

# Regularized Wavelet Estimation in Partially Linear Models

Leming Qu

Department of Statistics, Purdue University, W. Lafayette, IN 47906  
quleming@stat.purdue.edu

## Abstract

The estimates in Partially Linear Models have been studied previously in traditional smoothing methods such as smoothing spline, kernel and piecewise polynomial smoothers. Here, 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. Simulation results show that regularized wavelet approach performs well. The wavelet method makes less restrictive assumptions about the smoothness of the underlying function for nonparametric part. The computational time is linear.

**Key Words:** Soft thresholding; Wavelet; DWT; Universal threshold; Regularization.

## 1 Introduction

Assume that responses  $y_1, \dots, y_n$  are observed at non-stochastic points  $0 \leq t_1 < \dots < t_n \leq 1$  of a predictor variable  $t$ . The response and predictor values are connected by a Partially Linear Model:

$$y_i = x_i' \beta + f(t_i) + \sigma z_i \quad (1)$$

for  $i = 1, \dots, n$ , where  $x_i$  are fixed known  $p$ -dimensional vectors,  $\beta$  is an unknown  $p$ -dimensional parameter vector,  $f$  is an unknown function, and the  $z_i$  are i.i.d. normal random variables with mean zero and variance one.

In vector-matrix notation, the model (1) is

$$Y = X\beta + f + \sigma Z$$

where  $Y = (y_1, \dots, y_n)'$ ,  $X' = [x_1, \dots, x_n]$ ,  $f = (f(t_1), \dots, f(t_n))'$  and  $Z = (z_1, \dots, z_n)'$ .

The parameter vector  $\beta$  and the function  $f$  need to be estimated efficiently.

For this model a number of approaches based on different smoothing techniques have been proposed. Among the most important are smoothing splines by Green et al. [6], Wahba [16], Green and Silverman [7], Eubank [4], and Schimek [14]; smoothing kernels by Speckman [15] and Robinson [13]; piecewise polynomials by Chen [1]; and local linear smoother by Hamilton and Truong [8]. All these methods require the continuity and smoothness of  $f$  on  $[0,1]$ .

Wavelet smoothing has become a powerful tool for nonparametric regression, see, for example, Donoho and Johnstone [2], Donoho et al. [3]. There are several advantages to wavelet shrinkage smoothing. It is nearly minimax for a wide range of loss functions and for general function classes. It is simple, practical and fast. It is adaptable to frequency inhomogeneities. For Partially Linear Models, applying wavelet smoothing technique to estimating parameters is a natural extension of the traditional methods. Here we propose a method using wavelets to estimate  $\beta$  and  $f$ . The computation time is linear. Less restrictive hypotheses of smoothness of the underlying function  $f$  are made.

## 2 Regularized Wavelet Nonparametric Regression

Consider first the nonparametric regression problem of estimating  $f$  in

$$y_i = f(t_i) + \sigma z_i$$

obtained from model (1) with  $\beta = 0$  and assume equally spaced sample points  $t_i = i/n$  and  $n$  is power of 2. The discrete wavelet transform can be represented by an orthonormal matrix  $W$ . Then  $w = WY$  performs the discrete wavelet transform(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

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

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, the unbiasedness is sacrificed for a smaller variance. To take advantage the sparsity of  $\theta$ , one penalizes the  $l_1$  norm of  $\theta$ . For a given  $\lambda > 0$ , the solution to the regularized least squares minimization of

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

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 wavelet transform are kept unchanged.

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

For DWT and IDWT, a fast algorithm developed by Mallat [10] is used to perform the transform in  $O(n)$  operation and matrix multiplication is avoided. However, use of the fast DWT and IDWT requires  $n$  to be power of 2, i.e.,  $n = 2^J$  for some integer  $J$ . This requirement is not a real restriction. Methods exist to overcome the limitation, allowing the DWT to be applied on any length of data.

The algorithm is easily implemented in S-Plus, a statistical and graphical computing environment. We use the *WaveThresh3* Software developed by Nason [11] running under S-Plus for our simulation.

