Support Vector Regression with ANOVA Decomposition Kernels

Mark O. Stitson,
Alex Gammerman, Vladimir Vapnik,
Volodya Vovk, Chris Watkins,
Jason Weston

Technical Report

CSD-TR-97-22

November 27, 1997

Royal Holloway
University of London

Department of Computer Science
Egham, Surrey TW20 0EX, England

Abstract

Support Vector Machines using ANOVA Decomposition Kernels (SVAD) [Vapng] are a way of imposing a structure on multi-dimensional kernels which are generated as the tensor product of one-dimensional kernels. This gives more accurate control over the capacity of the learning machine (VC-dimension). SVAD uses ideas from ANOVA decomposition methods and extends them to generate kernels which directly implement these ideas.

SVAD is used with spline kernels and results show that SVAD performs better than the respective non ANOVA decomposition kernel. The Boston housing data set from UCI has been tested on Bagging [Bre94] and Support Vector methods before [DBK+97] and these results are compared to the SVAD method.

1 Introduction

In this paper we will introduce ANOVA kernels for support vector machines. We firstly introduce multiplicative kernels, which form the basis of the ANOVA kernels, then we introduce the general ANOVA decomposition idea. From this we derive ANOVA kernels and lastly show some experimental results on a well known data set.

2 Multiplicative Kernels

There is a whole set of kernels, the multi-dimensional case of which is calculated as the product of the one-dimensional case.

If the one-dimensional case is $k(x^i, y^i)$ then the n-dimensional case is

$$K(x, y) = \prod_{i=1}^{n} k(x^i, y^i)$$

Such kernels include Spline kernels with an infinite number of nodes, which allow approximation of a function using the support vectors as nodes [VGS97][1].

$$k(x, y) = \int_0^{min(x,y)} (x-t)^d (y-t)^d dt + \sum_{r=0}^{d} x^r y^r$$

In the linear case ($d = 1$):

$$k(x, y) = 1 + xy + \frac{1}{2}|x-y| \min(x, y)^2 + \frac{\min(x, y)^3}{3} \qquad (1)$$

[1] This kernel requires all values to be positive.

It is also possible to use RBF Kernels [Vap95]

$$k(x, y) = \exp^{-\gamma(x-y)^2}$$

or regularized Fourier transforms [VGS97][Vapng]

$$\phi(x) = \frac{a_0}{\sqrt{2}} + \sum_{i=1}^{\infty} q^i (a_i \sin ix + b_i \cos ix)$$

and

$$k(x, y) = \frac{1 - q^2}{2(1 - 2q\cos(x - y) + q^2)}$$

These kernels all consider the combination of all coordinates of two examples. This however might not always be appropriate as it can produce a too rich set of functions, from which the approximation is choosen. Quite often it might be preferable to analyse the combination of subsets of coordinates [Vapng]. In other words, the kernel is a sum of terms, each of which is the value of only a small number of input parameters.

3 ANOVA Decomposition

ANOVA Decomposition is a statistical idea of analysing the variances between different variables and finding certain dependencies on subsets of variables.

The general form of an ANOVA decomposition of a function of n variables $f(x) = f(x^1, \ldots, x^n)$ can be written as

$$f(x^1, \ldots, x^n) = F_0 + F_1(x^1, \ldots, x^n) + \ldots + F_n(x^1, \ldots, x^n) \tag{2}$$

where F_0 is a constant and

$$\begin{aligned}
F_1(x^1, \ldots, x^n) &= \textstyle\sum_{k=1}^{n} \alpha_{1_k} \phi_k(x^k) && (3)\\
F_2(x^1, \ldots, x^n) &= \textstyle\sum_{k_1 < k_2 \leq n} \alpha_{2_k} \phi_{k_1,k_2}(x^{k_1}, x^{k_2}) && (4)\\
&\ldots && (5)\\
F_r(x^1, \ldots, x^n) &= \textstyle\sum_{k_1 < k_2 < \ldots < k_r \leq n} \alpha_{r_k} \phi_{k_1,k_2,\ldots,k_r}(x^{k_1}, x^{k_2}, \ldots, x^{k_r}) && (6)\\
&\ldots && (7)\\
F_n(x^1, \ldots, x^n) &= \alpha_{n_k} \phi_{k_1,\ldots,k_n}(x^{k_1}, \ldots, x^{k_n}) && (8)
\end{aligned}$$

where the functions ϕ and coefficients α are choosen according to some algorithm analysing the variances.

This very general idea is usually modified to consider onlz sets of variables of size p less than n and called p^{th} order ANOVA decomposition. Thus with increasing order hopefully a more exact approximation is found.

4 ANOVA Decomposition Kernels

The ANOVA decomposition idea is very simplistically converted into kernels by ignoring the idea of analysing variances and concentrating on all possible subsets of variables up to a certain size. The heuristic here is that the SVM will choose coefficients such that analysis of variables by the kernel is not necessary.

Kernels are generated from this in the usual way by replacing the coefficients by an identical term for the second parameter/vector of the kernel.

To be able to use standard multiplicative kernels we have to make some restrictions on the functions we use. Assume all functions $\phi_{k_1,\dots,k_m}(x^{k_1},\dots,x^{k_m}) = \prod_{i=1}^{m}\phi_{k_i}(x^{k_i})$, so we are using a multiplicative form. Further assume $\forall\phi_{k_i}(x^{k_i}) = \phi(x^{k_i})$, this is our non-linear transformation into a different space, which is the same for all coordinates. Finally assume the one-dimensional kernel is $k(x,y) = \phi(x)\phi(y)$. Now we can create kernel functions from the functions F_1 to F_n:

$$K_1(x,y) = \sum_{k=1}^{n}\phi(x^k)\phi(y^k) = \sum_{k=1}^{n}k(x^k,y^k) \tag{9}$$

$$K_2(x,y) = \sum_{1\le k_1<k_2\le n}\phi(x^{k_1})\phi(x^{k_2})\phi(y^{k_1})\phi(y^{k_2}) \tag{10}$$

$$\begin{aligned}= \frac{1}{2}[\sum_{k_1=1}^{n}\sum_{k_2=1}^{n}\phi(x^{k_1})\phi(x^{k_2})\phi(y^{k_1})\phi(y^{k_2}) \\ -\sum_{k=1}^{n}\phi(x^k)\phi(x^k)\phi(y^k)\phi(y^k)]\end{aligned} \tag{11}$$

$$= \frac{1}{2}[\sum_{k_1=1}^{n}\sum_{k_2=1}^{n}k(x^{k_1},y^{k_1})k(x^{k_2},y^{k_2}) - \sum_{k=1}^{n}(k(x^k,y^k))^2] \tag{12}$$

and from (C. Burges and V. Vapnik, 95) the following recurrent procedure can be used.

