AMlet and GAMlet: Automatic Nonlinear Fitting of Additive Models and Generalized Additive Models with Wavelets

By SYLVAIN SARDY*,
Swiss Federal Institute of Technology, Switzerland

and

PAUL TSENG
University of Washington, U.S.A.

A simple and yet powerful method is presented to estimate nonlinearly and nonparametrically the components of additive models using wavelets. The estimator enjoys the good statistical and computational properties of the Waveshrink scatterplot smoother and it can be efficiently computed using the block coordinate relaxation optimization technique. A rule for the automatic selection of the smoothing parameters, suitable for data mining of large data sets, is derived.

The wavelet-based method is then extended to estimate generalized additive models. A primal-dual log-barrier interior point algorithm is proposed to solve the corresponding convex programming problem. Based on an asymptotic analysis, a rule for selecting the smoothing parameters is derived, enabling the estimator to be fully automated in practice. We illustrate the finite sample property with a Gaussian and a Poisson simulation.

Key Words: (Generalized) Additive Models; Convex Programming; Data mining; Interior-point algorithm; Regularized Likelihood; Relaxation algorithm; Signal denoising; Universal smoothing parameter; Wavelet.

*Address for correspondence: Department of Mathematics, Swiss Federal Institute of Technology, 1015 Lausanne, Switzerland. Email: Sylvain.Sardy@epfl.ch

1 Introduction

We first recall the standard regression problem. Available is a set of responses $\underline{Y} = (Y_1, \ldots, Y_N)'$ measuring a function $f(\cdot)$ at N locations $\underline{X}_1, \ldots, \underline{X}_N$ sampled from a distribution $F_{\underline{X}}$. The underscore denotes a vector and $\underline{X}'$ denotes the transpose of the vector $\underline{X}$; for instance $\underline{X}' = (X_1, \ldots, X_Q)'$ is the column vector of Q explanatory variables. Based on the training sample $\mathcal{T} = \{(Y_n, \underline{X}_n)\}_{n=1,\ldots,N}$, the primary goal of regression is to estimate the multivariate function $f(\cdot)$ which explains best the association between the predictors $\underline{X}$ and the noisy measurement Y. Assuming the noise is additive and unbiased conditioned on $\underline{x}$, i.e.,

$$Y = f(\underline{X}) + \epsilon \quad \text{with} \quad \mathrm{E}(\epsilon \mid \underline{x}) = 0,$$

the predictive performance of an estimate $\hat{f}(\cdot)$ is often measured by its mean squared error

$$\mathrm{MSE}(\hat{f}(\cdot)) = \mathrm{E}_{\mathcal{T}} \mathrm{E}_{\underline{X}^*} \left\{ f(\underline{X}^*) - \hat{f}(\underline{X}^*) \right\}^2, \tag{1}$$

where the inside expectation is taken over a new explanatory variable $\underline{X}^*$ with the same distribution $F_{\underline{X}}$ as that of $\underline{X}$ in the training sample $\mathcal{T}$. While prediction is the primary goal, interpretability of the estimated function $\hat{f}(\cdot)$ is also desirable.

A well studied model for such a regression problem is the parametric additive linear model $f(\underline{X}) = \alpha_0 + \sum_q \alpha_q X_q$, where the coefficients $\underline{\alpha}$ are typically estimated by least squares. If the parametric model is correct, then the L_2 rate of convergence is in N^{-1}. But, what if the true function $f(\cdot)$ is far from being linear in each covariate? To avoid bias, nonparametric techniques do not assume a straight line but let the data fit themselves. The price to pay is a slower optimal L_2 rate of convergence typically of the form $N^{-2\tau/(2\tau+Q)} > N^{-1}$, where τ is a measure of the smoothness of $f(\cdot)$ and Q is the number of covariates. Many scatterplot smoothers ($Q = 1$) have been developed and achieve an optimal rate of convergence for an appropriate selection of the smoothing parameter. To fit general high-dimensional surfaces when the number of predictors is larger ($Q > 1$), the number of observations N must grow exponentially with Q to achieve the same rate of convergence as when $Q = 1$. This is known as the curse of dimensionality.

The nonparametric additive model is a compromise between the rigid parametric linear model and the too flexible general high-dimensional nonparametric model. The nonparametric additive model looks for the closest approximation to $f(\cdot)$ of the form

$$f(\underline{x}) = \sum_{q=1}^{Q} f_q(x_q), \tag{2}$$

where each $f_q(\cdot)$ is a general univariate function. The advantage of nonparametric additive modeling is to avoid the curse of dimensionality, a result of Stone

(1985) who provided the striking result that the optimal L_2 rate of convergence of each additive component $\hat{f}_q(\cdot)$ to $f_q(\cdot)$ is the same as in the univariate case ($Q = 1$), namely of the form $N^{-2\tau/(2\tau+1)}$, when the multivariate function is indeed additive. Another advantage of the additive model is the ability to visualize the univariate trends in each covariate and provide a simple interpretation of the fitted model. For estimating additive models from data, two main approaches can be distinguished:

- The backfitting approach of Buja, Hastie, and Tibshirani (1989) iteratively applies a univariate linear smoother until convergence.

  Properties of the estimate have been studied in many papers when the univariate smoother used is *linear*. Buja, Hastie, and Tibshirani (1989) established the existence of a solution and convergence of a backfitting algorithm for linear smoothers having a symmetric 'hat' matrix with eigenvalues in $[0, 1]$. They also showed that uniqueness of solution is not guaranteed because of potential 'concurvity'. These results have been extended to other linear smoothers by Opsomer and Ruppert (1997) with local polynomials, by Mammen, Linton, and Nielsen (1999) with local polynomials and the Nadaraya-Watson kernel smoother, and by Amato and Antoniadis (2001) with a linear wavelet-based smoother.

  Conveniently, the backfitting approach uses the univariate smoothing technology, but, because the smoothers must be linear, spatially heterogeneous functions cannot be efficiently estimated. The selection of the smoothing parameters also remains an open question (Cantoni and Hastie, 2002).

- The Turbo knot selection approach of Friedman and Silverman (1989) is *nonlinear*. It assumes that each $f_q(\cdot)$ can be written as a parsimonious expansion on a few truncated power functions defined by their knot location. Turbo then performs a stepwise search for the optimal knot location. If the knots are located optimally, the procedure is efficient even when estimating spatially heterogeneous functions. However, the knot selection procedure is known to suffer from an instability that many practitioners have observed and that Breiman (1996) has studied.

In this paper, we propose to use the good features of both approaches by using a nonlinear smoother and a nonlinear backfitting algorithm. In particular, we consider the recently developed Waveshrink univariate smoother of Donoho and Johnstone (1994) that enjoys nice theoretical and computational properties (Donoho, Johnstone, Kerkyacharian, and Picard, 1995). As in Turbo, Waveshrink models each $f_q(\cdot)$ as a parsimonious expansion on a set of basis functions, but uses wavelets instead of splines. The selection of which wavelets to use is then carried out by nonlinear shrinkage.

Additive models assume that, not only the underlying multivariate function, but also the noise is additive. To handle a wider class of noise distributions, Hastie and Tibshirani (1986) proposed the generalized additive models in the spirit of the generalized linear model of Nelder and Wedderburn (1972). Under mild conditions, Stone (1986) extended his convergence results obtained for additive models, showing that generalized additive models do not suffer from the curse of dimensionality either. In this paper, we also generalize our wavelet-based estimator to estimate the components of generalized additive models.

The article is organized as follows. In Section 2, we review Waveshrink, the nonlinear wavelet-based univariate smoother of Donoho and Johnstone (1994). In Section 3, we define the *AMlet* wavelet-based estimator of additive models and study the convergence of a block coordinate relaxation (nonlinear backfitting) algorithm. In Section 4, we define the *GAMlet* estimator that generalizes *AMlet* to a wide class of noise distributions. Both *AMlet* and *GAMlet* require smoothing parameters as input. Based on the idea of universal threshold of Donoho and Johnstone (1994), Section 5 proposes a practical rule for choosing the smoothing parameters automatically, making *AMlet* and *GAMlet* fully implementable in practice. A Gaussian and a Poisson simulation show how well the universal rule works in practice. The final section discusses our results and suggests some further areas of research. For clarity, we postpone some mathematical derivations to the Appendix.

2 Review of Waveshrink

We consider the univariate situation when only one predictor is available ($Q = 1$). Waveshrink is a wavelet-based smoother and, as such, belongs to the class of expansion-based estimators: it assumes that the univariate function $f(\cdot)$ can be well represented by a linear combination of approximation $\phi(\cdot)$ and fine scale $\psi(\cdot)$ wavelets. The standard univariate wavelets are multi-resolution functions that are locally supported and indexed by a location parameter k and a scale parameter j. A father wavelet $\phi(\cdot)$ such that $\int_0^1 \phi(x)dx = 1$ generates $p_0 = 2^{j_0}$ approximation wavelets by means of the dilation and translation relation

$$\phi_{j_0,k}(x) = 2^{j_0/2}\phi(2^{j_0}x - k), \quad k = 0, 1, \ldots, 2^{j_0} - 1;$$

they capture the coarse features of the signal. Similarly, a mother wavelet $\psi(\cdot)$ such that $\int_0^1 \psi(x)dx = 0$ generates $N - p_0$ fine scale wavelets

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k), \quad j = j_0, \ldots, J; \; k = 0, 1, \ldots, 2^j - 1,$$

where $J = \log_2(N) - 1$. Because they are locally supported, the fine scale wavelets capture the local features of the signal. Unless otherwise stated, we will assume that the function we want to estimate has the following expansion

$$f(x) = \sum_{\kappa=0}^{2^{j_0}-1} \beta_\kappa \phi_{j_0,\kappa}(x) + \sum_{j=j_0}^{J} \sum_{\kappa=0}^{2^j-1} \gamma_{j,\kappa}\psi_{j,\kappa}(x), \tag{3}$$

where the wavelet functions $\{\phi_{j_0,\kappa}(\cdot), \psi_{j,\kappa}(\cdot)\}$ are orthonormal with respect to the L_2 norm. An orthonormal matrix Φ can be extracted from these functions that is appropriate for a fixed equispaced design for the sampling locations $x_1, \ldots, x_N$. Hence the function $f(\cdot)$ calculated at the design points $\underline{f} = (f(x_1), \ldots, f(x_N))$ has the following wavelet decomposition

$$\underline{f} = \Phi\underline{\alpha} = [\Phi_0 \;\; \Psi] \left[\begin{array}{c} \underline{\beta} \\ \hline \underline{\gamma} \end{array} \right],$$

where Φ_0 is the $N \times p_0$ matrix of approximation wavelets, Ψ is the $N \times (N - p_0)$ matrix of fine scale wavelets, and $\underline{\beta}, \underline{\gamma}$ are the corresponding coefficients. To simplify notation, for any two column vectors $\underline{u}$ and $\underline{v}$, we write $(\underline{u}, \underline{v})$ for $\left[\begin{array}{c} \underline{u} \\ \hline \underline{v} \end{array} \right]$.

