Adaptive wavelet series estimation in separable nonparametric regression models

Umberto Amato
Istituto per Applicazioni della Matematica CNR
Via Pietro Castellino 111, 80131 Napoli, Italy
e-mail: amato@iamna.iam.na.cnr.it

and

Anestis Antoniadis[1]
Laboratoire LMC - IMAG
Université Joseph Fourier
BP 53, 38041 Grenoble Cedex 09, France
e-mail: anestis.antoniadis@imag.fr

Abstract

It is well-known that multivariate curve estimation suffers from the "curse of dimensionality". However, reasonable estimators are possible, even in several dimensions, under appropriate restrictions on the complexity of the curve. In the present paper we explore how much appropriate wavelet estimators can exploit a typical restriction on the curve such as additivity. We first propose an adaptive and simultaneous estimation procedure for all additive components in additive regression models and discuss rate of convergence results and data-dependent truncation rules for wavelet series estimators. To speed up computation we then introduce a wavelet version of functional ANOVA algorithm for additive regression models and propose a regularization algorithm which guarantees an adaptive solution to the multivariate estimation problem. Some simulations indicate that wavelets methods complement nicely the existing methodology for nonparametric multivariate curve estimation.

Keywords: WAVELET SERIES; ADAPTIVE ESTIMATION; ADDITIVE MODEL; WAVELET REGULARIZATION

[1] Correspondig author.

1 Introduction

Multivariate curve estimation is often considered with some skepticism, because it is associated with the term of the "curse of dimensionality". This notion reflects the fact that nonparametric statistical methods lose much of their power if the dimension p of the design is large. In the presence of r bounded derivatives, the usual optimal rate of convergence in regression or density estimation is $n^{-2r/(2r+p)}$, where n denotes the number of observations. To get the same rate as in the one-dimensional case, one has to assume a smoothness of order rp rather than r. This phenomenon can also be explained by the sparsity of data in high dimensions. If we have a uniformly distributed sample over the hypercube $[-1;1]^p$, then we will find only a fraction of about 2^{-p} of the data in the hypercube $[0;1]^p$.

Often a remedy to problems with high dimensionality consists in imposing additional structural assumptions which restrict the complexity of the curve. A well-established case in this direction are additive models. In the statistical literature such models have been introduced in the early eighties and they have led to the development of a variety of theoretical results and practical implementations in modern software, see e.g. Venables and Ripley (1994). Buja, Hastie and Tibshirani (1989) and Hastie and Tibshirani (1990) give a good overview and analyze various estimation techniques of iterative nature. Additive regression models naturally generalize the linear regression models and allow interpretation of marginal changes i.e. the effect of one variable on the mean function. They are also interesting from a theoretical point of view since they combine flexible nonparametric modeling of many variables with statistical precision that is typical for just one explanatory variable. More precisely, it is known that one can estimate the mean curve with a rate of convergence corresponding to the one-dimensional case, see, e.g., Stone (1985, 1986). It is clear that such a strong structural assumption is almost always inadequate, and one actually estimates some kind of projection of the true function on the lower-dimensional functional class. Stone (1985) derived his results in this general setting of a possibly inadequate model.

In practice, the backfitting procedures proposed in Breiman and Friedman (1985) and Buja, Hastie and Tibshirani (1989) are widely used to estimate the additive components. While backfitting has proven very useful in application and simulation studies, it has been somewhat difficult to analyze theoretically, mainly because the technique is iterative and depends on starting values and a convergence criterion. This lack of understanding of the theoretical properties of backfitting has long been a drawback to its universal acceptance, until recently, when Opsomer (1995) and Opsomer and Ruppert (1996) provided conditional mean squared error expressions for backfitting albeit under rather strong conditions on the smoothing matrices and design. Another method for fitting additive regression models, called marginal integration, has been recently proposed by Linton and Nielsen (1995) (see also Tjøstheim and Auestad (1994) and Newey (1994)).

This method is easier to understand for non-statisticians since it only involves averaging rather than iterative solution of nonlinear equations. Its statistical properties are easy to obtain, and have been established in the aforementioned papers. Although tractable, marginal integration is not generally efficient. Linton (1996) showed how to improve on the efficiency of the marginal integration estimator in regression by carrying out one-step backfitting from this initial consistent starting point. However, the method is computationally intensive.

The present paper considers wavelet series estimators of additive regression models. First, a direct approach for adaptive simultaneous estimation for all nonparametric components is considered and the asymptotic rates of convergence for mean integrated squared error of the resulting estimator are investigated. Second, a fast adaptive approach relying on tensor wavelet series estimators and wavelet regularization is investigated and leads, as in Opsomer (1995), to precise conditional mean squared error expressions. Moreover, the 'hat matrix' of the proposed estimator is of less complicated form as in the direct approach. Both these methods complement existing methodology for modeling and fitting additive regression models.

The organization of the paper is as follows. Section 2 proposes an adaptive and simultaneous estimation procedure for all additive components in additive regression models and discusses rate of convergence results and data-dependent truncation rules for wavelet series estimators. Section 3 introduces a wavelet method that naturally allows a direct separation of the problem into independent 1d problems. It also introduces a regularization algorithm which guarantees an adaptive solution to the estimation problem. In this section the conditional mean squared error of the resulting estimator is also derived. A finite sample comparison between the two approaches using simulated and real data is provided in Section 4. The detailed conditions and proofs of the theorems are given in the Appendix.

2 Simultaneous wavelet series estimation of additive regression models.

Before introducing our estimator we need some preliminaries. A p-order additive nonparametric regression model is generally defined by

$$Y = m(\mathbf{X}) + U, \tag{1}$$

where $\mathbf{X} = (X^1, \ldots, X^p)^T$ is a vector of p, real valued, stochastic regressors $\{X^i, 1 \leq i \leq p\}$, Y is a real valued square integrable random variable, U is a zero-mean, square integrable random variable conditionally independent of $\mathbf{X}$ and where $m(\cdot)$ is known to be of the form

$$m(\mathbf{X}) = \sum_{i=1}^{p} m^i(X^i).$$

Here, $m^i(x)$ is an unknown univariate real-valued function that is usually assumed to be smooth.

Hereafter, we will observe an independent non identically distributed sample (Y_t, X_t), $t = 1, \ldots, n$ of size n, from model (1),

$$Y_t = \sum_{i=1}^{p} m^i(X_t^i) + U_t, \quad t = 1, \ldots, n, \tag{2}$$

and we shall assume that $\{m^i(\cdot); 1 \leq i \leq p\}$ is a collection of p unknown functions defined on $[0, 1]$, and that, conditionally on the X_t, the U_t are mean zero, bounded variance random variables, independent across t. All subsequent results hold conditionally on the X_t's, and therefore, if we assume as in Stone (1985), that the distribution of $\mathbf{X}$ is absolutely continuous with a density that is bounded away from 0 and ∞ on its support, there is no loss of generality in assuming that the regression design is deterministic.

The approach taken in this section is to approximate each $m^i(\cdot)$ by a wavelet series sum $\sum_{k=0}^{2^{J_i}-1} \theta_{J_i,k}\phi_{J_i,k}(\cdot)$ where $\{\phi_{J_i,k}, k = 0, \ldots, 2^{J_i}-1\}$ denotes the set of scaling functions forming a base of the approximating space V_{J_i} associated to a Cohen, Daubechies and Vial (1992) multiresolution analysis of $L^2([0, 1])$ and where $\theta_{J_i,k}$ are the corresponding scaling coefficients. The truncation parameter 2^{J_i} implicitly depends on the sample size n and a crucial problem is how to select J_i practically.

For the sake of simplicity, we present the main concepts in the bivariate regression case ($p = 2$). Let K_1 and K_2 denote respectively the integers 2^{J_1} and 2^{J_2}. Let $\boldsymbol{\theta}$ be the vector of dimension $K_1 + K_2$ defined by $\boldsymbol{\theta} = (\theta_{J_1,0}, \ldots, \theta_{J_1,K_1-1}, \theta_{J_2,0}, \ldots, \theta_{J_2,K_2-1})^T$ and let Z be the $n \times (K_1 + K_2)$ design matrix whose ith row is given by $(\phi_{J_1,0}(X_i^1), \ldots, \phi_{J_1,K_1-1}(X_i^1), \phi_{J_2,0}(X_i^2), \ldots, \phi_{J_2,K_2-1}(X_i^2))$. The least squares estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is given by

$$\hat{\boldsymbol{\theta}} = (Z^T Z)^- Z^T \mathbf{Y}, \tag{3}$$

provided that the right-hand side is well defined, where $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ and (-) denotes some generalized inverse. The corresponding series estimator $\hat{m}$ of m is

$$\hat{m}(\mathbf{X}) = \sum_{i=1}^{p} \sum_{k=0}^{2^{J_i}-1} \hat{\theta}_{J_i,k}\phi_{J_i,k}(X^i).$$

Various properties of $\hat{m}$ are now investigated. Following Andrews (1991) and for theoretical purposes only, let us consider the following "artificial" models:

$$Y_t^1 = m^1(X_t^1) + U_t^1, \quad t = 1, \ldots, n, \tag{4}$$
$$Y_t^2 = m^2(X_t^2) + U_t^2, \quad t = 1, \ldots, n, \tag{5}$$

with $Y_t = Y_t^1 + Y_t^2$, $m(\mathbf{X}_t) = m^1(X_t^1) + m^2(X_t^2)$ and $U_t = U_t^1 + U_t^2$. We shall also assume that U_t^1 and U_t^2 are independent and share the same properties as the variables

U_t, with variances (conditional variances when the design is random) σ_{1t}^2 and σ_{2t}^2 such that $\mathrm{Var}(U_t) = \sigma_t^2 = \sigma_{1t}^2 + \sigma_{2t}^2$. One may therefore consider that the artificial models given in expressions (4) and (5) generate the additive model (2) by summation.

As done previously with m, let $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ be the coefficients obtained by least squares for estimating the unknown functions m^1 and m^2 with "observations" Y_t^1 and Y_t^2 respectively. Before establishing the main results of this section, we first need to recall that the *mean average squared error* (MASE) of $\hat{m}$ is defined to be

$$\mathrm{MASE}(\hat{\mathbf{m}}, \mathbf{m}) = n^{-1} I\!\!E\left[\|\hat{\mathbf{m}} - \mathbf{m}\|_2^2\right] = n^{-1} \sum_{t=1}^{n} I\!\!E\left[(\hat{m}(\mathbf{X}_t) - m(\mathbf{X}_t))^2\right], \tag{6}$$

where $\|\cdot\|_2$ denotes the Euclidean norm of $I\!\!R^n$ and $I\!\!E$ is the expectation operator. The MASE of $\hat{m}^i$ for estimating m^i $(i = 1, 2)$ is defined analogously. With the above notation we now have:

Theorem 1
Assume that assumptions A.1–A.2 listed in the Appendix hold. Then,

$$MASE(\hat{\mathbf{m}}, \mathbf{m}) \leq 2\frac{\tau^*}{\tau_*}\left(MASE(\hat{\mathbf{m}}^1, \mathbf{m}^1) + MASE((\hat{\mathbf{m}}_2, \mathbf{m}_2)\right),$$

where $0 < \tau_* = \min_t\{\sigma_{1t}^2, \sigma_{2t}^2\}$ *and* $\max_t\{\sigma_{1t}^2, \sigma_{2t}^2\} = \tau^* < \infty$.

Theorem 1, which has clear implications for asymptotic rate of convergence results, shows that the MASE of $\hat{\mathbf{m}}$ in the full additive regression model (2) can be bounded above by a constant times the sum of the MASE of the estimates of each component. The latter MASE does not depend on the dimension p. In this sense, the wavelet series estimator circumvents the curse of dimensionality.

Next we present rate of convergence result for our wavelet series estimator. More precisely we have:

Theorem 2
Assume that assumptions A.1–A.3 listed in the Appendix hold. Then

(a) *If* $K_i/n \to 0$ *and* $K_i \to \infty$, $(i = 1, 2)$, *then* $MASE(\hat{\mathbf{m}}, \mathbf{m}) \to 0$.

(b) *If* $K_i \sim n^{r_i}$ *for some* $0 < r_i < 1$ $(i = 1, 2)$, *then* $MASE(\hat{\mathbf{m}}, \mathbf{m}) = \mathcal{O}(n^{-r})$ *where* $r = \min_{i=1,2}\{\min(1 - r_i, 2\beta_i r_i)\}$, β_i *denoting the Sobolev smoothness index of* m^i.

(c) *If* $r_i = 1/(2\beta_i + 1)$, $(i = 1, 2)$, *then* $r = \min_{i=1,2}\{2\beta_i/(2\beta_i + 1)\}$ *is the maximum convergence rate* r *of the MASE in part (b).*

Theorem 2 shows that the MASE rate of convergence for the wavelet series estimator is arbitrary close to the optimal rate given in Stone (1985) for additive regression models.

It remains now to derive a data-driven way to select optimum estimators $\hat{J}_i$ for the truncation parameters J_i. To do so we need to introduce the following definition.

