

Adaptive Splines and Genetic Algorithms

Jennifer Pittman, Department of Statistics
The Pennsylvania State University, `pittman@stat.psu.edu`

Key Words: genetic algorithms, adaptive splines, nonparametric regression

Abstract:

Most existing algorithms for fitting adaptive splines are based on non-linear optimization and/or stepwise selection. Although computationally fast and spatially adaptive, stepwise knot selection is necessarily suboptimal while determining the best model over the space of adaptive knot splines is a very poorly behaved non-linear optimization problem. A possible alternative is to use a genetic algorithm to perform knot selection.

A spatially adaptive modeling technique referred to as genetic adaptive splines is introduced which combines the optimization power of a genetic algorithm with the flexibility of polynomial splines. Preliminary simulation results comparing the performance of the genetic algorithm method to other current methods are presented.

1 Introduction

In the univariate case, the dataset $\{(x_i, y_i) : i = 1, \dots, N\}$ in \mathcal{R}^2 , $a \leq x_1 < x_2 < \dots < x_N \leq b$, $a, b \in \mathcal{R}$, is assumed to be a number of realizations of some underlying process combined with random noise, i.e. $\mathbf{y} = f(\mathbf{x}) + \epsilon$, where the ϵ_i s are assumed to be independent and follow some distribution with mean zero and finite variance. In the case where little is known about the function f , the modeling technique employed should be *flexible*. A popular and rapidly developing class of such modeling techniques are spatially adaptive smoothing methods. Traditionally, spatially adaptive smoothing methods could be classified into one of two groups. The first group consists of methods which use a locally adaptive smoothing parameter with smoothing splines or kernel techniques, e.g., Fan and Gijbels [3]. The second group consists of regression spline fitting algorithms where the knot locations are chosen adaptively, e.g., nonlinear optimization routines such as Schwetlick and Schütze's [5]. Other applications of statistical variable selection techniques to adaptive splines include Friedman's [6] MARS which

utilizes stepwise knot selection with a generalized cross-validation (GCV) model selection criteria, and the recent contributions to additive modeling made by Luo and Wahba [7] [HAS] and Stone, Hansen, Kooperberg, and Truong [8] [POLYMARS].

The algorithms mentioned are based on non-linear optimization and/or stepwise selection. Although computationally fast and spatially adaptive, stepwise knot selection is necessarily suboptimal while determining the best model over the space of adaptive knot splines is a very poorly behaved non-linear optimization problem [9]. A possible alternative is to use a genetic algorithm (GA) to perform knot selection. Genetic algorithms are stochastic search methods which, as discussed below, have the potential to find models that are more appropriate in comparison to models selected using stepwise or nonlinear optimization techniques.

2 Genetic Algorithms

Originally developed by Holland [11], *genetic algorithms* are stochastic search methods which provide a near optimal solution to the evaluation function of an optimization problem. In each iteration t , a GA starts with a population P^t of M potential solutions. Each solution x in the domain space D is encoded as a string or chromosome S , hence $P^t = \{S_1, \dots, S_M\}$. A string is built of elements from a finite alphabet $A = \{\alpha_1, \alpha_2, \dots, \alpha_a\}$; the alphabet is determined by the encoding scheme. The length L of each string is determined by the number of parameters that need to be determined and the desired precision. The GA seeks to optimize a specified evaluation function $fit(x)$, $x \in D$. If x is represented by string S , then $fit(x)$ is the *fitness value* of S .

2.1 The Basic Steps

Starting with a randomly generated population, at each iteration the algorithm applies chosen operators to the given population to yield a new population (the next *generation*). The standard operators are

- **Selection:** Selection gives members of the present population P^t with large fitness values

an increased chance of being present in an intermediate population P_1^t . For example, in ordinal (ranking) selection, the strings are sorted according to their fitness values and $p_s(S_i)$ is based on the rank of S_i using a non-increasing assignment function.