## 3 Regularized Estimator in Partially Linear Model

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

$$Y = X\beta + W'\theta + \sigma Z$$

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

$$\text{argmin}_{\{\beta, \theta\}} 2^{-1} \|Y - X\beta - W'\theta\|_2^2 + \lambda \|\theta\|_1$$

We only study the case of dimension  $p = 1$  in this paper for simplicity, which is ultimately to find the root of a univariate nonlinear function. Solving multivariate nonlinear equations is more complicated. The case of  $p > 1$  is left for future research. Let's denote

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

Then, by the orthonormality of the wavelet transform  $W$ ,

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

where  $u = WX$  is the wavelet transform of the vector  $X$ . Note that  $l(\beta, \theta)$  tends to infinity as  $|\beta| \rightarrow \infty$ . Thus, minimizers of  $l(\beta, \theta)$  do exist. Let  $\hat{\beta}$  and  $\hat{\theta}$  be a solution. The following theorem characterizes the estimator by necessary and sufficient conditions:

**Theorem 1**

$$\{\hat{\beta}, \hat{\theta}\} = \operatorname{argmin}_{\{\beta, \theta\}} 2^{-1} \sum_i (w_i - u_i\beta - \theta_i)^2 + \lambda \sum_i |\theta_i|$$

if and only if the following conditions hold:

$$\sum_i u_i(u_i\hat{\beta} + \hat{\theta}_i - w_i) = 0, \quad (4)$$

$$\hat{\theta}_i = \operatorname{sign}(w_i - u_i\hat{\beta})(|w_i - u_i\hat{\beta}| - \lambda)_+, \text{ for each } i, \quad (5)$$

The proof of the theorem is given by an equivalent constraint minimization problem and characterization of corresponding conditions, as given for example in chapter 3 of Gill et al. [5]

One question arising from the Theorem 1 is: Is there a unique solution to the equations (4) and (5)? The following theorem answers this question:

**Theorem 2** For any given vectors  $w$  and  $u \neq 0$ , let  $s_i(\beta) = u_i(u_i\beta + \theta_i - w_i)$ , where  $\theta_i = \operatorname{sign}(w_i - u_i\beta)(|w_i - u_i\beta| - \lambda)_+$ , and  $s(\beta) = \sum_i s_i(\beta)$ , then the function  $s(\beta)$  has the following properties:

$$s(\beta) = -\lambda \|u\|_1, \text{ if } \beta \leq \beta_L = \min_{u_i \neq 0} \{(w_i - \operatorname{sign}(u_i)\lambda)/u_i\},$$

$$s(\beta) = \lambda \|u\|_1, \text{ if } \beta \geq \beta_R = \max_{u_i \neq 0} \{(w_i + \operatorname{sign}(u_i)\lambda)/u_i\},$$

$$s(\beta) \text{ is monotonely increasing in } [\beta_L, \beta_R].$$

The proof of the theorem is given in the Appendix.

From this theorem we immediately see that there is a unique solution  $\hat{\beta}$  of  $s(\beta) = 0$  in  $(\beta_L, \beta_R)$ . And the root  $\hat{\beta}$  can be easily found by simple bisection search numerical method. In practice, we choose an initial value  $\beta_0 = u'w/u'u$ , i.e, the usual least square estimate of  $w$  on  $u$  without interception, since usually  $\beta_0$  is near  $\hat{\beta}$ , while  $\beta_L$  and  $\beta_R$  are far from the  $\hat{\beta}$  and need more computation.

An important question is how to choose the regularization parameter  $\lambda$ . Here, we choose  $\lambda$  as the universal threshold  $\lambda_{UV} = \sigma \sqrt{2 \log(n)}$ .

## 4 Simulation

A Monte Carlo simulation based on the algorithm was carried out. We generated the  $x_i$  from  $N(0, 1)$ . This was the same setting as in Heckman [9]. The sample sizes are  $n = 128, 256, 512$  and  $1024$ . The standard deviation  $\sigma = 1$ . We assumed the knowledge of  $\sigma$  in computing the regularization parameter  $\lambda_{UV}$ . For the nonparametric component we selected two functions  $f(t) = m_n f_0(t)$  with :

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

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

(F1) is a smoothing function that appeared in Schimek [14]. (F2) is a piecewise polynomial with discontinuity that appeared in Nason [12]. We chose  $m_n = 9$  such that  $\max|f(t)| = 9$ .

We used the values for the regression coefficient  $\beta = \{0.5, 1, 1.5, 2, 2.5, 3\}$ . For each setting and the four sample sizes 100 replicates were generated. 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.

The box plots of the estimated regression coefficient  $\hat{\beta}$  are in Figure 1. From the box plots, we see that on average, we get fairly good estimates of  $\beta$ . Reduction of the range of estimates with growing sample size is clearly identified in Figure 1.

## 5 Conclusion

Summing up the Monte Carlo simulation results above, we conclude that the regularized wavelet estimator approach for the partially linear models works well under reasonable signal-to-noise ratios.

Further developments include data driven methods such as cross-validation to choose the regularization parameter  $\lambda$ ; methods for non-equally spaced designs; theoretical properties about the asymptotic behavior of the estimates; methods for high dimensional data.