Let $\mathcal{K}_n$ be the collection of integer vectors $(K_1(n), K_2(n))$ from which the data-driven method chooses when the sample size is n and assume that there is a sequence $K(n) = (K_1(n), K_2(n))$ in $\mathcal{K}_n$ for which $K_i(n) \sim n^{1/(2\beta_i+1)}$ (see Assumption A.5. in the Appendix).

Definition 1

A data-driven truncation estimator $\hat{K} = (\hat{K}_1, \hat{K}_2)$ is asymptotically optimal if

$$\frac{\|\hat{\mathbf{m}}_{\hat{K}} - \mathbf{m}\|_2^2}{\min_{K \in \mathcal{K}_n} \|\hat{\mathbf{m}}_K - \mathbf{m}\|_2^2} \to_p 1, \tag{7}$$

and

$$\frac{MASE(\hat{\mathbf{m}}_{\hat{K}}, \mathbf{m})}{\min_{K \in \mathcal{K}_n} MASE(\hat{\mathbf{m}}_K, \mathbf{m})} \to 1. \tag{8}$$

To select an optimal truncation parameter we use the GCV criterion introduced by Craven and Wahba (1979). The GCV criterion selects the $\hat{K}$ that achieves

$$\mathrm{GCV}(\hat{K}) = \min_{K \in \mathcal{K}_n} \mathrm{GCV}(K) = \min_{K \in \mathcal{K}_n} \frac{n^{-1}\|\mathbf{Y} - \hat{\mathbf{m}}_K\|_2^2}{\left(1 - \frac{K_1+K_2}{n}\right)^2}. \tag{9}$$

Note that the GCV criterion as defined in eq. (9) does not depend on the "pseudo-models" (4) and (5).

We have the following asymptotic property for the GCV criterion and the estimated truncation parameter.

Theorem 3

Assume that assumptions A.1–A.6 listed in the Appendix hold. Then GCV is asymptotically optimal in the sense of Definition 1 and $MASE(\hat{\mathbf{m}}_{\hat{K}}, \mathbf{m}) = \mathcal{O}(n^{-r})$ with $r = \min_{i=1,2} 2\beta_i/(2\beta_i + 1)$.

Now we can define the adaptive and simultaneous estimation procedure as follows:

i) Find the least squares estimator $\hat{\theta}(K)$ by (3);

ii) Use the associated series estimator $\hat{\mathbf{m}}_K$ to compute the GCV criterion given in (9).

iii) Solve the GCV-based minimization to obtain $\hat{K}$.

iv) Define the adaptive estimator by $\hat{m}_{\hat{K}}$.

While this method seems reasonable and quite attractive, it has the drawback that it presents a big computational complexity. Its implementation requires the evaluation of the design matrix Z, followed by the GCV-based minimization. Due to this limitation, it can only be used for moderate sample size designs. The next section gives another wavelet-based method for fitting additive models which takes advantage of the fast wavelet transform and can be computed with $O(n)$ operations.

3 Wavelet direct separation

Throughout this section we consider again a p-order additive nonparametric regression model, as defined in (1), with β_i being the Sobolev smoothness index of m^i. To simplify the presentation we will assume that the main effect functions m^i, $i = 1, \dots, p$ are periodic. We will consider first the case of a deterministic regular design with an equispaced sampling on a grid of size a power of two for each covariate. These assumptions of periodicity and regular design will be later on relaxed. For the sake of simplicity, the main features of the method will be shown in the bivariate regression case ($p = 2$).

3.1 Wavelet transform for linearly additive models

The representation of additive models in terms of tensor Hilbert spaces plays an important role in the analysis that follows and is discussed below for the case $p = 2$. Moreover, this formulation is an appealing one because the abstract setting looks exactly as a finite dimensional one way ANOVA problem. Hereafter, if H_1 and H_2 are two real separable Hilbert spaces, the space $H_1 \hat{\otimes}_2 H_2$ denotes the tensor Hilbert space obtained by completing the pre-hilbertian space $H_1 \otimes H_2$ for the scalar product

$$\langle h_1 \otimes h_2, k_1 \otimes k_2 \rangle_{H_1 \otimes H_2} = \langle h_1, k_1 \rangle_{H_1} \langle h_2, k_2 \rangle_{H_2}.$$

When $H_i = H^{\beta_i}([0,1])$ $(i = 1, 2)$ are the Sobolev function spaces on $[0,1]$ with smoothness index β^i, the associated Hilbert space $H_1 \hat{\otimes}_2 H_2$ will be denoted by W. We will denote by $\mathcal{K}^1$ and $\mathcal{K}^2$ the linear closed subspaces of H^{β_1} and H^{β_2} spanned respectively by the constant functions on $[0,1]$, $\mathbf{1}^1$ and $\mathbf{1}^2$ defined by $\mathbf{1}^i(t) = 1$ for all $t \in [0,1]$. It is then easy to see that $W^{\beta_2} = \mathcal{K}^1 \hat{\otimes}_2 H^{\beta_2}$ and $W^{\beta_1} = H^{\beta_2} \hat{\otimes}_2 \mathcal{K}^1$ are closed linear subspaces of W, isomorphic to H^{β_1} and H^{β_2} respectively.

According to the above, each component $m^i(X^i)$ of our additive model can be represented by the corresponding function $M^i(\mathbf{X})$ in W^{β_i} $(i = 1, 2)$, defined by

$$M^1(\mathbf{X}) = m^1(X^1) \otimes \mathbf{1}^2(X^2) \quad \text{and} \quad M^2(\mathbf{X}) = \mathbf{1}^1(X^1) \otimes m^2(X^2).$$

The function $M^i(\mathbf{X})$ represents, in a functional ANOVA model, the main effect of the variable X^i on the mean.

With these notations, the mean $m(\mathbf{X})$ of observations following model (1) can be explicitly written as

$$m(\mathbf{X}) = M^1(\mathbf{X}) + M^2(\mathbf{X}). \tag{10}$$

Now let $\{\phi_{0,0},\ \psi_{j,k},\ j \geq 0,\ 0 \leq k < 2^j\}$ be a system of periodic wavelets of sufficient regularity (see Antoniadis (1996) for the discussion concerning the required regularity). Let us consider the separable wavelet basis for $L^2(\Re^2) = L^2(\Re) \hat{\otimes}_2 L^2(\Re)$ given by the separable tensor product of the wavelet bases for both directions,

$$B^1 \otimes B^2 := \{(\phi^1_{0,0}, \psi^1_{j,k}) \otimes (\phi^2_{0,0}, \psi^2_{j,k}),\ j \geq 0,\ 0 \leq k < 2^j\}$$

(note that the wavelet systems are not necessarily the same on the two coordinates).

Then each $M^i(\mathbf{X})$ can be decomposed as

$$M^i(\mathbf{X}) = \sum_{j_1,k_1;j_2,k_2} \mu^i_{j_1,k_1;j_2,k_2} \psi^1_{j_1,k_1}(X^1)\psi^2_{j_2,k_2}(X^2)$$

where both (j_1, k_1) and (j_2, k_2) span the set $\{(-1, 0) \cup (j, k),\ j \geq 0,\ 0 \leq k < 2^j\}$ and we set $\psi^i_{-1,0} := \phi^i_{0,0}$. For ease of notation, in the following we shall generally use the single-index notation

$$M^i(\mathbf{X}) = \sum_{i_1,i_2 \geq 0} \mu^i_{i_1,i_2} \psi^1_{i_1}(X^1)\psi^2_{i_2}(X^2) \tag{11}$$

where each i (that is, both i_1 and i_2) is related to the scale j and traslation parameter k of the basis functions by the relation

$$\begin{cases} i = 0 & \text{for } j = -1 \quad \text{(scale function)} \\ i = 2^j + k & \text{for } j \geq 0 \quad \text{(wavelet function)} \end{cases}$$

For the additive model (10) it is easy to see, since ψ has zero moment, that

$$\begin{cases} \mu^1_{i_1,i_2} = 0, \quad i_2 > 0, \\ \mu^2_{i_1,i_2} = 0, \quad i_1 > 0. \end{cases} \tag{12}$$

In other words the 2d wavelet transform of an additive model is simply given by the coefficients $\mu^1_{i_1,0}$ and μ^2_{0,i_2}, $i_1, i_2 \geq 0$, the other ones being null.

Let us now suppose that a sample of m of size $n = n_1 n_2$ is known over a regular lattice in $[0, 1] \times [0, 1]$, with n_1, n_2 points on each coordinate X^1, X^2, respectively. We assume $n_1 = 2^{J_1}$, $n_2 = 2^{J_2}$. The sample is denoted as $m_{i_1,i_2} = m^1_{i_1,i_2} + m^2_{i_1,i_2}$, $0 \leq i_1 < n_1$, $0 \leq i_2 < n_2$. Then let us consider the 2d (standard form) discrete wavelet transform of the sample. It can be fastly computed provided that an approximation of $< m, \phi^1_{J_1,k_1} \otimes \phi^2_{J_2,k_2} >$, $0 \leq k_1 < 2^{J_1}$, $0 \leq k_2 < 2^{J_2}$, is given. We assume that

$$< m, \phi^1_{J_1,k_1} \otimes \phi^2_{J_2,k_2} > \approx \frac{m_{k_1,k_2}}{\sqrt{n_1 n_2}}. \tag{13}$$

This approximation is usually made in applications and is more accurate if Coiflets are chosen as a wavelet basis (Beylkin, Coifmann and Rokhlin, 1991). For a more elaborate approximation we refer to the work of Delyon and Iouditski (1995).

Let M_{i_1,i_2} and $M^i_{i_1,i_2}$ be the 2d discrete wavelet transform (standard form) of the sample $2^{-(J_1+J_2)/2} m_{i_1,i_2}$ and its components. It follows by (13), that these "empirical" coefficients define a good approximation of the corresponding wavelet coefficients in the expansions (11) of the functions $M^i(\mathbf{X})$.

For the discrete wavelet transform it holds also

$$\begin{cases} M^1_{i_1,i_2} = 0, \quad 0 < i_2 < n_2, \\ M^2_{i_1,i_2} = 0, \quad 0 < i_1 < n_1, \end{cases}$$

analogously with the continuous case.
Then only the first row and column of the matrix of the wavelet coefficients are non null, all other elements being zero.

Remark 1
The elements of the matrix corresponding to the non null coefficients of the two wavelet systems do not overlap (except for the upper left element, where both $M^1_{0,0}$ *and* $M^2_{0,0}$ *are present, and which is due to the lack of orthogonality of the functions* $M^i(X^i)$ *with the constant functions* $\mathbf{1}^i(X^i)$*). In practice (except for the upper left element) the first row of the matrix* M_{i_1,i_2} *is related only to the discrete wavelet transform of the component* m_2*, while the first column of the same matrix is related only to the discrete wavelet transform of the component* m_1*. This has important consequences in the following.*

3.2 Regularization

Let us now consider the additive model (1). In the one-dimensional case ($p = 1$), several wavelet-based methods have been proposed to estimate the regression function m (e.g., thresholding, projection, regularization) and, as showm below, each of them can also be applied to multidimensional case as well.

The method proposed in this section is based on regularization. Wavelet regularization was introduced by the authors in Antoniadis (1996) and Amato and Vuza (1997) for one-dimensional functions, but the theory is easily extended to the general p-dimensional case. We defer to those papers for a full description of the method. Here we apply the regularization method for models with an additive structure; we shall see how properties of the wavelet transform will make it immediate to estimate all components of the unknown function m; moreover, we shall easily derive convergence results for the resulting estimators.

The stochastic setting for model (1) is the same as the one used in Section 1. The design will be regular in the sense described in the previous subsection. The noisy observations will be denoted by Y_{i_1,i_2}, $0 \le i_1 < n_1$, $0 \le i_2 < n_2$, and the 2d discrete wavelet transform coefficients will be denoted by y_{i_1,i_2}, $0 \le i_1 < n_1$, $0 \le i_2 < n_2$.

In order to estimate the smooth function m by the fast wavelet transform and since we can only observe a noisy version of a discretized version of m on a grid of size n, we will try to estimate by regularization the finite dimensional approximation of m, denoted hereafter by $\mathbf{m}$, onto the approximation space V_n. By approximation (13), the observations are equivalent to the "raw" estimator of $\mathbf{m}$ given by

$$Y_n(\cdot,\cdot) = \frac{1}{\sqrt{n_1 n_2}} \sum_{k_1,k_2} Y_{k_1,k_2} \psi^1_{n1,k_1}(\cdot)\psi^2_{n2,k_2}(\cdot),$$

denoted also by $\mathbf{Y}$, and which lies in the space W for wavelets with sufficient regularity.