To estimate the wavelet coefficients $\underline{\alpha} = (\underline{\beta}, \underline{\gamma})$ from the data

$$Y_n = f(x_n) + \epsilon_n,$$

the least squares estimate $\underline{\hat{\alpha}}^{\text{LS}} = \Phi'\underline{Y}$ must be regularized because the rank of Φ equals the number of observations N. Several regularization techniques have been proposed: linear (Antoniadis, Gregoire, and McKeague, 1994), nonlinear (Donoho and Johnstone, 1994) or nonlinear Bayesian (Abramovich, Sapatinas, and Silverman, 1998). The Waveshrink estimator of Donoho and Johnstone (1994) has a remarkable ability to estimate spatially inhomogeneous signals with near minimax results for a wide range of functions (Donoho, Johnstone, Kerkyacharian, and Picard, 1995). Waveshrink is moreover computationally efficient. We are particularly interested in soft-Waveshrink, defined as follows: Define the nonlinear soft-shrinkage function $\eta_\lambda^{\text{soft}}(\gamma) = \text{sign}(\gamma)(|\gamma| - \lambda)_+$, where $x_+ = x$ if $x \geq 0$, and $x_+ = 0$ if $x < 0$; then, for a given smoothing parameter λ, the (biased) estimate $\underline{\hat{\gamma}}_\lambda$ of the wavelet coefficients shrinks the least squares estimate componentwise toward zero with the soft shrinkage function

$$\underline{\hat{\gamma}}_\lambda = \eta_\lambda^{\text{soft}}(\underline{\hat{\gamma}}^{\text{LS}}). \tag{4}$$

Other shrinkage functions have been proposed, but the softshrink estimate has a useful penalized least squares interpretation: it is the closed form solution (Donoho, Johnstone, Hoch, and Stern, 1992) to

$$\min_{\underline{\alpha}=(\underline{\beta},\underline{\gamma})} \frac{1}{2}\|\underline{Y} - \Phi\underline{\alpha}\|_2^2 + \lambda \sum_{j=j_0}^{J} \sum_{\kappa=0}^{2^j-1} |\gamma_{j,\kappa}|. \tag{5}$$

This optimization problem defines more generally Basis Pursuit (Chen, Donoho, and Saunders, 1999) for non-orthonormal matrices, in particular when Φ has more columns than rows. We will be in that situation in the following section.

Waveshrink works well for equispaced observations, a situation which confers the orthonormality property to Φ. Unfortunately this situation is rare in statistical applications where the data are scattered randomly. Many generalization of Waveshrink have been proposed to handle the non-equispaced situation (see for instance Antoniadis and Fan (2001) and references therein). The isometric wavelets of Sardy, Percival, Bruce, Gao, and Stuetzle (1999) have the advantage of maintaining the convenient orthonormality of the wavelet matrix, a property we use later for AMlet and GAMlet. It is based on the idea that the mean squared error (1) is, on the one hand, the expected L_2 distance between $f(\cdot)$ and $\hat{f}(\cdot)$ weighted by the distribution $F_X(\cdot)$, and, on the other hand, the expected L_2 distance between $f \circ F_X^{-1}(\cdot)$ and $\hat{f} \circ F_X^{-1}(\cdot)$. Hence the observations $\underline{Y}$ can be regarded as equispaced measurements of $f \circ F^{-1}(\cdot)$; for more details, see Sardy, Percival, Bruce, Gao, and Stuetzle (1999).

3 AMlet

In this section, we implement Waveshrink as the scatterplot smoother for additive models. We study issues such as existence, uniqueness, and efficient computation of the *AMlet* estimate for a given smoothing parameter.

3.1 Definition

Let $\underline{f} = (f(\underline{X}_1), \ldots, f(\underline{X}_N))$ be the values of the function $f(\cdot)$ at the N locations. Similarly, let $\underline{f}_q$ be the values of the function $f_q(\cdot)$ at the N sampling points of the qth covariate X_q ordered in increasing order by the permutation matrix P_q. Using this notation, the sampled additive function writes

$$\underline{f} = \sum_{q=1}^{Q} P_q' \underline{f}_q, \tag{6}$$

where P_q' is the transpose of P_q. In turn, we represent each univariate function $\underline{f}_q$ as a linear combination of isometric wavelets,

$$\underline{f}_q = \Phi_q \underline{\alpha}_q, \tag{7}$$

where Φ_q is an orthonormal matrix. Combining equations (6) and (7) we obtain

$$\underline{f} = \sum_{q=1}^{Q} P_q' \Phi_q \underline{\alpha}_q =: \bar{\Phi}\underline{\alpha},$$

where $\bar{\Phi} = [P_1'\Phi_1, \ldots, P_Q'\Phi_Q]$ with corresponding coefficients $\underline{\alpha} = (\underline{\alpha}_1, \ldots, \underline{\alpha}_Q)$. Permuting the columns of $\bar{\Phi}$ and correspondingly the entries of $\underline{\alpha}$, we also write

$$\underline{f} =: [\bar{\Phi}_0 \ \bar{\Psi}] \begin{bmatrix} \underline{\beta} \\ \underline{\gamma} \end{bmatrix},$$

where $\bar{\Phi}_0$ is the $N \times Qp_0$ matrix of concatenated approximation wavelets with corresponding coefficients $\underline{\beta} = (\underline{\beta}_1, \ldots, \underline{\beta}_Q)$, and $\bar{\Psi}$ is the $N \times Q(N - p_0)$ matrix of fine scale wavelets with its corresponding coefficients $\underline{\gamma} = (\underline{\gamma}_1, \ldots, \underline{\gamma}_Q)$. Notice that the matrices $P_q' \Phi_q$ for $q = 1, \ldots, Q$ remain orthonormal and therefore $\bar{\Phi}$ is the concatenation of orthonormal blocks. Because $\bar{\Phi}$ has rank N, the least squares problem must be regularized to estimate $\underline{\alpha}$.

We propose the following regularization method. For given smoothing parameters $\underline{\lambda} = (\lambda_1, \ldots, \lambda_Q)$, we propose the *AMlet* penalized least squares estimator defined by $\underline{\hat{f}}_{\underline{\lambda}} = \bar{\Phi}\underline{\alpha}$, where $\underline{\alpha} = (\underline{\alpha}_1, \ldots, \underline{\alpha}_Q)$ is a solution to

$$\min_{\underline{\alpha}=(\underline{\beta},\underline{\gamma})} \frac{1}{2} \|\underline{Y} - \bar{\Phi}\underline{\alpha}\|_2^2 + \sum_{q=1}^{Q} \lambda_q \sum_{j=j_0}^{J} \sum_{\kappa=0}^{2^j-1} |\gamma_{q\,j,\kappa}|. \tag{8}$$

This regularization induces sparsity in the wavelet representation in the sense that there exists a solution vector $\underline{\alpha}$ of which at most N coefficients among the QN coefficients are nonzero. The optimization problem (8) is convex and the function $\|\underline{Y} - \cdot\|_2^2$ is strictly convex so $\underline{\hat{f}}_{\underline{\lambda}} = \bar{\Phi}\underline{\alpha}$ is unique, independent of the solution $\underline{\alpha}$. However, as for the 'concurvity' concern of Buja, Hastie, and Tibshirani (1989), $\underline{\alpha}$ itself is not unique, and the individual estimates $\underline{\hat{f}}_q = \Phi_q \underline{\alpha}_q$ are therefore not uniquely defined. In fact, the set of solutions $\underline{\alpha} = (\underline{\beta}, \underline{\gamma})$ is not bounded since adding to $\underline{\beta}$ any linear combination in the kernel of $\bar{\Phi}_0$ gives another solution. (The columns of $\bar{\Phi}_0$ are linearly dependent since $\bar{\Phi}_0 = [P_1'\Phi_{01}, \ldots, P_Q'\Phi_{0Q}]$ and $\underline{1} \in \text{Range}(P_q'\Phi_{0q})$ for $q = 1, \ldots, Q$.)

One way to define a unique solution $\underline{\alpha}$ is to choose the one of minimum ℓ_2 norm. This is motivated by the minimum ℓ_2 norm property of the often used Moore-Penrose generalized inverse (Albert, 1972) solution to the least squares problem when the normal equations do not have a unique solution. Hence, we define the *AMlet** estimator of additive models by taking the minimum ℓ_2 norm solution among all the solutions to (8), namely,

$$\underline{\alpha}^* = \arg\min_{\underline{\alpha}} \|\underline{\alpha}\|_2^2, \quad \text{where} \quad \underline{\alpha} \quad \text{solves} \quad (8). \tag{9}$$

Consequently, the *AMlet** component estimates $\underline{\hat{f}}_q^* = \Phi_q \underline{\hat{\alpha}}_q^*$ are uniquely defined for $q = 1, \ldots, Q$. Indeed, the solution set of (8) is a closed convex set, so it has a unique point whose ℓ_2 norm is minimum.

3.2 Solving AMlet's ℓ_1 penalized least squares problem

We must solve a nontrivial two-level optimization problem, namely (8) within (9).

To find a solution to (8), we propose to use the block coordinate relaxation (BCR) algorithm of Sardy, Bruce, and Tseng (2000) which is reminiscent of

Buja, Hastie, and Tibshirani (1989)'s backfitting algorithm, also called Gauss-Seidel algorithm. Convergence of the Gauss Seidel algorithm has been well studied for solving a symmetric positive semidefinite system of linear equations. Our convex quadratic program (8) is more difficult to solve, however, since a *nondifferentiable* ℓ_1 penalty is added to a convex quadratic function. For further studies of the BCR algorithm for solving such nondifferentiable optimization problems, see Sardy, Tseng, and Bruce (2001) and Tseng (2001).

The BCR algorithm exploits two properties: first, the matrix $\bar{\Phi}$ is orthonormal union-complete (here, the concatenation of orthonormal blocks $P_q'\Phi_q,\ q = 1, \ldots, Q$); second, (8) has a closed form solution (4) via the soft shrinkage function when $\bar{\Phi} = \Phi$ is orthonormal.

BCR algorithm for *AMlet*:

1. Choose an initial guess for $\underline{\alpha}$;
2. For $q \in \{1, \ldots, Q\}$, calculate $\underline{\mathrm{res}}_q = \underline{Y} - \sum_{i \neq q} P_i' \Phi_i \underline{\alpha}_i$ and solve:

$$\min_{\underline{\alpha}_q} \|\underline{\mathrm{res}}_q - (P_q'\Phi_q)\underline{\alpha}_q\|_2^2 + \lambda_q \sum_{j=j_0}^{J} \sum_{\kappa=0}^{2^j-1} |\gamma_{q\,j,\kappa}|.$$

   using the closed form solution (4) since $P_q'\Phi_q$ is an orthonormal matrix for all q;
3. update $\underline{\alpha}_q$ in $\underline{\alpha}$;
4. If convergence criterion not met, go to step 2;

In step 2 of the BCR algorithm, a block q must be chosen. A systematic cyclic rule and an optimal descent rule were proposed by Sardy, Bruce, and Tseng (2000). When Q is small and the wavelet matrices Φ_q are orthonormal, we recommend using the cyclic rule whose convergence is guaranteed by a result of Tseng (2001). The convergence is also rapid due to the blockwise relaxation that, at each step, involves Mallat (1989)'s $O(N)$ 'pyramid' algorithm.

To find the unique solution $\underline{\alpha}^*$ to (9), we propose for *AMlet** the following algorithm based on the idea of Tikhonov regularization. At the kth iteration ($k = 1, 2, \ldots$), a regularization parameter $\epsilon_k > 0$ and an accuracy tolerance δ_k are chosen, and the BCR algorithm is applied to

$$\min_{\underline{\alpha}=(\underline{\beta},\underline{\gamma})} \frac{1}{2}\|\underline{Y} - \bar{\Phi}\underline{\alpha}\|_2^2 + \sum_{q=1}^{Q} \lambda_q \sum_{j=j_0}^{J} \sum_{\kappa=0}^{2^j-1} |\gamma_{q\,j,\kappa}| + \epsilon_k \|\underline{\alpha}\|_2^2 \tag{10}$$

until it finds an approximate solution $\underline{\alpha}_k$ satisfying

$$\max_{q\in\{1,\ldots,Q\}} \|\underline{r}_q(\underline{\alpha};\epsilon_k)\|_2 \cdot \sum_{q=1}^{Q} \|\underline{\alpha}_q\|_2 \le \delta_k \quad \text{and} \quad \max_{q\in\{1,\ldots,Q\}} \|\underline{r}_q(\underline{\alpha};\epsilon_k)\|_2 \le \delta_k, \tag{11}$$

where $\underline{r}_q(\underline{\alpha};\epsilon_k)$ is the smallest (sub)gradient of the cost function of (10) with respect to the qth orthonormal block; see (26) for a definition. By solving (10) approximately and letting $\epsilon_k \to 0$ and $\delta_k \to 0$ at suitable rates, as stated in the following theorem, the approximate solutions $\underline{\alpha}_k$ are guaranteed to converge to $\underline{\alpha}^*$.