### Appendix

First, let's prove the following Lemma:

**Lemma:** For any given scalars  $x \neq 0$  and  $y$ , let  $g(\beta) = x(x\beta + \theta - y)$ , where  $\theta = \text{sign}(y - x\beta)(|y - x\beta| - \lambda)_+$ , and  $\lambda$  is a given positive constant. Then the function  $g(\beta)$  has the following properties:

$$\begin{aligned} g(\beta) &= -\lambda|x|, & \text{if } \beta \leq \beta_1 &= (y - \text{sign}(x)\lambda)/x, \\ g(\beta) &= \lambda|x|, & \text{if } \beta \geq \beta_2 &= (y + \text{sign}(x)\lambda)/x, \\ g(\beta) &\text{ is monotonely } & \text{increasing in } & [\beta_1, \beta_2] \end{aligned}$$

**Proof:** For  $x > 0$ ,

if  $y - x\beta \geq \lambda$ , i.e.,  $\beta \leq (y - \lambda)/x$ , then  $g(\beta) = -\lambda x$ ;

if  $y - x\beta \leq -\lambda$ , i.e.,  $\beta \geq (y + \lambda)/x$ , then  $g(\beta) = \lambda x$ ;

if  $-\lambda \leq y - x\beta \leq \lambda$ , i.e.,  $(y - \lambda)/x \leq \beta \leq (y + \lambda)/x$ , then  $g(\beta) = x(x\beta - y)$ , it is linearly increasing.

For  $x < 0$ ,

if  $y - x\beta \leq -\lambda$ , i.e.,  $\beta \leq (y + \lambda)/x$ , then  $g(\beta) = \lambda x$ ;

if  $y - x\beta \geq \lambda$ , i.e.,  $\beta \geq (y - \lambda)/x$ , then  $g(\beta) = -\lambda x$ ;

if  $-\lambda \leq y - x\beta \leq \lambda$ , i.e.,  $(y + \lambda)/x \leq \beta \leq (y - \lambda)/x$ , then  $g(\beta) = x(x\beta - y)$ , it is linearly increasing.

**Proof of Theorem 2:** By Lemma, each  $s_i(\beta) = u_i(u_i\beta + \theta_i - w_i)$  with  $u_i \neq 0$  has the following properties:

$$\begin{aligned} s_i(\beta) &= -\lambda|u_i|, & \text{if } \beta \leq \beta_{1i} &= (w_i - \text{sign}(u_i)\lambda)/u_i, \\ s_i(\beta) &= \lambda|u_i|, & \text{if } \beta \geq \beta_{2i} &= (w_i + \text{sign}(u_i)\lambda)/u_i, \\ s_i(\beta) &\text{ is linearly } & \text{increasing in } & [\beta_{1i}, \beta_{2i}] \end{aligned}$$

Then  $s(\beta) = \sum_i s_i(\beta)$ , has the following properties:

$$\begin{aligned} s(\beta) &= -\lambda\|u\|_1, & \text{if } \beta \leq \beta_L &= \min_i\{\beta_{1i}\}, \\ s(\beta) &= \lambda\|u\|_1, & \text{if } \beta \geq \beta_R &= \max_i\{\beta_{2i}\}, \\ s(\beta) &\text{ is monotonely } & \text{(more accurately, piecewise linearly) } & \text{increasing in } [\beta_L, \beta_R]. \end{aligned}$$

