ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2015)
77, *Part* 2, *pp.* 443–459

Sparse additive regression on a regular lattice

Felix Abramovich and Tal Lahav

Tel Aviv University, Israel

[Received July 2013. Revised March 2014]

Summary. We consider estimation in a sparse additive regression model with the design points on a regular lattice. We establish the minimax convergence rates over Sobolev classes and propose a Fourier-based rate optimal estimator which is adaptive to the unknown sparsity and smoothness of the response function. The estimator is derived within a Bayesian formalism but can be naturally viewed as a penalized maximum likelihood estimator with the complexity penalties on the number of non-zero univariate additive components of the response and on the numbers of the non-zero coefficients of their Fourer expansions. We compare it with several existing counterparts and perform a short simulation study to demonstrate its performance.

Keywords: Adaptive minimaxity; Additive models; Complexity penalty; Maximum *a posteriori* rule; Sparsity

1. Introduction

Consider a general non-parametric d-dimensional regression model, where the design points are on a regular lattice of size $n_1 \times \ldots \times n_d$ on $[0,1]^d$:

$$y(i_1/n_1,\ldots,i_d/n_d) = f(i_1/n_1,\ldots,i_d/n_d) + \epsilon(i_1/n_1,\ldots,i_d/n_d), \quad i_j = 0,\ldots,n_j-1, \quad j=1,\ldots,d, \tag{1}$$

$\epsilon(i_1/n_1,\ldots,i_d/n_d) \sim \mathcal{N}(0,\sigma^2)$ and are independent, and the unknown response function $f: \mathbb{R}^d \to \mathbb{R}$ is assumed to belong to a class of functions of certain smoothness. Let $N = \Pi_{j=1}^d n_j$ be the overall number of observations in model (1).

In particular, a regular grid can be useful for design of experiments when we have some prior belief on the relative relevance of predictors. Thus, we can use a finer grid (larger n_j) for more important variables and a coarse grid (smaller n_j) otherwise.

When d is large, estimation of f in model (1) suffers severely from the 'curse of dimensionality' problem. A typical remedy is to impose some additional structural constraints on f. One of the common approaches is to consider the class of *additive* models (Hastie and Tibshirani, 1990), where the unknown f can be decomposed into a sum of d univariate functions: $f(x_1,\ldots,x_d) = \Sigma_{j=1}^d f_j(x_j)$. The original model (1) becomes then

$$y(i_1/n_1,\ldots,i_d/n_d) = a_0 + \sum_{j=1}^d f_j(i_j/n_j) + \epsilon(i_1/n_1,\ldots,i_d/n_d), \quad i_j = 0,\ldots,n_j-1, \quad j=1,\ldots,d. \tag{2}$$

To make model (2) identifiable, we impose $\Sigma_{i=0}^{n_j-1} f_j(i/n_j) = 0$ for all $j=1,\ldots,d$. The goal is to estimate the unknown global mean a_0 and the functions f_j.

Address for correspondence: Felix Abramovich, Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv 69978, Israel.
E-mail: felix@post.tau.ac.il

 1369–7412/15/77443

Additive models have become a standard tool in multivariate non-parametric regression and can be efficiently fitted by the backfitting algorithm of Friedman and Stuetzle (1981). However, in a variety of modern high dimensional statistical set-ups the number of predictors d may still be large relatively to the amount of observed data. A key extra assumption then is *sparsity*, where it is assumed that only a small fraction of f_j in model (2) has a truly relevant influence on the response whereas the other $f_j = 0$. Let $\mathcal{J}_0$ and $\mathcal{J}_0^c$ be the (unknown) subsets of indices corresponding respectively to the zero and non-zero f_j. The *sparse additive model* (SPAM) is

$$y(i_1/n_1, \ldots, i_d/n_d) = a_0 + \sum_{j \in \mathcal{J}_0^c} f_j(i_j/n_j) + \epsilon(i_1/n_1, \ldots, i_d/n_d), \quad i_j = 0, \ldots, n_j - 1, \quad j = 1, \ldots, d, \tag{3}$$

and $\Sigma_{i=0}^{n_j-1} f_j(i/n_j) = 0$, $j \in \mathcal{J}_0^c$.

Expand each f_j, $j \in \mathcal{J}_0^c$, in the orthogonal discrete Fourier series assuming for simplicity of exposition that all n_j are odd:

$$f_j(i/n_j) = \sum_{k=-(n_j-1)/2}^{(n_j-1)/2} c_{kj} \exp(-2\pi \mathrm{I} k i/n_j),$$

with $I = \sqrt{-1}$ and discrete Fourier coefficients

$$c_{kj} = \frac{1}{n_j} \sum_{i=0}^{n_j-1} f_j\left(\frac{i}{n_j}\right) \exp\left(\frac{2\pi \mathrm{I} k i}{n_j}\right). \tag{4}$$

The identifiability condition $\Sigma_{i=0}^{n_j-1} f_j(i/n_j) = 0$ implies $c_{0j} = 0$.

One should make some assumptions on regularity properties of f_j. We assume that the vector of discrete Fourier coefficients c_j of f_j in equation (4) belongs to a Sobolev ellipsoid

$$\Theta_{n_j}(s_j, R_j) = \left\{ c_j : \sum_{k=-(n_j-1)/2}^{(n_j-1)/2} |c_{kj}|^2 |k|^{2s_j} \leqslant R_j^2;\ c_{0j} = 0 \right\},$$

where $s_j > \frac{1}{2}$ and $R_j < C_R$ for some constant $C_R > 0$, and denote the corresponding class of functions f_j by $\mathcal{F}_{n_j}(s_j, R_j)$. The class $\mathcal{F}_{n_j}(s_j, R_j)$ is a discrete analogue of a Sobolev ball of functions of smoothness s_j with a radius R_j (see, for example, Korostelev and Korosteleva (2011), section 10.5).

We establish the minimax rates of estimating f in model (3), where $f_j \in \mathcal{F}_{n_j}(s_j, R_j)$. The corresponding rates for the case of N distinct points for each predictor x_j were derived in Raskutti *et al.* (2012). However, we consider a design on the regular lattice, where there are N/n_j repeated observations at each of n_j grid points for every x_j. It turns out that this difference affects the resulting minimax rates.

In particular, we show that the average mean-squared error

$$\mathrm{AMSE}(\hat{f}_j, f_j) = \frac{1}{n_j} E \|\hat{f}_j - f_j\|_{n_j}^2$$

for estimating a single univariate function $f_j \in \mathcal{F}_{n_j}(s_j, R_j)$ in model (3) at the design points, where a general notation $\|\cdot\|_n$ is used for the Euclidean norm in $\mathbb{R}^n$, is of the order

$$\min(N^{-2s_j/(2s_j+1)}, n_j/N). \tag{5}$$

For sufficiently smooth f_j with $2s_j + 1 \geqslant \ln(N)/\ln(n_j)$, the rate in expression (5) is the standard minimax rate $N^{-2s_j/(2s_j+1)}$ for non-parametric estimation of a univariate function from $\mathcal{F}_{n_j}(s_j, R_j)$ (see, for example, Korostelev and Korosteleva (2011), section 10.5), but for $2s_j + 1 <$

$\ln(N)/\ln(n_j)$ it corresponds to the parametric rate of estimating f_j at each grid point i/n_j by simple averaging over the corresponding N/n_j replications. To understand this phenomenon recall that in a standard non-parametric regression set-up smoothing (local averaging over neighbour points) is necessary to reduce the variance. Although it introduces bias, the effect of the bias is negligible under smoothness assumptions on an unknown response function, whereas the benefits of variance reduction are essential. As we have mentioned above, in the case considered there are N/n_j repeated observations at each grid point i/n_j and the variance can already be reduced by their averaging without causing any bias. At the same time, the grid might be too coarse to use neighbour points in smoothing since the resulting bias becomes dominating in the bias–variance trade-off for non-smooth f_j, where $2s_j + 1 < \ln(N)/\ln(n_j)$.

In particular, when all $n_j = N^{1/d}$ are equal, the minimax $\mathrm{AMSE}(\hat{f}_j, f_j)$ in expression (5) is of the order N^{-r_j}, where $r_j = \max\{2s_j/(2s_j+1), 1-1/d\}$ and the parametric rate of averaging occurs when $2s_j + 1 < d$.