Theorem 1: If we choose ϵ_k and δ_k to tend to zero so that

$$\lim_{k\to\infty} \delta_k/\epsilon_k = 0, \tag{12}$$

then the sequence of approximate solutions $\{\underline{\alpha}_k\}$ will converge to the unique *AMlet** solution, i.e., the minimum ℓ_2 norm solution $\underline{\alpha}^*$ to (8).
Proof: See Appendix A.

We implemented this algorithm with $\epsilon_{k+1} = \epsilon_k/3$, $\delta_k = 10(\epsilon_k)^\nu$ and $\nu = 1.1$. Also, α_{k-1} is used as the starting point for the BCR algorithm at the kth iteration for $k > 1$. Rapid convergence is observed.

4 GAMlet

Hastie and Tibshirani (1986)'s generalized additive models extend additive models to a wider class of noise (see also Hastie and Tibshirani (1990)). Introducing the notation $\mu(\cdot)$ and $\eta(\cdot)$ in place of $f(\cdot)$ to match standard notation of generalized models, generalized additive models assume that the response variable Y_n is an unbiased measurement of $\mu_n = \mu(\underline{x}_n)$ conditioned on the covariates; that is $Y_n|\underline{x} = \underline{x}_n \sim \rho_Y(y;\mu_n,\phi)$, where the density function $\rho_Y(\cdot)$ is parameterized by its expectation μ_n and a nuisance parameter ϕ. We also assume that the model is additive in $\eta_n = \eta(\underline{x}_n) = \sum_{q=1}^Q \eta_q(x_{nq})$ and that μ_n and η_n are linked through

$$g(\mu_n) = \eta_n. \tag{13}$$

The reason for using a link function (13) is two-fold: First, it ensures the existence and uniqueness of the estimate by strict concavity of the log-likelihood $l(\eta;y) = \sum_n \log \rho_Y(Y_n;\mu_n,\phi)$. Second, it is often chosen to map the parameter estimate in its domain (e.g., the log link for Poisson). The canonical link is often chosen for computational convenience since the estimation problem is constraint-free. In some applications, however, other links should be considered. In the following, we propose an estimator and an algorithm capable of using non-canonical links, as long as the log-likelihood function is strictly concave, e.g., the identity link for Poisson.

4.1 Definition

We consider a penalized likelihood approach and generalize *AMlet* by replacing the quadratic term in (8) with the negative log-likelihood function of the assumed noise distribution. For given smoothing parameters $\underline{\lambda} = (\lambda_1, \ldots, \lambda_Q)$, we define the *GAMlet* estimator $\hat{\underline{\eta}}_{\underline{\lambda}} = \bar{\Phi}\underline{\alpha}$ as the solution to

$$\min_{\underline{\eta}, \underline{\alpha}=(\underline{\beta},\underline{\gamma})} -l(\underline{\eta}; \underline{Y}) + \sum_{q=1}^{Q} \lambda_q \sum_{j=j_0}^{J} \sum_{\kappa=0}^{2^j-1} |\gamma_{q_{j,\kappa}}| \quad \text{with} \quad \underline{\eta} = \bar{\Phi}\underline{\alpha} \text{ and } \underline{\eta} \in C, \qquad (14)$$

where $l(\underline{\eta}; \underline{Y}) = \sum_{n=1}^{N} l(\eta_n, Y_n)$ and $C = C_1 \times \cdots \times C_N$ and C_n denotes the domain of $l(\cdot; Y_n)$. Assuming that the negative log-likelihood function is strictly convex in $\underline{\eta}$ for a judicious choice of the link function $g(\cdot)$ in (13), we have the following two important properties: sparsity in the wavelet representation and uniqueness of the estimate $\hat{\underline{\eta}}_{\underline{\lambda}}$.

As with *AMlet*, the solution $\underline{\alpha}$ to (14) is not unique and the set of such solutions is not bounded. To achieve uniqueness, we define the *GAMlet** estimator of the components $\hat{\eta}_{q,\underline{\lambda}}$ of generalized additive models by taking the minimum ℓ_2 norm solution among all the solutions to (14):

$$\underline{\alpha}^* = \arg\min_{\underline{\alpha}} \|\underline{\alpha}\|_2^2, \quad \text{where} \quad \underline{\alpha} \quad \text{solves} \quad (14). \qquad (15)$$

The solution $\underline{\alpha}^*$ is defined and unique since the solution set of (14) is a closed convex set, so it has a unique point whose ℓ_2 norm is minimum. Hence the *GAMlet** component estimates $\hat{\underline{\eta}}_q^* = \Phi_q \hat{\underline{\alpha}}_q^*$ are uniquely determined.

4.2 Solving GAMlet's ℓ_1 penalized likelihood problem

Our definition of *GAMlet* allows the use of a non-canonical link function by means of constraints in (14). In the Poisson case, for instance, both the log link and the identity link satisfy the strict convexity assumption on the negative log-likelihood; the canonical link maps $\eta \in R$ to $\mu \in R^+$ without constraints, whereas the identity link requires constraints in (14) with $C = [0, \infty)^N$. To solve the constrained optimization (14) in the univariate situation, Sardy, Antoniadis, and Tseng (2003) employ a primal-dual log-barrier interior point algorithm. The multivariate situation ($Q > 1$) demands a new algorithm to find the minimum ℓ_2 norm solution. One approach for *GAMlet** is Tikhonov regularization used in Section 3.2 whereby, at each iteration, an interior point algorithm is applied to solve approximately a regularized version of (14) analogous to (10). The Tikhonov regularization parameter and the solution accuracy tolerance are decreased after each iteration. We describe such an algorithm and prove its convergence in Appendix B.

A conceptually simpler approach is the following two-stage algorithm: First, find *any* solution to (14) using, for instance, the algorithm of Sardy, Antoniadis, and Tseng (2003); second, find the unique ℓ_2 norm solution by solving a quadratic programming problem based on the solution found in the first stage.

Specifically, letting

$$\underline{\mu}^* = \bar{\Phi}\underline{\alpha} \quad \text{and} \quad \zeta^* = \sum_{q=1}^{Q} \lambda_q \|\underline{\gamma}_q\|_1 \tag{16}$$

for any solution $\underline{\alpha} = (\beta, \gamma)$ to (14), then the problem of finding the minimum ℓ_2 norm solution $\underline{\alpha}^* = (\underline{\beta}^*, \underline{\gamma}^*)$ can be set up as a convex quadratic program of the form:

$$\min_{\underline{\alpha}=(\underline{\beta},\underline{\gamma})} \|\underline{\beta}\|_2^2 \quad \text{with} \quad \bar{\Phi}\underline{\alpha} = \underline{\mu}^* \quad \text{and} \quad \sum_{q=1}^{Q} \lambda_q \|\underline{\gamma}_q\|_1 \le \zeta^*. \tag{17}$$

Indeed, by strict convexity of $-l(\cdot; \underline{Y})$, $\underline{\mu}^* = \bar{\Phi}\underline{\alpha}$ is unique, independent of the solution $\underline{\alpha}$ to (14). Since the minimum objective value of (14) is also unique, this means the real number ζ^* is unique, independent of the solutions $\underline{\alpha}$ to (14). In fact, (16) are necessary and sufficient conditions for $\underline{\alpha} = (\underline{\beta}, \underline{\gamma})$ to be a solution to (14).

The second stage optimization problem (17) can be solved by a primal-dual interior point algorithm of which we give the main steps. Letting $\underline{a}^* = (\zeta^*, \underline{\mu}^*)$, $A = \tilde{\Phi}_0\tilde{\Phi}_0'$ with $\tilde{\Phi}_0' = [\underline{0}, \bar{\Phi}_0']$ and $B = \begin{bmatrix} \underline{\lambda}' & \underline{\lambda}' \\ \bar{\Psi} & -\bar{\Psi} \end{bmatrix}$ with $\underline{\lambda} = (\lambda_1\underline{1}, \ldots, \lambda_Q\underline{1})$, the dual problem is

$$\max_{\underline{y}} \underline{y}'\underline{a}^* - \frac{1}{2}\underline{y}'A\underline{y} \quad \text{with} \quad B'\underline{y} \le \underline{0},$$

where $\underline{y}$ is the dual variable (not to be confused with the response variable $\underline{Y}$). The reason for solving the dual and the primal together is that the duality gap is zero, and a solution to the primal problem yields as a byproduct a solution to the dual and vice versa (Rockafellar, 1984, §11D). Consequently, the convergence can be monitored by measuring the gap between them. The log-barrier subproblem associated with it is

$$\min_{\underline{y}} \quad -\underline{y}'\underline{a}^* + \frac{1}{2}\underline{y}'A\underline{y} - \rho \sum_{p=1}^{2Q(N-p_0)} \log(-\underline{B}_p'\underline{y}),$$

where $\rho > 0$ and $\underline{B}_p$ is the pth column of B. By introducing the slack variable $\underline{z} = -B'\underline{y}$, the Karush–Kuhn–Tucker conditions are

$$\begin{aligned} -B'\underline{y} - \underline{z} &=: \underline{r}_x &= \underline{0}, \\ \underline{a}^* - A\underline{y} - B\underline{x} &=: \underline{r}_y &= \underline{0}, \\ \rho\underline{1} - X\underline{z} &=: \underline{r}_z &= \underline{0}, \end{aligned}$$

with $\underline{x} > 0$, $\underline{z} > 0$. The Newton directions are $\Delta\underline{z} = \underline{r}_x - B'\Delta\underline{y}$, $\Delta\underline{x} = Z^{-1}(\underline{r}_z - X\Delta\underline{z})$ with $\Delta\underline{y}$ being the solution to

$$(A + BDB')\Delta\underline{y} = \underline{r}_y - B(Z^{-1}\underline{r}_z - D\underline{r}_x),$$

where $D = Z^{-1}X$. Since the left-hand matrix is symmetric positive definite and involves fast matrix multiplications, we can solve the linear system with the conjugate gradient algorithm. There has been many convergence studies (Kojima, Megiddo, and Mizuno, 1993) of interior point algorithms, and rapid convergence is achieved if the initial point is close to the optimal solution.

5 Automatic selection of smoothing parameters

In the univariate situation $Q = 1$ and for Gaussian noise, Donoho and Johnstone (1994) proposed to select the smoothing parameter of Waveshrink with the *universal* rule that is based on an asymptotic consideration. Donoho, Johnstone, Kerkyacharian, and Picard (1995) proved that, with the universal smoothing parameter $\lambda = \sigma\sqrt{2 \log N}$, Waveshrink is nearly minimax for a wide variety of loss functions and for a wide range of smoothness classes. The definition of the universal rule has been generalized to other distributions in Sardy, Antoniadis, and Tseng (2003).

We now derive the universal rule for *AMlet* (assuming near Gaussian noise) and for *GAMlet* in the multivariate situation $Q > 1$. The universal smoothing parameters $\lambda_1(N), \ldots, \lambda_Q(N)$ control the amount of smoothing of each of the Q smoothers. Defined for any concave likelihood $l(\cdot; \underline{Y})$, the smoothing parameters must have the asymptotic property (18) of the following proposition under constraints (19). We present rules for selecting smoothing parameters having such a property for the cases of Gaussian and Poisson distributions.