Let $K^s(x,y) = \sum_{i=1}^{n}(k(x^i,y^i))^s$ and $K_0(x,y)=1$ then

$$K_p(x,y) = \sum_{1 \le k_1 < \ldots < k_p \le n} k(x^{k_1}, y^{k_1}) \times \ldots \times k(x^{k_p}, y^{k_p}) \tag{13}$$

$$K_p(x,y) = \frac{1}{p}\sum_{s=1}^{p}(-1)^{s+1}K_{p-s}(x,y)K^s(x,y) \tag{14}$$

This recurrent procedure is very efficient as all lower orders can be calculated at the same time.

There are two ways of using ANOVA decomposition to produce kernels of order p. The first method includes order p and all lower orders:

$$K(x,y) = \sum_{i=1}^{p} K_p(x,y) \tag{15}$$

The second method only includes order p:

$$K(x,y) = K_p(x,y) \tag{16}$$

In our experiments we used ANOVA decompositions of the latter type, only considering a term of exactly order p.

5 Experiments

5.1 Data

Our experiments were conducted on the Boston housing problem from StatLib at Carnegie Mellon University [HR78]. This is a well known data set for testing non-linear regression methods. Previous uses include Drucker et al. [DBK+97] and Breiman [Bre94].

The data set consists of 506 cases in which 12 continuous variables and 1 binary variable determine the median house price in a certain area of Boston in thousands of dollars. The prices lie between \$5000 and \$50000 in units of \$1000.

5.2 Method

As the data set is very small we decided to follow Drucker et al. [DBK$^+$97] and partition the data randomly 100 times for 100 trials into a training set of 401 cases, a validation set of 80 cases and a test set of 25 cases.

For each trial we created SVMs for linear spline kernels, ANOVA linear spline kernels and polynomial kernels.

For each kernel we selected a set of parameters which gave the smallest error on the validation set and then measured the error on the test set.

5.3 SVM Parameters

For the spline kernels it is important that all variables are greater than 0, so we scaled the data linearly and shifted it, such that the values of each variable lie between 0 and 1. For the polynomials we used the same technique to make all variables lie between -1 and 1.

The parameters that had to be chosen were the kernel specific parameters, i.e. the order of ANOVA decomposition and the degree of the polynomial kernel. For the ANOVA decomposition we choose 2, 4, 6, 8, 10 and 13, as the intermediate orders did not seem to improve the result much in previous experiments. For the polynomial kernel, we followed Drucker et al. [DBK$^+$97] in choosing 4 and 5 as the recommended degree.

The next parameter to be chosen is independent of the kernel and determines the accuracy of the regression. It defines the amount by which a training set point is allowed to diverge from the regression. This is the so-called ϵ-margin or ϵ-tube and was chosen as 1, 2, 3, 4 and 5 in turn [SWG$^+$96].

The final parameter is the upper bound C on the Lagrangian multipliers. This was chosen to minimize the error on the validation set using a heuristic technique which minimized C until the validation set error increased.

5.4 Results

After creating about 50,000 SVM for the ANOVA splines and selecting 100 from them on the basis of the minimal validation set error. Then creating about 15,000 SVM for polynomial kernels and again selecting 100 from them. Finally creating about 8,000 SVM for the spline kernel and selecting another 100 from them. We calculated the average error for every kernel over the 100 trials.

We produced all of the following results except for the Bagging on the same 100 datasets. Drucker's result for polynomial SVM is better than ours, but given the high variance, this is probably due to the random selection of training, validation and test sets. Breiman has a higher mean square error, which is probably due to the data selection, but can be viewed as an indication of bagging regression trees.

Error rates on the test set:

Kernel	Avg. Square Error	Variance
Bagging [Bre94]	11.7	
Polynomial	8.28	24.02
Splines	7.87	12.67
ANOVA Splines	7.56	8.70

Error rates on the validation set:

Kernel	Avg. Square Error	Variance
Polynomial	7.14	4.66
Splines	6.46	2.14
ANOVA Splines	6.07	1.61

The experiments show that Splines and ANOVA Splines can outperform Polynomials in fitting complicated functions to the Boston housing data, but what seems to be far more significant is that the variance is far lower for the ANOVA Splines than the other two methods.

The average order of ANOVA kernels choosen was approximately 8.5.

The parameter selection can be improved if the estimated VC dimension is found. This unfortunately involves finding the radius of the smallest circle enclosing the support vectors or training points in feature space, which is an optimisation problem of non-trivial scale; it has to be done for every SVM if the more accurate method using the support vectors is used.

6 Conclusion and Further research

ANOVA decomposition kernels provide far better result than previous kernels on the Boston housing data, which is a standard statistical test for non-linear regression estimation. They also provide better results than other known statistical methods.

The Boston housing data is a highly complex set of data, as shown by the high average order of ANOVA decomposition kernel choosen.

ANOVA decomposition is applicable to many kernels like Fourier expansions, Hermite polynomials and Radial Basis Functions.

Further research into the ANOVA decomposition kernels will investigate kernels which include lower order components and better ways of parameter selection using the estimated VC-dimension.