Furthermore, we prove that the overall minimax $\mathrm{AMSE}(\hat{f}, f) = (1/N)E\|\hat{f} - f\|_N^2$ for the SPAMs with $d_0 = |\mathcal{J}_0^c|$ non-zero f_j is of the order

$$\max\left\{ \sum_{j \in \mathcal{J}_0^c} \min\left(N^{-2s_j/(2s_j+1)}, \frac{n_j}{N} \right), \frac{d_0 \ln(d/d_0)}{N} \right\}. \tag{6}$$

The term $\Sigma_{j \in \mathcal{J}_0^c} \min(N^{-2s_j/(2s_j+1)}, n_j/N)$ in expression (6) is associated with the minimax rates of estimating d_0 non-zero univariate functions in $\mathcal{F}_{n_j}(s_j, R_j)$, $j \in \mathcal{J}_0^c$, whereas $d_0 \ln(d/d_0)/N$ corresponds to the error of selecting a subset of d_0 non-zero elements out of d and appears in various related model selection set-ups (e.g. Abramovich and Grinshtein (2010, 2013), Raskutti *et al.* (2011, 2012) and Rigollet and Tsybakov (2011)). For the design with N distinct points for each x_j, the similar rate $\max\{\Sigma_{j \in \mathcal{J}_0^c} N^{-r_j}, d_0 \ln(d/d_0)/N\}$, where $r_j = 2s_j/(2s_j+1)$, was derived in Raskutti *et al.* (2012).

We also propose a rate optimal estimator for estimating SPAMs (3) which is adaptive to the unknown parameters (s_j, R_j), $j \in \mathcal{J}_0^c$, of Sobolev ellipsoids and to the unknown sparsity d_0. The estimation is performed in the Fourier domain and is based on identifying non-zero vectors of (univariate) discrete Fourier coefficients c_j by imposing a penalty on the number of non-zero c_js and estimating their components by truncating the corresponding series of empirical Fourier coefficients of the data, and can be efficiently computed. The resulting estimator is developed within a Bayesian framework and can be viewed as a maximum *a posteriori* (MAP) sparse additive estimator. From a frequentist view, it corresponds to penalized maximum likelihood estimation of c_j with the complexity type of penalties on the number of non-zero c_j and numbers of their non-zero entries.

We compare the sparse additive MAP estimator with several existing counterparts proposed recently in the literature, e.g. the component selection and smoothing algorithm COSSO of Lin and Zhang (2006), the SPAM estimator of Ravikumar *et al.* (2009), the sparse additive estimator of Meier *et al.* (2009) and the M-estimator of Raskutti *et al.* (2012) (see also Koltchinskii and Yuan (2010) and Suzuki and Sugiyama (2013)). In the Fourier domain, these estimators also correspond to penalized maximum likelihood estimation of c_j but with penalties on the magnitudes of c_{kj} rather than on their cardinality. However, only the M-estimator has been proved to be rate optimal (in the minimax sense) for the case when there are N distinct observations for each predictor x_j. Moreover, all those procedures (except the SPAM estimator) are not adaptive to the smoothness s_j of f_j.

The paper is organized as follows. In Section 2 we derive the sparse additive MAP estimator. Its asymptotic adaptive minimaxity is established in Section 3, where we compare it also with its existing counterparts. The results of a simulation study are given in Section 4. Some

concluding remarks and possible extensions are discussed in Section 5. All the proofs are placed in Appendix A.

2. Maximum *a posteriori* estimator

2.1. Main idea

For any fixed $j=1,\dots,d$, averaging a general additive model (2) over all N/n_j observations at points i_j/n_j and using the identifiability conditions yields

$$\begin{aligned}\bar{y}_j(i_j/n_j) &= (n_j/N)\sum_{i_1=0}^{n_1-1}\dots\sum_{i_{j-1}=0}^{n_{j-1}-1}\sum_{i_{j+1}=0}^{n_{j+1}-1}\dots\sum_{i_d=0}^{n_d-1} y(i_1/n_1,\dots,i_j/n_j,\dots,i_d/n_d)\\ &= a_0 + f_j(i_j/n_j) + \epsilon'(i_j/n_j), \qquad i_j=0,\dots,n_j, \end{aligned} \tag{7}$$

where $\epsilon'(i/n_j)\sim\mathcal{N}\{0,(n_j/N)\sigma^2\}$ and are independent.

Equivalently, in the Fourier domain we have

$$\xi_{kj} = c_{kj} + \frac{\sigma^2}{N} z_{kj}, \qquad k=-(n_j-1)/2,\dots,(n_j-1)/2, \quad j=1,\dots,d, \tag{8}$$

where

$$\xi_{kj} = \frac{1}{n_j}\sum_{i=0}^{n_j-1}\bar{y}_j\left(\frac{i}{n_j}\right)\exp\left(\frac{2\pi \mathrm{I} k i}{n_j}\right)$$

are discrete (one-dimensional) Fourier coefficients of the vector $\bar{y}_j$, the c_{kj} are given in equation (4) and the z_{kj} are independent standard complex normal variates.

The goal now is to estimate the unknown discrete Fourier coefficients c_{kj} in expression (8) by some $\hat{c}_{kj}$. The resulting estimator $\hat{f}$ in the original domain will then be

$$\hat{f}(i_1/n_1,\dots,i_d/n_d) = \hat{a}_0 + \sum_{j=1}^{d}\hat{f}_j(i_j/n_j) = \hat{a}_0 + \sum_{j=1}^{d}\sum_{k=-(n_j-1)/2}^{(n_j-1)/2}\hat{c}_{kj}\exp(-2\pi \mathrm{I} k i_j/n_j).$$

Additivity of f and Parseval's equality imply that

$$\mathrm{AMSE}(\hat{f},f) = E|\hat{a}_0-a_0|^2 + \sum_{j=1}^{d}E\|\hat{c}_j-c_j\|_{n_j}^2$$

and the original dimensionality of the problem N is thus reduced to $\Sigma_{j=1}^{d}(n_j-1)+1$ in the Fourier domain (recall that $c_{0j}=0$ for all j).

Estimate the overall mean a_0 by the overall sample mean $\bar{y}$. Because of the identifiability conditions $\Sigma_{i=0}^{n-1} f_j(i/n)=0$, we have

$$\bar{y} = a_0 + \epsilon^*,$$

where $\epsilon^*\sim\mathcal{N}(0,\sigma^2/N)$, yielding $E|\bar{y}-a_0|^2=\sigma^2/N$. Furthermore, we naturally set $\hat{c}_{0j}=0$ for all j with no error and, therefore, $\Sigma_{i=0}^{n_j-1}\hat{f}_j(i/n_j)=0$.

Recall now that we consider a SPAM (3), where most f_j and, therefore, c_j are 0s. Under the assumption $f_j\in\mathcal{F}_{n_j}(s_j,R_j)$, $j\in\mathcal{J}_0^{\mathrm{c}}$, the corresponding c_{kj} decrease polynomially in k and c_j can be well approximated by several first c_{kj}. The algorithm proposed tries first to identify the set $\mathcal{J}_0^{\mathrm{c}}$ of non-zero vectors c_j and then estimates their entries by truncating the corresponding vectors ξ_j of empirical discrete Fourier coefficients in expression (8) at the properly adaptively chosen cut points.

2.2. Derivation

For non-zero vectors c_j in expression (8) we consider truncated estimators of the form $\hat{c}_{kj} = \xi_{kj}$, $|k| = 1, \ldots, k_j$, and $\hat{c}_{kj} = 0$ otherwise. Thus, if we knew the set of indices $\mathcal{J}_0^c$ of non-zero c_j and the cut points k_j, $j \in \mathcal{J}_0^c$, we would estimate c_{kj}, $|k| = 1, \ldots, k_j$, $j \in \mathcal{J}_0^c$, by the corresponding ξ_{kj} and set the others to 0. Since in reality they are unknown we should estimate them from the data.

We use a Bayesian framework. Consider the following hierarchical prior model on vectors c_j. Let $d_0 = |\mathcal{J}_0^c| = \#\{j : c_j \neq 0, j = 1, \ldots, d\}$ be the number of non-zero c_j, and assume some prior distribution $\pi(d_0) > 0$, $d_0 = 0, \ldots, d$, on d_0. For a given d_0, assume that all possible sets $\mathcal{J}_0^c$ of non-zero c_j with $|\mathcal{J}_0^c| = d_0$ are equally likely, i.e.

$$P(\mathcal{J}_0^c \,||\, \mathcal{J}_0^c| = d_0) = \binom{d}{d_0}^{-1}.$$

Obviously, $k_j | (j \in \mathcal{J}_0) \sim \delta(0)$ and, thus, $c_j | (j \in \mathcal{J}_0) \sim \delta(0)$. For non-zero c_j we assume some independent priors $\pi_j(k_j) | (j \in \mathcal{J}_0^c) > 0$, $k_j = 1, \ldots, (n_j - 1)/2$. To complete the prior we place independent normal priors for non-zero $c_{kj} \sim \mathcal{N}(0, \gamma\sigma^2/N)$, $j \in \mathcal{J}_0^c$, $|k| = 1, \ldots, k_j$, where $\gamma > 0$. One can also consider different γ_j.