Proposition 1: Suppose that the signal $\underline{Y}$ has a log-likelihood function $l(\cdot; \underline{Y})$ defined on a domain of the form $C = C_1 \times \cdots \times C_N$, where C_n denotes the domain of $l(\cdot\,; Y_n)$. Suppose that the parameters of interest are a linear combinations of the Qp_0 approximation wavelets only, i.e., $\underline{\eta}_0 = \bar{\Phi}_0 \underline{\beta}_0$ for some $\underline{\beta}_0$ such that $\bar{\Phi}_0 \underline{\beta}_0 \in C$, and suppose that the log-likelihood is concave and differentiable in $\underline{\eta}$ on C. Then the universal parameter $\underline{\lambda}_N = (\lambda_1, \ldots, \lambda_Q)$ is defined as the smallest $\underline{\lambda}(N) = (\lambda_1(N), \ldots, \lambda_Q(N))$ such that

$$\mathrm{P}\left\{\|\Psi_1' P_1 \underline{y}\|_\infty \le \lambda_1(N), \ldots, \|\Psi_Q' P_Q \underline{y}\|_\infty \le \lambda_Q(N)\right\} \stackrel{N\to\infty}{\longrightarrow} 1, \tag{18}$$

where the random vector $\underline{y}$ together with some $\underline{\eta} \in C$ and $\underline{\beta}$ satisfies

$$\begin{array}{rcl} -\nabla_{\underline{\eta}} l(\underline{\eta}; \underline{Y}) + \underline{y} &=& \underline{0}, \\ \bar{\Phi}_0' \underline{y} &=& \underline{0}, \\ \bar{\Phi}_0 \underline{\beta} &=& \underline{\eta}. \end{array} \tag{19}$$

Furthermore, we can bound from below the desired probability by

$$1 - \sum_{q=1}^{Q} P\{\|\Psi_q' P_q \underline{y}\|_\infty > \lambda_q(N)\}; \tag{20}$$

since the dimension Q of the predictor space does not increase with N, we obtain the desired asymptotic property (18) provided each term in the sum goes to zero as N goes to infinity.

Proof: The proof is similar to that of Sardy, Antoniadis, and Tseng (2003, Appendix B) after writing the Karush–Kuhn–Tucker conditions for (14):

$$\begin{aligned} -\nabla_{\underline{\eta}} l(\underline{\eta};\underline{Y}) + \underline{y} &= \underline{0}, \\ \Psi_q' P_q \underline{y} &\in [-\lambda_q \underline{1}, \lambda_q \underline{1}] \quad q = 1, \ldots, Q, \\ \bar{\Phi}_0' \underline{y} &= \underline{0}, \\ \bar{\Phi}_0 \underline{\beta} + \bar{\Psi} \underline{\gamma} &= \underline{\eta}. \end{aligned}$$

□

For the Gaussian distribution, a single universal parameter $\lambda_1 = \cdots = \lambda_Q = \sigma\sqrt{2\log N}$ provides *AMlet* with the desired convergence of each term of the sum in (20) as the sample size tends to infinity. The universal smoothing parameter is appropriate for obtaining estimates with good visual appearances, but it tends to oversmooth. In the univariate situation, the minimax threshold (Donoho and Johnstone, 1994) is known to be better than the universal threshold in terms of mean squared errors. We will investigate in the Monte Carlo simulation of Section 6 how the minimax threshold performs in the multivariate additive model situation. In practice, the standard deviation σ of the noise is rarely known. For univariate signals, Donoho and Johnstone (1995) propose to take the median absolute deviation (MAD) of the fine-scale wavelet coefficients rescaled by $1/0.6745$ for Gaussian noise. In the multivariate setting, this approach cannot be applied directly, but an estimate of σ can be obtained at each iteration of the BCR algorithm by taking the MAD of the fine-scale wavelet coefficients in the qth direction, namely

$$\hat{\sigma} = \text{MAD}\left(\Psi_q' P_q (\underline{Y} - \sum_{i \neq q} P_i' \Phi_i \underline{\alpha}_i) \right) /0.6745,$$

or the rescaled median absolute deviation of the residuals. While Theorem 1 guarantees the convergence of the Tikhonov-based BCR algorithm to the minimum ℓ_2 norm solution for a fixed smoothing parameter, such a convergence guarantee has yet to be shown when $\hat{\sigma}$ is updated at each iteration. In practice, it does not seem to prevent convergence.

For additive but long-tailed noise, a robust version of *AMlet* can be derived. Because it is based on an ℓ_2 loss function, *AMlet* is appropriate for symmetric noise distributions, but the tails should not depart drastically from that of a Gaussian distribution. For heavy tails distributions, e.g., ϵ-contamination, a robust extension of *AMlet* can be developed by replacing the ℓ_2 loss function in (8) by a robust loss function $\rho(\cdot)$: we define the *RAMlet** estimate as the minimum ℓ_2 norm solution $\underline{\alpha}^*$ to

$$\min_{\underline{\alpha}=(\underline{\beta},\underline{\gamma})} \|\underline{s} - \bar{\Phi}\underline{\alpha}\|_\rho + \sum_{q=1}^{Q} \lambda_q \sum_{j=j_0}^{J} \sum_{\kappa=0}^{2^j-1} |\gamma_{q\,j,\kappa}|, \tag{21}$$

where $\|\underline{w}\|_\rho = \sum_{n=1}^N \rho(w_n)$. For the Huber loss function (Huber, 1981) in the univariate situation ($Q = 1$), Sardy, Tseng, and Bruce (2001) show how to transform (21) into a problem for which the efficient BCR algorithm converges to the optimum. Their results generalize to the multivariate additive situation ($Q > 1$) as well. They also give rules on how to choose the smoothing parameters in that situation.

For the Poisson distribution, we use the level dependent universal threshold derived by Sardy, Antoniadis, and Tseng (2003) using results from Poisson processes, namely,

$$\lambda_{q\,N,j} = M(\eta_{0\,q}(\cdot), \Psi_j(\cdot))2^{j/2}\sqrt{2\log N}/\sqrt{N}, \tag{22}$$

where

$$M^2(\eta_{0\,q}(\cdot), \Psi_j(\cdot)) = \max_{u\in[0,1]}\{\psi_j^2(u)\} \int_0^1 1/\eta_{0\,q}(s)ds.$$

This selection of the regularization parameters, which assumes by (20) that each component is positive, provides the desired asymptotic property of each term of the sum in (20). A rough estimate of the constants $M(\eta_{0\,q}(\cdot), \Psi_j(\cdot))$ is required. We estimate $\underline{\eta_0}_q$ using *AMlet* on $\tilde{Y}_n = 2\sqrt{Y_n + 3/8}$. Indeed, the Anscombe (1948)'s variance stabilizing transformation makes the data approximately Gaussian $\tilde{Y}_n \dot{\sim} \mathrm{N}(2\sqrt{\eta(\underline{x}_n)}, 1)$.

6 Simulation

The goal of the simulation is to investigate the finite sample performances of the wavelet-based estimators and to compare them to existing estimators. In Section 6.1, we consider the most standard scenario where the measurements' noise is additive and Gaussian, while, in Section 6.2, we perform a simulation with Poisson noise. We considered the scenario where the covariates are uniformally and normally distributed (and rescaled) on $[0, 1]$, but, because we obtained comparable results, we only report here results with the uniform distribution.

AMlet* and *GAMlet** remove the indeterminacy of the *AMlet* and *GAMlet* penalized likelihood problems and define unique component estimates by taking the unique, but arbitrary, minimum ℓ_2 norm solution among all the solutions. However, by no means are the unique component estimates optimal in the mean squared error sense; it is only a convenient way of defining a unique estimate.

In the Monte Carlo simulation, we measure the performance of the *AMlet* and *GAMlet* component estimates $\hat{f}_q$ after mean-centering them around the mean $\bar{f}_q$ of the true component f_q to calculate the squared errors

$$\mathrm{SE}(\hat{f}_q(\cdot)) = \frac{1}{N}\sum_{n=1}^{N}\{\hat{f}_q(x_{nq}) - f_q(x_{nq}) - (\bar{\hat{f}}_q - \bar{f}_q)\}^2 \tag{23}$$

for $q = 1, \ldots, Q$. We then average these values over the Monte Carlo runs to estimate the natural mean squared error criterion, the L_2 distance weighted by the marginal density of the covariates.

6.1 Gaussian noise – Uniform covariates

We conduct in this section an extensive simulation based on the results obtained by Amato and Antoniadis (2001). They compared three estimators: the spline-based estimator *addreg* of Nychka, Bailey, Ellner, Haaland, and O'Connel (1993), the local polynomial-based estimator *addfit* of Opsomer and Ruppert (1997) and their wavelet-based estimator *Wavelet direct separation*. *addreg* gave the best results and did not have occasional convergence problems as *addfit* did. Moreover, *addreg* naturally provides an estimate at the sampled points to estimate the natural mean squared error (1), as *AMlet* does. Using the binning step to coerce the data on an equispaced grid (that must unfortunately be coarse to prevent some bins to have zero observations), *Wavelet direct separation* does not provide an estimate at the sampled points, and thus does not allow direct use of the natural mean squared error criterion. The signal-to-noise ratio of the simulation of Amato and Antoniadis is small, and consequently the numerical results reported in the tables have much uncertainty. Moreover, the functions used are rather smooth, which favors linear estimators. Nevertheless, Amato and Antoniadis (2001)'s simulation is insightful and provides a basis for comparison in our simulation, which takes the best estimator *addreg* of the three in their study as the reference estimator.

Based on these considerations, we design the following Monte Carlo experiment to compare the nonlinear wavelet-based *AMlet* estimator to the linear spline-based *addreg* estimator. We choose sample sizes N equal to $2^9, 2^{11}, 2^{13}, 2^{15}$ to study the evolution of the relative efficiency of the estimators when the sample size increases. The large sample size $2^{15} = 32768$ will highlight the storage and cpu time advantages of *AMlet*. Correspondingly, we repeat the Monte Carlo experiment M times with M equal to $320, 80, 20, 5$ to obtain an equal number of terms in the estimated mean squared errors across sample sizes. We choose the number of covariates $Q = 4$ for two reasons: first, additive models are typically used when the number of covariates is larger than two, and second, treating the case $Q = 2$ in simulations might hide computational or conceptual weaknesses of certain estimators in higher dimensions. We use the following $Q = 4$ functions defined on $[0, 1]$:

- $f_1(x)$ is the relatively smooth `heavisine` function (Donoho and Johnstone, 1994);
- $f_2(x) = 0$ is the smooth `zero` function which represents a non-significant variable as in the simulation of Friedman (1991, §4.3 p. 37);
- $f_3(x)$ is the piecewise constant `blocks` function (Donoho and Johnstone, 1994);
- $f_4(x)$ is the continuous but erratic `bumps` function (Donoho and Johnstone, 1994).

The nonzero functions are then scaled to have a 'standard error' equal to 3:

$$\int_0^1 (f(x) - \bar{f})^2 dx = 3^2, \quad \text{where} \quad \bar{f} = \int_0^1 f(x)dx.$$

We choose a standard deviation of $\sigma = 0.05$ for the Gaussian noise to obtain a large signal-to-noise ratio for the four test functions. We choose the wavelets used by *AMlet* to be the least asymmetric wavelets with 8 vanishing moments, and we fix the number of approximation wavelets in the linear expansion at $j_0 = 5$. We note that our results are not sensitive to these choices. A more important issue is the automatic selection of the smoothing parameters. For *AMlet*, we use either the universal or minimax threshold rules discussed in Section 5. For *addreg*, generalized cross validation is used.

Figure 1 illustrates typical outputs for the three estimators for $N = 8192$. For the two *AMlet* estimates, the universal smoothing parameter provides a smoother appearance, but a worse mean squared error, than the minimax one. The spline-based estimate performs well on smooth functions, but looks wiggly on erratic functions.

Figure 2 reports the estimated mean squared errors (on a log-scale) in a boxplot summary. The layout of the boxplots reveal some interesting features. First, the rate of convergence of the mean squared error is faster for *AMlet* than for *addreg* when estimating non-smooth functions. Second, the minimax rule gives better results for *AMlet* than the universal rule in terms of mean squared error. Third, observe that there is no results for *addreg* when $N = 2^{15}$ as the code crashed, possibly due to the high memory needed to build the smoothing splines' matrix. With wavelets, there is no need to build or store a matrix, since operations are performed with the $O(N)$ wavelet filter, making it suitable for data mining of large data sets.