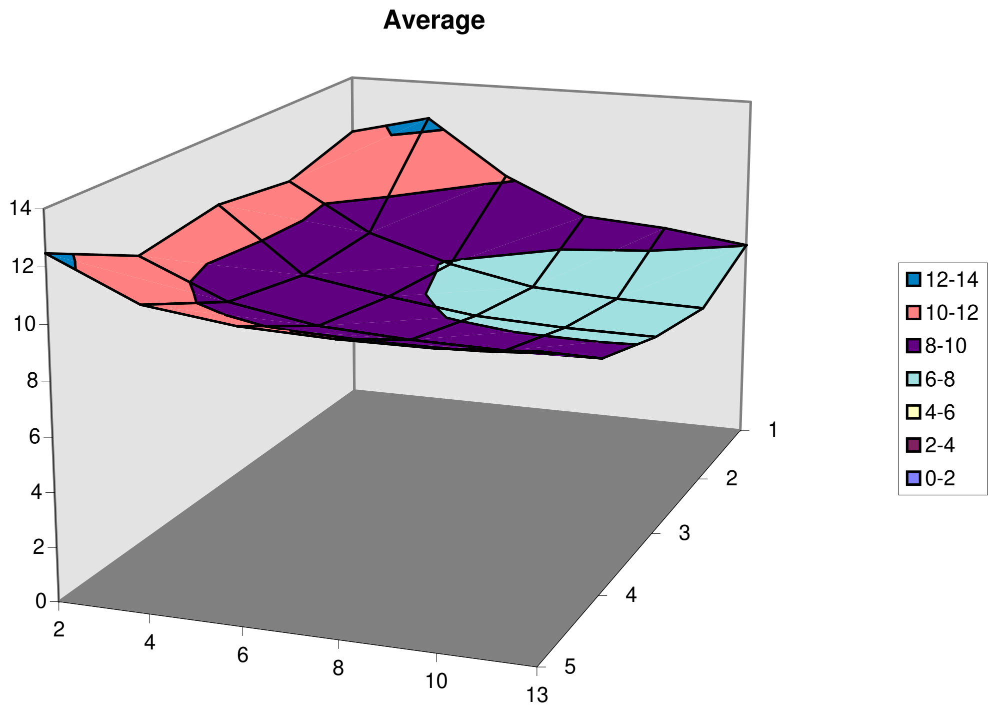
A Detailed results for ANOVA Splines

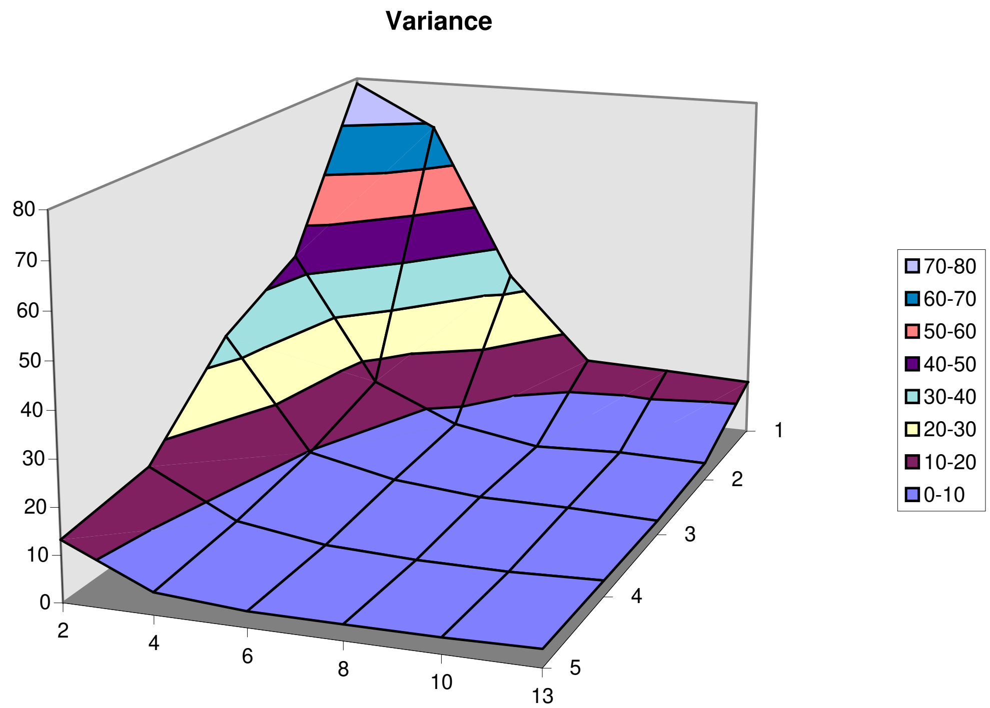
A.1 Validationset

Error Rates for ANOVA Splines
Validationset

Average	Order →							
Epsilon ↓	2	4	6	8	10	13	8.42	7.49
5	12.4686	10.9934	10.5647	10.4556	10.4595	10.6977		
4	10.8151	9.40033	8.8315	8.64783	8.59072	8.64192		
3	11.2732	8.78097	8.23963	7.79875	7.65453	7.62412		
2	10.8223	8.93634	7.92361	7.26806	7.08321	7.01415		
1	11.6426	12.4472	10.1781	8.68552	8.45711	8.01343		
Best Validation Choice								
1.74							6.07388	
Best Possible Choice								
1.86								7.15957

Variance	Order →							
Epsilon ↓	2	4	6	8	10	13	8.42	7.49
5	13.3099	4.7851	3.44014	3.4289	3.4397	3.8416		
4	16.3255	6.61926	3.69692	2.88847	2.8108	3.48514		
3	34.7403	9.80767	5.66673	3.84857	3.72945	3.14507		
2	44.0485	15.3163	7.00057	3.66771	4.37857	3.99546		
1	78.7126	69.0754	33.5665	13.8619	13.145	12.2989		
Best Validation Choice								
0.5924							1.60717	
Best Possible Choice								
0.7804								6.61306


Average
14
12
10
8
6
4
2
0
2
4
6
8
10
13
1
2
3
4
5
12-14
10-12
8-10
6-8
4-6
2-4
0-2

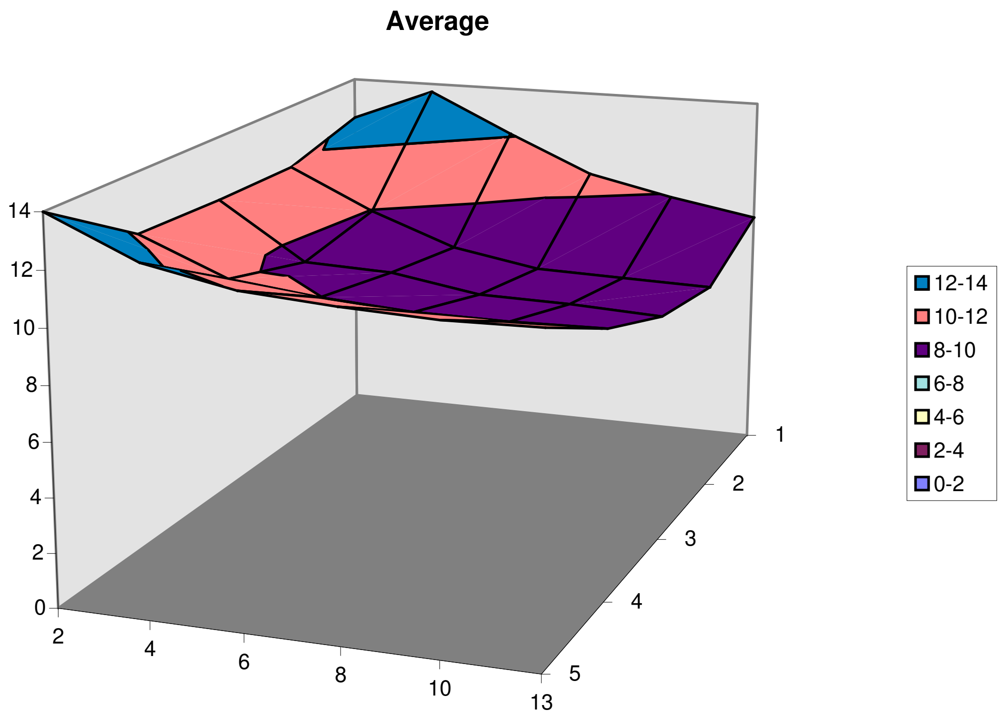

Variance
80
70
60
50
40
30
20
10
0
2
4
6
8
10
13
1
2
3
4
5
70-80
60-70
50-60
40-50
30-40
20-30
10-20
0-10

