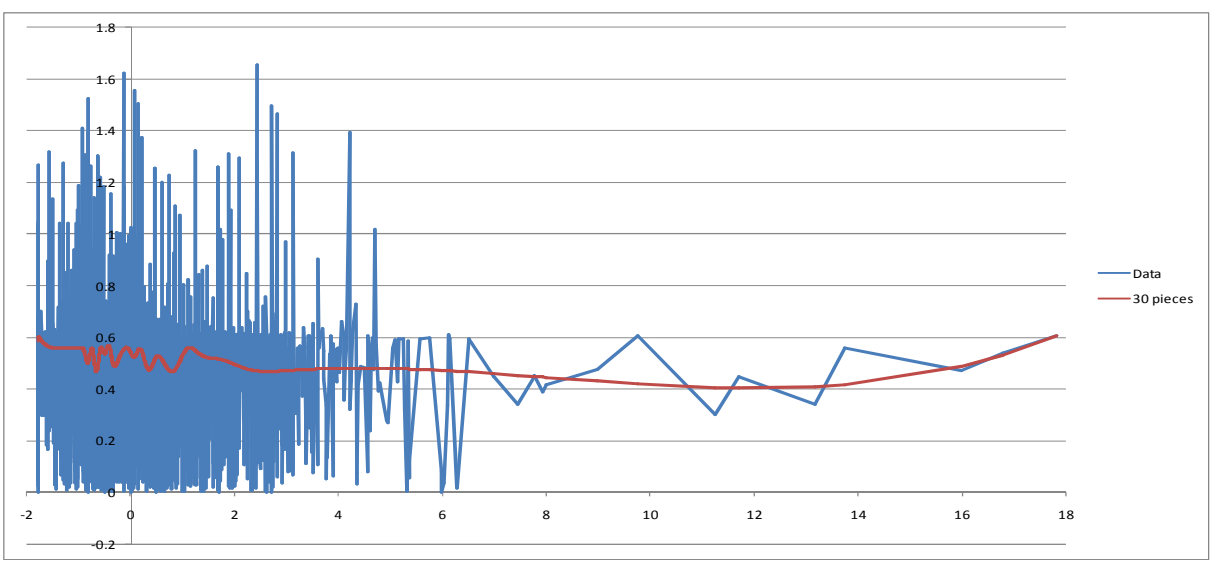

1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-2
0
2
4
6
8
10
12
14
16
18
Data
10 pieces