6.2 Poisson noise – Uniform covariates

A simulation for generalized additive models is less common in the literature, so we extend our simulation of Section 6.1 to Poisson noise to investigate the finite sample properties of *GAMlet*. However, we cannot compare the performance of

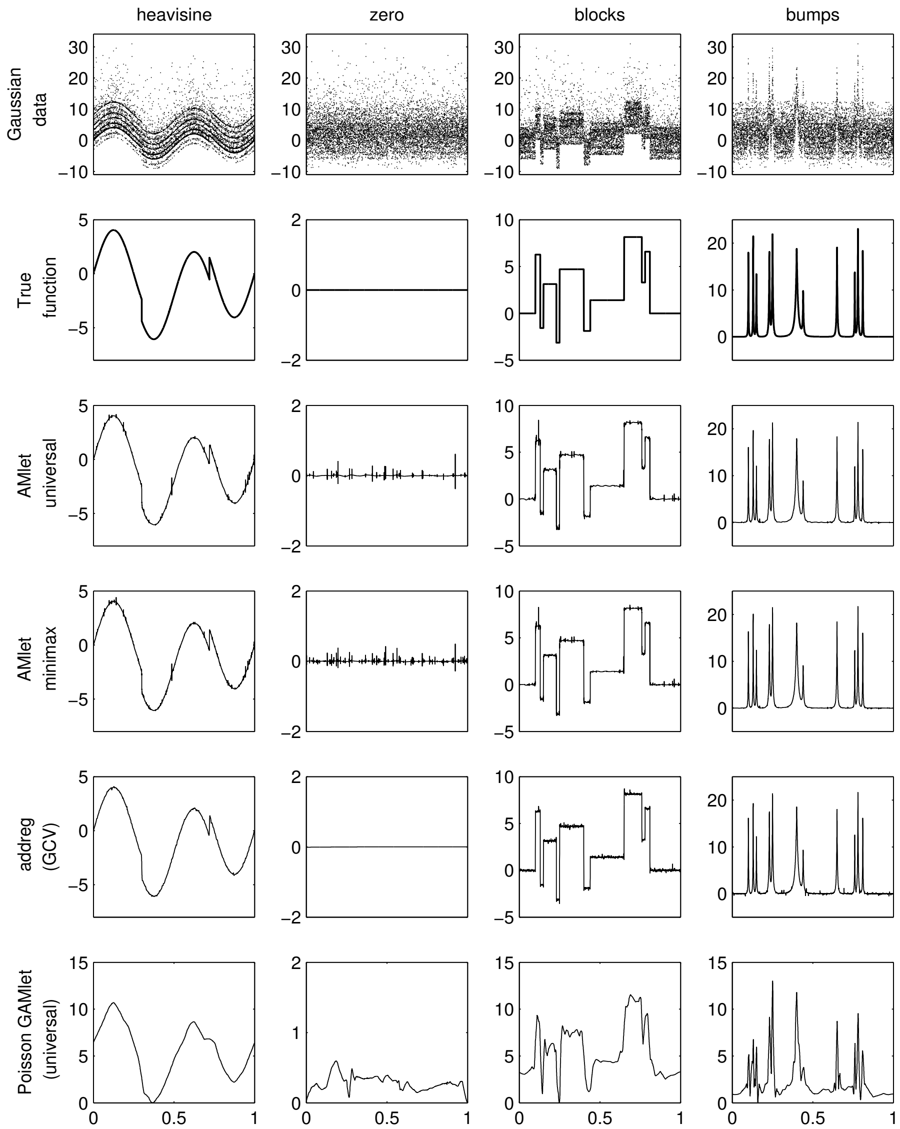

Figure 1: Typical estimates from the simulation of §6.1 and §6.2. Row-wise: raw Gaussian data, true function, *AMlet* universal and minimax, *addreg*, and *GAMlet* from Poisson data. (The raw Poisson data is not plotted since it is visually similar to the Gaussian data). Note: the striations in the upper left-hand plot is a plotting artifact.

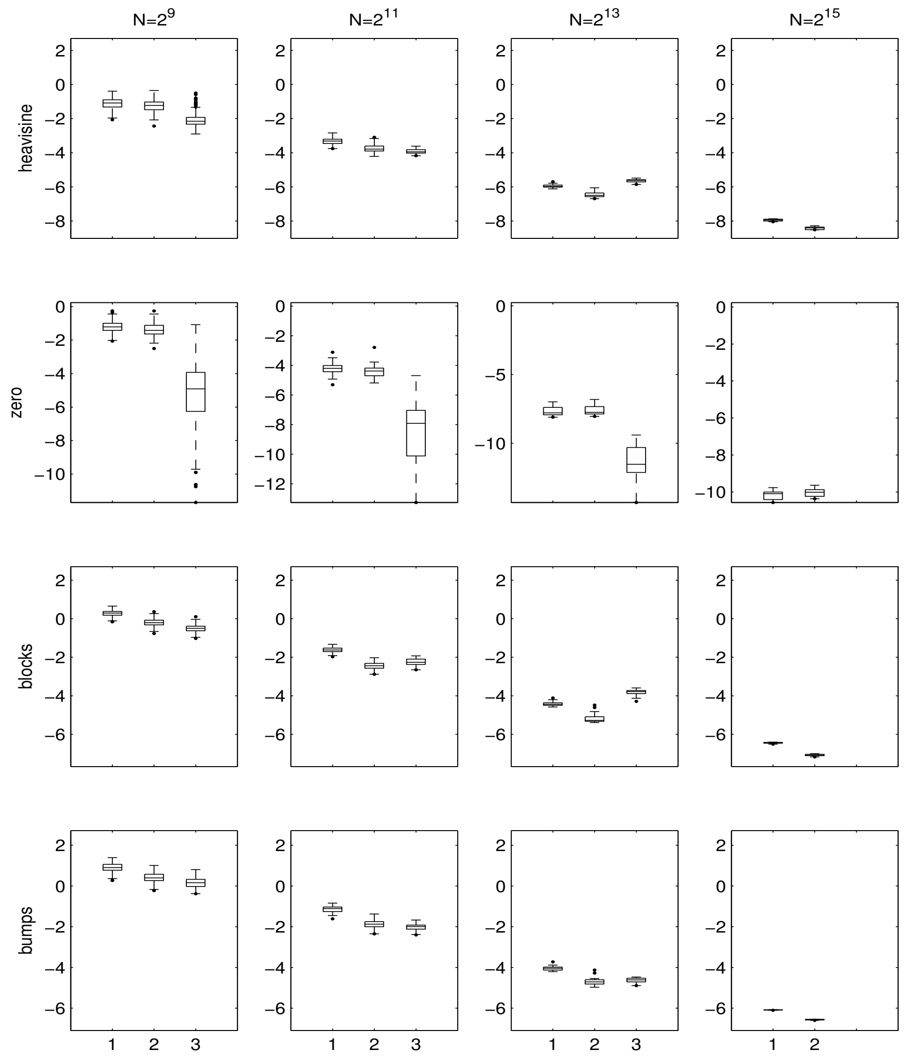

Figure 2: Gaussian simulation of §6.1. Boxplots of the squared errors on a log-scale. '1' is for *AMlet* universal threshold, '2' is for *AMlet* minimax threshold and '3' is for *addreg* GCV. Least asymmetric wavelet of order 8 with $j_0 = 5$.

GAMlet to existing generalized additive model estimators because the latter all use a link function (e.g., log-link) to make a constraint-free estimation (Hastie and Tibshirani, 1986). In contrast, *GAMlet* can handle constraints such as positivity of the Poisson parameters by using the identity link. Using the log-link or the identity link creates two different models since the log-linear model effectively transforms the additive model into a multiplicative model for the Poisson parameter:

$$\mu_n = \exp\left\{\sum_{q=1}^{Q} \eta_q(x_{nq})\right\} = \prod_{q=1}^{Q} \exp\{\eta_q(x_{nq})\}.$$

Because the underlying models are different, a simulation to compare estimators with two different link functions on the basis of mean squared errors criterion is not meaningful. Instead, we report a simulation based on the assumption of a true additive model (i.e., identity link) to investigate how *GAMlet* estimates smooth and non-smooth functions. For the Gaussian simulation, we generate $Q = 4$ independently uniformally distributed covariates. The Poisson random responses are generated according to the model assumed by *GAMlet*, i.e., with Poisson parameters modeled as the linear combination of the four components directly. The linear combination is therefore positive by construction. For the automatic selection of the smoothing parameters, we use the universal threshold (22) that we derived for the identity link.

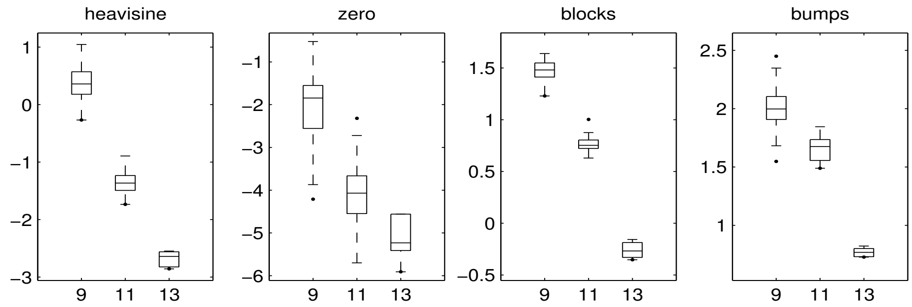

Figure 3: Poisson simulation of §6.2. Boxplots of the squared errors on a log-scale for *GAMlet* with smoothing parameters chosen *automatically*. Least asymmetric wavelet of order 4 with $j_0 = 2$.

The bottom row of Figure 1 shows typical *GAMlet* estimates for $N = 8192$. Despite the oversmoothing effect that we impute to the universal threshold, the estimates reproduce important features of the underlying smooth and non-smooth functions. The complete results of the simulation are presented in Figure 3, where boxplots of the estimated mean squared errors are plotted against increasing sample sizes. As for *AMlet*, we observe that the rate of convergence of *GAMlet* for Poisson noise is faster for erratic functions than for smooth functions.

6.3 CPU time

In terms of computational efficiency, the block coordinate relaxation BCR algorithm for *AMlet* is fast and easy to program in a high level language such as `S-Plus` or `Matlab`. As opposed to spline-based techniques which require the storage of matrices, *AMlet* uses the discrete wavelet transform filter and can therefore be used on massive data sets for data mining purposes. In contrast, we were unable to run *addreg* for sample sizes of $N = 2^{15}$ and larger (`NA` entry in *addreg* column of Table 1). By using `Matlab`'s 'cputime' function, we can make a rough comparison of the computational complexity of the spline-based and the wavelet-based estimators. In particular, *addreg* is programmed in `FORTRAN` while *AMlet* is programmed in `Matlab`, a slower language. Despite this disadvantage, *AMlet* compares favorably with *addreg*, especially for the universal rule. The interior point algorithm for *GAMlet* is CPU intensive when programmed in `Matlab`, but could run in a timely manner if programmed in `C`, for example.

Table 1: CPU time comparison in seconds as a function of sample size.

N	addreg† GCV	AMlet‡ minimax	AMlet‡ universal	GAMlet‡ universal
512	0.8	1.0	0.4	50
2048	1.4	3.0	1.3	150
8192	6.6	8.5	4.1	800
32768	`NA`	23	14	`NA`

† FORTRAN code. ‡ Matlab code

7 Conclusion

To fit or not to fit additive models well might be answered by *AMlet* and *GAMlet*: the wavelet-based *AMlet* and *GAMlet* estimators fit nonlinearly and nonparametrically the smooth and nonsmooth components of additive models and generalized additive models. The *AMlet* estimator is particularly useful when the data set is massive since no existing estimators can estimate additive models automatically in that situation. *AMlet* is also useful when some components of the additive model are believed to be erratic. Moreover, it is easy to program as all it requires is an ordering of the covariates, a wavelet transform, and a 'while' loop until convergence. We recommend using the universal threshold to select significant covariates or to obtain a visually pleasing model, and the minimax threshold for good prediction in the mean squared error sense. *GAMlet* has a higher computational complexity, but has the advantage of handling constraints for generalized additive models.

The definition of *AMlet* and *GAMlet* can also be generalized to semiparametric models, and their corresponding optimization problem can readily be solved by the proposed algorithms. We also see a possible extension of *AMlet* and *GAMlet* to the more general Projection Pursuit (Friedman and Stuetzle, 1981) model that is also additive, but not necessarily in the canonical directions. Using wavelet-based scatterplot smoothers within the Projection Pursuit framework is an interesting area of research.

