

# Estimating Partially Linear Models Using Wavelets: A Nonlinear Backfitting Algorithm

Leming Qu  
Department of Mathematics  
Boise State University  
Boise, ID, 83725  
qu@math.boisestate.edu

## Abstract

Partially linear models have a linear part as in the linear regression and a nonlinear part similar to that in the nonparametric regression. The estimates in partially linear models have been studied previously in traditional smoothing methods such as smoothing spline, kernel and piecewise polynomial methods. In this paper, we apply the regularized wavelet estimators by penalizing the  $l_1$  norm of the wavelet coefficients of the nonparametric function. The regularization parameter is chosen by universal threshold. When the linear part has multivariate predictors, we developed an iterative algorithm similar to backfitting based on the necessary and sufficient conditions of the minimum point. Simulation results confirmed the good performance of the regularized wavelet approach.

## Keywords

Soft thresholding; Wavelet; DWT; Universal threshold; Regularization; Partially Linear Models.

## 1 Introduction

Considering observations  $\{(y_i, x_i, t_i), i = 1, \dots, n\}$  from the following model:

$$y = x'\beta + f(t) + \varepsilon \quad (1)$$

where  $x$  is a fixed known  $p$ -dimensional vector,  $\beta$  is an unknown  $p$ -dimensional parameter vector,  $f$  is an unknown function,  $t$  is a scalar such as the time at which the observation is made,  $\varepsilon$  is the random error usually assumed to be  $N(0, \sigma^2)$  distributed. Without loss of generality, assume  $t \in [0, 1]$ .

For the  $n$  observed data points, in vector-matrix notation, the model is written as

$$Y = X\beta + F + \epsilon$$

where  $Y = (y_1, \dots, y_n)'$ ,  $X' = [x_1, \dots, x_n]$ ,  $F = (f(t_1), \dots, f(t_n))'$  and  $\epsilon = (\varepsilon_1, \dots, \varepsilon_n)'$ .

This model has received a considerable amount of research in the past two decades. One reason is that it is much more flexible than the standard linear model since it combines both parametric and nonparametric components. Another reason is that it allows easier interpretation of the effect of each variable compared

to a completely nonparametric regression. Because of its relation to the classical linear model, this model is called “partially linear model”. Engle et. al. [6] were among the first to apply this model in analyzing the relationship between weather and electricity sales. Most recently, a whole new book [12] is dedicated to this model.

All the existing approaches are based on different nonparametric regression procedures. Among the most important are spline method by Green et al. [10], Wahba [25], Green and Silverman [11], Eubank [7], and Schimek [23]; kernel method by Speckman [24] and Robinson [22]; piecewise polynomials method by Chen [3]; and local linear smoothing method by Hamilton and Truong [13]. Important assumptions by all the existing approaches for  $f(t)$  are its continuity and smoothness. But in reality, these assumptions may not be satisfied. In areas like signal and image processing, objects are frequently inhomogeneous.

Wavelet nonparametric regression methods have become a powerful tool in the past decade since Donoho and coauthors’ pionnering work [4], [5]. There are several advantages for wavelet shrinkage estimator. It is nearly minimax for a wide range of loss functions and for general function classes. It is adaptable to various ranges of unknown smoothness. It is simple, practical and fast due to efficient algorithms.

For Partially Linear Models, applying wavelet nonparametric regression methods to the estimation of  $\beta$  and  $f(t)$  is a natural extension to the traditional methods. Less restrictive hypotheses of degrees of smoothness of the underlying function  $f(t)$  are made. Qu [21] proposed regularized wavelet estimation in partially linear models for the special case when dimension  $p = 1$ . Here we deal with the general case for  $p > 1$ .

The paper is organized as follows. In section 2 we give a brief review about the wavelet nonparametric regression and Discrete Wavelet Transform (DWT). We discuss the necessary and sufficient conditions for the regularized wavelet estimator in partially linear models in section 3. In section 4, we propose a nonlinear back-fitting algorithm based on the necessary and sufficient conditions and present some simulation results. We conclude the paper with some comments and suggestions for future research in the last section.

## 2 Wavelet Nonparametric Regression

The classical nonparametric regression problem is to recover  $f(t)$  after observing data

$\{(y_i, t_i), i = 1, \dots, n\}$  from the standard Gaussian “signal-plus-noise” model

$$y = f(t) + \epsilon$$

where  $f(t)$  is an unknown function and  $\epsilon \sim N(0, \sigma^2)$ . If  $\beta = 0$  in model (1), then the above model is obtained. For the sake of DWT, it is further assumed that  $t_i$ ’s are equally spaced and  $n$  is power of 2.