By a straightforward Bayesian calculus, the posterior probability of a given set $\mathcal{J}_0^c$ and the corresponding k_js is

$$P(\mathcal{J}_0^c; k_1, \ldots, k_{d_0} | \xi) \propto \pi_0(d_0) \binom{d}{d_0}^{-1} \prod_{j \in \mathcal{J}_0^c} \left\{ \pi_j(k_j)(1+\gamma)^{-k_j} \exp\left(\frac{\gamma}{1+\gamma} \frac{\sum_{|k|=1}^{k_j} |\xi_{kj}|^2}{2\sigma^2/N} \right) \right\}.$$

Given the posterior distribution $P(\mathcal{J}_0^c; k_1, \ldots, k_{d_0} | \xi)$ we apply the MAP rule to find the most likely set of non-zero vectors $\mathcal{J}_0^c$ and the corresponding cut points k_j, $j \in \mathcal{J}_0^c$:

$$\max_{\mathcal{J}_0^c; k_1, \ldots, k_{d_0}} \left(\sum_{j \in \mathcal{J}_0^c} \left[\sum_{|k|=1}^{k_j} |\xi_{kj}|^2 + 2\frac{\sigma^2}{N} \left(1 + \frac{1}{\gamma}\right) \log\{\pi_j(k_j)(1+\gamma)^{-k_j}\} \right] + 2\frac{\sigma^2}{N} \left(1 + \frac{1}{\gamma}\right) \log\left\{ \pi_0(d_0) \binom{d}{d_0}^{-1} \right\} \right). \tag{9}$$

To solve problem (9), define $\hat{k}_j$ by

$$\begin{aligned} \hat{k}_j &= \arg \min_{1 \leqslant k_j \leqslant (n_j - 1)/2} \left[\sum_{k : |k| > k_j} |\xi_{kj}|^2 + 2\frac{\sigma^2}{N} \left(1 + \frac{1}{\gamma}\right) \log\{\pi_j^{-1}(k_j)(1+\gamma)^{k_j}\} \right] \\ &= \arg \min_{1 \leqslant k_j \leqslant (n_j - 1)/2} \left[-\sum_{|k|=1}^{k_j} |\xi_{kj}|^2 + 2\frac{\sigma^2}{N} \left(1 + \frac{1}{\gamma}\right) \log\{\pi_j^{-1}(k_j)(1+\gamma)^{k_j}\} \right] \end{aligned} \tag{10}$$

for each $j = 1, \ldots, d$. The MAP rule in problem (9) is then equivalent to minimizing

$$\sum_{j \in \mathcal{J}_0^c} \left[-\sum_{|k|=1}^{\hat{k}_j} |\xi_{kj}|^2 + 2\frac{\sigma^2}{N} \left(1 + \frac{1}{\gamma}\right) \log\{\pi_j^{-1}(\hat{k}_j)(1+\gamma)^{\hat{k}_j}\} + 2\frac{\sigma^2}{N} \left(1 + \frac{1}{\gamma}\right) \log\left\{ \pi_0^{-1}(d_0) \binom{d}{d_0} \right\} \right] \tag{11}$$

over all subsets of indices $\mathcal{J}_0^c \subseteq \{1, \ldots, d\}$, where $d_0 = |\mathcal{J}_0^c|$, and the resulting algorithm for solving problem (9) is then as follows.

Step 1: for each $j=1,\ldots,d$, find $\hat{k}_j$ in equation (10) and calculate

$$W_j=-\sum_{|k|=1}^{\hat{k}_j}|\xi_{kj}|^2+2\frac{\sigma^2}{N}\left(1+\frac{1}{\gamma}\right)\log\{\pi_j^{-1}(\hat{k}_j)(1+\gamma)^{\hat{k}_j}\}.$$

Step 2: order W_j in ascending order $W_{(1)}\leqslant\ldots\leqslant W_{(d)}$ and find $\hat{d}_0$:

$$\hat{d}_0=\arg\min_{0\leqslant d_0\leqslant d}\sum_{j=1}^{d_0}\left[W_{(j)}+2\frac{\sigma^2}{N}\left(1+\frac{1}{\gamma}\right)\log\left\{\pi^{-1}(d_0)\binom{d}{d_0}\right\}\right].$$

Step 3: let $\hat{\mathcal{J}}_0^{\mathrm{c}}$ be the set of indices corresponding to the $\hat{d}_0$ smallest W_j. Set $\hat{c}_j=0$ for all $j\in\hat{\mathcal{J}}_0$ and $\hat{c}_{kj}=\xi_{kj}\,\mathbb{I}(1\leqslant|k|\leqslant\hat{k}_j), k=0,\ldots,n_j, j\in\hat{\mathcal{J}}_0^{\mathrm{c}}$ (recall that, because of the identifiability conditions, $\hat{c}_{0j}=0$ for all j).

One can easily verify that the resulting MAP estimators $\hat{c}_j$ can be equivalently viewed as penalized likelihood estimators of c_j in expression (8) of the form

$$\min_{\tilde{c}_1,\ldots,\tilde{c}_d}\left[\sum_{j=1}^{d}\{\|\xi_j-\tilde{c}_j\|_{n_j}^2+\mathrm{Pen}_j(k_j)\}+\mathrm{Pen}_0(d_0)\right] \tag{12}$$

with the complexity penalty

$$\mathrm{Pen}_0(d_0)=2\frac{\sigma^2}{N}\left(1+\frac{1}{\gamma}\right)\log\left\{\pi_0^{-1}(d_0)\binom{d}{d_0}\right\} \tag{13}$$

on the number of non-zero $\tilde{c}_j$ and the complexity penalties

$$\mathrm{Pen}_j(k_j)=2\frac{\sigma^2}{N}\left(1+\frac{1}{\gamma}\right)\log\{\pi_j^{-1}(k_j)(1+\gamma)^{k_j}\},\qquad k_j=1,\ldots,(n_j-1)/2, \tag{14}$$

on the number of non-zero entries $2k_j$ of $\tilde{c}_j$.

3. Theoretical properties

3.1. Upper bound

In this section we establish theoretical properties of the proposed sparse additive MAP estimator and establish its adaptive minimaxity with respect to $\mathrm{AMSE}(\hat{f},f)=\Sigma_{j=1}^{d}\mathrm{AMSE}(\hat{f}_j,f_j)$. As we have mentioned, owing to Parseval's equality,

$$\mathrm{AMSE}(\hat{f},f)=\frac{\sigma^2}{N}+\sum_{j=1}^{d}E\|\hat{c}_j-c_j\|_{n_j}^2,$$

where $\hat{c}_j$ and c_j are discrete Fourier coefficients of $\hat{f}_j$ and f_j respectively (see expression (8)).

We start from a general upper bound on $\mathrm{AMSE}(\hat{f},f)$. Recall that $N=\Pi_{j=1}^{d}n_j$.

Proposition 1 (general upper bound). Consider the SPAM (3). Let $\hat{c}_1,\ldots,\hat{c}_d$ be the sparse additive MAP estimators (12) of the Fourier coefficients vectors $c_1,\ldots,c_d$ in equation (4) with the complexity penalties (13) and (14). Assume that $\pi_j(k)\leqslant\exp\{-c(\gamma)k\}$, $k=1,\ldots,(n_j-1)/2$, for all $j=1,\ldots,d$, where $c(\gamma)=8(\gamma+3/4)^2>9/2$. Then,

$$\mathrm{AMSE}(\hat{f},f)\leqslant C_1(\gamma)\min_{\mathcal{J}_0\subseteq\{1,\ldots,d\}}\left[\sum_{j\in\mathcal{J}_0^{\mathrm{c}}}\min_{1\leqslant k_j\leqslant(n_j-1)/2}\left\{\sum_{|k|=k_j+1}^{(n_j-1)/2}|c_{kj}|^2+\mathrm{Pen}_j(k_j)\right\}\right.$$

$$+\sum_{j\in\mathcal{J}_0}\sum_{k=-(n_j-1)/2}^{(n_j-1)/2}|c_{kj}|^2+\mathrm{Pen}_0(|\mathcal{J}_0^{\mathrm{c}}|)\Bigg]+C_2(\gamma)\frac{\sigma^2}{N}\{1-\pi_0(0)\},$$

where $C_1(\gamma)$ and $C_2(\gamma)$ depend only on γ.

Proposition 1 holds without any regularity conditions on non-zero f_j. Now we consider $f_j\in\mathcal{F}_{n_j}(s_j,R_j)$, $j\in\mathcal{J}_0^{\mathrm{c}}$.

Theorem 1 (upper bound over $\mathcal{F}_{n_j}(s_j,R_j)$). Consider model (3), where $\mathcal{J}_0^{\mathrm{c}}\neq\emptyset$. Assume that $f_j\in\mathcal{F}_{n_j}(s_j,R_j)$ for all $j\in\mathcal{J}_0^{\mathrm{c}}$.