In a similar way we shall denote by $\mathbf{M}^i$ $(i = 1, 2)$ the finite dimensional projection of M^i onto the one-dimensional scaling space V_{n_i}. Moreover let $\overline{f} :=< f, \mathbf{1} >^2_L$. Then $\mathbf{m}$ can be written according to Eq. (10) as $\mathbf{m} = \mathbf{M}^1 + \mathbf{M}^2$. To estimate $\mathbf{m}$ by regularization one must solve the following problem:

$$\min_{\mathbf{G}=\mathbf{G}_1+\mathbf{G}_2} \quad \begin{array}{l} \|\mathbf{G} - \mathbf{Y}\|^2_{L^2[0,1]^2} + \lambda_0 \left((\overline{\mathbf{G}}^1)^2 + (\overline{\mathbf{G}}^2)^2 \right) + \\ \lambda_1 \|\mathbf{G}^1 - \overline{\mathbf{G}}^1\|^2_{W^{\beta_1}} + \lambda_2 \|\mathbf{G}^2 - \overline{\mathbf{G}}^2\|^2_{W^{\beta_2}}, \end{array} \tag{14}$$

where λ_0, λ_1 and λ_2 are regularization parameters to be chosen properly.

Let us expand $\mathbf{G}$ and $\mathbf{Y}$ in terms of separable wavelet bases (11); we indicate by $\theta^1_{i_1,i_2}$ and $\theta^2_{i_1,i_2}$ the wavelet coefficients of $\mathbf{G}^1$ and $\mathbf{G}^2$, respectively. Following Antoniadis (1996) or Amato & Vuza (1997) and using the equivalent, in term of wavelet coefficients, sequence Sobolev norms (Meyer, 1990), the continuous regularization problem (14) can be expressed in terms of wavelet coefficients as follows:

$$\min_{\substack{\theta^1_{i_1,i_2}, \theta^2_{i_1,i_2} \\ 0 \le i_1 < n_1 \\ 0 \le i_2 < n_2}} \quad \begin{array}{l} \displaystyle\sum_{i_1=0}^{n_1-1} \sum_{i_2=0}^{n_2-1} \left(w^1_{i_1,i_2} \theta^1_{i_1,i_2} + w^2_{i_1,i_2} \theta^2_{i_1,i_2} - y_{i_1,i_2} \right)^2 + \\ \displaystyle\lambda_0 \theta_{0,0} + \lambda_1 \sum_{i_1=1}^{n_1-1} \sum_{i_2=1}^{n_2-1} 2^{2(s(i_1)\beta_1 + s(i_2)\beta_2)} (\theta^1_{i_1,i_2})^2 + \\ \displaystyle\lambda_2 \sum_{i_1=1}^{n_1-1} \sum_{i_2=1}^{n_2-1} 2^{2(s(i_1)\beta_1 + s(i_2)\beta_2)} (\theta^2_{i_1,i_2})^2, \end{array} \tag{15}$$

where $s(i)$ is the scale corresponding to element i, $s(i) = [\log_2 i]$, $i > 0$ ($[\cdot]$ denotes integral part), $\theta_{0,0} := \theta^1_{0,0} + \theta^2_{0,0}$ and from Eq. (12) we can set

$$\begin{cases} w^1_{i_1,i_2} &= \delta_{i_2,0} \\ w^2_{i_1,i_2} &= \delta_{i_1,0}. \end{cases} \tag{16}$$

It is easy to prove that the solution of the regularization problem (15) is explicitly given by

$$\begin{cases} \hat{\theta}_{0,0} &= \dfrac{y_{0,0}}{1 + \lambda_0}, \\ \hat{\theta}^1_{i_1,0} &= \dfrac{y_{i_1,0}}{1 + 2^{2s(i_1)\beta_1} \lambda_1}, \quad 0 < i_1 < 2^{J_1}, \\ \hat{\theta}^2_{0,i_2} &= \dfrac{y_{0,i_2}}{1 + 2^{2s(i_2)\beta_2} \lambda_2}, \quad 0 < i_2 < 2^{J_2}, \end{cases} \tag{17}$$

all other coefficients being estimated by 0.

Apart from the simplicity of the explicit solution of the regularization problem (15) two important remarks are due at this point.

Remark 2
The solution (17) of the regularization problem (14) is obtained by processing separately each coordinate and, in particular, an optimal regularization parameter can be derived for each direction.

Remark 2 is a consequence of the particular structure of the matrix of coefficients for the 2d wavelet transform (standard form) of univariate functions, see Remark 1.

Remark 3
The solution (17) of the regularization problem (15) has exactly the same expression as the 1d case for each coordinate.

Remark 3 has an immediate consequence in that the curse of dimensionality is trivially circumvented and all theorems developed for the 1d case easily follow for the 2d case.

In particular, convergence of the regularization method can be proved by the following (here stated for the general p-dimensional case) Theorem. For simplicity's sake we drop dependence of the regularized solution on the regularization parameter from the notation when there is no ambiguity.

Theorem 4
Let $m^i(X^i) : [0,1] \to H^{\beta_i}$, $\beta_i > 1/4$ and $\hat{\mathbf{M}}^i(X^i)$ be the regularized solution of the regularization problem (15) along the ith coordinate, $i \le p$ for regularization parameter λ^i. Then

$$MASE(\hat{\mathbf{M}}^i, \mathbf{M}^i) = O(n^{-2\beta_i/(2\beta_i+1)})$$

for $\lambda^i = O(n^{-2\beta_i/(2\beta_i+1)})$.

Note that convergence holds when $n \to \infty$, that is any of $n_i \to \infty$. In order that $\hat{\mathbf{M}}^i$ converges to m^i, n_i must tend to ∞.

To estimate the optimal value of the regularization parameter, two different methods have been considered in the literature of wavelet regularization. The first one is the GCV criterion developed by Craven and Wahba (1979) and extended to the wavelet case by Amato and Vuza (1997). There, the estimate λ^i_{GCV} $(i = 1, 2)$ is given by the minimum over λ of the functional $\mathrm{GCV}^i(\lambda)$,

$$\min_{\lambda \ge 0} \mathrm{GCV}^i(\lambda), \text{ where } \mathrm{GCV}^i(\lambda) = \frac{\displaystyle\sum_{k=1}^{n_i-1} \left[\left(1 - \frac{1}{1 + 2^{2s(k)\beta_i}\lambda} \right) y^i_k \right]^2}{\left[\displaystyle\frac{1}{n_i - 1} \sum_{k=1}^{n_i-1} \left(1 - \frac{1}{1 + 2^{2s(k)\beta_i}\lambda} \right) \right]^2}, \tag{18}$$

with y^i_k denoting the wavelet coefficients of the ith component.
For the GCV estimate the following theorem can be proved (in the general p-dimensional case). For a proof see Amato and Vuza (1997), Theorem 9. In Amato and Vuza (1997)

assumption A.7 of the appendix is made, but it is easy to show that the theorem holds also under the weaker assumption A.2 of the appendix.

Theorem 5
Let $\beta_i > 1/2$. Then the choice of λ^i_{GCV} provided by the GCV criterion is asymptotically optimal in the average, in the sense that if λ_{n_i} is any minimizer of $GCV^i(\lambda)$, then

$$\lim_{n_i \to \infty} \frac{MASE(\hat{\mathbf{M}}^i(X^i; \lambda_{n_i}), \mathbf{M}^i)}{\min\limits_{\lambda \geq 0} MASE(\hat{\mathbf{M}}^i(X^i; \lambda), \mathbf{M}^i)} = 1.$$

In Antoniadis (1996) another criterion has been considered, based on an unbiased estimate of $\Lambda_i(\lambda) := \|\hat{\mathbf{M}}^i - \mathbf{M}^i\|^2$. The optimal value λ^i_{ue} of the regularization parameter is obtained by minimizing the following functional

$$\begin{aligned} &\min_{\lambda \geq 0} \tilde{\Lambda}_i(\lambda), \text{ where} \\ &\tilde{\Lambda}_i(\lambda) = \frac{\sigma^2_{\mathrm{MS}}}{n_i}\left(\frac{1 - n_i^{1-2\beta_i}}{1 - 2^{1-2\beta_i}} + \lambda n_i\right) + \sum_{k=1}^{n_i - 1}\left[\left(\frac{\lambda}{\lambda + 2^{-2s(k)\beta_i}}\right)^2 y^i_k\right], \end{aligned} \tag{19}$$

where σ^2_{MS} is the Muller and Stadtmuller (1987) estimate of the noise variance; $\tilde{\Lambda}_i(\lambda)$ is an unbiased estimate of $\Lambda_i(\lambda)$.

From the proof of Theorem 3.3 in Antoniadis (1996), one can easily show that the following theorem holds:

Theorem 6
Let $m^i : [0,1] \to H^{\beta_i}$, $\beta_i > 1/2$. Then

$$E\left(\tilde{\Lambda}_i(\lambda)\right) = \Lambda_i(\lambda) + O\left(\frac{\log_2 n_i}{n_i}\right).$$

Setting $\lambda_0 = 0$ is natural since both criteria can not be used to estimate λ_0 from the datum $y_{0,0}$.

Summarizing, the algorithm for solving the 2d additive model in the case of regular design (equispaced data) is the following

Algorithm 1

Step 1. Calculate the 2d finite wavelet transform (standard form) of data $2^{-(J_1+J_2)/2} Y_{i_1,i_2}$, $0 \leq i_1 < 2^{J_1}$, $0 \leq i_2 < 2^{J_2}$ and let y_{i_1,i_2}, $0 \leq i_1 < 2^{J_1}$, $0 \leq i_2 < 2^{J_2}$, be the matrix of wavelet coefficients.

Step 2. Estimate the regularization parameters λ_1 and λ_2 using one of the two criteria (18) or (19) applied to the first row and the first column, y_{0,i_2}, $0 < i_2 < 2^{J_2}$ and $y_{i_1,0}$, $0 < i_1 < 2^{J_1}$, respectively, of the wavelet matrix.

Step 3. Compute the estimates of the wavelet transform coefficients of the components m^1 and m^2 of m ($\hat{\theta}^1_{i_1,0}$, $0 < i_1 < 2^{J_1}$, and $\hat{\theta}^2_{0,i_2}$, $0 < i_2 < 2^{J_2}$) by Eq. (17) correponding to the regularization parameters chosen in Step 2.

Step 4. Calculate the inverse discrete wavelet transform of $2^{J_1/2}\hat{\theta}^1_{i_1,0}$, $0 < i_1 < 2^{J_1}$ and of $2^{J_2/2}\hat{\theta}^2_{0,i_2}$, $0 < i_2 < 2^{J_2}$; they are the nonparametric regression estimates of the functions $m^1(X^1) - \overline{m}^1(X^1)$ and $m^2(X^2) - \overline{m}^2(X^2)$, respectively; together with the estimate $\overline{m}^1(X^1) + \overline{m}^2(X^2) = 2^{(J_1+J_2)/2}y_{0,0}$, they give the final estimate of $m(\mathbf{X})$.

Note that in the limit of approximation (13), due to the linearity of averaging and of the wavelet transform, the whole procedure is equivalent to make the 1d wavelet transform for all rows (or for all columns), to average them and then to smooth the wavelet coefficients so obtained by regularization. This gives an intuitive insight about why the method results so easy. In addition, it also follows that the method is equivalent to average directly data Y_{i_1,i_2} over columns (or over rows), and then to perform an ordinary wavelet regularization on the 1d averaged row (column, respectively). This equivalence gives a physical interpretation of the method and allows one to extend it to the case of irregular or random design in a natural way. Moreover, when only an estimate of the function is needed, then the procedure is exceptionally fast, since only averages and 1d wavelet transforms are required. The overall computational cost of the method proposed is absolutely noteworthy. The most time consuming step is the 2d discrete wavelet transform (in the case the full 2d wavelet matrix is required), to be performed only once, that is in any case very fast ($O(n_1 n_2)$ operations).

Finally we remark that the method proposed here has a strong ressemblance with marginal integration; however in our case the averaging procedure is directly performed by the fast wavelet transform, so that the method has an overall computational efficiency much better than marginal integration.

3.3 Random design

Assume that we have n independent observations $(Y_1, \underline{X}_1), \ldots, (Y_n, \underline{X}_n)$, where $\underline{X}_t$ is a p-dimensional density. For

$$f(\underline{x}) = E(Y_t/\underline{X}_t = \underline{x})$$

we obtain the usual nonparametric regression model

$$Y_t = f(\underline{X}_t) + U_t, \qquad t = 1, \ldots, n,$$

where $E(U_t/\underline{X}_t = \underline{x}_t) = 0$.

In contrast with the deterministic regular design case we have to take care of the bias when constructing our empirical wavelet coefficients.