8 Software availability

The *AMlet* and *GAMlet* Matlab codes are downloadable at
`http://statwww.epfl.ch/people/sardy/Tar/AMlet.tar.gz`
and
`http://statwww.epfl.ch/people/sardy/Tar/GAMlet.tar.gz`
The Figures are also reproducible using the functions provided. We make use of the Wavelab toolbox developed at Stanford
`http://www-stat.stanford.edu/~wavelab/`
The *addreg* code is available on
`http://www.cgd.ucar.edu/stats/Software/Funfits/`

9 Acknowledgments

We are very grateful to the Associate Editor and the three anonymous referees for their constructive comments that led to a substantial improvement of the paper. The first author also wishes to thank Anestis Antoniadis for an invitation to the Université Joseph Fourier in 1999 during which this work was initiated. We thank Werner Stuetzle for comments on an earlier version of the article. We also thank Umberto Amato for his help with using *addreg.* This work was partially supported by the Swiss National Science Foundation and by the National Science Foundation Grant CCR-9731273.

A Convergence proof for AMlet*

For simplicity and without much loss of generality, we consider the case of a common smoothing parameter $\lambda_q = \lambda$ for all $q = 1, \ldots, Q$. Let C be our original cost function (8), namely,

$$C(\underline{\alpha}) = \frac{1}{2}\|\underline{Y} - \bar{\Phi}\underline{\alpha}\|_2^2 + \lambda \sum_{q=1}^{Q} \sum_{j=j_0}^{J} \sum_{\kappa=0}^{2^j-1} |\gamma_{q_{j,\kappa}}|,$$

and let $\underline{\alpha}_k^*$ denote the unique solution to (10). Then

$$C(\underline{\alpha}_k^*) + \epsilon_k \|\underline{\alpha}_k^*\|_2^2 \leq C(\underline{\alpha}^*) + \epsilon_k \|\underline{\alpha}^*\|_2^2. \tag{24}$$

Since $\underline{\alpha}^*$ solves (8) so that $C(\underline{\alpha}^*) \leq C(\underline{\alpha}_k^*)$, the above inequality yields

$$\|\underline{\alpha}_k^*\|_2^2 \leq \|\underline{\alpha}^*\|_2^2. \tag{25}$$

Also, either by direct calculation or by using the fact that $\underline{r}(\underline{\alpha};\epsilon)$ is the smallest subgradient of the convex function $C(\underline{\alpha}) + \epsilon\|\underline{\alpha}\|_2^2$ at $\underline{\alpha}$, we have for any $\underline{\alpha}$ and $\underline{\tilde{\alpha}}$ that

$$C(\underline{\alpha}) + \epsilon\|\underline{\alpha}\|_2^2 + \underline{r}(\underline{\alpha};\epsilon)'(\underline{\tilde{\alpha}} - \underline{\alpha}) \leq C(\underline{\tilde{\alpha}}) + \epsilon\|\underline{\tilde{\alpha}}\|_2^2,$$

where the ith element of the subgradient $\underline{r}(\underline{\alpha};\epsilon)$ is

$$r_i(\underline{\alpha};\epsilon) = \begin{cases} \nabla_i c(\underline{\alpha};\epsilon) & \text{if } \alpha_i = \beta_i \\ \nabla_i c(\underline{\alpha};\epsilon) + \lambda\gamma_i/|\gamma_i| & \text{if } |\gamma_i| \neq 0 \\ \nabla_i c(\underline{\alpha};\epsilon) + \eta_i & \text{if } |\gamma_i| = 0, \end{cases} \qquad i = 1,\ldots,NQ, \tag{26}$$

with $\eta_i = \arg\min_{0\leq|\eta|\leq\lambda} |\nabla_i c(\underline{\alpha};\epsilon) + \eta|$ and $c(\underline{\alpha};\epsilon) = \frac{1}{2}\|\underline{Y} - \bar{\Phi}\underline{\alpha}\|_2^2 + \epsilon\|\underline{\alpha}\|_2^2$.

Then, letting $\epsilon = \epsilon_k$, $\underline{\tilde{\alpha}} = \underline{\alpha}_k^*$ and using (24), we obtain

$$\begin{aligned} C(\underline{\alpha}) + \epsilon_k\|\underline{\alpha}\|_2^2 &\leq C(\underline{\alpha}^*) + \epsilon_k\|\underline{\alpha}^*\|_2^2 + \underline{r}(\underline{\alpha};\epsilon_k)'(\underline{\alpha} - \underline{\alpha}_k^*) \\ &= C(\underline{\alpha}^*) + \epsilon_k\|\underline{\alpha}^*\|_2^2 + \sum_{q=1}^{Q} \underline{r}_q(\underline{\alpha};\epsilon_k)'(\underline{\alpha}_q - (\underline{\alpha}_k^*)_q) \\ &\leq C(\underline{\alpha}^*) + \epsilon_k\|\underline{\alpha}^*\|_2^2 + \sum_{q=1}^{Q} \|\underline{r}_q(\underline{\alpha};\epsilon_k)\|_2\|\underline{\alpha}_q - (\underline{\alpha}_k^*)_q\|_2 \\ &\leq C(\underline{\alpha}^*) + \epsilon_k\|\underline{\alpha}^*\|_2^2 \\ &\quad + \max_{q\in\{1,\ldots,Q\}} \|\underline{r}_q(\underline{\alpha};\epsilon_k)\|_2 \cdot \sum_{q=1}^{Q}(\|\underline{\alpha}_q\|_2 + \|(\underline{\alpha}_k^*)_q\|_2). \end{aligned}$$

Letting $\underline{\alpha} = \underline{\alpha}_{k+1}$ and using the fact that $\underline{\alpha}$ satisfies (11) yields

$$\begin{aligned} C(\underline{\alpha}_{k+1}) + \epsilon_k\|\underline{\alpha}_{k+1}\|_2^2 &\leq C(\underline{\alpha}^*) + \epsilon_k\|\underline{\alpha}^*\|_2^2 + \delta_k + \delta_k\sum_{q=1}^{Q}\|(\underline{\alpha}_k^*)_q\|_2 \\ &\leq C(\underline{\alpha}^*) + \epsilon_k\|\underline{\alpha}^*\|_2^2 + \delta_k + \delta_k Q\|\underline{\alpha}^*\|_2, \end{aligned} \tag{27}$$

where the last inequality uses (25). Since $\underline{\alpha}^*$ solves (8) so that $C(\underline{\alpha}^*) \leq C(\underline{\alpha}_{k+1})$, the above inequality yields

$$\|\underline{\alpha}_{k+1}\|_2^2 \leq \|\underline{\alpha}^*\|_2^2 + (\delta_k/\epsilon_k)(1 + Q\|\underline{\alpha}^*\|_2). \tag{28}$$

This together with (12) implies that $\|\underline{\alpha}_{k+1}\|_2$ is bounded as $k \to \infty$, so $\{\underline{\alpha}_{k+1}\}$ has cluster points. By (27) and $\{\epsilon_k\} \to 0$, $\{\delta_k\} \to 0$, we see that any cluster point $\underline{\bar{\alpha}}$ satisfies

$$C(\underline{\bar{\alpha}}) \leq C(\underline{\alpha}^*).$$

Thus $\underline{\bar{\alpha}}$ is a solution to (8). Moreover, (28) and (12) yields in the limit

$$\|\underline{\bar{\alpha}}\|_2^2 \le \|\underline{\alpha}^*\|_2^2.$$

Since $\underline{\alpha}^*$ is the unique minimum ℓ_2 norm solution to (8) (the solution set of (8) is a closed convex set, so it has a unique point whose ℓ_2 norm is minimum), this implies $\underline{\bar{\alpha}} = \underline{\alpha}^*$. This shows that $\{\underline{\alpha}_{k+1}\}$ is a bounded sequence of points with $\underline{\alpha}^*$ as its only cluster point. Hence $\{\underline{\alpha}_{k+1}\} \to \underline{\alpha}^*$. □

B Algorithm and convergence proof for GAMlet*

For simplicity and without much loss of generality, we consider the case of a common smoothing parameter $\lambda_q = \lambda$ for all $q = 1, \dots, Q$. We wish to solve the two-level optimization problem (15), i.e., find the minimum ℓ_2 norm solution to

$$\min_{\underline{\alpha}=(\underline{\beta},\underline{\gamma})} -l(\bar{\Phi}\underline{\alpha}; \underline{Y}) + \lambda\|\underline{\gamma}\|_1. \tag{29}$$

We consider a solution approach based on Tikhonov regularization whereby, at each iteration, we apply a primal-dual interior point method to solve approximately a regularized problem of the form ($\epsilon > 0$):

$$\min_{\underline{\alpha}=(\underline{\beta},\underline{\gamma})} -l(\bar{\Phi}\underline{\alpha}; \underline{Y}) + \lambda\|\underline{\gamma}\|_1 + \frac{\epsilon}{2}\|\underline{\alpha}\|_2^2. \tag{30}$$

To ensure convergence, ϵ is decreased towards zero after each iteration and the solution accuracy tolerance is decreased towards zero sufficiently fast relative to ϵ and the log-barrier parameter ρ. As in Sardy, Antoniadis, and Tseng (2003), we assume for all $n = 1, \dots, N$ that $-l(\cdot; Y_n)$ is twice differentiable, its second derivative is always positive, and its first derivative tends to $-\infty$ and $+\infty$ at the left and right endpoints of its domain C_n. This assumption is satisfied for the Gaussian and the exponential distribution, as well as for the Poisson distribution with positive counts.

First, we describe how to apply a primal-dual interior point method to solve (30). Following the derivation in Section 4.2 on *GAMlet*, we rewrite the primal problem (30) as

$$\min_{\underline{\eta}\in C, \underline{\alpha}=(\underline{\beta},\underline{\gamma}), \underline{\zeta}} \max_{\underline{y},\underline{v}} -l(\underline{\eta}; \underline{Y}) + \lambda\|\underline{\gamma}\|_1 + \underline{y}'(\underline{\eta} - \bar{\Phi}\underline{\alpha}) + \underline{v}'(\underline{\zeta} - \underline{\gamma}) + \frac{\epsilon}{2}\|\underline{\beta}\|_2^2 + \frac{\epsilon}{2}\|\underline{\zeta}\|_2^2.$$

Interchanging "max" and "min" gives the dual problem

$$\max_{\underline{y},\underline{v}} h(\underline{y}; \underline{Y}) - \frac{1}{2\epsilon}\|\bar{\Phi}_0'\underline{y}\|_2^2 - \frac{1}{2\epsilon}\|\underline{v}\|_2^2 \text{ with } -\lambda\underline{1} \le \bar{\Psi}'\underline{y} + \underline{v} \le \lambda\underline{1},\ \underline{y} \in K,$$

where K is the domain of $h(\cdot;\underline{Y})$. Notice that, as $\epsilon \to 0^+$, the above dual problem reduces to the dual problem (5) of Sardy, Antoniadis, and Tseng (2003). The corresponding log-barrier subproblem is

$$\min_{\underline{y},\underline{v}} -h(\underline{y};\underline{Y})+\frac{1}{2\epsilon}\|\bar{\Phi}_0'\underline{y}\|_2^2+\frac{1}{2\epsilon}\|\underline{v}\|_2^2-\rho\sum_p \log(\lambda-\bar{\Psi}_p'\underline{y}-v_p)-\rho\sum_p \log(\lambda+\bar{\Psi}_p'\underline{y}+v_p),$$

where $\rho > 0$ and $h(\underline{y};\underline{Y}) = \min_{\eta\in C} \underline{y}'\underline{\eta} - l(\underline{\eta};\underline{Y})$. By introducing the slack variables $\underline{z} = (\underline{z}_+,\underline{z}_-) = (\lambda\underline{1} - \bar{\Psi}'\underline{y} - \underline{v}, \lambda\underline{1} + \bar{\Psi}'\underline{y} + \underline{v})$ and $\underline{x} = (\underline{x}_+,\underline{x}_-) = \rho((Z_+)^{-1}\underline{1},(Z_-)^{-1}\underline{1})$, the Karush–Kuhn–Tucker conditions for the subproblem can be written as

$$\begin{aligned}
-\nabla h(\underline{y};\underline{Y}) + \frac{1}{\epsilon}\bar{\Phi}_0\bar{\Phi}_0'\underline{y} + \bar{\Psi}\underline{x}_+ - \bar{\Psi}\underline{x}_- &= \underline{0}\\
\frac{1}{\epsilon}\underline{v} + \underline{x}_+ - \underline{x}_- &= \underline{0}\\
-\underline{v} - \bar{\Psi}'\underline{y} + \lambda\underline{1} - \underline{z}_+ &= \underline{0}\\
\underline{v} + \bar{\Psi}'\underline{y} + \lambda\underline{1} - \underline{z}_- &= \underline{0}\\
\rho\underline{1} - X\underline{z} &= \underline{0}.
\end{aligned}$$