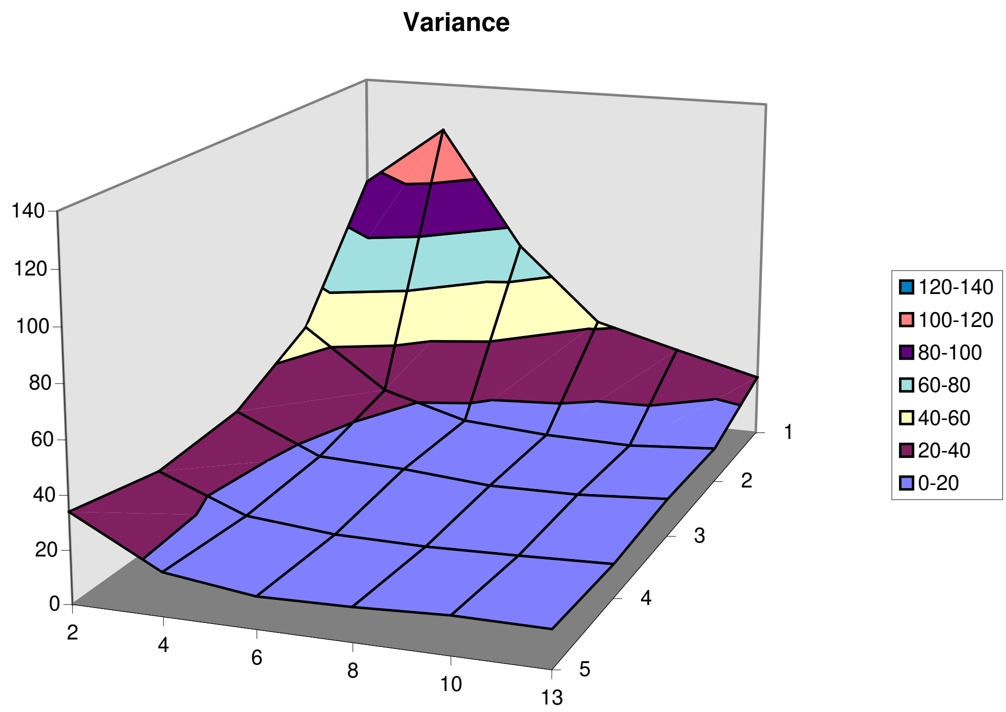
A.2 Test set

Error Rates for ANOVA Splines
Testset

Average	Order →							
Epsilon ↓	2	4	6	8	10	13	8.42	7.49
5	13.9972	12.5535	11.903	11.6752	11.5372	11.6018		
4	11.7445	10.3886	10.0143	9.79649	9.76744	9.82794		
3	11.5712	9.42329	9.31112	8.76714	8.69904	8.54496		
2	11.5214	9.955	8.74698	8.15614	8.07987	7.99673		
1	12.3519	13.673	12.0951	10.6326	9.89504	9.3088		
Best Validation Choice								
1.74							7.56228	
Best Possible Choice								
1.86								5.95217

Variance	Order →							
Epsilon ↓	2	4	6	8	10	13	8.42	7.49
5	33.9849	16.3176	12.0308	12.7184	14.278	14.112		
4	27.3846	14.2055	11.1405	10.7225	11.5706	12.7901		
3	30.7944	16.1182	14.7702	12.0793	12.3369	14.6536		
2	47.5293	23.8555	14.4079	11.9633	11.2007	13.6755		
1	95.4371	120.429	72.3163	41.9806	32.829	24.3272		
Best Validation Choice								
0.5924							8.70254	
Best Possible Choice								
0.7804								5.86091