- **Crossover:** Crossover allows pairs of strings from P_1^t to combine features to create improved strings for the next intermediate population P_2^t . For example, let $S_1 = (\gamma_{1,1}, \dots, \gamma_{1,L})$ and $S_2 = (\gamma_{2,1}, \dots, \gamma_{2,L})$ be two strings from P_1^t selected for crossover, where the probability of undergoing crossover for any pair of strings is p_c . In *simple* crossover [11], a position $i \in \{1, 2, \dots, L-1\}$ is randomly chosen and two new chromosomes are built

$$S'_1 = (\gamma_{1,1}, \dots, \gamma_{1,i}, \gamma_{2,i+1}, \dots, \gamma_{2,L})$$

$$S'_2 = (\gamma_{2,1}, \dots, \gamma_{2,i}, \gamma_{1,i+1}, \dots, \gamma_{1,L})$$

S'_1 and S'_2 are placed in P_2^t and S_1 and S_2 are discarded.

- **Mutation:** Mutation gives the algorithm an opportunity to branch into previously unexplored regions of the domain space by arbitrarily altering one or more characters of a selected string. Each character of every string will undergo mutation with probability p_m . In the simplest case, for each character $\gamma_{i,j}$, $i = 1, \dots, M$; $j = 1, \dots, L$ from P_2^t a random number rnd is generated from $[0, 1]$. If $rnd > p_m$, $\gamma_{i,j}$ is unchanged; otherwise, $\gamma_{i,j}$ is replaced by a randomly selected member of $\{A - \gamma_{i,j}\}$.

P_2^t is now relabeled as $P^{(t+1)}$ and the cycle of operations is repeated until some termination criterion is met, at which time the best string achieved is generally taken as the solution to the optimization problem. An additional selection strategy referred to as an *elitist* step [10], where the best individual from the present population is included in the subsequent population, is often incorporated.

2.2 Why GAs?

In the present problem, it is necessary to find not only the proper knot placement but also the proper number of knots. For a fixed number of knots, the sum of squares error is a nonconvex function of the knot sequence and hence a traditional derivative-based approach may get stuck in a local minimum. With GAs, one does have theoretical convergence to the global optimum. In order for this convergence to occur, good initial estimates of the knot locations are not required.

Determining the proper ‘free’ knot spline for a given dataset can be viewed as a variable selection problem: given a large set of candidate knot locations and a criterion of fit, find the subset of knots of a certain size which yields the best fit. Exhaustive enumeration is not an option while stepwise procedures are necessarily suboptimal [12]. Hence GAs, by performing a directed search over the model space without resorting to stepwise methods, have the potential to better address this sort of problem.

Given the above remarks a note of caution about GAs is in order. Although their convergence to the global optimum has been proven, the choice of the various algorithm parameters does affect the rate of convergence. The nature of this dependence is generally unknown so the selection of parameter values to achieve the best performance can be a difficult task. GAs are very computer intensive and hence slower than most existing methods. Finally, different runs of the same GA can lead to different results (although the results are usually reasonable solutions). Hence the nature of the current problem may motivate the development of a genetic algorithm based method, but GAs may not be an appropriate optimization tool in other modeling contexts.

3 Current Problem

Recall the univariate dataset described in Section 1. The goal will be to approximate the function f so that a criterion based on SSE or weighted SSE is minimized. Suppose $\{\tau_i\}_{i=1}^{k+2}$ is a strictly increasing sequence of points in $\{x_i\}_{i=1}^N$ with $\tau_1 = a$, $\tau_{k+2} = b$, $k_{\min} \leq k \leq k_{\max}$, k_{\min} , k_{\max} given. Let $\{P_i\}_{i=1}^{k+1}$ be a sequence of polynomials of order m . The model space may be represented as

$$\mathcal{P}_{m,\tau} = \left\{ g : g(x) = P_i(x) \text{ if } \tau_i < x < \tau_{i+1}, \right. \\ \left. i = 1, \dots, k+1; \text{ the first } (m-1) \text{ derivatives} \right. \\ \left. \text{of } g \text{ at } \tau_i \text{ are continuous; } \{\tau_i\}_{i=1}^{k+2} \text{ and } \{P_i\}_{i=1}^{k+1} \right. \\ \left. \text{are given sequences, } k_{\min} \leq k \leq k_{\max} \right\}$$