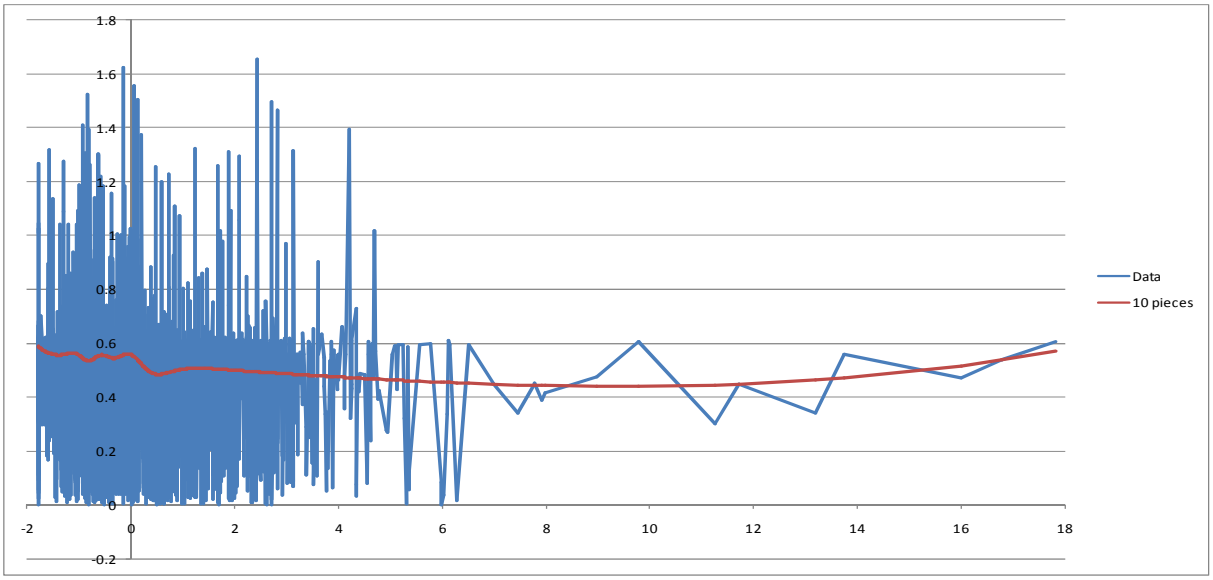

1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-2
0
2
4
6
8
10
12
14
16
18
Data
5 pieces