A first possibility we envisage here is to compute first a bivariate local polynomial estimate of f with some small bandwidth h_n and then to interpolate this estimator on a regular grid for obtaining the values needed in (13). To be more specific we assume:

i) The marginal density f of the $\underline{X}_t$ is bounded away from 0 on $[0, 1]^2$.

ii) f is $C^r([0,1]^2)$, with $r > 1$.

iii) For all $M < +\infty$, there exist constants $c_M < +\infty$ such that

$$\sup_{\underline{x} \in [0,1]^p} E(|U_t|^M / \underline{X}_t = \underline{x}) \leq c_M.$$

Let $\tilde{f}(x)$ be a multivariate local polynomial estimator of f of order r with some bandwidth h_n (see Fan and Gijbels, 1995). Conditioned on $\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_n$, $\tilde{f}(\underline{x})$ can be written as

$$\tilde{f}(\underline{x}) = \sum_{t=1}^{n} \omega_t(\underline{x}) Y_t.$$

Assume that h_n is such that, for any $\delta > 0$, $h^{2r} = O(n^{-1-\delta})$ and $h^{-p} = O(n^{1-\delta})$ ($p = 2$ here). It may be shown that for C large enough, it holds that, for arbitrary $\alpha > 0$

$$P\left(\sup_{\underline{x} \in [0,1]^2} \left\{ \left| \sum_{t=1}^{n} \omega_t(\underline{x}) f(\underline{X}_t) - f(\underline{x}_t) \right| \right\} \geq C h_n^r \right) = O(n^{-\alpha}),$$

that is, the maximal bias of $\tilde{f}(\underline{x})$ over $[0,1]^2$ is of order h_n^r with a probability exceeding $1 - O(n^{-\alpha})$.
If we set

$$\tilde{\theta}_{k_1,k_2,J_1,J_2} =< \phi^1_{J_1,k_1} \otimes \phi^2_{J_2,k_2}, \tilde{f} >,$$

then

$$E(\tilde{\theta}_{k_1,k_2,J_1,J_2} / \underline{X}_1, \ldots, \underline{X}_n) - \theta_{k_1,k_2,J_1,J_2} = O(h_n^r)$$

is also satisfied with the above probability.

Of course the $\tilde{\theta}_{k_1,k_2,J_1,J_2}$ are no longer independent, but their conditional variances $\sigma^2_{k_1,k_2,J_1,J_2}$ remain bounded and the one-dimensional results of Antoniadis and Pham (1998) with the BINWAV estimator carry over. Note however that $\mathrm{var}\tilde{f}(\underline{x}) \leq O(n^{-1} h_n^{-1})$ and this will lead to slower convergence rates for our wavelet estimator.

Another possibility could be to directly extend BINWAV to the p-dimensional case and directly reduce original data to a uniform grid, that can be handled by the tools described in the previous sections for the regular design. Indeed it is well known that wavelet analysis is not suitable as it is when data are not equispaced, since the wavelet transform cannot be applied and ordinary, slow methods for estimating the coefficients of the transform have to be used. BINWAV (Antoniadis and Pham, 1998) is able to face random design of data and even number of data different from a power of 2 and naturally extends to the case of general additive models, thanks to the particular structure of the wavelet transform and to the comment made at the end of Section 3.2. We shall give a sketch of the method, omitting the technicalities that coincide or can be trivially desumed from Antoniadis and Pham (1998).

According to Antoniadis and Pham (1998), a preliminary binning of the 2d data is performed with B_i bins over each coordinate i (we suppose for simplicity B_i a power of 2, but this assumption can be dropped; different numbers of bins are allowed for each coordinate). In Antoniadis and Pham (1998) it is shown for the 1d case that there is no need to chose a large value of B_i, and usually $B_i \ll n_i$; the same arguments also apply to the general additive models considere here. Then a scaling function is chosen; such a choice can be quite arbitrary, but in the case of particular functions, e.g., first or second order cardinal B-splines, better approximation properties are known (Daubechies, 1992).

The scaling coefficients at the highest resolution

$$C_{J_1,k_1;J_2,k_2} = \iint m(X^1, X^2)\varphi^1(2^{J_1}X^1 - k_1)\varphi^2(2^{J_2}X^2 - k_2)dX^1dX^2$$

are approximated as

$$C_{J_1,k_1;J_2,k_2} \approx \frac{1}{\sqrt{M_1 M_2}} Y^B_{i_1,i_2} \tag{20}$$

where apex B stands for binned values and $Y^B_{i_1,i_2}$ are the binned values at the coordinates $X^{1,B}_{i_1}$, $X^{2,B}_{i_2}$ (bin centers).

Starting from Eq. (20), all other coefficients of the 2d wavelet transform can be estimated following the algorithm described in Antoniadis and Pham (1998) for the 1d case and easily generalized to two dimensions. As in the fixed design case, only the first row and column of the wavelet coefficient matrix are retained: the common element is an estimate of the average value of Y; the row and column can be handled according to one of the methods known in the literature for reducing noise (e.g., thresholding, projection, regularization). When inverse transformed, they give the final estimate of the components of the general additive model.

3.4 Not necessarily periodic functions

When the functions modelling the main effects are not periodic, the finite interval wavelet transform of Cohen, Daubechies and Vial (1992) can be used. A minor complication arises then in the additive structure since that the wavelet transform cannot be resolved up to the coarsest scale, for reasons connected to the definition of orthogonal wavelets on the boundaries.

Let L_i be the minimum level of the transform for each coordinate i and let $\ell_i + 1$ be the number of corresponding coefficients of the scaling functions, $\ell_i = 2^{L_i} - 1$. Now, in the matrices of wavelet coefficients of M^1 and M^2 the non-null overlapping elements belong to the upper left submatrix of size $(\ell_1 + 1, \ell_2 + 1)$.

As far as the regularization problem (14) is concerned, the right setting of the coefficients $w^1_{i_1,i_2}$ and $w^2_{i_1,i_2}$ is

$$
\begin{aligned}
w^1_{i_1,i_2} &= \begin{cases} 0, & i_2 > \ell_2 \\ 1, & i_2 \le \ell_2 \end{cases} \\
w^2_{i_1,i_2} &= \begin{cases} 0, & i_1 > \ell_1 \\ 1, & i_1 \le \ell_1. \end{cases}
\end{aligned}
$$

It is possible to prove in this case that the solution of the regularization problem (14) is

$$
\begin{cases}
\hat\theta_{0,0} = \dfrac{y_{0,0}}{1+\lambda_0} & \\
\hat\theta^1_{i_1,i_2} = \dfrac{y_{i_1,i_2}}{1+2^{2(s(\ell_1)\beta_1+s(i_2)\beta_2)}\lambda_2}, & i_1 \le \ell_1,\ i_2 > \ell_2 \\
\hat\theta^2_{i_1,i_2} = \dfrac{y_{i_1,i_2}}{1+2^{2(s(i_1)\beta_1+s(\ell_2)\beta_2)}\lambda_1}, & i_1 > \ell_1,\ i_2 \le \ell_2 \\
\begin{cases}
\hat\theta^1_{i_1,i_2} = \dfrac{y_{i_1,i_2}}{1+2^{2(s(\ell_1)\beta_1+s(\ell_2)\beta_2)}\lambda_1+\lambda_1/\lambda_2}, \\
\hat\theta^2_{i_1,i_2} = \dfrac{y_{i_1,i_2}}{1+2^{2(s(\ell_1)\beta_1+s(\ell_2)\beta_2)}\lambda_2+\lambda_2/\lambda_1},
\end{cases} & i_1 \le \ell_1,\ i_2 \le \ell_2,\ i_1+i_2 > 0,
\end{cases}
$$

all other $\hat\theta$ being null.

This solution generalizes to the p-dimensional case for each covariate d as

$$
\begin{cases}
\hat\theta_0 = \dfrac{y_{0,\dots,0}}{1+\lambda_0} & \\
\hat\theta^d_{i_1,\dots,i_p} = \dfrac{y_{i_1,\dots,i_p}}{1+2^{2(s(\ell_d)\beta_d+\sum_{k\neq d}s(i_k)\beta_k)}\lambda_d}, & i_d \le \ell_d,\ i_k > \ell_k,\ 1 \le k \le p,\ k \neq d, \\
\hat\theta^d_{i_1,\dots,i_p} = \dfrac{y_{i_1,\dots,i_p}}{1+2^{2\sum_k s(\ell_k)\beta_k}\lambda_d+\sum_{k\neq d}\lambda_d/\lambda_k}, & i_k \le \ell_k,\ 1 \le k \le p,\ \sum_k i_k > 0.
\end{cases}
\tag{21}
$$

Note that this solution differs from the full wavelet transform case only by a few coefficients in general; however some dependence among the one-dimensional problems arises from the presence of mixed terms of the form $\sum_{k\neq d}\lambda_d/\lambda_k$. To overcome this problem without degrading computational efficiency, we may consider two different ways: first, from a practical point of view, we may neglect the mixed terms, as far as the choice of the regularization parameter is concerned, i.e., the regularization parameter is still chosen independently for each direction. Once the choice of the regularization parameter is done, full Eq. (21) (i.e., including mixed terms) can then be considered for the evaluation of $\hat\theta$. Secondly, the elements of the overlapping submatrix can be excluded from the beginning from the minimization so that the corresponding $\hat\theta$ equal the input y. This is accomplished by setting:

$$
w^d_{i_1,\dots,i_p} = \begin{cases} 1, & i_d > \ell_d,\ i_k \le \ell_k,\ 1 \le k \le p,\ k \neq d \\ 0, & \text{otherwise} \end{cases}
$$

3.5 Testing additivity

The additive structure of the regression function is important in terms of its interpretability and its ability to deliver fast rates of convergence with minimal algorithmic complexity. Unfortunately, the additive model is not robust. A natural question is wether or not the additive structure provides a good approximation of the true regression function. It is therefore important to be able to test such an hypothesis statistically.

The wavelet analysis of Section 3.2 shows that, in the case that an additive structure holds, only the first row and column of the wavelet coefficient matrix contribute to the method, the other elements representing only noise. In other words information concerning the two univariate functions is compacted in the first row and column, even not overlapped, giving rise to the extraordinary simplicity of the method. This represents a small fraction $(n_1^{-1} + n_2^{-1} - (n_1 n_2)^{-1})$ of the whole matrix. Let S be the set of elements belonging to this part of the matrix. When additivity holds, the remaining (major) part of the wavelet coefficient matrix (let us call T the set of elements belonging to it) represents noise and contains no information on $m^1(X^1)$ and $m^2(X^2)$.

By orthogonality of the wavelet transform and since the elements of T are not affected by the regularization procedure, the components of T will be independent zero-mean normally distributed random variables, even in the presence of binning in the nonregular design case. If binning is performed, the variances of the components of T are not anymore equal, but conditionally on the X's, they can be linearly transformed to be homoscedastic. Let U be the r-dimensional Gaussian vector obtained after transforming T, with r being the number of elements of T. Under the additivity assumption, it is easy to see that U is distributed as a r-dimensional multivariate normal $N_r(\mathbf{a}, \sigma^2 \mathbf{I}_r)$ with $\mathbf{a} = 0$. To test whether or not the model is additive one can then use the data-driven modifications of the Neyman smooth test as derived in Fan (1996) to test the hypothesis that $\mathbf{a} = 0$, whose test statistic is given by

$$R_{\mathrm{Fan}} = \max_{1 \leq k \leq r} \left\{ \frac{1}{\sqrt{2k}} \sum_{j=1}^{k} \left(\frac{2n}{\sigma^2} U_j^2 - 1 \right) \right\}.$$

Fan's test, as described above, assumes that the scale of the errors σ is known and fixed. When σ^2 is unknown in the above expression, we may replace it by a consistent estimator. One way to achieve this, especially in the absence of high frequency alternatives to the null hypothesis, is to take the median absolute deviation of the empirical wavelet coeffcients in T associated to the finest scales and divide by 0.6745, as it is usually done in the wavelet regression context. The comparison of this approach with other possible approaches for testing additivity in multivariate functional ANOVA models is out of the scope of the present paper but presents interesting problems for future statistical work.

4 Numerical experiments

In the present section we show results of numerical experiments worked out on the method of wavelet direct separation proposed in the paper; the results are compared with those of other methods available in the literature. The experiments were conducted on some test examples considered in the literature. To further illustrate our method we also consider an application on a real data set.

The algorithm of wavelet direct separation was implemented in MATLAB on a PC. For the simulation study we have used several bivariate models (either purely additive or with an interaction term) and several types of noise. For each setting of the experiments we replicated 200 times the simulation of the response variable by drawing for each replication a random sample for the error term ε according to the noise model chosen. To fairly compare the smoothers, they were computed on the same set of simulated data. To describe the behaviour of the additive smoothers several criteria were calculated. As an overall measure to the true pattern defined by the conditional mean of Y we have used the mean square error of the estimates. If $\hat{y}_t^{(k)}$ is the smoothed value at the design point (X_t^1, X_t^2) for replication k of the simulation ($k = 1, \ldots, 200$), the m.s.e. for replication k is

$$s_k = \frac{1}{n}\sum_{t=1}^{n}\left(\hat{y}_t^{(k)} - m^1(X_t^1) - m^2(X_t^2)\right)^2 .$$

Notice however that this criterion does not indicate the quality of the fit for a single predictor. To measure the quality of the fit of each predictor (when possible) separately we have used the MSE and MASE criteria for each component separately, that is

$$s_k^j = \frac{1}{n}\sum_{t=1}^{n}\left(\hat{m}^j(X_t^j) - m^j(X_t^j)\right)^2, \qquad j = 1, 2,$$

as well as their averages on replications.
Recalling that the $\hat{m}^j$ and m^j are only unique up to some additive constant and to make sure that the comparisons between methods only reflect substantial differences in shape between the true and the estimated function we have used

$$s_k^j = \frac{1}{n}\sum_{t=1}^{n}\left(\hat{m}^j(X_t^j) - m^j(X_t^j) - \frac{1}{n}\sum_{\tau=1}^{n}\left(\hat{m}^j(X_\tau^j) - m^j(X_\tau^j)\right)\right)^2 \tag{22}$$

as a quality of fit criterion.