Here $X = \mathrm{diag}(\underline{x})$ and similarly for Z, Z_+ and Z_-. Eliminating $\underline{v}$ and letting $\underline{\beta} = -\frac{1}{\epsilon}\bar{\Phi}_0'\underline{y}$ and $A = [\bar{\Psi}\ \ -\bar{\Psi}]$, $B = [I\ \ -I]$, $\underline{c} = \lambda(\underline{1},\underline{1})$ and letting $\mu_{\min}(\underline{y};\underline{Y}) = -\nabla h(\underline{y};\underline{Y})$, this can be written equivalently as

$$\begin{array}{ll}
-\mu_{\min}(\underline{y};\underline{Y}) + \bar{\Phi}_0\underline{\beta} - A\underline{x} & =: \underline{r}_y = \underline{0}\\
\epsilon B'B\underline{x} - A'\underline{y} + \underline{c} - \underline{z} & =: \underline{r}_x = \underline{0}\\
\rho\underline{1} - X\underline{z} & =: \underline{r}_z = \underline{0}\\
-\bar{\Phi}_0'\underline{y} - \epsilon\underline{\beta} & =: \underline{r}_\beta = \underline{0}
\end{array} \qquad (31)$$

The corresponding Newton equation has the form:

$$\begin{aligned}
Q\Delta\underline{y} - \bar{\Phi}_0\Delta\underline{\beta} &= \underline{r}\\
\bar{\Phi}_0'\Delta\underline{y} + \epsilon\Delta\underline{\beta} &= \underline{r}_\beta,
\end{aligned}$$

where $P = \mathrm{diag}(-\mu'_{\min}(\underline{y};\underline{Y}))$ and $Q = P + A(ZX^{-1} + \epsilon B'B)^{-1}A'$. The above system of equations can be solved by applying the conjugate gradient method to its least square reformulation. Then $(\underline{y},\underline{\beta})$ is updated by moving it in the direction $(\Delta\underline{y},\Delta\underline{\beta})$ and similarly for $(\underline{x},\underline{z})$, ρ is decreased, and the process is reiterated (Sardy, Antoniadis, and Tseng, 2003), (Kojima, Megiddo, and Mizuno, 1993).

Now we describe the algorithm for finding the unique minimum ℓ_2 norm solution to (29). At the kth iteration ($k = 1, 2, \ldots$), a regularization parameter $\epsilon_k > 0$ and an accuracy tolerance $\delta_k > 0$ are chosen, and we apply a primal-dual interior point method, with $(\underline{x},\underline{y},\underline{z},\underline{\beta})$ and ρ suitably initialized (either from scratch or using values generated from the previous iteration), to solve (30) with $\epsilon = \epsilon_k$. We terminate the method when the current $(\underline{x}, \underline{y}, \underline{z}, \underline{\beta})$, together with the current log-barrier parameter ρ, solves (30) with accuracy δ_k in the sense that

$$\underline{r}_z \geq -\theta\rho\underline{1}, \tag{32}$$

$$\begin{array}{ll} \|\underline{r}_x\| \leq \delta_k, & \|\nabla l(\bar{\Phi}\underline{\alpha} + \underline{r}_y; \underline{Y}) - \nabla l(\bar{\Phi}\underline{\alpha}; \underline{Y})\| \leq \delta_k, \\ \|\underline{r}_\beta\| \leq \delta_k, & \sqrt{\rho}\|\underline{\gamma}\| \leq \delta_k, \quad \sqrt{\rho} \leq \delta_k, \end{array} \tag{33}$$

where we let $\underline{\gamma} = B\underline{x}$, $\underline{\alpha} = (\underline{\beta}, \underline{\gamma})$, $\theta > 0$ is some constant (independent of k), and the norm $\|\cdot\|$ can be any ℓ_p norm $(1 \leq p \leq \infty)$. The condition (32) is satisfied provided the interior point method maintains $\underline{r}_z/\rho$ to be uniformly bounded below componentwise. The condition (33) is satisfied after a finite number of interior point iterations provided the method maintains $\bar{\Phi}\underline{\alpha}$ and $\underline{\gamma}$ to be bounded and drives $\underline{r}_x, \underline{r}_y, \underline{r}_z, \underline{r}_\beta$ and ρ to zero, as is generally the case. Then, we choose an $\epsilon_{k+1} < \epsilon_k$ (e.g., $\epsilon_{k+1} = \epsilon_k/3$) and a $\delta_{k+1} < \delta_k$, and set $\underline{\alpha}_k$ to be the $\underline{\alpha}$ corresponding to the $(\underline{x}, \underline{y}, \underline{z}, \underline{\beta})$ satisfying (32) and (33).

We claim that, by choosing ϵ_k to tend to zero and by choosing δ_k so that

$$\lim_{k\to\infty} \delta_k/\epsilon_k = 0 \tag{34}$$

(e.g., $\delta_k = (\epsilon_k)^\nu$, with $\nu > 1$), then $\{\underline{\alpha}_k\}$ would converge to the unique minimum 2-norm solution $\underline{\alpha}^*$ of (29). To see this, note that since $\underline{\alpha}^* = (\underline{\beta}^*, \underline{\gamma}^*)$ solves (29), then

$$\underline{0} = -\bar{\Phi}'\nabla l(\bar{\Phi}\underline{\alpha}^*; \underline{Y}) + (\underline{0}, \lambda\underline{\eta}^*), \tag{35}$$

for some subgradient $\underline{\eta}^*$ of $\|\cdot\|_1$ at $\underline{\gamma}^*$, i.e.,

$$\eta_p^* \in \Gamma(\gamma_p^*) \; := \; \begin{cases} 1 & \text{if } \gamma_p^* > 0 \\ [-1, 1] & \text{if } \gamma_p^* = 0 \\ -1 & \text{if } \gamma_p^* < 0 \end{cases} .$$

Fix any k and any $(\underline{x}, \underline{y}, \underline{z}, \underline{\beta})$ and ρ satisfying (32) and (33). Let

$$M = Q(N - p_0), \qquad \underline{\gamma} = -B\underline{x} = \underline{x}_- - \underline{x}_+, \qquad \underline{\alpha} = (\underline{\beta}, \underline{\gamma}).$$

Then we have from (31) that

$$\begin{aligned} \nabla h(\underline{y}; \underline{Y}) + \bar{\Phi}\underline{\alpha} &= \underline{r}_y & (36) \\ -\epsilon\underline{\gamma} - \bar{\Psi}'\underline{y} + \lambda\underline{1} - \underline{z}_+ &= \underline{r}_{x+} & (37) \\ \epsilon\underline{\gamma} + \bar{\Psi}'\underline{y} + \lambda\underline{1} - \underline{z}_- &= \underline{r}_{x-} & (38) \\ \rho\underline{1} - X\underline{z} &= \underline{r}_z & (39) \\ \bar{\Phi}_0'\underline{y} + \epsilon\underline{\beta} &= -\underline{r}_\beta & (40) \end{aligned}$$

Fix any p. We consider three cases:

Case 1. $\gamma_p \geq \sqrt{\rho}$: Since $\gamma_p = x_{M+p} - x_p$ and $\underline{x} > 0$, this implies $x_{M+p} > \sqrt{\rho}$ so (39) and (32) yield

$$z_{M+p} = \frac{\rho - (r_z)_{M+p}}{x_{M+p}} \leq \frac{\rho + \theta\rho}{x_{M+p}} \leq (1+\theta)\sqrt{\rho}.$$

In this case, let $\eta_p = 1 - z_{M+p}/\lambda$ and $\hat{\eta}_p = 1$. Then

$$|\eta_p - \hat{\eta}_p| = z_{M+p}/\lambda \leq (1+\theta)\sqrt{\rho}/\lambda, \tag{41}$$

and we have from (38) that

$$\epsilon\gamma_p + \bar{\Psi}'_p y + \lambda\eta_p = (r_x)_{M+p}. \tag{42}$$

Case 2. $\gamma_p \leq -\sqrt{\rho}$: Since $\gamma_p = x_{M+p} - x_p$ and $\underline{x} > 0$, this implies $x_p > \sqrt{\rho}$ so (39) and (32) yield

$$z_p = \frac{\rho - (r_z)_p}{x_p} \leq \frac{\rho + \theta\rho}{x_p} \leq (1+\theta)\sqrt{\rho}.$$

In this case, let $\eta_p = -1 + z_p/\lambda$ and $\hat{\eta}_p = -1$. Then

$$|\eta_p - \hat{\eta}_p| = z_p/\lambda \leq (1+\theta)\sqrt{\rho}/\lambda \tag{43}$$

and we have from (37) that

$$\epsilon\gamma_p + \bar{\Psi}'_p y + \lambda\eta_p = -(r_x)_p. \tag{44}$$

Case 3. $|\gamma_p| < \sqrt{\rho}$: By summing (37) and (38), we obtain

$$2\lambda 1 - (z_p + z_{M+p}) = (r_x)_p + (r_x)_{M+p}$$

so that

$$\frac{(z_p + z_{M+p})}{2} = \lambda - \frac{(r_x)_p + (r_x)_{M+p}}{2}.$$

In this case, let $\eta_p = (z_p - z_{M+p})/(2\lambda)$ and $\hat{\eta}_p = \max\{-1, \min\{1, \eta_p\}\}$. Since $z_p > 0, z_{M+p} > 0$, then

$$|\eta_p| \leq \frac{(z_p + z_{M+p})}{2\lambda} = 1 - \frac{(r_x)_p + (r_x)_{M+p}}{2\lambda}.$$

It follows from the definition of $\hat{\eta}_p$ that

$$|\eta_p - \hat{\eta}_p| \;=\; \begin{cases} |\eta_p| - 1 & \text{if } |\eta_p| > 1 \\ 0 & \text{else} \end{cases} \;\leq\; \frac{|(r_x)_p + (r_x)_{M+p}|}{2\lambda}. \tag{45}$$

Also, subtracting (37) from (38) and then dividing both sides by 2 yields

$$\epsilon\gamma_p + \bar{\Psi}'_p y + \lambda\eta_p = ((r_x)_{M+p} - (r_x)_p)/2. \tag{46}$$

Then, we have from (36) and the observation that $\underline{\eta} = -\nabla h(\underline{y};\underline{Y})$ if and only if $-\nabla l(\underline{\eta};\underline{Y}) = \underline{y}$ (i.e., $-\nabla h(\cdot;\underline{Y})$ is the inverse function of $-\nabla l(\cdot;\underline{Y})$) that

$$\underline{y} = -\nabla l(\bar{\Phi}\underline{\alpha} + \underline{r}_y;\underline{Y}) = -\nabla l(\bar{\Phi}\underline{\alpha};\underline{Y}) - \underline{r}_0,$$

where we let $\underline{r}_0 = \nabla l(\bar{\Phi}\underline{\alpha}+\underline{r}_y;\underline{Y}) - \nabla l(\bar{\Phi}\underline{\alpha};\underline{Y})$. Substituting this into (40) yields

$$-\bar{\Phi}_0'\nabla l(\bar{\Phi}\underline{\alpha};\underline{Y}) + \epsilon\underline{\beta} = \underline{r}_1.$$

where we let $\underline{r}_1 = \bar{\Phi}_0'\underline{r}_0 - \underline{r}_\beta$. Also, we have from (42), (44), (46) that

$$-\bar{\Psi}'\nabla l(\bar{\Phi}\underline{\alpha};\underline{Y}) + \epsilon\underline{\gamma} + \lambda\underline{\eta} = \bar{\Psi}'\underline{r}_0 + \underline{r}_2,$$

where $(r_2)_p$ is either $(r_x)_{M+p}$ or $-(r_x)_p$ or $((r_x)_{M+p} - (r_x)_p)/2$, depending on which case. Then, letting $\underline{r} = (\underline{r}_1, \bar{\Psi}'\underline{r}_0 + \underline{r}_2)$, we obtain

$$-\bar{\Phi}'\nabla l(\bar{\Phi}\underline{\alpha};\underline{Y}) + \epsilon\underline{\alpha} + (0, \lambda\underline{\eta}) = \underline{r}.$$