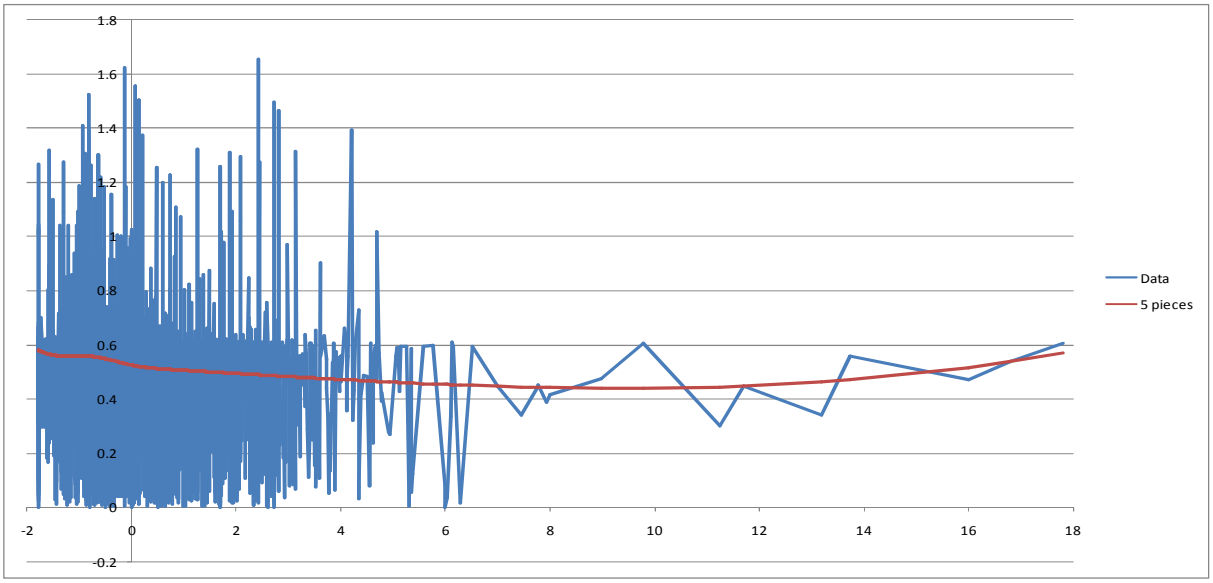


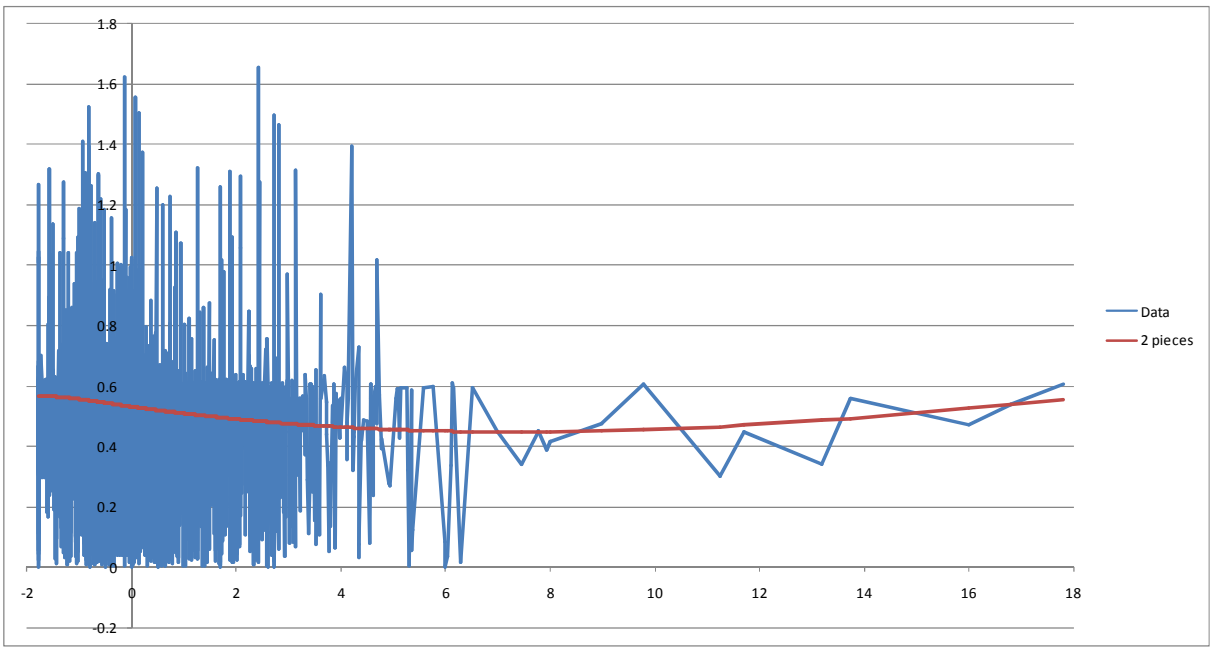
Solved Problem 2

Here arefourgraphs presenting data and spline approximation obtained with *Optimization Problem 2*. Data include 4,371 points connected by blue lines. Majority of points lie in the segment [-2, 4]. The graphs show splines with different number of spline pieces: 30, 10, 5, 2.




1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-2
0
2
4
6
8
10
12
14
16
18
Data
10 pieces