These 200 values of s_k^j for each smoother are displayed graphically by means of boxplots and summarized by their average with a standard error (MASE) in suitable tables.

Moreover, in order to show the behaviour of the smoothers over the functions, we have computed bias and mean square error of each predictor as a function of covariate for each component. Their average over the samples is shown in suitable error plots.

Finally, in order to rank our smoothers we derived an efficiency measure. Of course the notion of efficiency in nonparametric models is not as clear and well understood as it is in parametric models. In particular the MASE comparisons do not usually provide a simple ranking between estimators. Our approach is to measure the efficiency of the procedures against a similar procedure estimating $m^1(X^1)$ based on a knowledge of $m^2(\cdot)$ (and conversly). Linton (1996) already defined such an efficiency measure. In our case the efficiency is to compare the ratio of the s_k^j by the MSE of the estimated m^j when one smooths the partial errors $Y_t - m^2(X_t^2)$ on the direction of interest X^1.

The following test examples have been considered:

Example 1 *(Sperlich, Tjøstheim and Yang, 1998)*

$$m = m^1 + m^2 + \varepsilon$$

$$\begin{aligned} m^1 &= 1.5\sin(-1.5X^1), \qquad -2 \leq X^1 \leq 2 \\ m^2 &= -(X^2)^2 + E(X^2)^2, \qquad -2 \leq X^2 \leq 2. \end{aligned}$$

X^1 and X^2 come from a regular design or from a uniform distribution on $[-2,2]$. Noise ε is independent Gaussian with zero mean and standard deviation 0.5.

Example 2

$$m = m^1 + m^2 + X^1X^2 + \varepsilon$$

where X^1, X^2, m^1, m^2 and ε are the same as Example 1.

Example 3 *(Sperlich, Linton, and Härdle, 1997)*

$$m = m^1 + m^2 + \varepsilon$$

where m^1 and m^2 are two different functions from the set

$$\begin{aligned} f_1(X) &= 2X \\ f_2(X) &= X^2 - EX^2 \\ f_3(X) &= \exp(X) - E\exp(X) \\ f_4(X) &= 0.5\sin(-1.5X). \end{aligned}$$

X can come from a regular design or from a uniform random distribution, both on $[-3,3]$. Then the test example is individualized by the code 4pq, where p and q ($1 \leq p,q \leq 4$) are the functions chosen among the four listed. Noise ε is independent Gaussian with zero mean and standard deviation 0.5.

Example 4 *Same as Example 3, but for X^1 and X^2 having uncorrelated Gaussian distributions with mean 0 and variance 1.*

In order to compare the wavelet direct separation method, we considered `funfits` code (Nychka et al., 1993), based on spline smoothing, and `addfit` code (Opsomer and Ruppert, 1997), based on local polynomial regression, both adapted to additive models.

Experiments were worked out under different settings, that is sample sizes ($n = 64$, 256 and 1024), and designs (fixed and random).

Tables 1 and 2 summarize the results in all experiments in terms of average and standard error of (22) for `funfits` and wavelet direct separation methods. Ratio with respect to the values obtained in the unidimensional fit as explained above is also shown. In addition, for a fair comparison with respect to both components and across functions, the MASE is normalized with respect to the power of the true function, so that values shown in the tables are to be read as the square root of the actual relative MASE.

***** TABLE 1 ABOUT HERE *****

***** TABLE 2 ABOUT HERE *****

Note that the results for the `addfit` code are not shown because convergence was obtained only for a small number of experiments.

A visual representation of these results is shown in the boxplots of Fig. 1 for fixed designs and Figs. 2 for random designs, both for $n = 256$. Figs. 3 and 4 show plots of bias and mean square error as a function of covariate for $n = 256$ and for fixed and random design, respectively, in the case of `funfits` predictor. Figures 5 and 6 show the same results for the wavelet direct separation predictor.

***** FIGURE 1 ABOUT HERE *****

***** FIGURE 2 ABOUT HERE *****

***** FIGURE 3 ABOUT HERE *****

***** FIGURE 4 ABOUT HERE *****

***** FIGURE 5 ABOUT HERE *****

***** FIGURE 6 ABOUT HERE *****

Analysis of tables and figures, as well as results on further experiments available from the authors, show that the `funfits` predictor generally performs better than its competitors; in particular `addfit` very often fails to find a solution even for small number of data. Also note that in the case of fixed design wavelet direct separation method estimates functions with less bias than `funfits`, even though variance is larger. Performance of wavelet direct separation can be quite poor in the case of random design when the distribution of data is not uniform over the covariates, as expected from the method.

Finally Tab. 3 shows CPU execution times for the three predictors under different settings of the experiments (MATLAB code was run on a 100MHz Pentium PC). It is clear that wavelet direct separation outperforms its competitors.

***** TABLE 3 ABOUT HERE *****

To end this section and to give a comparison of the wavelet smoother to the other methods Figure 7 reports the fitting results on a real data example. The `minitri` data (available in the `funfits` package) reports the results from a mini triathlon sponsored by Bud Lite, held in Cary, NC, June 1990. Times are in minutes for the male 30-34 group. Components are: swim: (1/2 mile) bike: (15 miles) run: (4 miles). We have applied a bi-variate additive model to this dataset in order to explain the run record times as a function of the swim and bike record times. The important practical message is that the wavelet method behaves quite well even in the case of random and sparse designs.

***** FIGURE 7 ABOUT HERE *****

Acknowledgments

A. Antoniadis would like to thank the Istituto per Applicazioni della Matematica CNR for support and excellent hospitality whilst visiting Naples to carry out this work. Authors wish to thank Professor Opsomer and Professor Nychka for having provided `addfit` and `funfits` codes, respectively.

Appendix

We first state below the main assumptions used for asserting the results given in the paper.

Assumption A.1 The distribution of the random vector $\mathbf{X}$ is absolutely continuous with a density that is bounded away from 0 and ∞ on its support.

Assumption A.2 Conditionally on the X_t, the U_t are mean zero square integrable random variables with $0 < \inf_{t\geq 1} \mathrm{Var}(U_t) \leq \sup_{t\geq 1} \mathrm{Var}(U_t) < \infty$. Moreover the U_t are, conditionally on $\{X_t\}$, independent across t.

Assumption A.3 For each $i = 1, \ldots, p$, the additive components g_i are Hölder continuous on $[0,1]$ with Hölder exponent $1 \leq \beta_i$.

Assumption A.4 $\mathcal{K}_n = \{(K_1(n), K_2(n)),\ n \geq 1\}$ is a sequence of integer vectors such that, for all n there is a $(K_1(n), K_2(n))$ in $\mathcal{K}_n$ which satisfies $K_i(n) \sim n^{1/(2\beta_i+1)}$, where β_i are as in Assumption A.3.

Assumption A.5 There exists an absolute constant $D \geq 4$ such that

$$\sup_{t\geq 1} \sup_{\mathbf{x}} \mathbb{E}(|Y_t - \mathbb{E}(Y_t|X_t)|^{2D} | X_t = \mathbf{x}) < \infty.$$

Assumption A.6 For all $n \geq 1$, $\max_{K\in\mathcal{K}_n} (K_1 + K_2)/n \leq \gamma$ for some $\gamma < 1$.

Assumption A.7 The variance of U_t is constant in t.

Assumptions A.1 and A.2 are quite common in nonparametric regression. Assumption A.3 is some smoothness condition on the additive components. Assumptions A.4 and A.6 are not restrictive and are easy to verify (or to impose).

Proof of Theorem 1

Consider the following decomposition of MASE:

$$\begin{aligned}
\mathrm{MASE}(\hat{\mathbf{m}}, \mathbf{m}) &= n^{-1}\mathbb{E}\{\|\hat{\mathbf{m}} - \mathbf{m}\|_2^2\} = n^{-1}\mathbb{E}\{\|P_Z\mathbf{Y} - \mathbf{m}\|_2^2\} \\
&= n^{-1}\|P_Z\mathbf{m} - \mathbf{m}\|_2^2 + n^{-1}\mathbb{E}\{\|P_Z\mathbf{U}\|_2^2\},
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{MASE}(\hat{\mathbf{m}}^1, \mathbf{m}^1) + \mathrm{MASE}(\hat{\mathbf{m}}^2, \mathbf{m}^2) &= n^{-1}\|P_{Z_1}\mathbf{m}^1 - \mathbf{m}^1\|_2^2 + n^{-1}\mathbb{E}\{\|P_{Z_1}\mathbf{U}^1\|_2^2\} \\
&\quad + n^{-1}\|P_{Z_2}\mathbf{m}^2 - \mathbf{m}^2\|_2^2 + n^{-1}\mathbb{E}\{\|P_{Z_2}\mathbf{U}^2\|_2^2\},
\end{aligned}$$

where $P_A = A(A^TA)^{-}A^T$ denotes the orthogonal projector onto the space generated by the columns of A.

Using the triangle inequality and the fact that each matrix Z_i (i=1,2) is made by selecting the appropriate columns of Z, we have:

$$\|\mathbf{m} - P_Z\mathbf{m}\|_2 \leq \|\mathbf{m}^1 - P_{Z_1}\mathbf{m}^1\|_2 + \|\mathbf{m}^2 - P_{Z_2}\mathbf{m}^2\|_2.$$

Therefore,

$$\|\mathbf{m} - P_Z\mathbf{m}\|_2^2 \leq 2\left(\|\mathbf{m}^1 - P_{Z_1}\mathbf{m}^1\|_2^2 + \|\mathbf{m}^2 - P_{Z_2}\mathbf{m}^2\|_2^2\right), \tag{23}$$

which controls the bias term. As for the variance terms we have:

$$\begin{aligned} \mathbb{E}\left(\|P_Z\mathbf{U}\|_2^2\right) &= \operatorname{tr}\left(P_Z \operatorname{var}(\mathbf{U})\right) && (24)\\ &= \operatorname{tr}\left(P_Z\{\operatorname{var}(\mathbf{U^1}) + \operatorname{var}(\mathbf{U^2})\}\right) && (25)\\ &\leq 2\tau^*\operatorname{tr}(P_Z) = 2\tau^*(K_1+K_2). && (26) \end{aligned}$$

On the other hand,

$$\mathbb{E}\left(\|P_{Z_1}\mathbf{U}^1\|_2^2\right) + \mathbb{E}\left(\|P_{Z_2}\mathbf{U}^2\|_2^2\right) \geq \tau_*(K_1 + K2). \tag{27}$$

Combining expressions (23)-(27) proves the assertion of the Theorem. ∎

Proof of Theorem 2

Let $m_{K_i} = P_{V_{J_i}} m^i$ and let $m^r_{K_i} = m^i - P_{V_{J_i}} m^i$ denote the remainder from approximating m^i by its projection onto V_{J_i}. Let $m_K = m_{K_1} + m_{K_2}$ and $m^r_K = m^r_{K_1} + m^r_{K_2}$ be the analogous projection and remainder functions for m. Note that $P_Z\mathbf{m}_K = \mathbf{m}_K$ since the columns of Z generate the same linear subspace as the columns of Z_1 and Z_2. We now have

$$\begin{aligned} \frac{1}{n}\|(I-P_Z)\mathbf{m}\|_2^2 &= \frac{1}{n}\|(I-P_Z)\mathbf{m}^r_K\|_2^2 \leq \frac{1}{n}\|\mathbf{m}^r_K\|_2^2 \\ &\leq \frac{2}{n}\left(\|\mathbf{m}^r_{K_1}\|_2^2 + \|\mathbf{m}^r_{K_2}\|_2^2\right) \leq 2\left(\|m^r_{K_1}\|^2_{L^2([0,1])} + \|m^r_{K_2}\|^2_{L^2([0,1])}\right). \end{aligned}$$

We already have seen in the proof of Theorem 1 that

$$\frac{1}{n}\mathbb{E}\left(\|P_Z\mathbf{U}\|_2^2\right) \leq \frac{\tau^*(K_1+K_2)}{n}.$$

Combining the above inequalities we have

$$\operatorname{MASE}(\hat{m}, m) \leq \frac{\tau^*(K_1+K_2)}{n} + 2\left(\|m^r_{K_1}\|^2_{L^2([0,1])} + \|m^r_{K_2}\|^2_{L^2([0,1])}\right). \tag{28}$$

Assertion (a) follows from (28) and Assumption A.2. Assertion (b) follows from (24), Assumption A.2 and $K_i \sim n^{r_i}$, $(i = 1, 2)$, since the latter imply that for all $0 < \beta_i' < \beta_i$:

$$\begin{aligned} \operatorname{MASE}(\hat{m}, m) &\leq \mathcal{O}(n^{r_1-1}) + \mathcal{O}(n^{r_2-1}) \\ &\quad + \mathcal{O}(n^{-2\beta_1' r_1})K_1^{2\beta_1'}\|m^1 - P_{V_{J_1}}m^1\|^2_{L_2} + \mathcal{O}(n^{-2\beta_2' r_2})K_2^{2\beta_2'}\|m^2 - P_{V_{J_2}}m^2\|^2_{L_2} \\ &= \mathcal{O}(n^{-r}). \end{aligned}$$

Assertion (c) follows because r is minimized by taking r_i such that $1 - r_i = 2\beta_i r_i$. ∎

Proof of Theorem 3

The first assertion of the Theorem is true by Theorem 3.1* of Andrew (1990) since the given assumptions imply those of Theorem 3.1*. The second assertion holds by Theorems 1 and 2 of this paper. ∎

Proof of Theorem 4

For the proof see Antoniadis (1996), Theorem 3.1, and Amato and Vuza (1997), Theorem 3. Indeed in Amato and Vuza (1997) assumption A.7 is made, but it can be easily released so that theorem holds also under the weaker assumption A.2. ∎

References

Amato, U. and Vuza, D.T. (1997) Wavelet approximation of a function from samples affected by noise. Revue Roumaine de Mathématiques Pures et Appliquées, 42, 481–493.