Subtracting (35) from this yields

$$\underline{r} - \epsilon\underline{\alpha} = \bar{\Phi}'(-\nabla l(\bar{\Phi}\underline{\alpha};\underline{Y}) + \nabla l(\bar{\Phi}\underline{\alpha}^*;\underline{Y})) + (0, \lambda(\underline{\eta} - \underline{\eta}^*))$$

so that

$$\begin{aligned}
& (\underline{\alpha} - \underline{\alpha}^*)'(\underline{r} - \epsilon\underline{\alpha}) \\
= \;& (\bar{\Phi}\underline{\alpha} - \bar{\Phi}\underline{\alpha}^*)'(-\nabla l(\bar{\Phi}\underline{\alpha};\underline{Y}) + \nabla l(\bar{\Phi}\underline{\alpha}^*;\underline{Y})) + \lambda(\underline{\gamma} - \underline{\gamma}^*)'(\underline{\eta} - \underline{\eta}^*) \\
\geq \;& \lambda(\underline{\gamma} - \underline{\gamma}^*)'(\underline{\eta} - \underline{\eta}^*) \\
= \;& \lambda \sum_{|\gamma_p| \geq \sqrt{\rho}} (\gamma_p - \gamma_p^*)(\eta_p - \hat{\eta}_p) + \lambda \sum_{|\gamma_p| \geq \sqrt{\rho}} (\gamma_p - \gamma_p^*)(\hat{\eta}_p - \eta_p^*) \\
& + \lambda \sum_{|\gamma_p| < \sqrt{\rho}} \gamma_p(\eta_p - \eta_p^*) - \lambda \sum_{|\gamma_p| < \sqrt{\rho}} \gamma_p^*(\eta_p - \hat{\eta}_p) \\
& + \lambda \sum_{|\gamma_p| < \sqrt{\rho}} (0 - \gamma_p^*)(\hat{\eta}_p - \eta_p^*) \\
\geq \;& \lambda \sum_{|\gamma_p| \geq \sqrt{\rho}} (\gamma_p - \gamma_p^*)(\eta_p - \hat{\eta}_p) + \lambda \sum_{|\gamma_p| < \sqrt{\rho}} \gamma_p(\eta_p - \eta_p^*) \\
& - \lambda \sum_{|\gamma_p| < \sqrt{\rho}} \gamma_p^*(\eta_p - \hat{\eta}_p) \\
\geq \;& - \sum_{|\gamma_p| \geq \sqrt{\rho}} |\gamma_p - \gamma_p^*|(1+\theta)\sqrt{\rho} - \lambda \sum_{|\gamma_p| < \sqrt{\rho}} \sqrt{\rho}|\eta_p - \eta_p^*| \\
& - \frac{1}{2} \sum_{|\gamma_p| < \sqrt{\rho}} |\gamma_p^*||(r_x)_p + (r_x)_{M+p}| \\
=: \;& -\hat{r},
\end{aligned}$$

where the first inequality uses the convexity of $-l(\cdot;\underline{Y})$ so that $-\nabla l(\cdot;\underline{Y})$ is monotone. The second inequality uses the monotone property of $\Gamma(\cdot)$ and the observation that $\eta_p^* \in \Gamma(\gamma_p^*)$ for all p and $\hat{\eta}_p \in \Gamma(\gamma_p)$ for $|\gamma_p| \geq \sqrt{\rho}$ and $\hat{\eta}_p \in \Gamma(0)$ for $|\gamma_p| < \sqrt{\rho}$. The third inequality uses (41), (43), (45). This implies

$$\begin{aligned} \|\underline{\alpha}^*\|_2^2 &= \|\underline{\alpha}^* - \underline{\alpha}\|_2^2 + \|\underline{\alpha}\|_2^2 + 2(\underline{\alpha}^* - \underline{\alpha})'\underline{\alpha} \\ &\geq \|\underline{\alpha}^* - \underline{\alpha}\|_2^2 + \|\underline{\alpha}\|_2^2 + 2(\underline{\alpha}^* - \underline{\alpha})'r/\epsilon - \hat{r}/\epsilon. \end{aligned} \tag{47}$$

Also, using (45) and the definition of $\hat{r}$, we see that

$$0 \leq \hat{r} \leq C(\|\underline{\gamma}\|\sqrt{\rho} + \sqrt{\rho} + \|\underline{r}_x\|\sqrt{\rho} + \|\underline{r}_x\|) \tag{48}$$

for some constant $C > 0$ depending on $\underline{\lambda}, \underline{\gamma}^*, \underline{\eta}^*$ only.

By our choice of α_k, we obtain from (47) that

$$\|\underline{\alpha}^*\|_2^2 \geq \|\underline{\alpha}_k\|_2^2 + 2(\underline{\alpha}^* - \underline{\alpha}_k)'\underline{r}_k/\epsilon_k - \hat{r}_k/\epsilon_k,$$

where $\underline{r}_k$ and $\hat{r}_k$ are the corresponding $\underline{r}$ and $\hat{r}$. We see from (33), (34) and (48) that $\hat{r}_k/\epsilon_k \to 0$. Similarly, we see from the definition of r_k and (33), (34) that $\|\underline{r}_k\|_2/\epsilon_k \to 0$. Using the Cauchy-Schwartz inequality $(\underline{\alpha}_k)'\underline{r}_k \leq \|\underline{\alpha}_k\|_2\|\underline{r}_k\|_2$ and rearranging terms, we obtain the inequality

$$\|\underline{\alpha}_k\|_2^2 - 2\|\underline{\alpha}_k\|_2 b_k - \|\underline{\alpha}^*\|_2^2 + c_k \leq 0,$$

where we let $c_k = 2(\underline{\alpha}^*)'\underline{r}_k/\epsilon_k - \hat{r}_k/\epsilon_k \to 0$ and $b_k = \|\underline{r}_k\|_2/\epsilon_k$. This is a quadratic inequality of the form $t^2 - 2bt - c \leq 0$ with $t \geq 0$, which has solutions $0 \leq t \leq b + \sqrt{b^2 + c}$. Thus

$$\|\underline{\alpha}_k\|_2 \leq \sqrt{(b_k)^2 + \|\underline{\alpha}^*\|_2^2 - c_k}.$$

Since $b_k \to 0$ and $c_k \to 0$, this implies $\{\alpha_k\}$ is bounded and any cluster point $\bar{\alpha}$ satisfies

$$\|\underline{\bar{\alpha}}\|_2 \leq \|\underline{\alpha}^*\|_2.$$

Moreover, it can be argued using (31), (32), (33), (34) that $\underline{\bar{\alpha}}$ satisfies the 1st-order optimality condition for (29). Since (29) is a convex program, this implies $\underline{\bar{\alpha}}$ is a solution to (29). Since $\underline{\alpha}^*$ is the unique minimum 2-norm solution to (29), this implies $\underline{\bar{\alpha}} = \underline{\alpha}^*$. This shows that $\{\underline{\alpha}_k\}$ is a bounded sequence of points with $\underline{\alpha}^*$ as its only cluster point. Hence $\{\underline{\alpha}_k\} \to \underline{\alpha}^*$. □

References

Abramovich, F., Sapatinas, T., and Silverman, B. W. (1998). Wavelet thresholding via a Bayesian approach. *Journal of the Royal Statistical Society, Series B* **60**, 725–749.

Albert, A. (1972). *Regression and the Moore-Penrose Pseudoinverse.* New York: Academic Press.

Amato, U. and Antoniadis, A. (2001). Adaptive wavelet series estimation in separable nonparametric regression models. *Statistics and Computing* **11**, 373–394.

Anscombe, F. J. (1948). The transformation of Poisson, binomial and negative-binomial data. *Biometrika* **35**, 246–254.

Antoniadis, A. and Fan, J. (2001). Regularization of wavelet approximations. *Journal of the American Statistical Association* **96**, 939–967.

Antoniadis, A., Gregoire, G., and McKeague, I. W. (1994). Wavelet methods for curve estimation. *Journal of the American Statistical Association* **89**, 1340–1353.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Annals of Statistics* **24**, 2350–2383.

Buja, A., Hastie, T. J., and Tibshirani, R. J. (1989). Linear smoothers and additive models (with discussion). *Annals of Statistics* **17**, 453–555.

Cantoni, E. and Hastie, T. J. (2002). Degrees of freedom tests for smoothing splines. *Biometrika* **89**, 251–263,.

Chen, S. S., Donoho, D. L., and Saunders, M. A. (1999). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing* **20(1)**, 33–61.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika* **81**, 425–455.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association* **90**, 1200–1224.

Donoho, D. L., Johnstone, I. M., Hoch, J. C., and Stern, A. S. (1992). Maximum entropy and the nearly black object (disc: P67-81). *Journal of the Royal Statistical Society, Series B, Methodological* **54**, 41–67.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1995). Wavelet shrinkage: Asymptopia? (with discussion). *Journal of the Royal Statistical Society, Series B* **57**, 301–369.

Friedman, J. H. (1991). Multivariate adaptive regression with splines (with discussion). *Annals of Statistics* **19**, 1–141.

Friedman, J. H. and Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modeling (with discussion). *Technometrics* **31**, 3–39.

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association* **76**, 817–823.

Hastie, T. J. and Tibshirani, R. J. (1986). Generalized additive models. *Statistical Science* **1**, 295–318.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models.* Washington, D.C.: Chapman and Hall.

Huber, P. J. (1981). *Robust Statistics.* New York: John Wiley.

Kojima, M., Megiddo, N., and Mizuno, S. (1993). A primal-dual exterior point algorithm for linear programming. *Mathematical Programming* **61**, 261–280.

Mallat, S. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11**, 674–693.

Mammen, E., Linton, O., and Nielsen, J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Annals of Statistics* **27**, 1443–1490.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A* **135**, 370–384.

Nychka, D., Bailey, B., Ellner, S., Haaland, P., and O'Connel, P. (1993). Funfits: Data analysis and statistical tools for estimating functions. Technical report, Raleigh, North Carolina State University, Raleigh.

Opsomer, J. D. and Ruppert, D. (1997). Fitting a bivariate additive model by local polynomial regression. *Annals of Statistics* **25**, 186–211.

Rockafellar, R. T. (1984). *Network Flows and Monotropic Programming.* New-York: Wiley-Interscience; republished by Athena Scientific, Belmont, 1998.

Sardy, S., Antoniadis, A., and Tseng, P. (2003). Automatic smoothing with wavelets for a wide class of distributions. *Tentatively accepted in the Journal of Computational and Graphical Statistics.*

Sardy, S., Bruce, A. G., and Tseng, P. (2000). Block coordinate relaxation methods for nonparametric wavelet denoising. *Journal of Computational and Graphical Statistics* **9**, 361–379.

Sardy, S., Percival, D. B., Bruce, A. G., Gao, H.-Y., and Stuetzle, W. (1999). Wavelet de-noising for unequally spaced data. *Statistics and Computing* **9**, 65–75.

Sardy, S., Tseng, P., and Bruce, A. G. (2001). Robust wavelet denoising. *IEEE Transactions on Signal Processing* **49**, 1146–1152.

Stone, C. J. (1985). Additive regression and other nonparametric models. *Annals of Statistics* **13**, 689–705.

Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Annals of Statistics* **14**, 590–606.

Tseng, P. (2001). Convergence of block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications* **109**, 475–494.