For the  $n$  observed data points, in vector-matrix notation, the model is written as

$$Y = F + \epsilon$$

where  $Y = (y_1, \dots, y_n)'$ ,  $F = (f(t_1), \dots, f(t_n))'$  and  $\epsilon = (\epsilon_1, \dots, \epsilon_n)'$ .

The DWT can be represented by an orthonormal matrix  $W$ . Then  $w = WY$  performs the DWT on the noisy data. Let  $\theta = WF$  be the wavelet transform of  $F$ , then  $F = W'\theta$  with  $W'$  the inverse discrete wavelet transform (IDWT). Then, the

observed data can be expressed as a linear model on the wavelet domain

$$Y = W'\theta + \epsilon \quad (2)$$

The ordinary Least Squares estimate is simply  $\hat{\theta}_{LS} = WY$ , the empirical wavelet coefficients. The  $\hat{\theta}_{LS}$  is an unbiased estimate of  $\theta$  and its covariance matrix is  $\sigma^2 I$ . For a smaller Mean Squared Error of  $\hat{\theta}$ , the unbiasedness is sacrificed to obtain a better tradeoff between biasness and variance. Since wavelet coefficients of  $f(t)$  in a wide range of function space are usually sparse, it is preferred to penalize the  $l_1$  norm of  $\theta$ . For a given  $\lambda > 0$ , the solution to the minimization of regularized least squares:

$$\min_{\theta} 2^{-1} \|Y - W'\theta\|_2^2 + \lambda \|\theta\|_1$$

is the soft thresholding of  $w$ :

$$\hat{\theta}_S = \text{sign}(w)(|w| - \lambda)_+$$

where  $x_+$  is  $x$  for  $x > 0$  and zero otherwise. Usually, the scaling coefficients from the DWT are kept unchanged.

The choice of the regularization parameter or the threshold  $\lambda$  is crucial. There are several approaches in the literature. It can be chosen by universal threshold  $\lambda_{UV} = \sigma \sqrt{2 \log(n)}$ , or by minimizing Stein Unbiased Risk Estimate(SURE), or by the method of cross validation.

For DWT and IDWT, a fast algorithm developed by Mallat [18] is used to perform the transform in  $O(n)$  operation and matrix multiplication is avoided. However, use of the fast DWT and IDWT requires equally spaced  $t_i$ 's and  $n$  to be power of 2. This requirement is not a real restriction. Methods exist to overcome these limitations, allowing the DWT to be applied on unequally spaced data with any length.

The algorithm is easily implemented in S-Plus, a statistical and graphical computing environment. We use the *WaveThresh3* Software developed by Nason [19] running under S-Plus for our simulation.

### 3 Wavelet Estimation for Partially Linear Model

The idea of wavelet nonparametric regression in the above section can be generalized to the estimation of model (1). We assume equally spaced time points  $t_i = i/n$  and  $n$  is power of 2 in model (1). In the wavelet domain, the observed data can be expressed as a linear model:

$$Y = X\beta + W'\theta + \epsilon$$

If  $\beta$  is known, the model is the same as model (2) in the above section. So our focus here is to estimate  $\beta$ . By penalizing the  $l_1$  norm of  $\theta$ , for a given  $\lambda$ , one finds  $\beta$  and  $\theta$  which minimize the quantity:

$$l(\beta, \theta) = 2^{-1} \|Y - X\beta - W'\theta\|_2^2 + \lambda \|\theta\|_1$$

Then, by the orthogonality of the matrix  $W$ ,

$$\begin{aligned} l(\beta, \theta) &= 2^{-1} \|WY - WX\beta - \theta\|_2^2 + \lambda \|\theta\|_1 \\ &= 2^{-1} \|w - u\beta - \theta\|_2^2 + \lambda \|\theta\|_1 \end{aligned} \quad (3)$$

where  $u = WX$  is the DWT of the matrix  $X$ , which is the transformation of each column of  $X$ . Note that  $l(\beta, \theta)$  tends to infinity as  $|\beta| \rightarrow \infty$  or  $|\theta| \rightarrow \infty$ . Thus, minimizers of  $l(\beta, \theta)$  do exist. Obviously,  $l(\beta, \theta)$  is a convex functional of  $\beta$  and  $\theta$ , since it is the sum of two terms with each of them being a norm. Because of the triangular inequality, any norm is convex.