Let $\hat{c}_1,\ldots,\hat{c}_d$ be the sparse additive MAP estimators (12) of the Fourier coefficients vectors $c_1,\ldots,c_d$ in equation (4) with the complexity penalties (14)–(13). Assume that there are constants $C_0,C_1>0$ such that

(a) $\pi_0(h)\geqslant(h/d)^{C_0h}$, $h=1,\ldots,\lfloor d/e\rfloor$ and $\pi_0(d)\geqslant\exp(-C_0d)$ and
(b) $\exp(-C_1k)\leqslant\pi_j(k)\leqslant\exp\{-c(\gamma)k\}$, $k=1,\ldots,(n_j-1)/2$, $j=1,\ldots,d$.

Then, for any $\mathcal{J}_0^{\mathrm{c}}\subseteq\{1,\ldots,d\}$ with $|\mathcal{J}_0^{\mathrm{c}}|=d_0$ and all $\mathcal{F}_{n_j}(s_j,R_j)$, $j\in\mathcal{J}_0^{\mathrm{c}}$,

$$\sup_{f_j\in\mathcal{F}_{n_j}(s_j,R_j),\,j\in\mathcal{J}_0^{\mathrm{c}}}\mathrm{AMSE}(\hat{f},f)\leqslant C_1(\gamma)\max\Bigg\{\sum_{j\in\mathcal{J}_0^{\mathrm{c}}}\min\Bigg(N^{-2s_j/(2s_j+1)},\frac{n_j}{N}\Bigg),\frac{d_0\ln(d/d_0)}{N}\Bigg\},\tag{15}$$

where $C_1(\gamma)$ is some constant depending on γ only.

One can easily verify that the conditions on priors $\pi(\cdot)$ and $\pi_j(\cdot)$ required in theorem 1 are satisfied for the (truncated) geometric priors $\pi_0(h)\propto q^h$, $h=1,\ldots,d$, and $\pi_j(k)\propto q_j^k$, $k=1,\ldots,(n_j-1)/2$, for some $0<q,q_j<1$ corresponding respectively to the complexity penalties

$$\mathrm{Pen}_0(h)\sim 2C(\gamma)\frac{\sigma^2}{N}h\Bigg\{\ln\Bigg(\frac{d}{h}\Bigg)+1\Bigg\}$$

of the $2h\ln(d/h)$ type and the Akaike information criterion type

$$\mathrm{Pen}_j(k)\sim 2\,C(\gamma)\frac{\sigma^2}{N}k$$

for some $C(\gamma)>1$.

3.2. Asymptotic minimaxity

To assess the goodness of the upper bound for the AMSE of the MAP estimator that was established in theorem 1 we derive the corresponding minimax lower bounds.

We start from the following proposition establishing the minimax lower bound for estimating a single $f_j\in\mathcal{F}_{n_j}(s_j,R_j)$ in model (7).

Proposition 2 (minimax lower bound for a single $f_j\in\mathcal{F}_{n_j}(s_j,R_j)$). Consider model (7), where $f_j\in\mathcal{F}_{n_j}(s_j,R_j)$. There is a constant $C_2>0$ such that

$$\inf_{\tilde{f}_j}\sup_{f_j\in\mathcal{F}_{n_j}(s_j,R_j)}\mathrm{AMSE}(\tilde{f}_j,f_j)\geqslant C_2\min\Bigg(N^{-2s_j/(2s_j+1)},\frac{n_j}{N}\Bigg),$$

where the infimum is taken over all estimators $\tilde{f}_j$ of f_j.

We now use this result to obtain the minimax lower bound for the AMSE in estimating f in the SPAM (3).

Theorem 2 (minimax lower bound). Consider model (3), where $f_j \in \mathcal{F}_{n_j}(s_j, R_j)$, $j \in \mathcal{J}_0^c$. There is a constant $C_2 > 0$ such that

$$\inf_{\tilde{f}} \sup_{f_j \in \mathcal{F}_{n_j}(s_j, R_j),\, j \in \mathcal{J}_0^c} \mathrm{AMSE}(\hat{f}, f) \geqslant C_2 \max\left\{ \sum_{j \in \mathcal{J}_0^c} \min\left(N^{-2s_j/(2s_j+1)}, \frac{n_j}{N} \right), \frac{d_0 \ln(d/d_0)}{N} \right\}, \tag{16}$$

where the infimum is taken over all estimators $\tilde{f}$ of f.

Theorems 1 and 2 show that, as both the sample sizes n_j and the dimensionality d increase, the asymptotic minimax convergence rate is either of order $\Sigma_{j \in \mathcal{J}_0^c} \min(N^{-2s_j/(2s_j+1)}, n_j/N)$ or $N^{-1} d_0 \ln(d/d_0)$. The former corresponds to the optimal rates of estimating d_0 single $f_j \in \mathcal{F}_{n_j}(s_j, R_j)$, whereas the latter is due to error in selecting a subset of d_0 non-zero f_j out of d and commonly appears in various related model selection set-ups (see, for example, Abramovich and Grinshtein (2010, 2013), Raskutti *et al.* (2011, 2012) and Rigollet and Tsybakov (2011)). The dominating term depends on the smoothness of the f_js (relatively to the sample sizes n_js) and the sparsity of the problem.

Furthermore, the proposed sparse additive MAP estimator with the priors $\pi_0(\cdot)$ and $\pi_j(\cdot)$ corresponding to $2d_0 \ln(d/d_0)$ type and Akaike information criterion type of penalties respectively is simultaneously minimax rate optimal over the entire range of sparse and dense amalgams of Sobolev balls $\mathcal{F}_{n_j}(s_j, R_j)$.

3.3. Comparison with other existing estimators

As we have already mentioned, various estimators for the SPAM (3) have been recently proposed in the literature. It can be shown that, as they are adapted to the set-up considered, they can be also equivalently formulated in the Fourier domain as penalized maximum likelihood estimators of c_j but with penalties on the magnitudes of c_{kj} rather than complexity-type penalties as for the proposed sparse additive MAP estimator.

Thus, the additive COSSO method of Lin and Zhang (2006), section 4, in this case can be written as

$$\arg \min_{\tilde{c}_1, \ldots, \tilde{c}_d; \theta_1 > 0, \ldots, \theta_d > 0} \left(\sum_{j=1}^{d} \|\xi_j - \tilde{c}_j\|_{n_j}^2 + \sum_{j=1}^{d} \theta_j^{-1} \sum_{k=-(n_j-1)/2}^{(n_j-1)/2} |k|^{2s_j} |\tilde{c}_{kj}|^2 + \lambda \sum_{j=1}^{d} \theta_j \right). \tag{17}$$

The form of estimator (17) is very similar to common spline smoothing which is equivalent to linear shrinkage in the Fourier domain (e.g. Wahba (1990)) with smoothing parameters θ_j but with the additional penalty on their sum. The latter makes the set of optimal θ_j sparse and, therefore, yields zero components $\hat{c}_j$ in the resulting COSSO estimators. To the best of our knowledge, there are no results on the convergence rates for the COSSO estimator.

Similarly, the sparse additive estimator of Meier *et al.* (2009) can be presented as

$$\arg \min_{\tilde{c}_1, \ldots, \tilde{c}_d} \left\{ \sum_{j=1}^{d} \|\xi_j - \tilde{c}_j\|_{n_j}^2 + \lambda_1 \sum_{j=1}^{d} \sqrt{\left(\|\tilde{c}_j\|_{n_j}^2 + \lambda_2 \sum_{k=-(n_j-1)/2}^{(n_j-1)/2} |k|^{2s_j} |\tilde{c}_{kj}|^2 \right)} \right\}, \tag{18}$$

where penalizing $\|\tilde{c}_j\|_{n_j}$ encourages sparsity, whereas the additional penalty term controls the smoothness of the estimators. For N distinct observations for each x_j, from the results of Meier *et al.* (2009), remark 2, it follows that their estimator has a suboptimal rate $O[\Sigma_{j \in \mathcal{J}_0^c} \{\ln(d)/N\}^{2s_j/(2s_j+1)}]$.

Applied to $f_j \in \mathcal{F}_{n_j}(s_j, R_j)$, a regularized M-estimator of Raskutti *et al.* (2012) is

$$\arg \min_{\tilde{c}_1,\ldots,\tilde{c}_d} \left\{ \sum_{j=1}^{d} \|\xi_j - \tilde{c}_j\|_{n_j}^2 + \lambda_1 \sum_{j=1}^{d} \|\tilde{c}_j\|_{n_j} + \lambda_2 \sum_{j=1}^{d} \sqrt{\left(\sum_{k=-(n_j-1)/2}^{(n_j-1)/2} |k|^{2s_j} |\tilde{c}_{kj}|^2 \right)} \right\} \tag{19}$$

which is similar to estimator (18) but separates the penalties on sparsity and smoothness into two additive terms. For the design with N distinct observations for each x_j, estimator (19) achieves the minimax rate

$$O\left\{ \min\left(\sum_{j \in \mathcal{J}_0^c} N^{-2s_j/(2s_j+1)} \right), \frac{d_0 \ln(d/d_0)}{N} \right\}.$$

Similar results for the M-estimator (19) were obtained in Koltchinskii and Yuan (2010) and Suzuki and Sugiyama (2013) under some additional conditions.