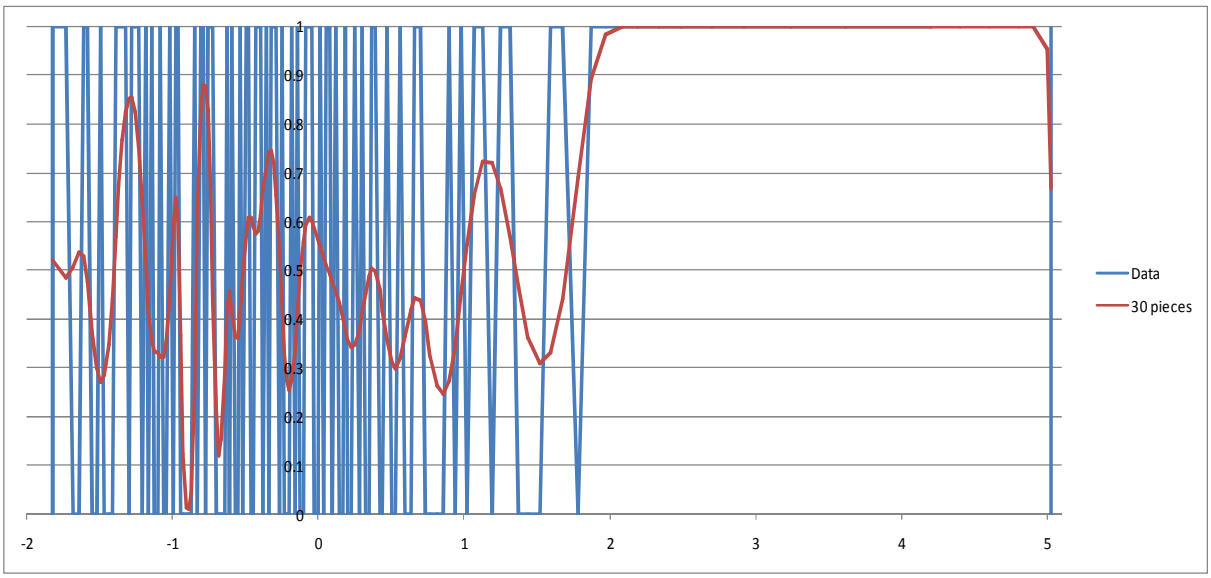

1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-2
0
2
4
6
8
10
12
14
16
18
Data
5 pieces

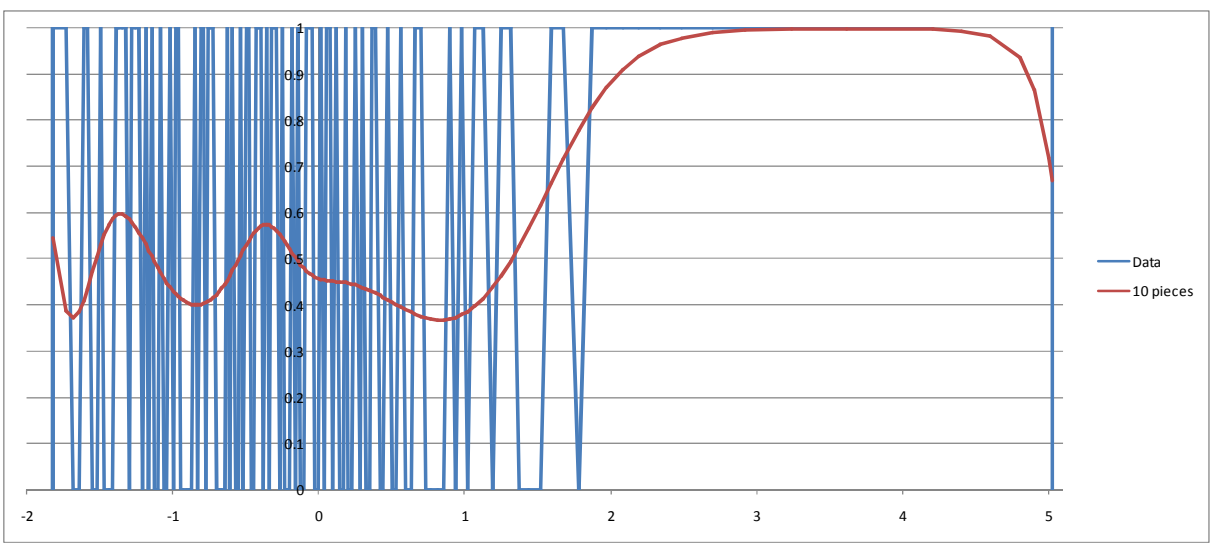

1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-2
0
2
4
6
8
10
12
14
16
18
Data
2 pieces