But the question whether  $l(\beta, \theta)$  is a strictly convex functional of both  $\beta$  and  $\theta$  is not so straightforward. It is well known that a convex functional has at least one minimum point, whereas its uniqueness is granted only in the case of strict convexity.

Taking the partial derivative of  $l(\beta, \theta)$  with respect to  $\beta$  and  $\theta_i$  at  $\theta_i \neq 0$  and setting them to zero leads to:

$$\begin{aligned} U'(w - U\beta - \theta) &= 0, \\ w_i - u'_i\beta - \theta_i - \lambda \text{sgn}(\theta_i) &= 0, \text{ for } \theta_i \neq 0, i = 1, \dots, n. \end{aligned}$$

The above equations characterize the minimum point(s) of  $l(\beta, \theta)$  according to classical results from differential calculus.

Recall the definition of soft threshold function

$$T_\lambda^s(d) = \begin{cases} d + \lambda & \text{if } d < -\lambda, \\ 0 & \text{if } |d| \leq \lambda, \\ d - \lambda & \text{if } d > \lambda. \end{cases}$$

It is understood that if  $d$  is a vector, then  $T_\lambda^s(d)$  is a vector operating componentwise.

Let  $\hat{\beta}$  and  $\hat{\theta}$  be a solution of the minimization problem. The following theorem characterizes the estimator by necessary and sufficient conditions:

**Theorem 1**

$$\{\hat{\beta}, \hat{\theta}\} = \text{argmin}_{\{\beta, \theta\}} l(\beta, \theta) = \text{argmin}_{\{\beta, \theta\}} 2^{-1} \|w - U\beta - \theta\|_2^2 + \lambda \|\theta\|_1$$

if and only if the following conditions hold:

$$U'(w - U\hat{\beta} - \hat{\theta}) = 0, \tag{4}$$

$$\hat{\theta} = T_\lambda^s(w - U\hat{\beta}) \tag{5}$$

The proof is similar to the *Theorem 2* of Alliney and Ruzinsky [1]. We omit the details here.

**Remark 1:** A slightly different result from Theorem 1 appeared in [8] on the context of penalized regressions comparing bridge versus Lasso, where the theorem is proved by mathematical induction on dimension  $p$ .

**Remark 2:** A more general result in the context of high dimensional generalized linear models appeared in [17], where the detailed proof of the theorem is not given, but it can be proved by formulating the target function as an equivalent constraint maximum likelihood problem and characterization of corresponding conditions.

**Remark 3:** The key difference between our result and the ones in [8] and [17] is that the dimension of parameters in our case (Theorem 1) is  $p + n$  which is greater than sample size  $n$  by a constant  $p$ , whereas the dimension of parameters in [8] and [17] is a fixed constant  $p$ .

## 4 A nonlinear backfitting algorithm

From the structure of equations of (4) and (5), we easily see that we can use an iterative scheme to find the solutions. Here we propose the following algorithm:

---

(i) start with  $\theta^0 = 0$ .

(ii) At step  $k$ ,

$$\beta^k = (U'U)^{-1}U'(w - \theta^{k-1}), \text{ i.e., the ordinary least squares estimator given } \theta^{k-1},$$

$$\theta^k = T_\lambda^s(w - U\beta^k)$$

(iii) Repeat (ii) until  $\beta^k$  converges.

---

The above algorithm is in spirit similar to the iterative *backfitting algorithm* [14] which is a general algorithm that enables one to fit an additive model using any regression-type fitting mechanisms. In the usual backfitting algorithm, the traditional linear smoother is used, while in our algorithm, the  $\theta$  is a nonlinear function of the data, hence the nonlinear backfitting.

The proof of the convergence of the above nonlinear backfitting algorithm needs further efforts. But numerical experiments as in the following simulation have been very promising. Most of the time, the algorithm converges within 10 iterations.

Note that  $\beta^1$  is simply the ordinary least squares estimator. In the trivial case, if  $|w - U\beta^1| < \lambda 1$ , then  $\theta^1 = 0$  and  $\beta^2 = \beta^1$ , and the  $\hat{\beta}$  is simply the OLS estimator. It means the  $\theta$  part should not be included into the model or  $\lambda$  is so big that the  $\theta$  is penalized too much.