For computational efficiency, each element $g \in \mathcal{P}_{m,\tau}$ will be represented as a linear combination of normalized B-spline basis functions of order m . For given $\{\tau_i\}_{i=1}^{k+2}$ define $\mathbf{t} = (t_1, t_2, \dots, t_{n+m})$ as

$$t_1 = \dots = t_m = \tau_1 = a \leq t_{m+1} \leq \dots$$

$$\dots \leq b = \tau_{k+2} = t_{n+1} = \dots = t_{n+m},$$

$n = m + k$. Then $\mathcal{P}_{m,\tau}$ may be expressed as

$$\mathcal{S}_{m,\mathbf{t}} = \left\{ g \in \mathcal{P}_{m,\tau} : g = \sum_{j=1}^n \alpha_j B_{j,m,\mathbf{t}}, \right. \\ \left. \alpha_j \in \mathcal{R}, j = 1, \dots, n \right\}$$

where $\{B_{j,m,\mathbf{t}}\}_{j=1}^n$ represents the sequence of normalized B-splines of order m with respect to the knot sequence \mathbf{t} .

3.1 Model Criterion

Traditionally, the amount of fit associated with a given model is represented by the number of ‘degrees of freedom’ it employs. In spline fitting, the definition $df = \text{tr}(\mathbf{S})$ for ‘degrees of freedom’ is popular in the spline literature; this choice yields the original GCV criteria,

$$GCV(g) = (RSS_g/N)/(1 - \text{tr}(\mathbf{S})/N)^2$$

where RSS_g is the residual sum of squares from the fit of the model g to the data and \mathbf{S} is the smoothing matrix corresponding to g . However, if the ‘degrees of freedom’ is viewed as a measure of the amount of fit or the cost of the estimation process, then adaptively selected knots should employ more ‘degrees of freedom’ relative to nonadaptively selected knots. This motivates adjusting the above criteria by altering the definition of ‘degrees of freedom’. This leads to a statistic of the form

$$GCV(g_n) = (RSS_{g_n}/N)/(1 - (n_1 + (n_2 * d))/N)^2$$

as used in MARS modeling [6] with $d = 3$, where n_2 of the n basis functions of which g is composed are adaptively selected, $n = n_1 + n_2$. Luo and Wahba [7] provide arguments for the use of $d = 1.2$ for reproducing kernel cubic spline bases. Although the procedure is not stepwise, genetic algorithms also employ an adaptive procedure for the selection of basis functions. For this reason, and for the purpose of comparison with existing methods, a similar criterion will be utilized here.

Whether adjusted GCV is an appropriate model criterion is open to further examination. If one accepts the choice of adjusted GCV, a value for d must be determined. The previously proposed values are based on arguments that depend upon the choice of basis functions and the nature of the modeling procedure [6, 7]. Thus they cannot be directly applied to the proposed method. A possible solution is the concept of generalized degrees of freedom (GDF) of Ye [13] for complex modeling procedures which may be utilized to determine d in GCV for the GA.

3.2 Convergence

Bhandari, Murthy, and Pal [10] modeled a GA with elitist step (EGA) as a finite state Markov chain and proved, under various conditions, that an EGA will converge to the global optimal solution if that solution is contained in the search space.

For an outline of their proof, assume a GA with finite population size M , string length L , and alphabet A of cardinality a . Let P represent a possible