Andrews, D.W.K. (1991) Asymptotic optimality of generalized C_L, cross-validation, and generalized cross-validation in regression with heteroskedastic errors. Journal of Econometrics, 47, 359–377.

Andrews, D.W.K. (1991) Asymptotic normality of series estimators for nonparametric and semiparametric regression models. Econometrica, 59, 307–345.

Antoniadis, A. (1996) Smoothing noisy data with tapered coiflet series. Scandinavian Journal of Statistics, 23, 313–330.

Antoniadis, A. and Pham, D.T. (1998) Wavelet regression for random or irregular design. Computational and Statistical Data Analysis, 28, 353–370.

Beylkin, G., Coifmann, R. and Rokhlin, V. (1991) The fast wavelet transform and numerical algorithms. Communications in Pure and Applied Mathematics, 44, 141–163.

Breiman, L. and Friedman, J.H. (1985) Estimating optimal transformations for multiple regression and correlation (with discussion). Journal of the American Statistical Association, 80, 580–619.

Buja, A., Hastie, T.J. and Tibshirani, R.J. (1989). Linear smoothers and additive models. Annals of Statistics, 17, 453–510.

Cohen, A., Daubechies, I. and Vial, P. (1993) Wavelets on the interval and fast wavelet transforms. Applied and Computational Harmonic Analysis, 1, 54–82.

Craven, P. and Wahba, G. (1979) *Smoothing Noisy Data with Spline Functions.* Numerische Mathematik **59**, 377–403.

Daubechies, I. (1992) Ten lectures on wavelets, SIAM, Philadelphia.

Delyon, B. and Ioudistski, A. (1995) Estimating wavelet coefficients. In Wavelets and statistics, Lecture Notes in Statistics, 103 (A. Antoniadis and G. Oppenheim, eds.), Springer-Verlag, 151–168.

Fan, J. (1996) Test of significance based on wavelet thresholding and Neyman's truncation. Journal of the American Statistical Association, 91, 674–688.

Fan, J. and Gijbels, I. (1995) Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. Journal of the Royal Statistical Society Series B, 57, 371–394.

Hastie, T.J. and Tibshirani, R.J. (1990) Generalized additive models, Chapman & Hall, London.

Linton, O. (1996) Efficient estimation of additive nonparametric regression models. Biometrika, 84, 469–473.

Linton, O. and Nielsen, J.P. (1995) A kernel method of estimating structured nonparametric regression based on marginal integration. Biometrika, 82, 93–100.

Meyer, Y. (1990) Ondelettes et opérateurs, Vol. I, Hermann, Paris.

Muller, H.G. and Stadtmuller, U. (1987) Estimation of heteroscedasticity in regression analysis. Annals of Statistics, 15, 610–625.

Newey, W.K. (1994) Kernel estimation of partial means. Econometric Theory, 10, 233–253.

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.

Opsomer, J.D. (1995) Optimal bandwidth selection for fitting an additive model by local polynomial regression. Ph. D. Dissertation, Cornell Univ.

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

Sperlich, S., Linton, O. B. and Härdle, W. (1997) A simulation comparison between integration and backfitting methods of estimating separable nonparametric regression models. Technical Report. Humboldt, Univ. Berlin.

Sperlich, S., Tjøstheim, D. and Yang, L. (1998) Nonparametric estimation and testing of interaction additive models. Technical Report. Humboldt, Yniv. Berlin.

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.

Tjøstheim, D. and Auestad, B. (1994) Nonparametric identification of nonlinear time series: projections. Journal of the American Statistical Association, 89, 1398–1409.

Venables, W.N. and Ripley, B. (1994) Modern applied statistics with S-Plus, Springer Verlag, New York.

Fixed design							
Test	$n = 64$		$n = 256$		$n = 1024$		
	Error	Ratio	Error	Ratio	Error	Ratio	
1	0.14±0.04	1.03±0.08	0.07±0.02	1.04±0.09	0.038±0.009	1.04±0.08	m^1
	0.11±0.04	1.02±0.09	0.06±0.02	1.0±0.1	0.035±0.009	1.05±0.08	m^2
2	0.4±0.1	1.5±0.3	0.14±0.03	1.20±0.05	0.07±0.01	1.18±0.03	m^1
	0.23±0.05	1.23±0.07	0.12±0.02	1.16±0.04	0.07±0.01	1.16±0.04	m^2
312	0.02±0.01	0.9±0.2	0.008±0.006	0.9±0.2	0.004±0.003	0.9±0.2	m^1
	0.06±0.02	1.0±0.1	0.033±0.009	1.05±0.09	0.018±0.004	1.04±0.07	m^2
334	0.04±0.01	1.04±0.06	0.023±0.005	1.04±0.06	0.013±0.002	1.06±0.06	m^1
	0.4±0.1	1.03±0.08	0.23±0.07	1.0±0.1	0.12±0.03	1.04±0.08	m^2
Random design							
1	0.14±0.05	1.1±0.2	0.07±0.02	1.1±0.1	0.04±0.01	1.04±0.08	m^1
	0.11±0.05	1.1±0.3	0.06±0.02	1.1±0.1	0.035±0.009	1.1±0.1	m^2
2	0.4±0.2	1.2±0.3	0.21±0.07	1.0±0.1	0.11±0.04	1.0±0.1	m^1
	0.3±0.2	1.1±0.3	0.17±0.07	1.0±0.2	0.09±0.03	1.0±0.1	m^2
312	0.02±0.01	1.0±0.9	0.009±0.006	1.0±0.4	0.005±0.003	0.9±0.2	m^1
	0.06±0.02	1.1±0.2	0.032±0.009	1.05±0.09	0.018±0.004	1.04±0.08	m^2
334	0.03±0.01	1.1±0.2	0.022±0.005	1.06±0.09	0.013±0.002	1.05±0.06	m^1
	0.5±0.2	1.2±0.3	0.24±0.07	1.1±0.1	0.12±0.03	1.1±0.1	m^2
412	0.05±0.03	1.1±0.7	0.020±0.002	1.0±0.5	0.012±0.009	1.0±0.4	m^1
	0.18±0.08	1.2±0.2	0.15±0.06	1.2±0.2	0.09±0.03	1.2±0.2	m^2
434	0.16±0.09	1.2±0.3	0.13±0.06	1.2±0.2	0.08±0.04	1.2±0.1	m^1
	0.6±0.3	1.3±0.5	0.4±0.2	1.1±0.2	0.24±0.07	1.0±0.1	m^2

Table 1: Summary of average and standard error of (22) for all experiments with the `funfits` predictor.

Fixed design							
Test	$n = 64$		$n = 256$		$n = 1024$		
	Error	Ratio	Error	Ratio	Error	Ratio	
1	0.15±0.04	1	0.11±0.02	1	0.08±0.01	1	m^1
	0.15±0.04	1	0.11±0.02	1	0.07±0.01	1	m^2
2	0.15±0.04	1	0.11±0.02	1	0.08±0.01	1	m^1
	0.15±0.04	1	0.11±0.02	1	0.07±0.01	1	m^2
312	0.05±0.01	1	0.036±0.006	1	0.025±0.003	1	m^1
	0.07±0.02	1	0.047±0.008	1	0.033±0.004	1	m^2
334	0.04±0.01	1	0.027±0.004	1	0.019±0.002	1	m^1
	0.5±0.1	1	0.36±0.06	1	0.25±0.04	1	m^2
Random design							
1	0.3±0.1	2±1	0.21±0.04	2.5±0.7	0.15±0.02	2.5±0.4	m^1
	0.28±0.09	1.7±0.7	0.19±0.04	2.1±0.5	0.13±0.02	2.1±0.4	m^2
2	0.4±0.2	1.4±0.7	0.31±0.07	1.3±0.3	0.22±0.04	1.3±0.2	m^1
	0.4±0.2	1.2±0.3	0.28±0.09	1.3±0.3	0.19±0.04	1.2±0.2	m^2
312	0.20±0.07	3±2	0.14±0.03	4±1	0.10±0.02	5±1	m^1
	0.3±0.1	2±1	0.24±0.05	4±1	0.17±0.02	5±1	m^2
334	0.13±0.05	1.01±0.07	0.06±0.01	2.4±0.6	0.034±0.006	1.05±0.07	m^1
	3±1	9±4	1.02±0.08	9±3	1.7±0.2	9±2	m^2
412	0.3±0.1	1.9±0.9	0.2±0.1	3±1	0.19±0.07	3±1	m^1
	0.6±0.2	2±1	0.5±0.2	3±1	0.4±0.1	3±1	m^2
434	0.3±0.2	1.1±0.3	0.2±0.2	1.1±0.2	0.2±0.3	1.1±0.1	m^1
	2±1	4±3	2±1	4±3	1.5±0.8	4±3	m^2

Table 2: Summary of average and standard error of (22) for all experiments with the wavelet direct separation predictor.

		$n = 64$	$n = 256$	$n = 1024$
`funfits`		0.05	0.17	0.81
`addfit`		5.1	N.A.	N.A.
Wavelet	Fixed design	0.08	0.14	0.35
direct separation	Random design	0.23	0.27	0.50

Table 3: CPU execution times (sec) of `funfits`, `addfit` and wavelet direct separation for different sizes of the sample and fixed and random design.

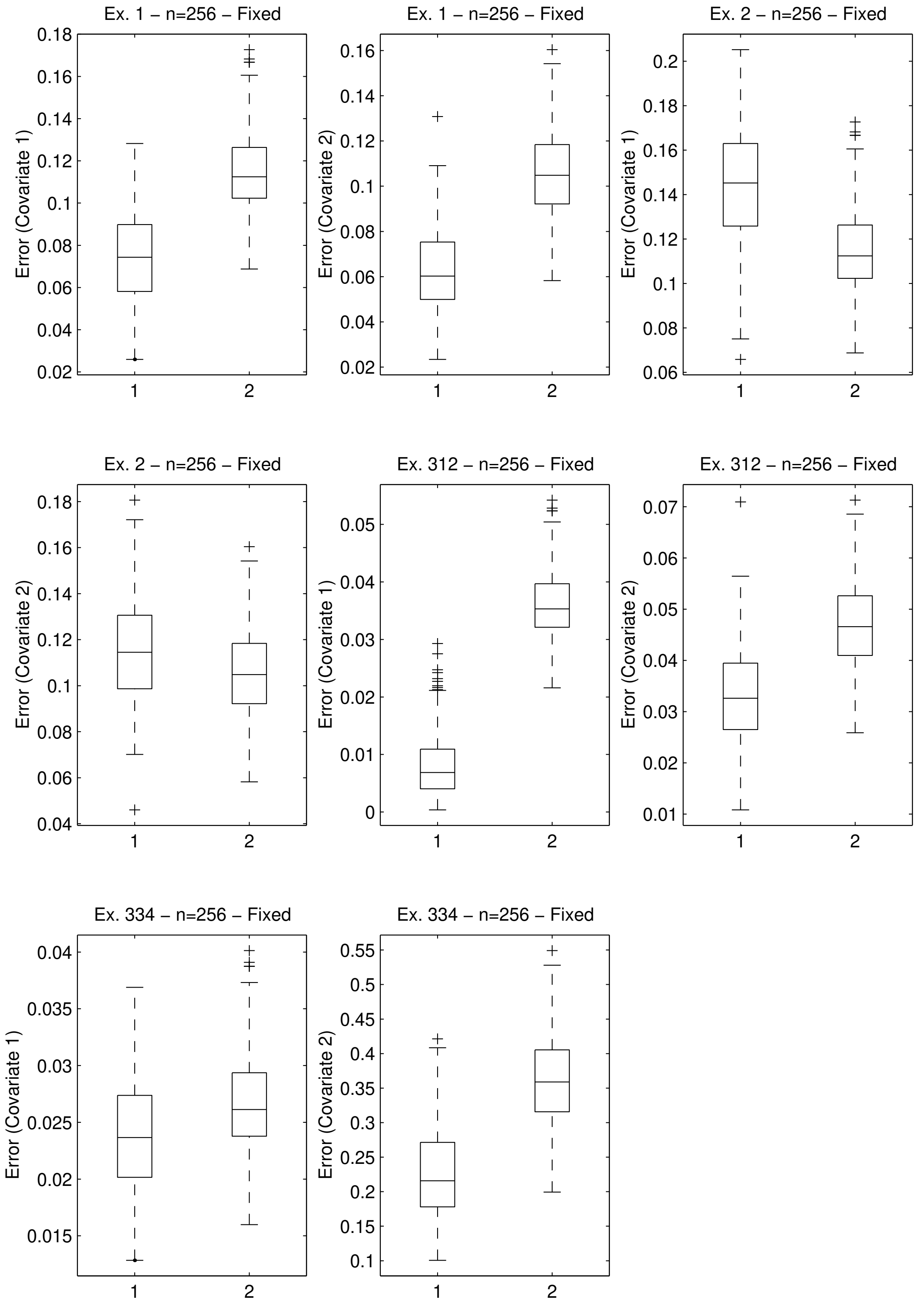

Figure 1: Boxplot of error (22) for fixed design, $n = 256$, Examples 1,2,312,334 and the predictors (1: `funfits`; 2: wavelet direct separation).