A Monte Carlo simulation based on the algorithm using the universal threshold was carried out. For DWT, we used the Daubechies least-asymmetric orthonormal compactly supported wavelet with 10 vanishing moments which is the default wavelet for wavelet transform in *WaveThresh3*. All the calculations were carried out in S-plus 3.4 for Unix on IBM RS/6000.

For the nonparametric component we select different functions  $f(t) = 9f_0(t)$  with  $\max_{t \in [0,1]} f_0(t) = 1$  and

$$(F1) f_0(t) = 4.26(\exp(-3.25t) - 4\exp(-6.5t) + 3\exp(-9.75t)),$$

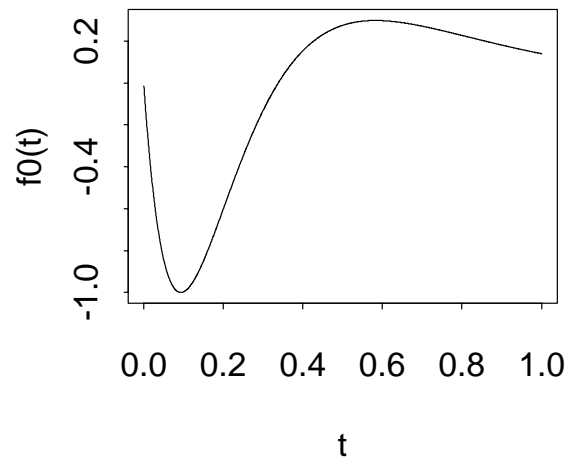
$$(F2) f_0(t) = \begin{cases} 4x^2(3 - 4x), & \text{if } 0 \leq x \leq 0.5 \\ \frac{4}{3}x(4x^2 - 10x + 7) - 1.5, & \text{if } 0.5 < x \leq 0.75 \\ \frac{16}{3}x(x - 1)^2, & \text{if } 0.75 < x \leq 1 \end{cases}$$

(F1) is a smoothing function that appeared in Schimek [23]. (F2) is a piecewise polynomial with discontinuity that appeared in Nason [20]. They are plotted in Fig (1)

For  $p = 2$ , we generated the  $x_{i1}$  from  $N(0, 1)$  and  $x_{i2}$  independently from  $N(0, 1)$ . This was the same setting as in Heckman [15]. For each set of  $x_i$ 's so chosen, we simulated a white noise with  $\sigma = 1$ . The sample sizes are  $n = 128, 256, 512$  and  $1024$ . We assumed the knowledge of  $\sigma$  in computing the regularization parameter  $\lambda_{UV}$ . Each simulated  $\{(y_i, x_i, t_i), i = 1, \dots, n\}$  was then used to estimate the true  $\beta_1, \beta_2$  and  $f(t)$  using the nonlinear backfitting algorithm. To eliminate effects that might be attributable to a particular choice of the  $x_i$ 's, we reselected the  $x_i$ 's for each simulation. We used the values for the regression coefficients  $\beta_1 = 0.5$  and  $\beta_2 = 1$ . For each setting and the four sample sizes 100 replicates were generated. Thus we obtained 100 estimated  $\beta_1$ 's and  $\beta_2$ 's in each setting.

The box plots of the estimated regression coefficients  $\hat{\beta}_1$  and  $\hat{\beta}_2$  are in Figure 2. From the box plots, we see that on average, we get fairly good estimates of  $\beta_1$  and  $\beta_2$ . Reduction of the range of estimates with growing sample size is clearly identified in Figure 2.

### Function: (F1)



### Function: (F2)

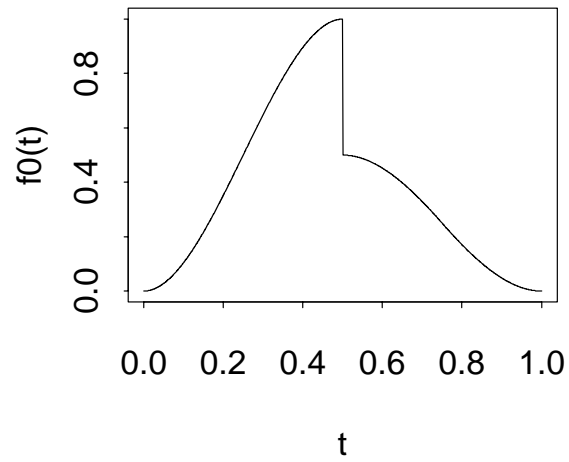


Figure 1: The Plots of  $f_0(t)$

The iteration usually finished in 3 or 4 steps for the simulated data. The convergence criteria we used was that the norm of the absolute error (i.e.,  $\|\beta^{k+1} - \beta^k\|$ ) was less than the prespecified precision  $10^{-4}$ .