The serious disadvantage of all the above estimators is that they are defined for penalties involving s_j and, hence, are inherently not adaptive to the smoothness of f_j which can rarely be assumed known.

The SPAM estimator of Ravikumar *et al.* (2009) for the set-up considered becomes

$$\arg \min_{\tilde{c}_1,\ldots,\tilde{c}_d} \left\{ \sum_{j=1}^{d} \|\xi_j - \tilde{c}_j\|_{n_j}^2 + \lambda \sum_{j=1}^{d} \sqrt{(2k_j)} \|\tilde{c}_j\|_{k_j} \right\} \tag{20}$$

for the fixed truncation cut points k_j. In this form, the SPAM is closely related to the group lasso estimator of Yuan and Lin (2006) and can be obtained explicitly:

$$\hat{c}_j = \left(1 - \frac{(\lambda/2)\sqrt{(2k_j)}}{\|\tilde{\xi}_j\|_{k_j}} \right)_+ \tilde{\xi}_j, \tag{21}$$

where $\tilde{\xi}_j$ is ξ_j truncated at k_j. Ravikumar *et al.* (2009) showed persistency of their estimator but did not provide results on convergence rates of its AMSE.

Finally, we mention Guedj and Alquier (2013) who considered a Bayesian model which was similar to that proposed in this paper with geometric priors $\pi_0(\cdot)$ and $\pi_j(\cdot)$. They estimated c_j by the corresponding posterior means and, for the case of N distinct observations for each x_j, showed that the resulting estimator is asymptotically nearly minimax (up to an additional log-factor) over Sobolev classes. A similar Bayesian estimator of Suzuki (2012) achieves the optimal rate but for smaller functional classes. The practical implementation of these procedures involves, however, high dimensional Markov chain Monte Carlo algorithms.

4. Simulation study

To illustrate the performance of the sparse additive MAP estimator proposed we conducted a simulation study. Similarly to example 1 of Lin and Zhang (2006), example 3 of Meier *et al.* (2009) and example 3 of Guedj and Alquier (2013), we considered the SPAM (3) with $d = 50$ and four non-zero components f_j ($d_0 = 4$):

$$f_1(x) = x,$$
$$f_2(x) = (2x-1)^2,$$
$$f_3(x) = \frac{\sin(2\pi x)}{2 - \sin(2\pi x)},$$
$$f_4(x) = 0.1 \sin(2\pi x) + 0.2 \cos(2\pi x) + 0.3 \sin^2(2\pi x) + 0.4 \cos^3(2\pi x) + 0.5 \sin^3(2\pi x)$$

but on the regular lattice $[0,1]^{50}$. We used $n = 101$ and, therefore, $N = 101^{50}$. Each non-zero f_j was standardized to have

Table 1. AMSE averaged over 1000 replications for various SNRs

SNR	*Method*	*AMSE*	$AMSE_1$	$AMSE_2$	$AMSE_3$	$AMSE_4$	$AMSE_0$	$\hat{d}_0$
1	MAP	0.6242	0.3083	0.1023	0.0926	0.1209	0.0000	4.0
	SPAM($\lambda=0.26$)	0.8007	0.3371	0.1283	0.1178	0.1467	0.0015	19.3
5	MAP	0.1937	0.1334	0.0285	0.0157	0.0161	0.0000	4.0
	SPAM($\lambda=0.10$)	0.2632	0.1492	0.0373	0.0238	0.0282	0.0005	25.7
10	MAP	0.1285	0.0936	0.0182	0.0099	0.0067	0.0000	4.0
	SPAM($\lambda=0.06$)	0.1686	0.1021	0.0220	0.0131	0.0114	0.0004	32.3

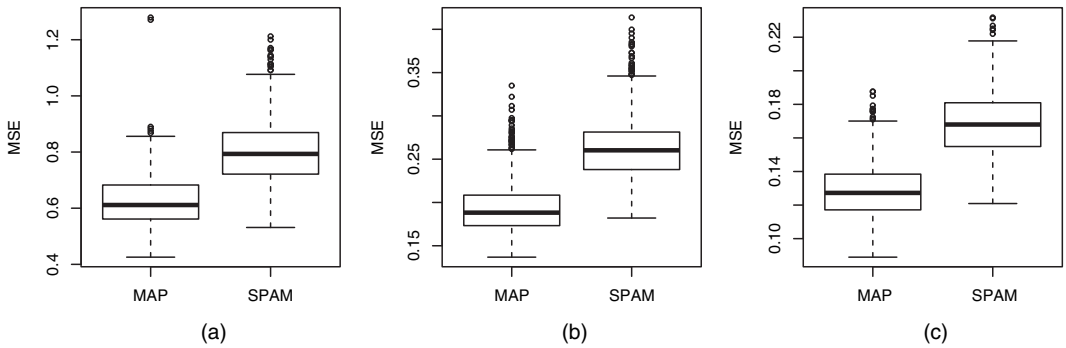

Fig. 1. Boxplots for (global) AMSE for various SNRs: (a) SNR = 1; (b) SNR = 5; (c) SNR = 10

$$(1/n)\sum_{i=0}^{n-1} f_j(i/n)=0,$$
$$(1/n)\sum_{i=0}^{n-1} f_j^2(i/n)=1.$$

The noisy data were generated according to model (7) by adding independent random Gaussian variates $\mathcal{N}\{0,(n/N)\sigma^2\}$ to $f_j(i/n)$, $i=0,\ldots,n-1$, $j=1,\ldots,d$. The values of the noise variance σ^2 were chosen to correspond to values 1, 5 and 10 for the signal-to-noise ratio SNR defined as

$$\mathrm{SNR}=\mathrm{var}(f_j)\Big/\left(\frac{n}{N}\sigma^2\right)=\frac{N}{\sigma^2 n}.$$

Performing the discrete Fourier transform of the noisy data yielded the equivalent model (8) in the Fourier domain. We then applied the proposed MAP algorithm to corresponding noisy Fourier coefficients ξ_{kj} by using truncated geometric priors for $\pi_0(\cdot)$ and $\pi_j(\cdot)$ with $q=q_j=0.5$ and $\gamma=5$. The noise level σ was assumed unknown and estimated from the data. Since the vector of the true Fourier coefficients c_j in expression (8) lies in a Sobolev ellipsoid, the sequence $|c_{kj}|$ decays to 0 polynomially with k. Thus, for large k, the empirical Fourier coefficients ξ_{kj} in expression (8) are mostly pure noise. To correct for the bias due to the possible presence of several large coefficients, we robustly estimated $\sigma/\sqrt{N}$ as

$$\frac{\hat{\sigma}}{\sqrt{N}}=\frac{\sqrt{2}\,\mathrm{MAD}[\{\mathrm{Re}(\xi_{kj}),\mathrm{Im}(\xi_{kj})\},\ k=0.8(n_j-1)/2,\ldots,(n_j-1)/2;\ j=1,\ldots,50]}{0.6745}.$$