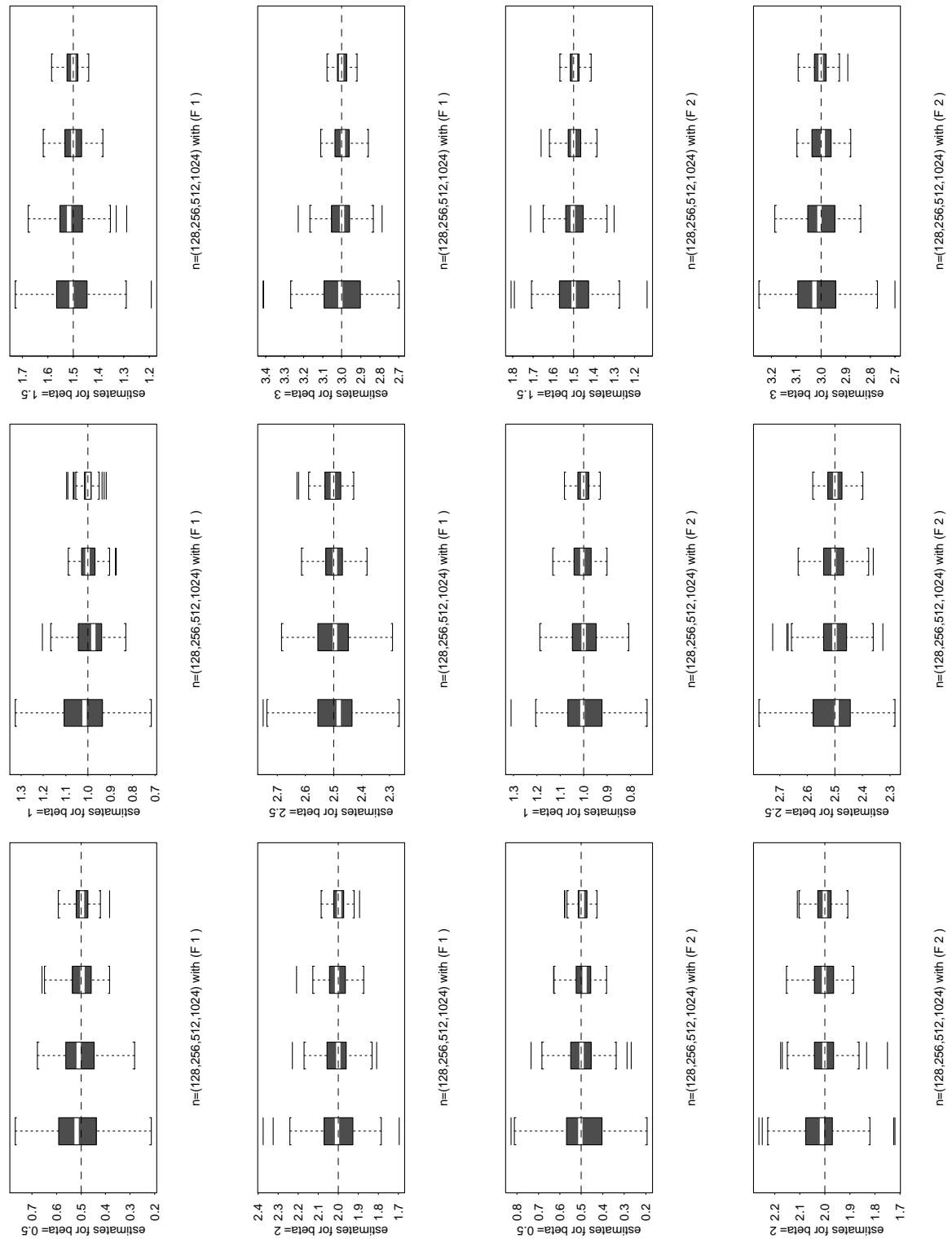


Figure 1: The Boxplots of  $\hat{\beta}$

**Acknowledgments** I would like to thank my advisor Prof. Mary Ellen Bock for guidance, support and encouragement.

## References

- [1] Chen, H., (1988), “Convergence rates for parametric components in a partly linear model”, *Ann. Statist.*, 16, 136-146.
- [2] Donoho, David L. , and Johnstone, Iain M. (1994), “Ideal spatial adaptation by wavelet shrinkage” , *Biometrika*, 81, 425-455
- [3] 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
- [4] Eubank, R.L., (1988), *Spline Smoothing and Nonparametric Regression*, Marcel Dekker, New York.
- [5] Gill, P.E., Murray, W. and Wright, M.H., (1981), *Practical Optimization*, San Diego: Academic Press.
- [6] Green, P., Jennison, C. and Seheult, A., (1985), “Analysis of field experiments by least squares smoothing”, *J. Roy. Statist. Soc. Ser. B*, 47, pp. 299-315.
- [7] Green, P. and Silverman, B.W., (1994), *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman & Hall, London.
- [8] Hamilton, Scott A.; Truong, Young K., (1997), “Local linear estimation in partly linear models”, *J. Multivariate Anal.*, 60, no. 1, 1-19.
- [9] Heckman, N., (1986), “Spline smoothing in a partly linear model”, *J. Roy. Statist. Soc. Ser. B*, 48, 244-248.
- [10] Mallat, S.G., (1989), “A theory for multiresolution signal decomposition: the wavelet representation”, *IEEE Trans. Pattn Anal. Mach. Intell.*, 11, 674-693
- [11] Nason, G.P. (1998), *WaveThresh3 Software*, Department of Mathematics, University of Bristol, Bristol, UK.
- [12] Nason, G.P. (1996), “Wavelet shrinkage using cross-validation”, *J. Roy. Statist. Soc. Ser. B*, 58, 463-479.
- [13] Robinson, P.M., (1988), “Root-n-consistent semiparametric regression”, *Econometrica*, 56, 931-954.
- [14] Michael G. Schimek , (2000), “Estimation and inference in partially linear models with smoothing splines” , *Journal of Statistical Planning and Inference*, 91, 525-540
- [15] Speckman, P., (1988), “Kernel smoothing in partial linear models”, *J. Roy. Statist. Soc. Ser. B*, 50, 413-436.
- [16] Wahba, G., (1990), *Spline Models for Observational Data*, SIAM, Philadelphia, PA.