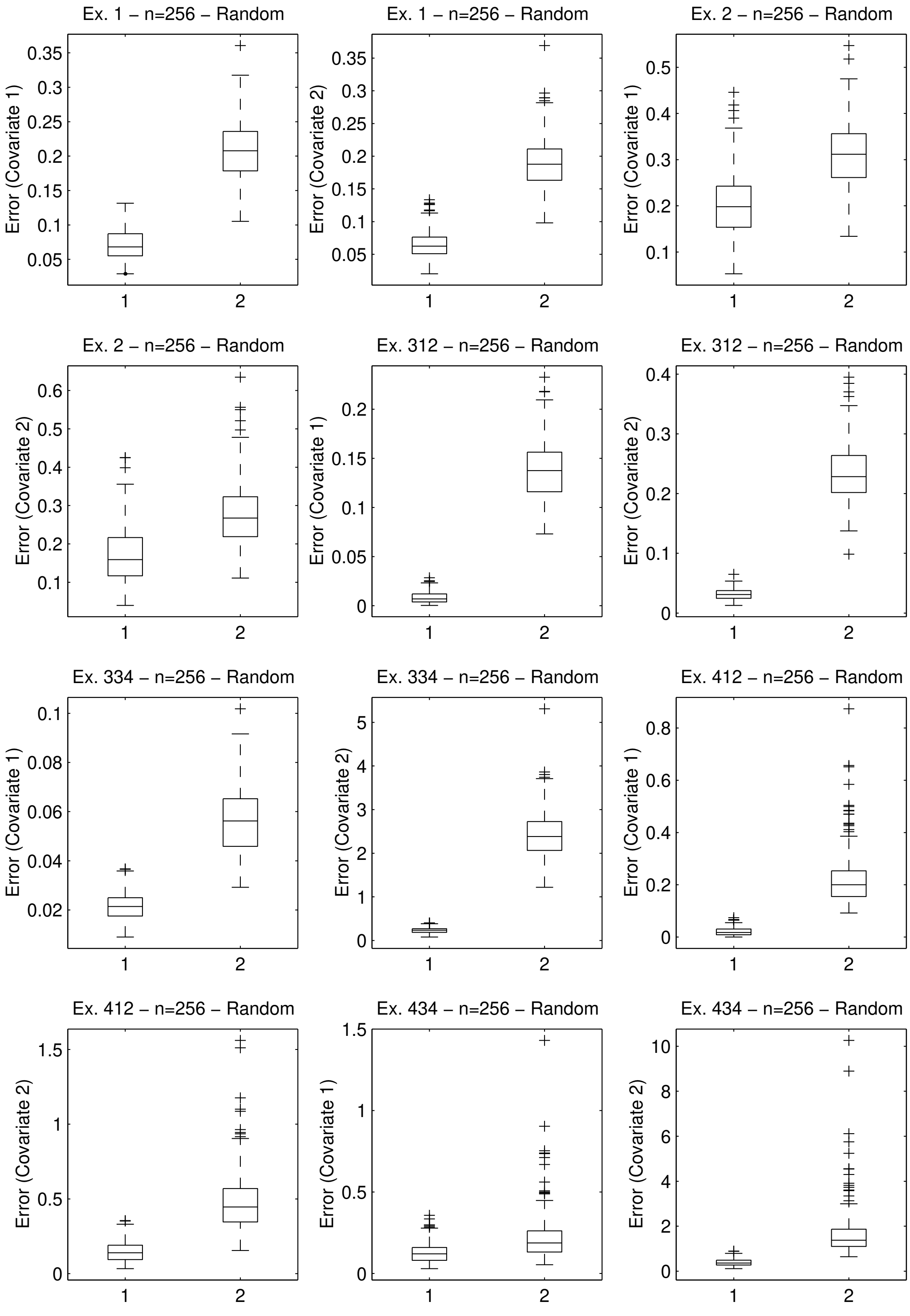

Figure 2: Boxplot of error (22) for random design, $n = 256$, Examples 1,2,312,334,412,434 and the predictors (1: `funfits`; 2: wavelet direct separation).

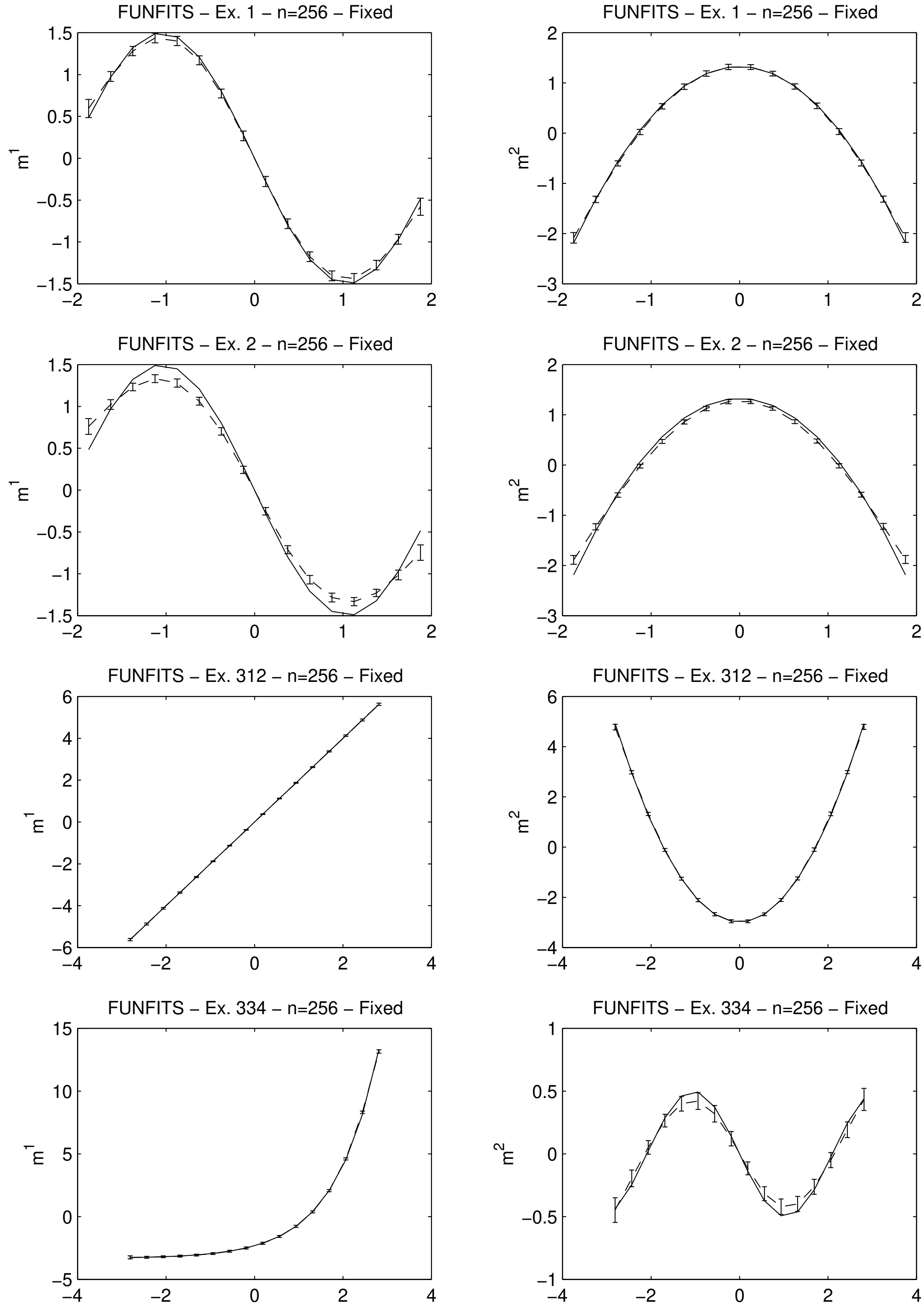

Figure 3: Average of predicted functions and mean square error (bars) for Examples 1,2,312,334, $n = 256$, fixed design and the `funfits` predictor. True functions are shown by a solid line.

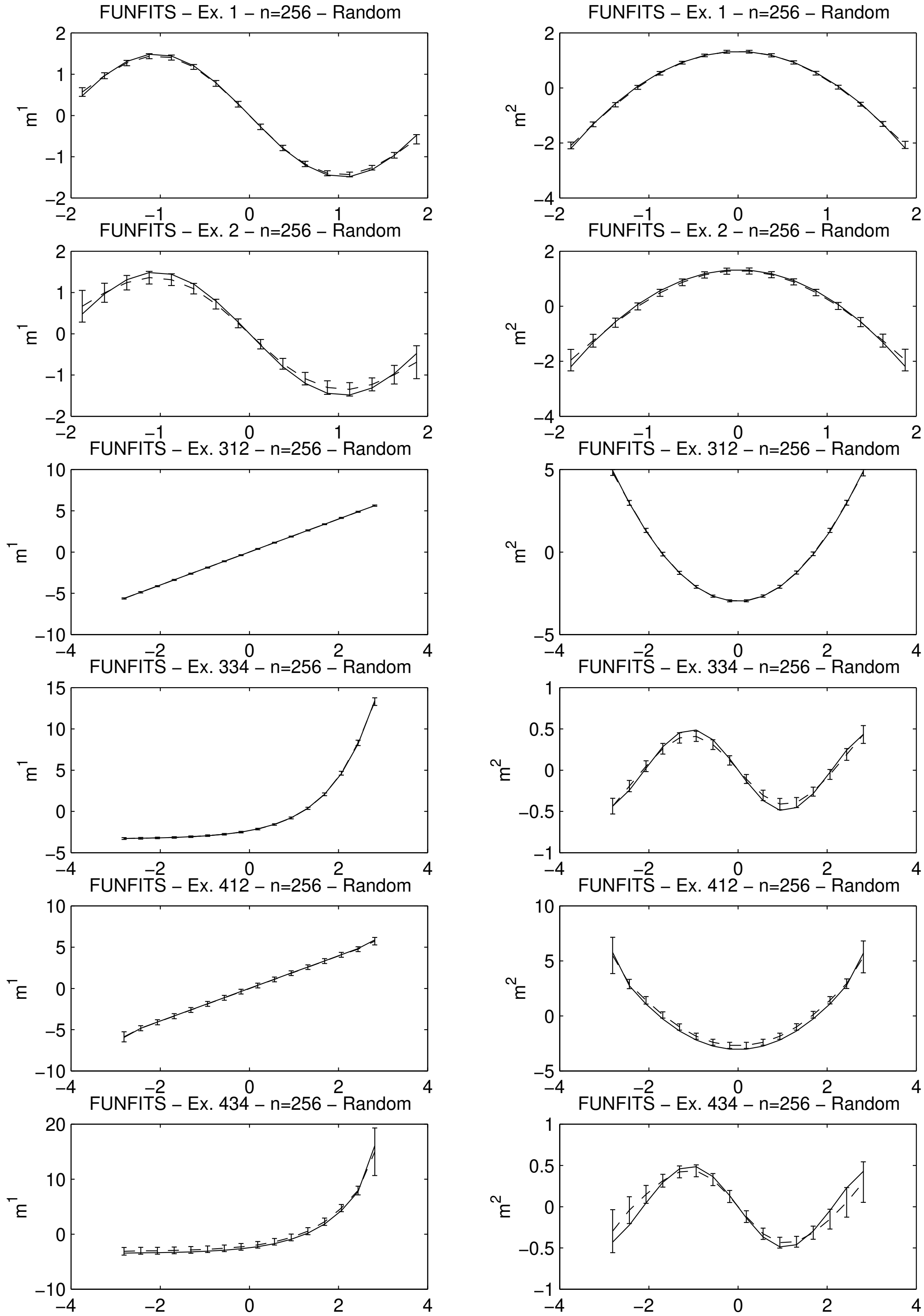

Figure 4: Average of predicted functions and mean square error (bars) for Examples 1,2,312,334,412,434, $n = 256$, random design and the `funfits` predictor. True functions are shown by a solid line.