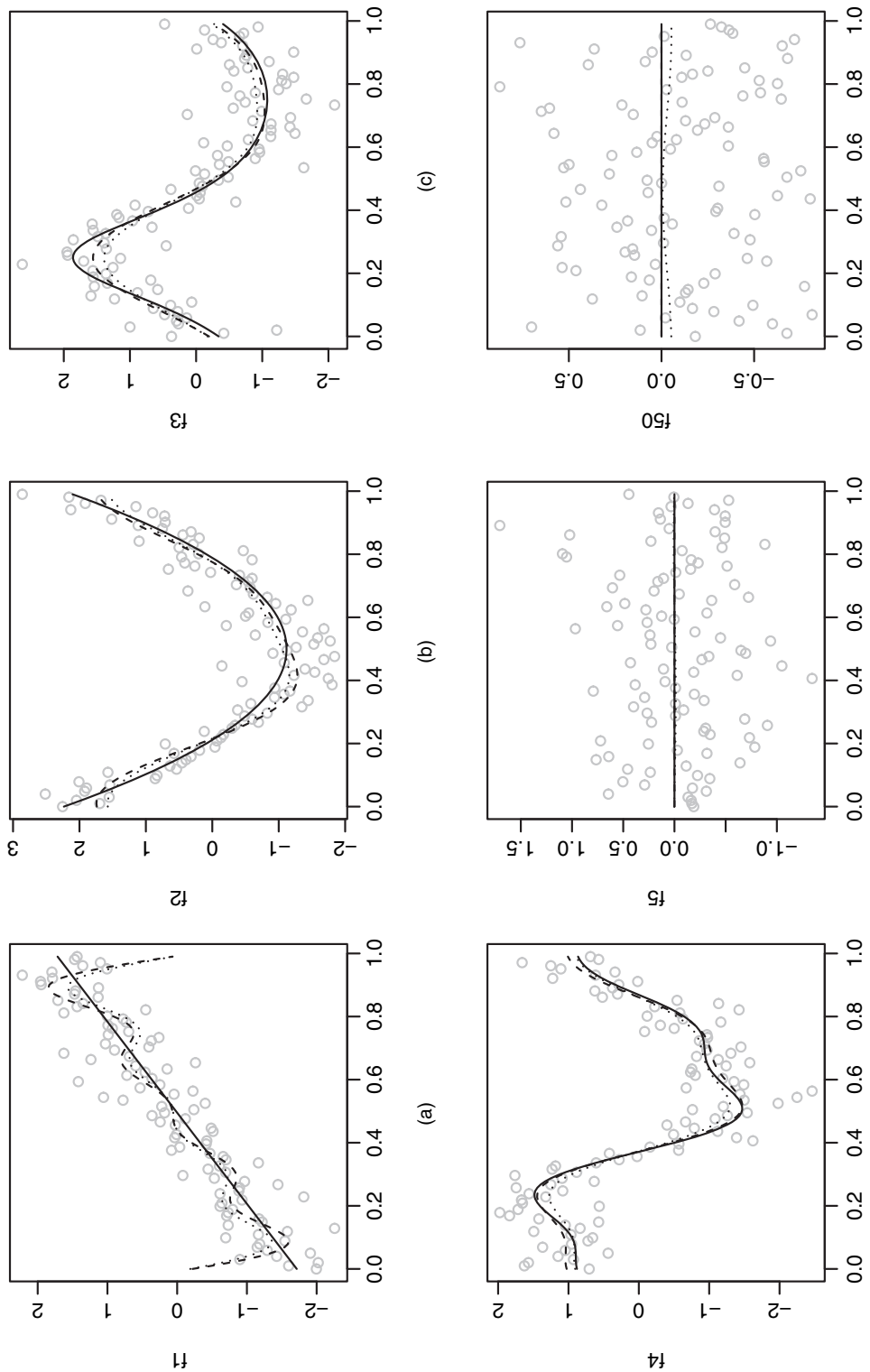

Fig. 2. Examples of MAP (- - - - - -) and SPAM (· · · · · ·) estimators for various f_j (———): (a) f_1; (b) f_2; (c) f_3; (d) f_4; (e) f_5 (= 0); (f) f_{50} (= 0) (SNR = 5)

This is similar to a standard practice for estimating σ from wavelet coefficients at the finest resolution level in wavelet-based methods (see, for example, Donoho and Johnstone (1994)). The resulting estimates for σ were very precise for all SNRs.

We compared also the resulting sparse additive MAP estimator with the SPAM estimator (20) of Ravikumar *et al.* (2009) which for the model considered is essentially the group lasso estimator of Yuan and Lin (2006) and is available in closed form in the Fourier domain—see expression (21). For the SPAM estimator we used the same cut points $\hat{k}_j$ from expression (10) as for the MAP estimator and the oracle-chosen threshold λ that minimizes

$$\mathrm{AMSE}(f, \hat{f}^{\mathrm{SPAM}}) = \sum_{j=1}^{d} \|\hat{c}_j^{\mathrm{SPAM}}(\lambda) - c_j\|_n^2$$

estimated by averaging over a series of 1000 replications for each value of λ by using a grid search. The resulting choices were $\lambda = 0.26$ for $\mathrm{SNR} = 1$, $\lambda = 0.10$ for $\mathrm{SNR} = 5$ and $\lambda = 0.06$ for $\mathrm{SNR} = 10$. Thus, the oracle λ decreased with increasing SNR.

For each SNR-level we calculated the (global) AMSE for both methods and analysed also their performance for each individual f_j. Thus, AMSE_1, AMSE_2, AMSE_3 and AMSE_4 are the AMSEs for the corresponding four non-zero f_j, $j = 1, \ldots, 4$, whereas AMSE_0 is the average AMSE over all 46 zero f_j. In addition, we compared the two methods for identifying non-zero f_j though it is a somewhat different problem from our original goal of estimating functions in quadratic norm and we calculated $\hat{d}_0 = \#\{j : \hat{f}_j \neq 0,\ j = 1, \ldots, 50\}$. The results are summarized in Table 1. Fig. 1 gives the corresponding boxplots. Fig. 2 gives typical examples of estimators obtained by both methods for non-zero and zero f_j.

The results in Table 1 show that the MAP consistently outperforms the SPAM estimator both globally and for each individual component f_j. For both methods the main contribution to the global AMSE came from estimating non-zero f_j. The MAP estimator almost perfectly identified the set of non-zero f_j whereas the oracle choices for λ in the SPAM estimator were quite small and, as a result, too many $\hat{f}_j$ were non-zero (see, for example, Fig. 2(f)). In fact, it is a known common phenomenon for lasso-type estimators.

5. Concluding remarks

We considered sparse additive regression on a regular lattice, where the univariate components f_j of the unknown response function f belong to Sobolev balls. We established the minimax convergence rates for estimating f and proposed an adaptive Fourier-based estimator which is rate optimal over the entire range of Sobolev classes of different sparsity and smoothness. The resulting estimator was developed within a Bayesian formalism but can also be viewed, in fact, as a penalized maximum likelihood estimator of the Fourier coefficients of f with certain complexity penalties on the number of non-zero f_j and on the numbers of non-zero entries of their Fourier coefficients c_j. It can be efficiently computed and the simulation study presented demonstrates its good performance.

The results of the paper can be extended to more general Besov classes of functions by using the wavelet series expansions of f_j. The corresponding vectors of wavelet coefficients will lie then within weak l_p-balls (e.g. Johnstone (2013), section 9.7) and one can apply the results of Abramovich and Grinshtein (2013) for estimating a sparse group of sparse vectors from weak l_p-balls. The extension is quite straightforward though the details should be worked out. In particular, the resulting MAP estimator should mimic (hard) thresholding within each non-zero vector of wavelet coefficients instead of truncation as in the case of Fourier series considered (see Abramovich and Grinshtein (2013)).

Acknowledgements

The work was supported by the Israel Science Foundation, grant ISF-820/13. We are grateful to Anestis Antoniadis, Alexander Goldenshluger and Vadim Grinshtein for fruitful discussions and valuable remarks. Helpful comments by the Joint Editor and a referee are gratefully acknowledged.

Appendix A

Throughout the proofs we use C to denote a generic positive constant, which is not necessarily the same each time that it is used, even within a single equation. Similarly, $C(\gamma)$ is a generic positive constant depending on γ.

A.1. Proof of proposition 1

As we have mentioned before, the sparse additive MAP estimator (12) proposed can be equivalently viewed as a penalized maximum likelihood estimator with complexity penalties (13) and (14). We can apply then the general results of Birgé and Massart (2001) for complexity-penalized estimators.

Rewrite first model (8) in a different form. Set $\xi = (\xi_{-(n_1-1)/2,1}, \ldots, \xi_{(n_1-1)/2,1}, \ldots, \xi_{-(n_d-1)/2,d}, \ldots, \xi_{(n_d-1)/2,d})^{\mathrm{T}}$ to be an amalgamated vector of length $N_0 = \sum_{j=1}^{d} n_j$ of d vectors $\xi_1, \ldots, \xi_d$. Similarly, define N_0-dimensional amalgamated vectors $c = (c_{-(n_1-1)/2,1}, \ldots, c_{(n_1-1)/2,1}, \ldots, c_{-(n_d-1)/2,d}, \ldots, c_{(n_d-1)/2,d})^{\mathrm{T}}$ and $z = (z_{-(n_1-1)/2,1}, \ldots, z_{(n_1-1)/2,1}, \ldots, z_{-(n_d-1)/2,d}, \ldots, z_{(n_d-1)/2,d})^{\mathrm{T}}$. The original model (8) can be rewritten then as

$$\xi_i = c_i + \frac{\sigma^2}{N} z_i, \qquad i = 1, \ldots, N_0, \tag{22}$$

where z_i are independent standard complex normal variates. Define an indicator vector v by $v_i = \mathbb{I}(c_i \neq 0)$, $i = 1, \ldots, N_0$. Thus, in terms of model (22), $k_j = \frac{1}{2}\sum_{i=S_{j-1}+1}^{S_j} v_i$, where $S_j = \sum_{l=1}^{j-1} n_l$, and $d_0 = \#\{j : k_j > 0\}$. For a given v, let $D_v = 2\sum_{j=1}^{d} k_j = \#\{i : v_i = 1, i = 1, \ldots, N_0\}$ be the overall number of non-zero entries of c, and define

$$L_v = \begin{cases} \dfrac{1}{D_v}\left[\displaystyle\sum_{j=1}^{d} \log\left\{\pi_j^{-1}(k_j)\right\} + \log\left\{\pi_0^{-1}(d_0)\binom{d}{d_0}\right\}\right] & \text{if } v \neq 0, \\ \log\{\pi_0^{-1}(0)\} & \text{if } v = 0. \end{cases} \tag{23}$$