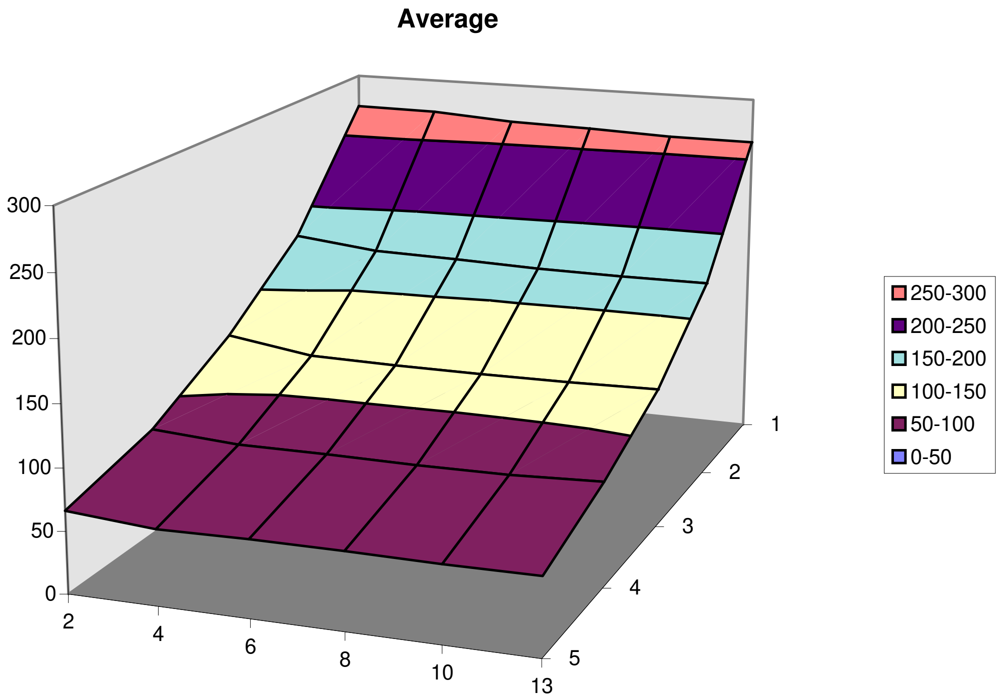


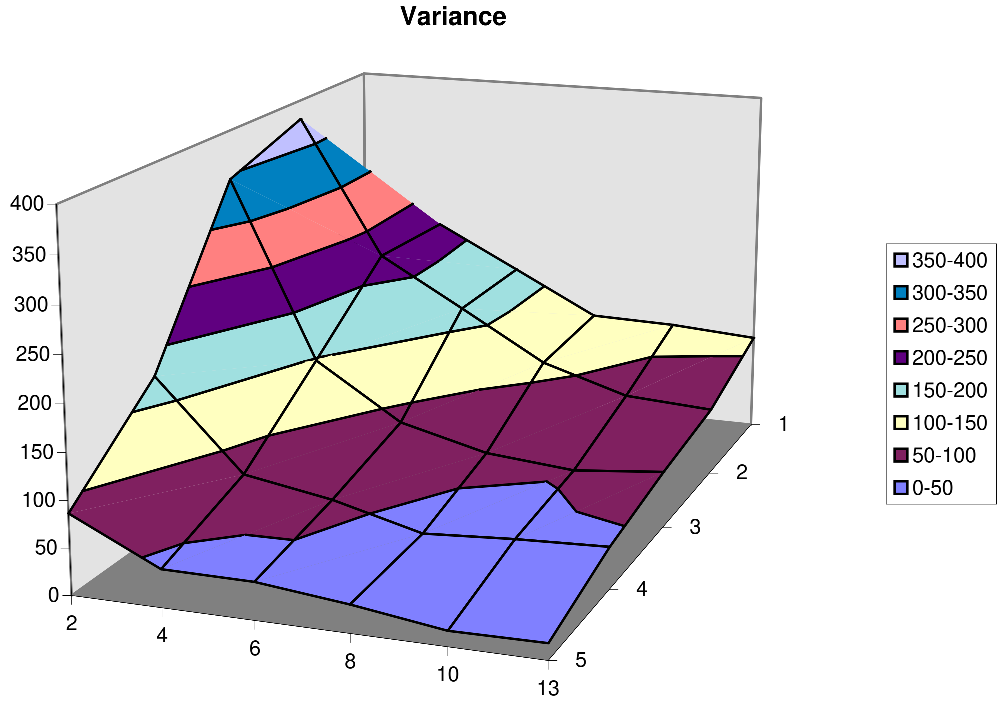
A.3 Support Vectors

SVs for ANOVA Splines

Average	Order →							
Epsilon ↓	2	4	6	8	10	13	8.42	7.49
5	66.52	60.8	62.07	62.2	61.47	62.24		
4	86.69	81.77	81.97	81.58	82.18	85.07		
3	125.91	115.37	114.23	114.04	114.58	116.02		
2	179.69	172.14	170.49	168.47	167.75	168.14		
1	271.42	270.84	266.2	264.66	262.16	262.04		
Best Validation Choice								
1.74							197.1	
Best Possible Choice								
1.86								191.77

Variance	Order →							
Epsilon ↓	2	4	6	8	10	13	8.42	7.49
5	86.5896	40.4	40.0451	29.68	16.1091	17.3424		
4	173.994	76.7571	60.2491	35.5036	41.5076	45.4451		
3	345.702	147.833	86.4571	62.1584	53.3236	62.7796		
2	377.974	222.1	167.31	111.509	81.8475	75.6204		
1	95.4371	217.034	168.06	118.444	115.734	108.898		
Best Validation Choice								
0.5924							3432.55	
Best Possible Choice								
0.7804								3937.64


Average
300
250
200
150
100
50
0
2
4
6
8
10
13
1
2
3
4
5
250-300
200-250
150-200
100-150
50-100
0-50


Variance
400
350
300
250
200
150
100
50
0
2
4
6
8
10
13
1
2
3
4
5
350-400
300-350
250-300
200-250
150-200
100-150
50-100
0-50