population and let \mathcal{P} represent the class of all possible populations. The fitness value of a population, $fit(P)$, is defined as $fit(P) = \max_{S \in \mathcal{P}} fit(S)$. Let $\{F_1, \dots, F_s\}$ represent the set of possible fitness values, $s \leq a^L$, where $F_1 > F_2 > \dots > F_s$. Define $E_i = \{P : P \in \mathcal{P} \text{ and } fit(P) = F_i\}$ $i = 1, \dots, s$; note $E_i \cap E_j = \emptyset$ and $\bigcap_{i=1}^s E_i = \mathcal{P}$. Let P_{ij} be the j th population of E_i , $i = 1, \dots, e_i$ and $i = 1, \dots, s$. Denote as $p_{ij.kl}$ the probability that the genetic operators, in one generation, result in a population $P_{k,l} \in E_k$ from $Q_{i,j} \in E_i$, and let $p_{i.k} = \sum_{l=1}^{e_k} p_{ij.kl}$, $j = 1, \dots, e_i$, $i, k = 1, \dots, s$. Under the assumption that $\min_{i,j} p_{ij.1} > 0$, it was proved that

$$\lim_{n \rightarrow \infty} p_{ij.1}^{(n)} = 1 \quad \forall j = 1, \dots, e_i \text{ and } i = 1, \dots, s$$

where $p_{ij.1}^{(n)}$ denotes the probability of reaching a population in E_1 in n generations with the starting population $P_{i,j}$.

A topic of future research is to verify that this proof holds for the specific genetic algorithm used in GAS.

3.3 Genetic Adaptive Splines

The genetic adaptive spline program was designed to determine the model from the space $\mathcal{S}_{m,t}$ which minimizes an adjusted GCV criterion. Given k , $k_{\min} \leq k \leq k_{\max}$, the GA will adaptively select candidate interior knot sequences of size k from $\{x_i\}$ for consideration. Each candidate knot sequence has a corresponding set of B-spline basis functions; the (weighted) least squares coefficients $\{\alpha_j\}_{j=1}^n$ of these basis functions are determined by the algorithm L2MAIN of de Boor [4]. The result of the GA is the least squares spline model of size k with the smallest (weighted) RSS. The execution of the GA for each value of k yields a series of models $\{g_k : k = k_{\min}, \dots, k_{\max}\}$; the appropriate model from this group is chosen by minimizing the GCV score described above.

In GAS, integer coding will be used, e.g., the character 2 will represent an interior knot located at x_2 . The version of the integer coded genetic algorithm (ICGA) which has been implemented uses linear ranking selection with stochastic sampling with replacement. The performance of standard linear ranking selection of Baker [14] with the efficient technique of stochastic universal sampling will be tested in future studies.

In the initial implementation, simple crossover will be utilized; crossover operators developed for RC-GAs (real coded genetic algorithms) [16] will be tested in future simulations.

The works of several authors [15, 16] suggest an adaptive mutation scheme, e.g., applying different mutation densities and varying the probability of mutation during the GA’s generations. Hence a triangular mutation scheme was utilized [17] and the mutation probability p_m^t varied from 0.9 to $1/L$ (this choice is supported by the work of Bäck [18]). An elitist step is included to ensure that the current best solution is retained as the algorithm progresses.

3.4 Simulation studies

Simulated examples were used to examine the performance of GAS compared to the MARS, HAS, POLYMARS, and wavelet shrinkage methods. The simulation studies performed here were modeled directly after those designed by Luo and Wahba [7] to facilitate comparison with existing methods. Due to space constraints, only results for 3 of 5 examples will be shown. Table 1 gives information about the simulated datasets. Example 1 is taken from Donoho and Johnstone [1] and shows considerable spatial inhomogeneity, Example 3 is from Schwetlick and Schütze [5], and Example 4 is from Fan and Gijbels [3] and shows less spatial variability. Gaussian noise was used for all examples except Example 3 where the noise was Uniform. To facilitate comparison with wavelet methods, the sample sizes were all powers of 2 and the designs were equally spaced (except for Example 2).

For wavelet shrinkage the SUREShrink method of Donoho and Johnstone [2] was selected with a “primary resolution level” of 5. Computations were performed with the *wavethresh* software of Nason and Silverman in S-PLUS, with the S-PLUS commands provided by Luo and Wahba. The family of wavelets was *DaubLeAsymm* with *filter number* 8.