In the above notation the sparse additive MAP estimator $\hat{c} = (\hat{c}_{-(n_1-1)/2,1}, \ldots, \hat{c}_{(n_1-1)/2,1}, \ldots, \hat{c}_{-(n_d-1)/2,d}, \ldots, \hat{c}_{(n_d-1)/2,d})^{\mathrm{T}}$ is the penalized maximum likelihood estimator of c with complexity penalty

$$\begin{aligned} \mathrm{Pen}(v) &= 2\frac{\sigma^2}{N}\left(1 + \frac{1}{\gamma}\right)\left[\sum_{j=1}^{d} \log\left\{\pi_j^{-1}(k_j)(1+\gamma)^{k_j}\right\} + \log\left\{\pi_0^{-1}(d_0)\binom{d}{d_0}\right\}\right] \\ &= 2\frac{\sigma^2}{N}\left(1 + \frac{1}{\gamma}\right) D_v\left\{L_v + \frac{1}{2}\log(1+\gamma)\right\} \end{aligned} \tag{24}$$

for $v \neq 0$, and

$$\mathrm{Pen}(0) = 2\frac{\sigma^2}{N}\left(1 + \frac{1}{\gamma}\right) L_0.$$

One can easily verify that

$$\sum_{v \neq 0} \exp(-D_v L_v) = \sum_{k=1}^{d} \pi_0(k) = 1 - \pi_0(0).$$

Furthermore, straightforward calculus similar to that in the proof of theorem 1 of Abramovich *et al.* (2007) implies that, under the conditions on the priors $\pi_j(\cdot)$ of proposition 1, the complexity penalty $\mathrm{Pen}(v)$ in equation (24) satisfies

$$\mathrm{Pen}(v) \geqslant C(\gamma)\frac{\sigma^2}{N} D_v \left\{1 + \sqrt{(2L_v)}\right\}^2,$$

for some $C(\gamma) > 1$. One can then apply theorem 2 of Birgé and Massart (2001) to have

$$\sum_{j=1}^{d} E(\|\hat{c}_j - c_j\|_2^2) \leqslant c_1(\gamma) \min_{\mathcal{J}_0 \subseteq \{1,\ldots,d\}} \left[\sum_{j \in \mathcal{J}_0^c} \min_{1 \leqslant k_j \leqslant (n_j-1)/2} \left\{ \sum_{k:|k|>k_j} |c_{kj}|^2 + \mathrm{Pen}_j(k_j) \right\} + \sum_{j \in \mathcal{J}_0} \sum_{|k|=1}^{(n_j-1)/2} |c_{kj}|^2 + \mathrm{Pen}_0(d_0) \right] + c_2(\gamma) \frac{\sigma^2}{N} \{1 - \pi_0(0)\}.$$

Parseval's equality

$$\mathrm{AMSE}(\hat{f}, f) = \sum_{j=1}^{d} E(\|\hat{c}_j - c_j\|_{n_j}^2) + \frac{\sigma^2}{N}$$

completes the proof.

A.2. Proof of theorem 1

Let $\mathcal{J}_0^{c*}$ be the true (unknown) subset of non-zero c_j and $d_0^* = |\mathcal{J}_0^{c*}|$. Consider separately two cases.

Case 1: $d_0^* \leqslant \lfloor d/e \rfloor$. Applying the general upper bound that was established in proposition 1 for $\mathcal{J}_0 = \mathcal{J}_0^*$ yields

$$\mathrm{AMSE}(\hat{f}, f) \leqslant C_1(\gamma) \left[\sum_{j \in \mathcal{J}_0^{c*}} \min_{1 \leqslant k_j \leqslant (n_j-1)/2} \left\{ \sum_{|k|=k_j+1}^{(n_j-1)/2} |c_{kj}|^2 + \mathrm{Pen}_j(k_j) \right\} + \mathrm{Pen}_0(d_0^*) \right] + C_2(\gamma) \frac{\sigma_2}{N} \{1 - \pi_0(0)\}. \tag{25}$$

Choose the cut points $k_j = \left\lfloor \frac{1}{2} \min(N^{1/(2s_j+1)}, n_j - 1) \right\rfloor$ for $j \in \mathcal{J}_0^{c*}$. If $k_j < (n_j - 1)/2$, for $c_j \in \Theta_{n_j}(S_j, R_j)$ we have $\sum_{|k|=k_j+1}^{(n_j-1)/2} |c_{kj}|^2 = O(k_j^{-2s_j}) = O(N^{-2s_j/(2s_j+1)})$, whereas for $k = (n_j - 1)/2$ this term obviously disappears. Furthermore, under the conditions on the priors $\pi_j(\cdot)$, the corresponding penalties $\mathrm{Pen}_j(\cdot)$ in equation (14) are of the Akaike information criterion type, where

$$\mathrm{Pen}_j(k_j) \sim 2C(\gamma) \frac{\sigma^2}{N} k_j = O\left\{ \min\left(N^{-2s_j/(2s_j+1)}, \frac{n_j}{N} \right) \right\}.$$

Hence, the first term $\sum_{j \in \mathcal{J}_0^{c*}}$ on the right-hand side of equation (25) is of the order $\sum_{j \in \mathcal{J}_0^{c*}} \min(N^{-2s_j/(2s_j+1)}, n_j/N)$.

Finally,

$$\binom{d}{d_0^*} \leqslant \left(\frac{d}{d_0^*} \right)^{2d_0^*}$$

for $d_0^* \leqslant \lfloor d/e \rfloor$ (see, for example, lemma A1 of Abramovich *et al.* (2010)) and, therefore, the conditions on $\pi_0(\cdot)$ imply that

$$\mathrm{Pen}_0(d_0^*) \leqslant C(\gamma) \frac{\sigma^2}{N} d_0^* \log\left(\frac{d}{d_0^*} \right).$$

Case 2: $\lfloor d/e \rfloor < d_0^* \leqslant d$. In this case we apply proposition 1 for $\mathcal{J}_0 = \emptyset$. Evidently, $|\mathcal{J}_0^c| = d$ and $\mathcal{J}_0^c = \mathcal{J}_0^* \cup \mathcal{J}_0^{c*}$. Choose the cut points $k_j = \left\lfloor \frac{1}{2} \min(N^{1/(2s_j+1)}, n_j - 1) \right\rfloor$ for $j \in \mathcal{J}_0^{c*}$ as before and $k_j = 1$ for $j \in \mathcal{J}_0^*$. Then,

$$\mathrm{AMSE}(\hat{f}, f) \leqslant C_1(\gamma) \left[\sum_{j \in \mathcal{J}_0^{c*}} \left\{ \sum_{|k|=k_j+1}^{(n_j-1)/2} |c_{kj}|^2 + \mathrm{Pen}_j(k_j) \right\} + \sum_{j \in \mathcal{J}_0^*} \mathrm{Pen}_j(1) + \mathrm{Pen}_0(d) \right] + C_2(\gamma) \frac{\sigma_2}{N} \{1 - \pi_0(0)\}.$$

We have already shown that the first term $\sum_{j \in \mathcal{J}_0^{c*}}$ on the right-hand side of equation (26) is

$$O\left\{ \sum_{j \in \mathcal{J}_0^{c*}} \min(N^{-2s_j/(2s_j+1)}, n_j/N) \right\}.$$

The conditions of $\pi_j(1)$ and $\pi_0(d)$ imply that both $\Sigma_{j\in J_0^*}\mathrm{Pen}_j(1)$ and $\mathrm{Pen}_0(d)$ are $O(d/N)$, and, therefore, the first term in equation (26) is dominating when $d_0^* \sim d$.

A.3. Proof of proposition 2

Consider model (7) and the equivalent Gaussian sequence model (8) in the Fourier domain. Evidently, $\inf_{\tilde f_j}\sup_{f_j\in\mathcal{F}_{n_j}(s_j,R_j)}\mathrm{AMSE}(\tilde f_j, f_j) = \inf_{\tilde c_j}\sup_{c_j\in\Theta_{n_j}(s_j,R_j)} E\|\tilde c_j - c_j\|_{n_j}^2$, where $\tilde c_j$ are discrete Fourier coefficients of $\tilde f_j$.

Most of the proof is a direct consequence of the standard techniques for establishing minimax lower bounds in the Gaussian sequence model over Sobolev ellipsoids (see, for example, Tsybakov (2009), section 3.2) but, unlike the standard set-up, the variance in the considered model (8) depends on the sample size N that may affect the minimax rates.