B Detailed results for Polynomials

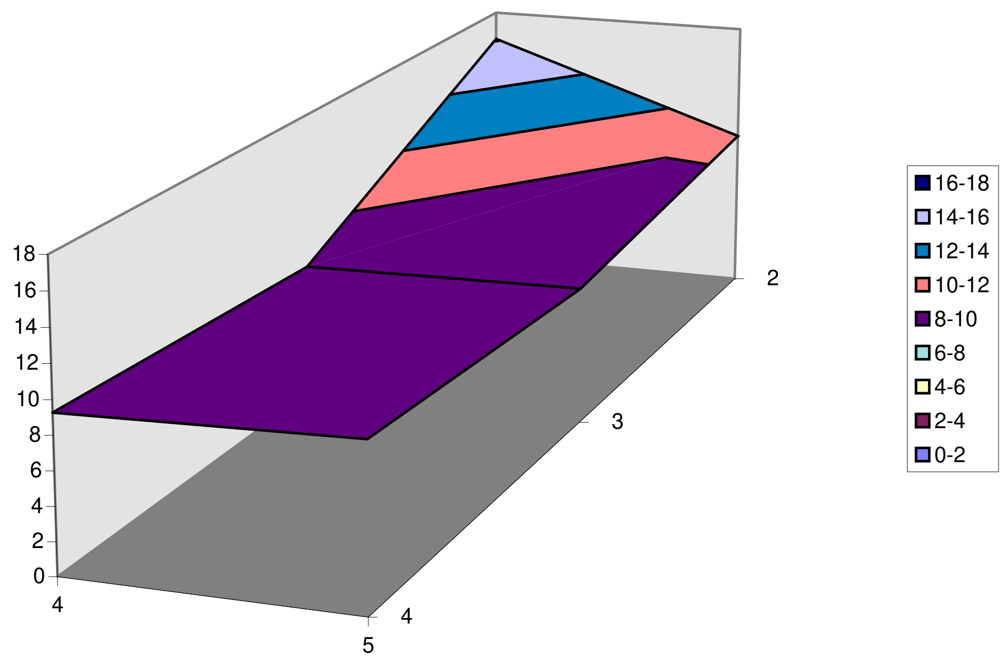
B.1 Validationset

Error Rates for Polynomials
Validationset

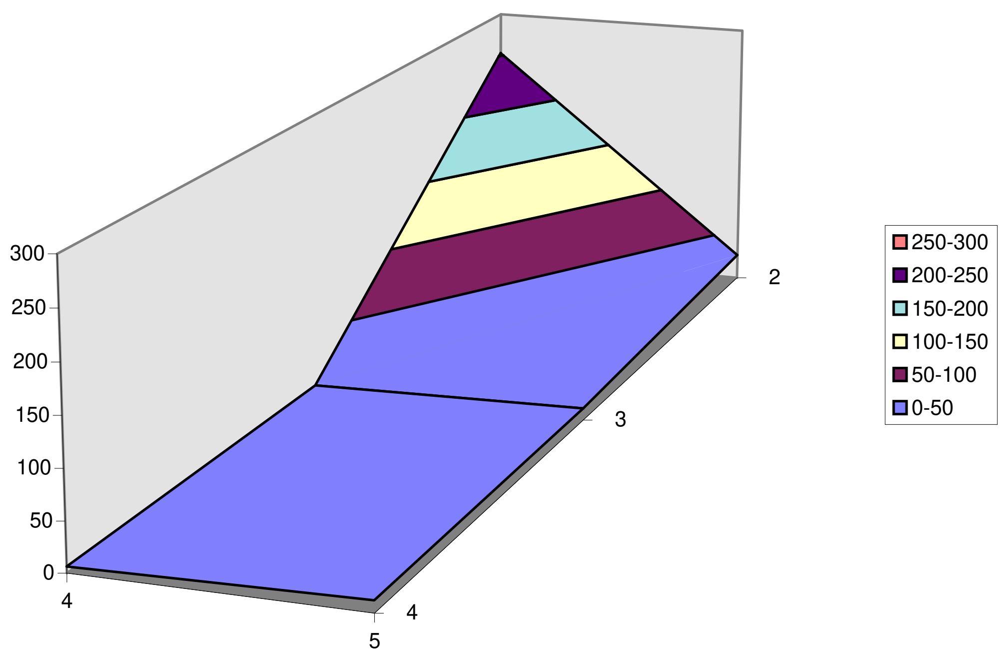
Average	Degree →			
Epsilon ↓	4	5	4.46	4.6
4	9.26585	9.7689		
3	8.28579	8.54029		
2	16.1042	10.3822		
Best Validation Choice				
2.56			7.13695	
Best Possible Choice				
2.8				8.1878

Variance	Degree →			
Epsilon ↓	4	5	4.46	4.6
4	6.27531	12.2083		
3	6.39339	12.6696		
2	252.663	28.2594		
Best Validation Choice				
2.56			4.663	
Best Possible Choice				
2.8				14.5635

Average



Variance

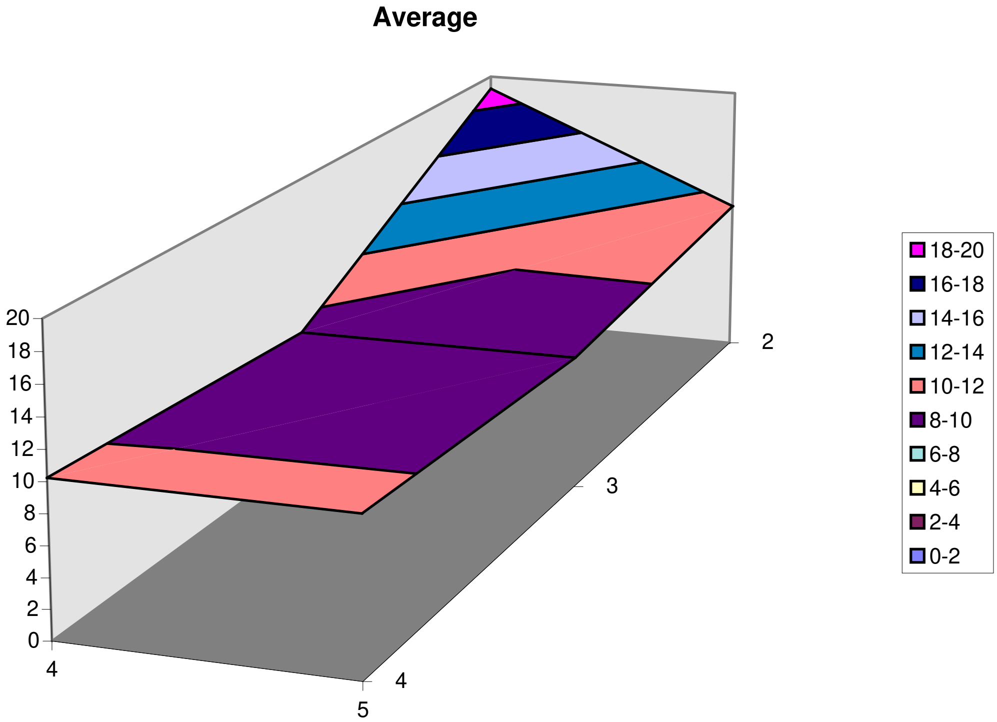


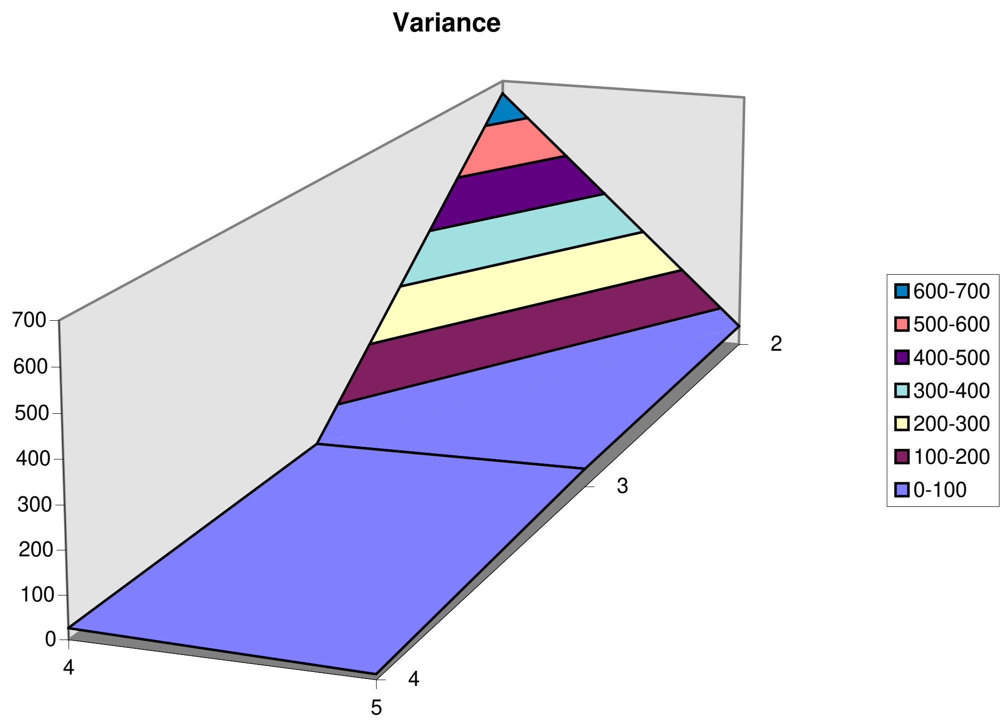
B.2 Test set

Error Rates for Polynomials
Testset

Average	Degree →				
	Epsilon ↓	4	5	4.46	4.6
	4	10.2415	10.2481		
	3	9.07575	9.13913		
	2	19.0249	11.0433		
Best Validation Choice					
	2.56			8.27855	
Best Possible Choice					
	2.8				7.564