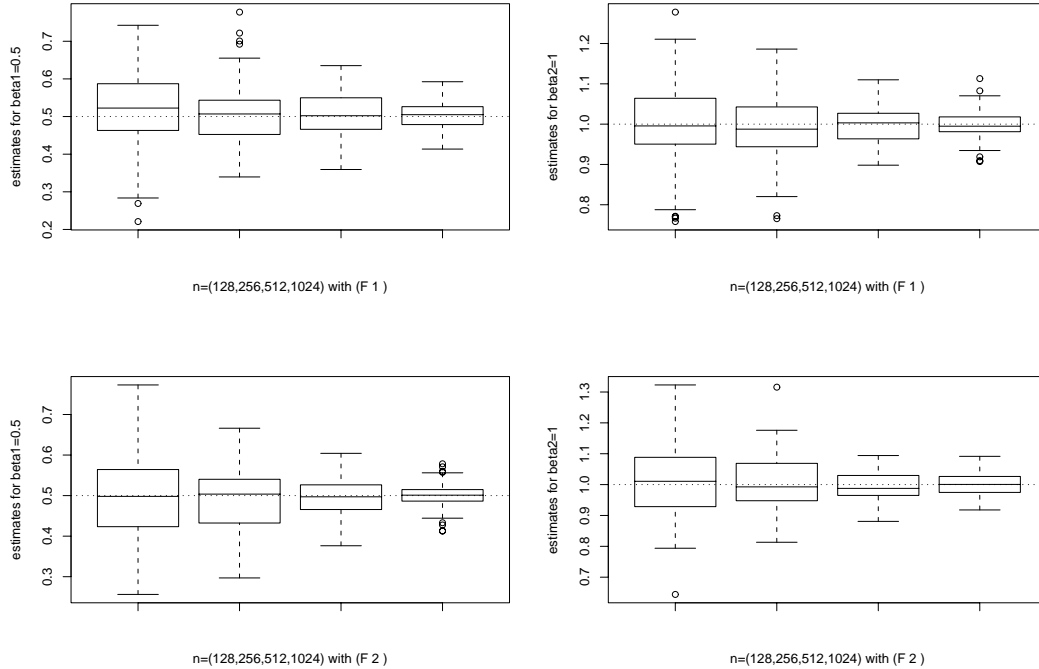


Figure 2: Plots of the  $\hat{\beta}$  by nonlinear backfitting algorithm

## 5 Conclusion

To formalize the notion of sparsity of the discrete wavelet transform of functions in a wide range of function spaces (such as Besov spaces), wavelet nonparametric regression usually penalizes the  $l_1$  norm of the population wavelet coefficients. For the regularized wavelet regression estimates of partially linear model, we derived the necessary and sufficient conditions for the minimum solution. Based on these conditions, we developed an iterative algorithm similar to backfitting. The regularization parameter is chosen by universal threshold or can be chosen by data-driven method such as cross validation. The Monte Carlo simulation results confirmed that the regularized wavelet estimators have good performance.

Further developments include methods for non-equally spaced designs. In the wavelet nonparametric regression context, many approaches have been developed to handle this situation. These methods are mostly based on interpolation or approximation, either in the original function domain or in the wavelet domain. Most recently, Antoniadis and Fan [2] introduced the nonlinear regularized wavelet estimators by using a large class of penalty functions. It will be of interest to extend this regularized Sobolev interpolator to the partially linear models.

In the real world application, we are most likely to encounter correlated data. Johnstone and Silverman [16] proposed level-dependent thresholdings for data with correlated noise in the wavelet nonparametric regression. This can be naturally extended to the partially linear regression settings. But it will be more computationally intensive.

In order to make inference about the linear coefficients, it is important to derive the asymptotic distribution of the regularized estimators. It is also of interest to study the asymptotic behaviour of the estimator for the nonparametric part. The rate of convergence of these estimators is also an important research topic.