The maximum number of basis functions in HAS, MARS, and POLYMARS was set at 150 for ex.1 and at 60 for ex.3 and ex.4. The number of basis functions considered by GAS was [18, 34] for ex.1 and [7, 16] for ex.3 and ex.4 (The number of basis functions fit by GAS was near 21 for ex.1 and 9 for ex.3 and ex.4). The remaining parameters in HAS, MARS, and POLYMARS were set at their default values except for the parameter *gcv* in POLYMARS, which was set at 2.5 as in Stone, Hansen, Kooperberg, and Troung [8]. The IDF factor for GAS was set at 3.

GAS was set to fit cubic splines with a maximum of 400 generations, a crossover probability of 0.8, and the same mutation probability density (as described above) for all examples. The population size M was 50 for ex.1 and 40 for ex.3 and ex.4.

The median MSE and the difference between the 1st and 3rd quartiles of the MSE for ex.1, ex.3 and ex.4 and each method are given in Table 2; the median results for ex.1 and ex.3 and all methods are shown in Figures 1 and 2.

The performance of GAS compares quite favorably with the results of SUREShrink and the other adaptive spline algorithms. On all examples GAS was best, most likely due to the directed global selection of basis functions. For ex.1 and ex.4, the results of HAS, SUREShrink, and MARS reflect those shown in Luo and Wahba [7]. For ex.3, MARS and GAS show the best results, followed closely by HAS and POLYMARS and then by SUREShrink. The selection of a lower “primary resolution level” for SUREShrink did not yield a better performance.

Despite the relatively superior performance of GAS on the above examples, the method does have several implementation issues to be resolved.

- Due to the global nature of the GA search, it is relatively slow and computationally expensive (e.g., for ex.1 with $N=2,048$, each model size took approximately 90 sec.; this limits the number of model sizes which can be considered. Hopefully its efficiency can be improved by the use of more ICGA specific genetic operators (as mentioned above) and improved programming.
- Crossover and mutation can create models with coincident knots. It is possible that this can be avoided by incorporating constraints into the GA model as in Michalewicz and Janikow [19].

4 Summary

A modeling technique has been proposed for fitting adaptive splines. The basis functions are B-spline basis functions of order m ($m < 20$); the method can be used to fit adaptive splines which minimize a weighted or unweighted adjusted GCV criterion. For each candidate model size, the search for the minimum (weighted) SSE model is performed by a genetic algorithm which adaptively selects the appropriate knot sequence.

References

- [1] Donoho, D.L. and Johnstone, I.M. (1994), “Ideal spatial adaptation by wavelet shrinkage”, *Biometrika*, **81**, pp. 425-455.
- [2] Donoho, D.L. and Johnstone, I.M. (1995), “Adapting to unknown smoothness via wavelet shrinkage”, *JASA*, **90**, pp. 1200-1224.

Table 1: Simulated Examples

<i>Example</i>	<i>f</i>	<i>Sample</i>		<i>Number of replicates</i>	
		σ	size (N)		SD(<i>f</i>)/ σ
1	DJ(1994) Heavisine*2.2	1.0	2,048	6.54	31
3	$10(4x - 2)/(1 + (100 * ((4x - 2)^2)))$	0.06	128	3.33	400
4	$\sin(2(4x - 2)) + 2 \exp(-16(4x - 2)^2)$	0.3	256	2.80	400

Table 2: Median MSE (Difference between First and Third Quartiles of MSE)

<i>Example</i>	GAS	HAS	SUREShrink	MARS	POLYMARS
1	.0195 (.0065)	.0412 (.0085)	.0702 (.0397)	.1518 (.0134)	.2483 (.0067)
3	.0004 (.0002)	.0006 (.0003)	.0013 (.0003)	.0009 (.0009)	.0016 (.0002)
4	.0052 (.0035)	.0071 (.0059)	.0178 (.0038)	.0070 (.0038)	.0087 (.0034)