Variance	Degree →				
	Epsilon ↓	4	5	4.46	4.6
	4	26.3263	12.2083		
	3	36.3981	45.9214		
	2	665.442	53.4141		
Best Validation Choice					
	2.56			24.0213	
Best Possible Choice					
	2.8				22.5756

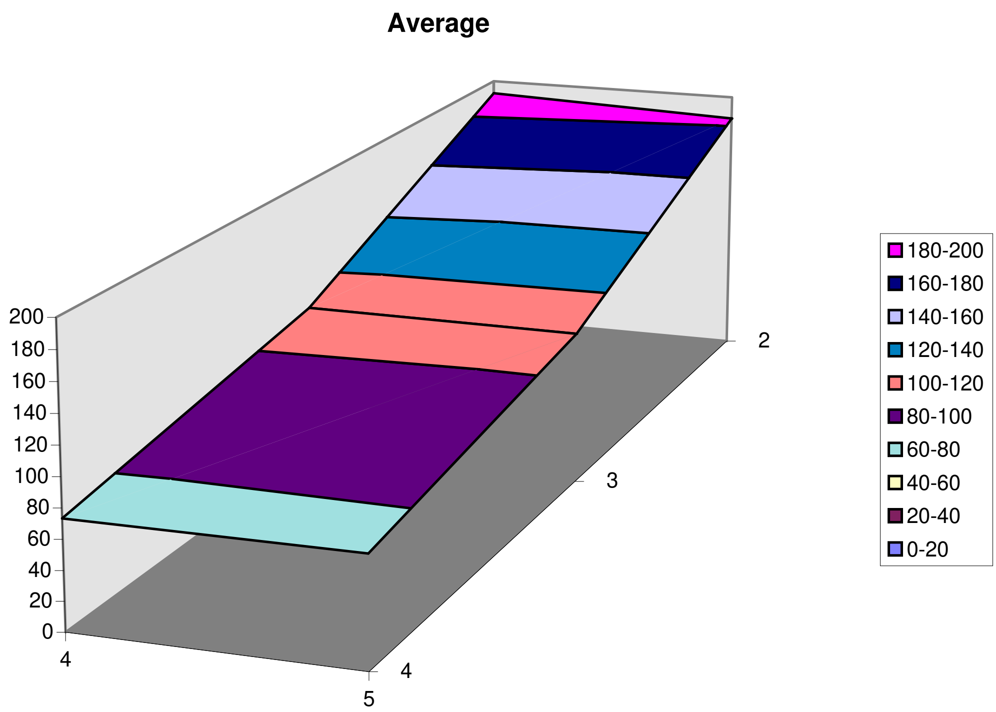


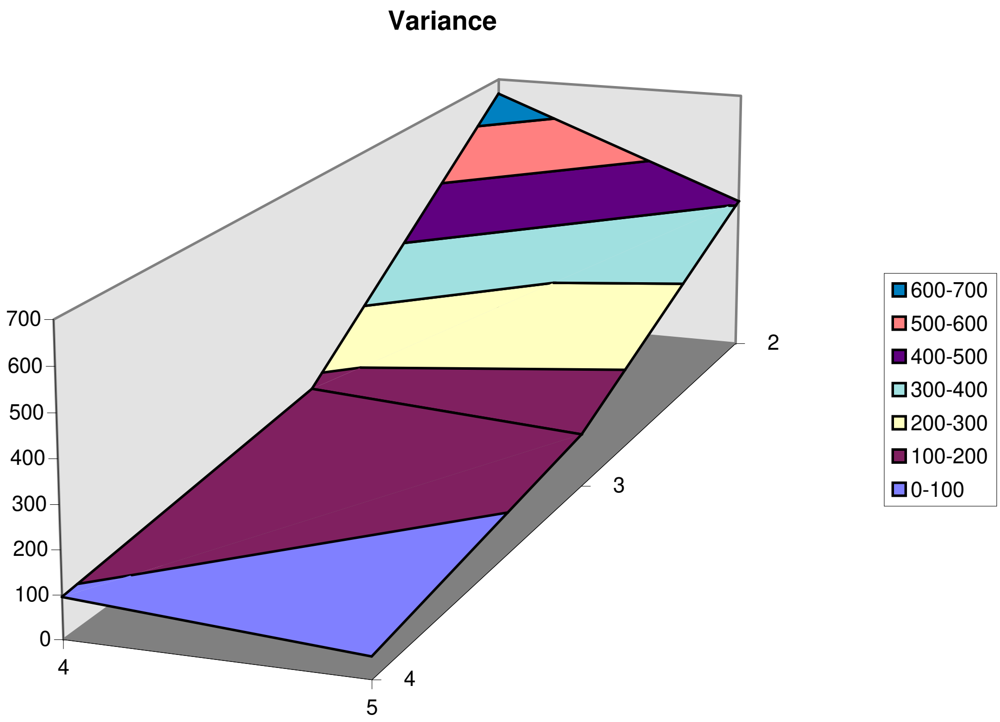


B.3 Support Vectors

Average	Degree →			
Epsilon ↓	4	5	4.46	4.6
4	73.5	74.02		
3	107.96	107.17		
2	190.06	182.97		
Best Validation Choice				
2.56			133.46	
Best Possible Choice				
2.8				122.85

Variance	Degree →			
Epsilon ↓	4	5	4.46	4.6
4	95.75	51.0596		
3	176.138	131.261		
2	659.376	405.809		
Best Validation Choice				
2.56			1325.83	
Best Possible Choice				
2.8				1337.35


Average
200
180
160
140
120
100
80
60
40
20
0
4
5
4
3
2
180-200
160-180
140-160
120-140
100-120
80-100
60-80
40-60
20-40
0-20