**Acknowledgments** This paper is part of the author’s Ph.D dissertation at the Statistics Department of Purdue University. I would like to thank my advisor Prof. Mary Ellen Bock for guidance, support and encouragement.

## References

- [1] Alliney, Stefano and Ruzinsky, S.A. (1994) “An Algorithm for the Minimization of Mixed  $l_1$  and  $l_2$  norms with application to Bayesian Estimation”, *IEEE Transactions on Signal Processing*, 42, 3
- [2] Antoniadis, Anestis and Fan, Jianqing (2001) “Regularization of wavelets approximations”, *Journal of the American Statistical Association*, 96, 455
- [3] Chen, H., (1988), “Convergence rates for parametric components in a partly linear model”, *Ann. Statist.*, 16, 136-146.
- [4] Donoho, David L. , and Johnstone, Iain M. (1994), “Ideal spatial adaptation by wavelet shrinkage”, *Biometrika*, 81, 425-455
- [5] Donoho, David L. , Johnstone, Iain M. , Kerkyacharian, G. and Picard, D. (1995), “Wavelet shrinkage: Asymptopia? (Disc: p337-369)”, *J. Roy. Statist. Soc. Ser. B*, 57, 301-337
- [6] Engle, Robert F., Granger, C. W. J., Rice, John and Weiss, Andrew, (1986), “Semiparametric Estimates of the Relation Between Weather and Electricity Sales”, *Journal of the American Statistical Association*, 81, 310-320,
- [7] Eubank, R.L., (1988), *Spline Smoothing and Nonparametric Regression*, Marcel Dekker, New York.
- [8] Fu, Wenjiang, (1998), “Penalized regressions: the bridge versus the lasso”, *J. Comput. Graph. Statist.*, V7, 3, 397-416.
- [9] Gill, P.E., Murray, W. and Wright, M.H., (1981), *Practical Optimization*, San Diego: Academic Press.
- [10] Green, P., Jennison, C. and Seheult, A., (1985), “Analysis of field experiments by least squares smoothing”, *J. Roy. Statist. Soc. Ser. B*, 47, 299-315.
- [11] Green, P. and Silverman, B.W., (1994), *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman & Hall, London.
- [12] Hardle, Wolfgang, Liang, Hua and Gao, Jiti, (2000), *Partially linear models*, Physica-Verlag, Heidelberg.

- [13] Hamilton, Scott A.; Truong, Young K., (1997), “Local linear estimation in partly linear models”, *J. Multivariate Anal.*, 60, no. 1, 1-19.
- [14] Hastie, T. J. and Tibshirani, R. J., (1990) *Generalized Additive Models*, Chapman & Hall.
- [15] Heckman, N., (1986), “Spline smoothing in a partly linear model”, *J. Roy. Statist. Soc. Ser. B*, 48, 244-248.
- [16] Johnstone, Iain M. and Silverman, Bernard W., (1997), “Wavelet Threshold Estimators for Data With Correlated Noise”, *Journal of the Royal Statistical Society, Series B, Methodological*, 59, 319-351.
- [17] Klinger, Artur, (2001) “Inference in high dimensional generalized linear models based on soft thresholding”, *J. Roy. Statist. Soc. Ser. B*, 63, 377-392.
- [18] Mallat, S.G., (1989), “A theory for multiresolution signal decomposition: the wavelet representation”, *IEEE Trans. Pattn Anal. Mach. Intell.*, 11, 674-693
- [19] Nason, G.P. (1998), *WaveThresh3 Software*, Department of Mathematics, University of Bristol, Bristol, UK.
- [20] Nason, G.P. (1996), “Wavelet shrinkage using cross-validation”, *J. Roy. Statist. Soc. Ser. B*, 58, 463-479.
- [21] Qu, Leming (2001), “Regularized wavelet estimation in partially linear models”, *referred paper, Interface 2002*
- [22] Robinson, P.M., (1988), “Root-n-consistent semiparametric regression”, *Econometrica*, 56, 931-954.
- [23] Michael G. Schimek , (2000), “Estimation and inference in partially linear models with smoothing splines”, *Journal of Statistical Planning and Inference*, 91, 525-540
- [24] Speckman, P., (1988), “Kernel smoothing in partial linear models”, *J. Roy. Statist. Soc. Ser. B*, 50, 413-436.
- [25] Wahba, G., (1990), *Spline Models for Observational Data*, SIAM, Philadelphia, PA.