- [3] Fan, J., and Gijbels, I. (1995), "Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaptation", *JRSS, Ser. B.*, **57**, pp. 371-394.
- [4] De Boor, C. (1978), *A practical guide to splines*, 1st Edition. New York: Springer-Verlag.
- [5] Schwetlick, H., and Schütze, T. (1995) "Least squares approximation by splines with free knots", *BIT*, **35**, pp. 361-384.
- [6] Friedman, J.H. (1991), "Multivariate adaptive regression splines (with discussion)", *Ann. Stat.*, **19**, pp. 1-141.
- [7] Luo, Z., and Wahba, G. (1997), "Hybrid adaptive splines", *JASA*, **92**, pp. 107-115.
- [8] Stone, C.J., Hansen, M., Kooperberg, C., and Troung, Y. (1997), "Polynomial splines and their tensor products in extended linear modeling (with discussion)", *Ann. Stat.*, **25**, 4, pp. 1371-1470.
- [9] Wahba, G. (1988), in discussion to J. Ramsay, "Monotone regression splines in action", *Stat. Sci.*, **3**, pp. 425-462.
- [10] Bhandari, D., Murthy, C.A., and Pal, S.K. (1996), "Genetic algorithm with elitist model and its convergence", *Int. J. Patt. Recog. Art. Intell.*, **10**, 6, pp. 731-747.
- [11] Holland, J.M. (1975) *Adaptation in natural and artificial systems*, Ann Arbor, MI: The University of Michigan Press.
- [12] S. Chatterjee, M. Laudato, and L.A. Lynch, "Genetic algorithms and their statistical applications: an introduction", *Comp. Statist. and Data Anal.*, **22**, 6, pp. 633-651, Oct. 1996.
- [13] Ye, J. (1998), "On measuring and correcting the effects of data mining and model selection", *JASA*, **93**, 441, pp. 120-131.
- [14] Baker, J.E. (1985), "Adaptive selection methods for genetic algorithms", *Proc. of an Int. Conf. on Genetic Algorithms*, L. Erlbaum Associates (eds.), Hillsdale, MA. pp. 101-111.
- [15] Bramlette, M.F. (1991), "Initialization, mutation, and selection methods in genetic algorithms for function optimization". *Proc. of Fourth Int. Conf. on Genetic Algorithms*, Belew, R., and Booker, L.B. (eds.). San Mateo. CA: Morgan Kaufmann Publishers, pp. 100-107.
- [16] Herrera, F., Lozano, M., and Verdegay, J.L.(1998), "Tackling real-coded genetic algorithms: Operators and tools for behavioral analysis", *Artificial Intelligence Review*, to appear.
- [17] Murthy, C.A. (1998), *personal communication*.
- [18] Bäck, T. (1993), "Optimal mutation rates in genetic search", *Proc. of Fifth Int. Conf. on Genetic Algorithms*, Forrest, S. (Ed.), San Mateo. CA: Morgan Kaufmann Publishers, pp. 2-8.
- [19] Michalewicz, Z., and Janikow, C. (1991), "Handling constraints in genetic algorithms", *Proc. of Fourth Int. Conf. on Genetic Algorithms*, Belew, R., and Booker, L.B. (eds.). San Mateo. CA: Morgan Kaufmann Publishers, pp. 151-157.
- [20] Rogers, D. (1991), "G/SPLINES: A hybrid of Friedman's Multivariate Adaptive Regression Splines (MARS) Algorithm with Holland's Genetic Algorithm". In R.K. Belew & L.B. Hooker (eds.), *Proceedings of the Fourth International Conference on Genetic Algorithms*, San Mateo, CA: Morgan Kaufmann.