Variance
700
600
500
400
300
200
100
0
4
5
4
3
2
600-700
500-600
400-500
300-400
200-300
100-200
0-100

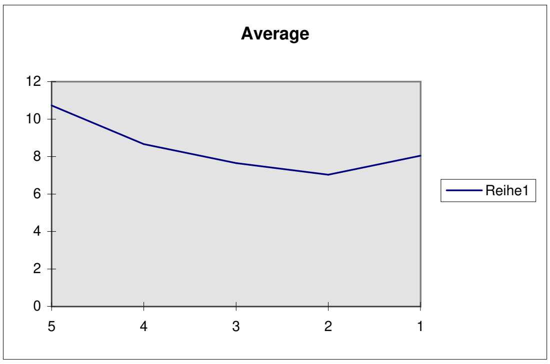
C Detailed results for Splines

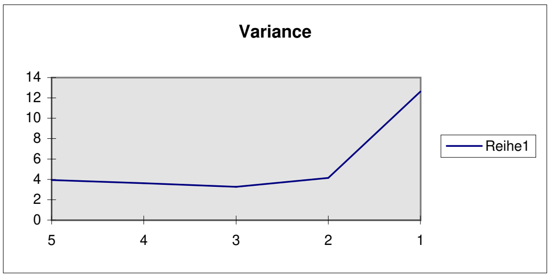
C.1 Validationset

Spline Validationset

Error Rates for Splines
Validationset

Average	Order
Epsilon ↓	13
5	10.7295
4	8.66898
3	7.65495
2	7.03404
1	8.04386
Best Validation Choice	
1.85	6.45866
Best Possible Choice	
2.02	7.21282
Variance	Order
Epsilon ↓	13
5	3.9285
4	3.61521
3	3.27452
2	4.13599
1	12.6179
Best Validation Choice	
1.85	2.13933
Best Possible Choice	
2.02	7.56974



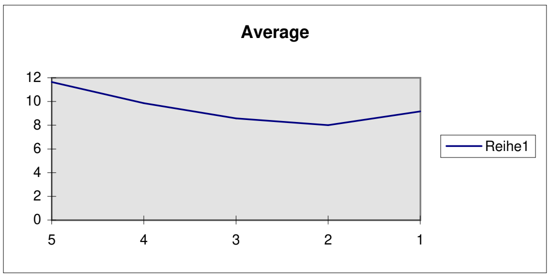


C.2 Test set

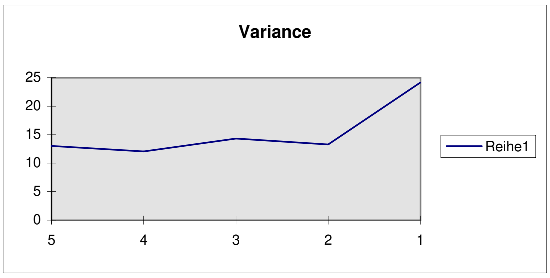
Spline Testset

Error Rates for Splines
Testset

Average	Order
Epsilon ↓	13
5	11.6367
4	9.86512
3	8.57635
2	8.01061
1	9.17272
Best Validation Choice	
1.85	7.87379
Best Possible Choice	
2.02	6.9979



Variance	Order
Epsilon ↓	13
5	13.0117
4	12.0531
3	14.3085
2	13.2818
1	24.1464
Best Validation Choice	
1.85	12.6747
Best Possible Choice	
2.02	9.81015

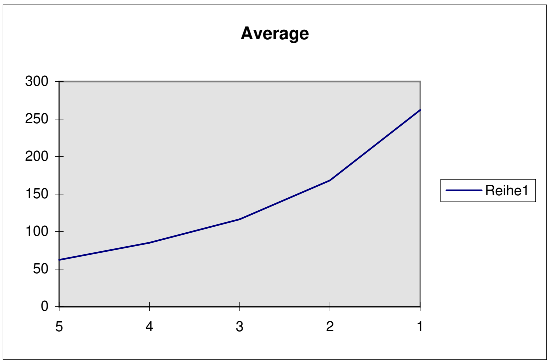


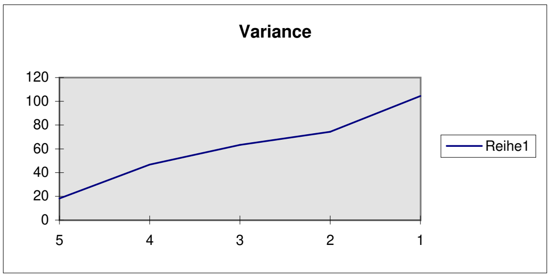
C.3 Support Vectors

Spline SVs

SVs for Splines

Average	Order
Epsilon ↓	13
5	62.3438
4	85.1667
3	116.312
2	168.188
1	261.927
Best Validation Choice	
1.85	187.83
Best Possible Choice	
2.02	181.91
Variance	Order
Epsilon ↓	13
5	18.1423
4	46.6389
3	63.2982
2	74.3607
1	104.547
Best Validation Choice	
1.85	2762.54
Best Possible Choice	
2.02	3882.34





References

[Bre94] Leo Breiman. Bagging predictors. Technical report, Department of Statistics, University of California, Berkeley, California 94720, September 1994. Also at ftp://ftp.stat.berkeley.edu/pub/tech-reports/421.ps.Z.

[DBK+97] Harris Drucker, Chris J.C. Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. Support vector regression machines. In Michael C. Mozer, Michael I. Jordan, and Thomas Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, page 155. The MIT Press, 1997.

[HR78] D. Harrison and D. L. Rubinfeld. Hedonic prices and the demand for clean air. In *J. Environ. Economics & Management*, volume 5, pages 81–102, 1978. Supposed original source, data actually from ftp://ftp.ics.uci.com/pub/machine-learning-databases/housing.

[SWG+96] M. O. Stitson, J. A. E. Weston, A. Gammerman, V. Vovk, and V. Vapnik. Theory of support vector machines. Technical Report CSD-TR-96-17, Royal Holloway, University of London, December 1996.

[Vap82] Vladimir N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer, New York, 1982.

[Vap95] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

[Vapng] Vladimir N. Vapnik. *Statistical Learning Theory*. J. Wiley, forthcoming.

[VC79] Vladimir N. Vapnik and A. J. Cervonenkis. *Theorie der Zeichenerkennung*. Akademie-Verlag, Berlin, 1979. Translated from the Russian original (1974).

[VGS97] Vladimir Vapnik, Steven E. Golowich, and Alex Smola. *Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing*. private communications, 1997.