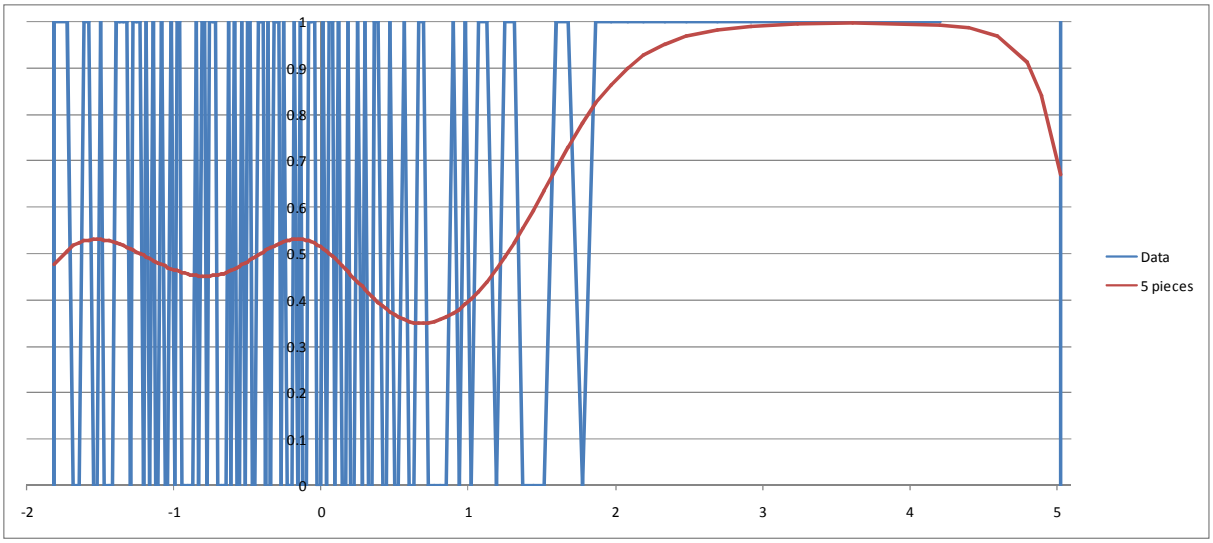
Solved Problem 3

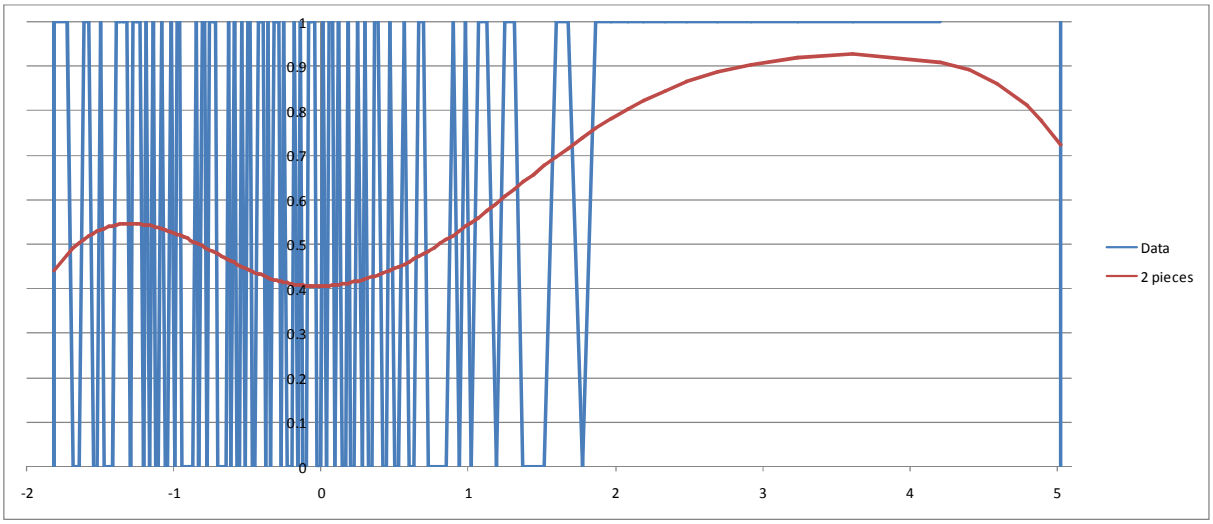
Here arefourgraphs presenting data and spline approximation obtained with *Optimization Problem 3*. Data include 149 points with values 0 and 1 connected by blue lines. Majority of points lie in the segment [-2, 2]. The graphs show spline approximations with different number of spline pieces: 30, 10, 5, 2.

Red curve showsthe functionof x calculated by usinglogistic function$e^{S(x)}/(1+e^{S(x)})$, where$S(x)$is the polynomial piecewisespline. This function (read curve) approximates probability depending on the independent factor x according to values of dependent variable in the dataset.






1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-2
-1
0
1
2
3
4
5
Data
5 pieces


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-2
-1
0
1
2
3
4
5
Data
2 pieces