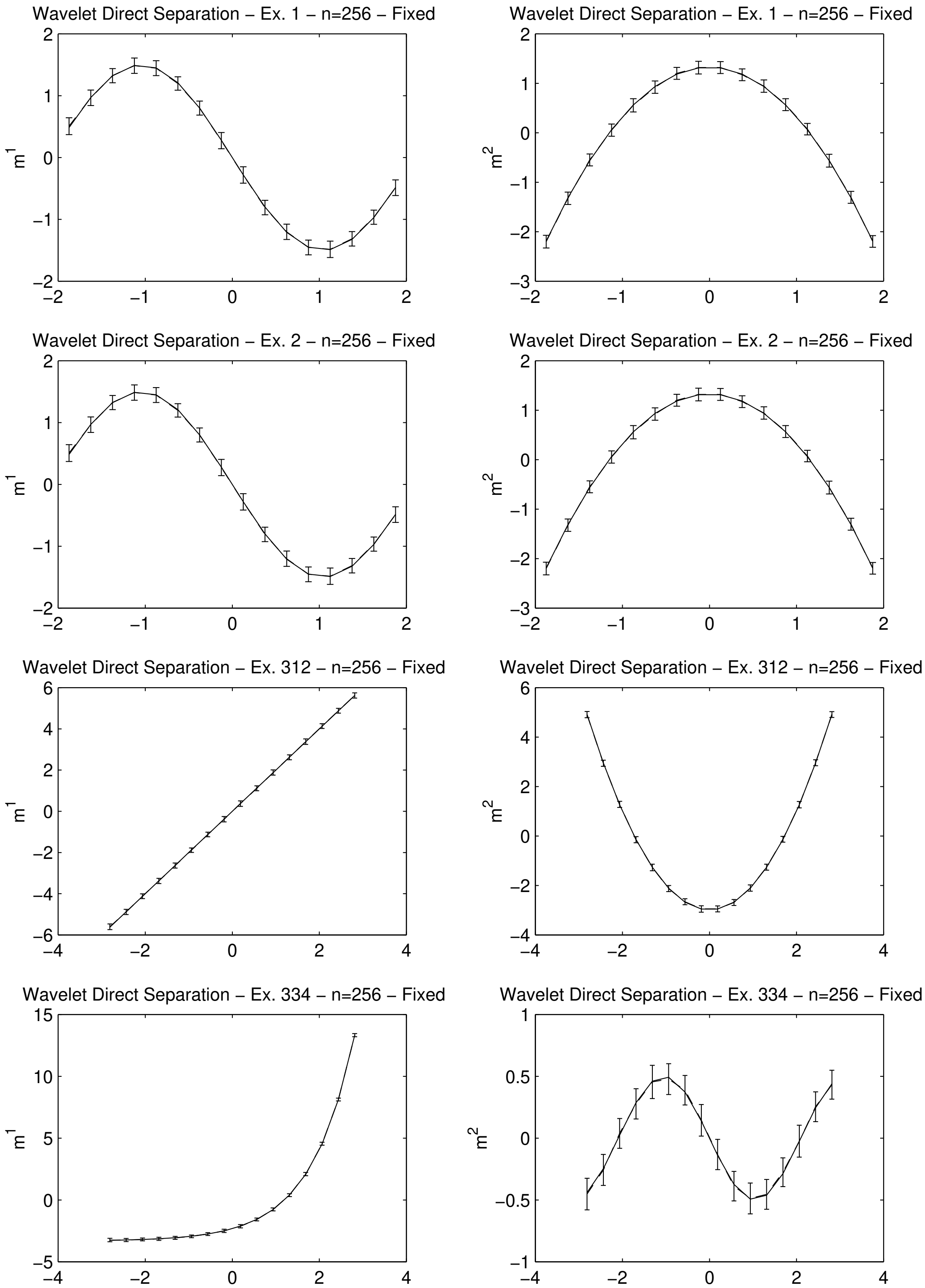

Figure 5: Average of predicted functions and mean square error for Examples 1,2,312,334, $n = 256$, fixed design and the wavelet direct separation predictor. True functions are shown by a solid line.

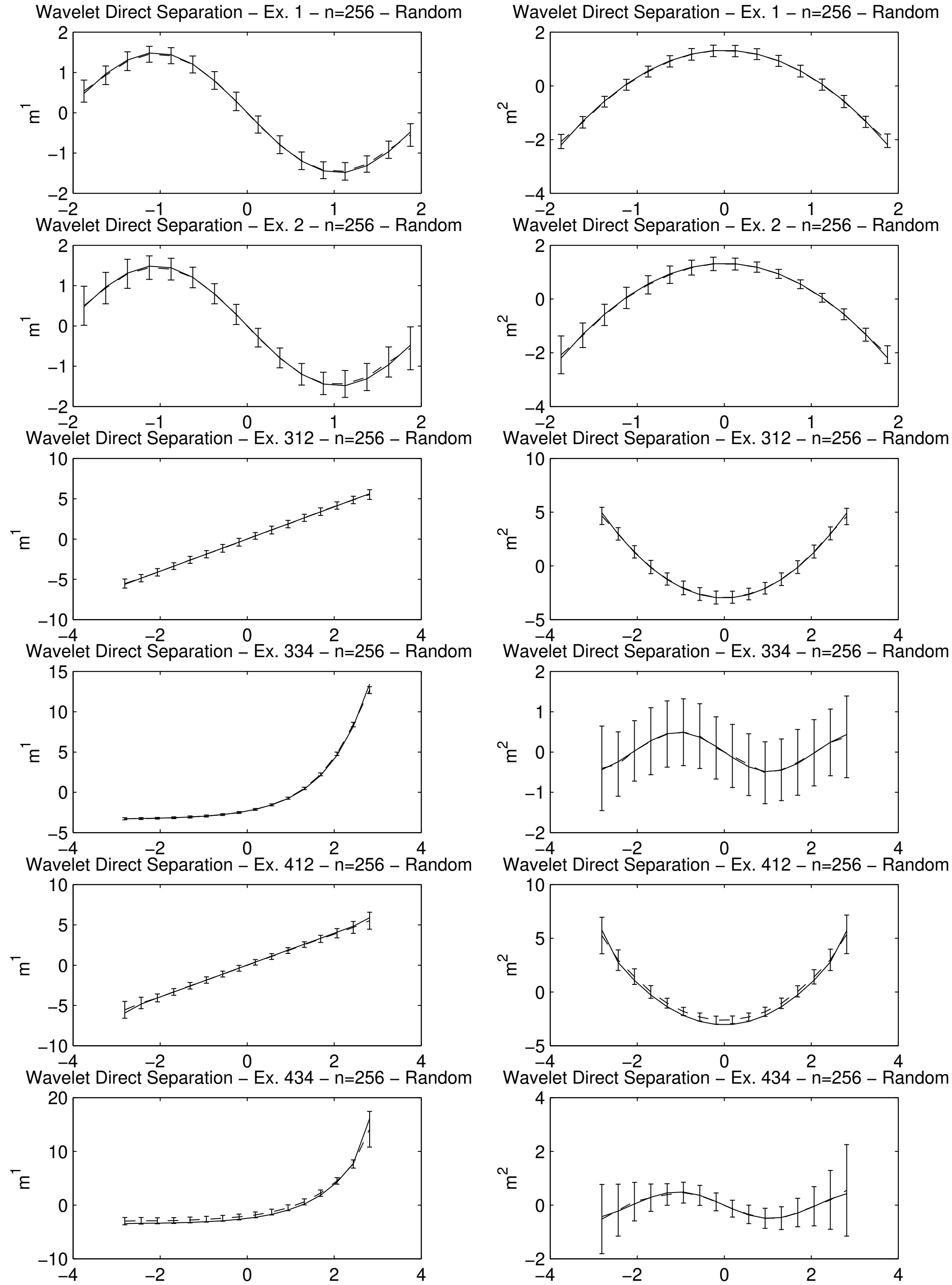

Figure 6: Average of predicted functions and mean square error for Examples 1,2,312,334,412,434, $n = 256$, random design and the wavelet direct separation predictor. True functions are shown by a solid line.

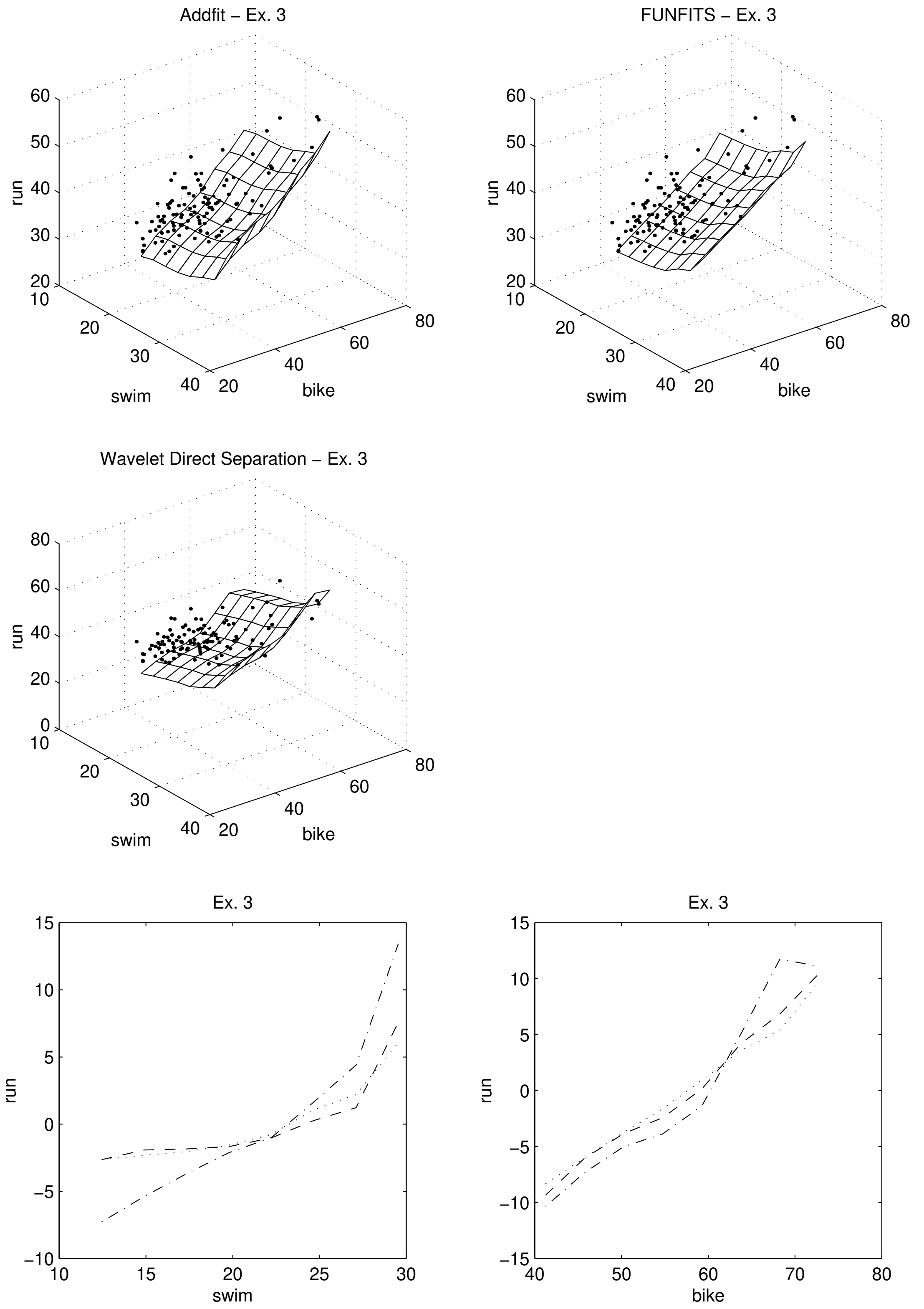

Figure 7: 3D scatterplot and surfaces predicted by the fitting methods on the `minitri` thriathlon data. The lower panel represents plots of the predicted components by `addfit` (dashed line), `funfits` (dotted line) and wavelet direct separation (dotted-dashed line).

List of Figures