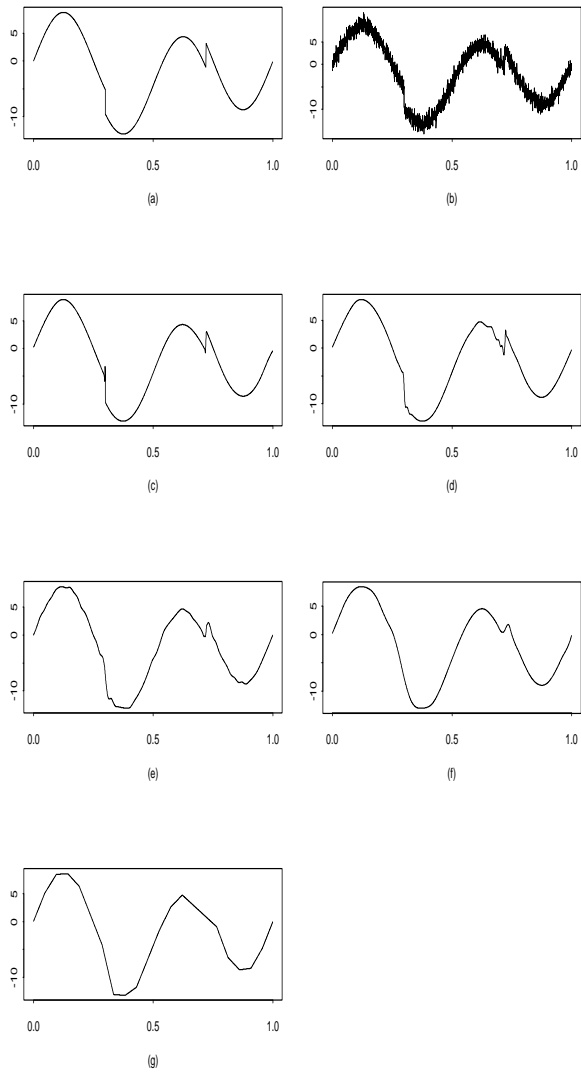


Figure 1: Ex1 (a) original function; (b) sample dataset; (c) GAS fit; (d) HAS fit; (e) SUREShrink fit; (f) MARS fit; (g) POLYMARS fit

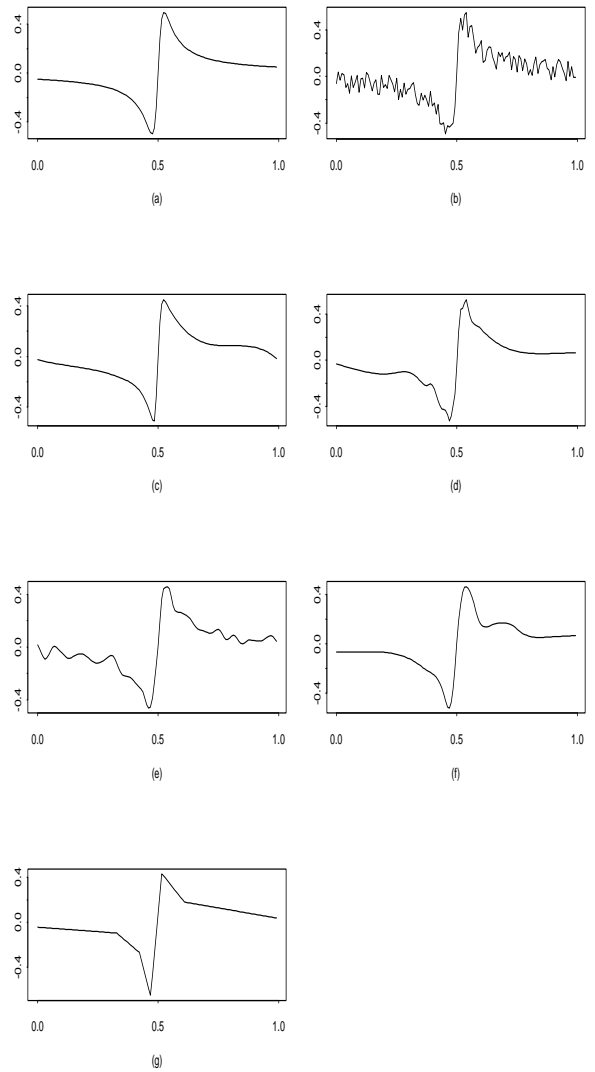


Figure 2: Ex3 (a) original function; (b) sample dataset; (c) GAS fit; (d) HAS fit; (e) SUREShrink fit; (f) MARS fit; (g) POLYMARS fit