Consider the class of diagonal linear estimators $\tilde c_j(\lambda)$ of the form $\tilde c_{kj} = \lambda_k \xi_{kj}$, $k = -(n-1)_j/2, \ldots, -1, 1, \ldots, (n_j-1)/2$ and $\tilde c_{0j} = 0$ (see Section 2.1). It is well known (see, for example, Tsybakov (2009), section 3.2) that, as $n_j \to \infty$, the minimax linear diagonal estimator is asymptotically minimax over all estimators of f_j:

$$\begin{aligned}\inf_{\tilde c_j}\sup_{c_j\in\Theta_{n_j}(s_j,R_j)} E\|\tilde c_j - c_j\|_{n_j}^2 &\sim \inf_{\lambda}\sup_{c_j\in\Theta_{n_j}(s_j,R_j)} E\|\tilde c_j(\lambda) - c_j\|_{n_j}^2 \\ &= \sup_{c_j\in\Theta_{n_j}(s_j,R_j)}\inf_{\lambda} E\|\tilde c_j(\lambda) - c_j\|_{n_j}^2.\end{aligned}$$

By standard calculus (see, for example, Tsybakov (2009), section 3.2),

$$\inf_{\lambda} E\|\tilde c_j(\lambda) - c_j\|_{n_j}^2 = \frac{\sigma^2}{N}\sum_{k=-(n_j-1)/2}^{(n_j-1)/2}\frac{|c_{kj}|^2}{|c_{kj}|^2 + \sigma^2/N} \tag{26}$$

and the minimax linear estimator $\hat c_j^{\mathrm{L}}$ is then of the form

$$\hat c_{kj}^{\mathrm{L}} = (1 - k^{s_j}\kappa_j)_+\xi_{kj},$$

where κ_j is the solution of the equation

$$\frac{\sigma^2}{N}\sum_{k=1}^{(n_j-1)/2}(2k)^{s_j}(1-(2k)^{s_j}\kappa_j)_+ = \kappa_j R_j^2.$$

Consider two cases.

(a) $2s_j + 1 \geqslant \ln(N)/\ln(n_j)$. In this case we can follow Tsybakov (2009), section 3.2, to obtain

$$\frac{\sigma^2}{N}\sum_{k=1}^{(n_j-1)/2}(2k)^{s_j}(1-(2k)^{s_j}\kappa_j)_+ = \frac{\sigma^2}{N}\sum_{k=1}^{k_j}(2k)^{s_j}\{1-(2k)^{s_j}\kappa_j\}, \tag{27}$$

where $k_j = \lfloor \frac{1}{2}\kappa_j^{-1/s_j}\rfloor$, and, neglecting the constants, $\kappa_j^2 = N^{-2s_j/(2s_j+1)}$ and $E\|\hat c_j^{\mathrm{L}} - c_j\|_{n_j}^2 = O(N^{-2s_j/(2s_j+1)})$.

The condition $2s_j + 1 \geqslant \ln(N)/\ln(n_j)$ is necessary to ensure that the resulting $k_j = \frac{1}{2}N^{1/(2s_j+1)} \leqslant (n_j-1)/2$ in equation (27).

(b) $2s_j + 1 < \ln(N)/\ln(n_j)$. In this case one can easily see that

$$\frac{\sigma^2}{N}\sum_{k=1}^{(n_j-1)/2}(2k)^{s_j}(1-(2k)^{s_j}\kappa_j)_+ = \frac{\sigma^2}{N}\sum_{k=1}^{(n_j-1)/2}(2k)^{s_j}\{1-(2k)^{s_j}\kappa_j\},$$

$\kappa_j^2 = n_j/N$ and $E\|\hat c_j^{\mathrm{L}} - c_j\|_{n_j}^2 = O(n_j/N)$.

A.4. Proof of theorem 2

No estimator $\tilde{f}$ of f in model (3) can obviously perform better than that of an oracle that knows the true subsets $\mathcal{J}_0$ and $\mathcal{J}_0^c$ of zero and non-zero components f_j of f. In this ideal case, one would certainly set $\hat{f}_j = 0$ for all $j \in \mathcal{J}_0$ with no error and, therefore, because of the additivity of the AMSE, proposition 2 yields

$$\inf_{\tilde{f}} \sup_{f_j \in \mathcal{F}_{n_j},\,(S_j R_j) j \in \mathcal{J}_0^c} \mathrm{AMSE}(\tilde{f}, f) = \sum_{j \in \mathcal{J}_0^c} \inf_{\tilde{f}_j} \sup_{f_j \in \mathcal{F}_{n_j}(s_j,\,R_j)} \mathrm{AMSE}(\tilde{f}_j, f_j)$$

$$\geqslant C_2 \sum_{j \in \mathcal{J}_0^c} \min\left(N^{-2s_j/(2s_j+1)}, \frac{n_j}{N} \right)$$

(see proposition 4.16 of Johnstone (2013)).

Furthermore, since $\min(N^{-2s_j/(2s_j+1)}, n_j/N) > N^{-1}$, $j \in \mathcal{J}_0^c$, for $d_0 > d/2$ we have

$$\frac{d_0 \ln(d/d_0)}{N} \leqslant \ln(2) \frac{d_0}{N} \leqslant \ln(2) \sum_{j \in \mathcal{J}_0^c} \min\left(N^{-2s_j/(2s_j+1)}, \frac{n_j}{N} \right)$$

and the first term on the right-hand side of equation (16) is dominating. Thus, to complete the proof we need to show that, for $d_0 \leqslant d/2$,

$$\inf_{\tilde{f}} \sup_{f_j \in \mathcal{F}_{n_j}(s_j,\,R_j),\, j \in \mathcal{J}_0^c} \mathrm{AMSE}(\tilde{f}, f) = \inf_{\tilde{c}} \sup_{c_j \in \Theta_{n_j}(s_j,\,R_j) j \in \mathcal{J}_0^c} \|\tilde{c} - c\|_{N_0}^2 \geqslant C_2 \frac{d_0 \ln(d/d_0)}{N}, \tag{28}$$

where $N_0 = \Sigma_{j=1}^d n_j$ and c is an N_0-dimensional amalgam of d n_j-dimensional vectors of discrete Fourier coefficients c_j of f_j.

The proof is based on finding a subset $\mathcal{C}_{d_0}$ of N_0-dimensional amalgamated vectors c with d_0 non-zero components $c_j \in \Theta_{n_j}(s_j, R_j)$ such that, for any pair $c^1, c^2 \in \mathcal{C}_{d_0}$ and some constant $C > 0$,

$$\|c^1 - c^2\|_{N_0}^2 \geqslant C \frac{\sigma^2}{N} d_0 \ln\left(\frac{d}{d_0} \right)$$

and the Kullback–Leibler divergence

$$K(\mathbb{P}_{c^1}, \mathbb{P}_{c^2}) = \frac{\|c^1 - c^2\|_{N_0}^2}{2\sigma^2/N} \leqslant \frac{1}{16} \ln\{\mathrm{card}(\mathcal{C}_{d_0})\}.$$

The required result in expression (28) then follows immediately from lemma A.1 of Bunea *et al*. (2007).

Define the subset $\tilde{\mathcal{V}}_{d_0}$ of all d-dimensional indicator vectors with d_0 entries of 1s: $\tilde{\mathcal{D}}_{d_0} = \{v : v \in \{0,1\}^d, \|v\|_0 = d_0\}$. Lemma A.3 of Rigollet and Tsybakov (2011) implies that, for $d_0 \leqslant d/2$, there is a subset $\mathcal{V}_{d_0} \subset \tilde{\mathcal{V}}_{d_0}$ such that, for some constant $C_0 > 0$, $\ln\{\mathrm{card}(\mathcal{V}_{d_0})\} \geqslant C_0 d_0 \ln(d/d_0)$ and, for any pair $v_1, v_2 \in \mathcal{V}_{d_0}$, the Hamming distance $\rho(v_1, v_2) = \Sigma_{j=1}^d \mathbb{I}(v_{1j} \neq v_{2j}) \geqslant C_0 d_0$.

To any indicator vector $v \in \mathcal{V}_{d_0}$ assign the corresponding vector $c \in \mathcal{C}_{d_0}$ as follows. Let

$$\tilde{C}^2 = \frac{1}{16} C_0 \frac{\sigma^2}{N} \ln\left(\frac{d}{d_0} \right).$$

Define c_j to be a zero vector if $v_j = 0$ and to have two non-zero entries $c_{-1j} = c_{1j} = \tilde{C}/\sqrt{2}$ otherwise. Evidently, $c_j \in \mathcal{F}(s_j, \tilde{C}) \subset \mathcal{F}_{n_j}(s_j, R_j)$ for $v_j = 1$ and $\mathrm{card}(\mathcal{C}_{d_0}) = \mathrm{card}(\mathcal{V}_{d_0})$.

For any pair $c^1, c^2 \in \mathcal{C}_{d_0}$ and the corresponding $v_1, v_2 \in \mathcal{V}_{d_0}$, we then have

$$\|c^1 - c^2\|_{N_0}^2 = \tilde{C}^2 \sum_{j=1}^d \mathbb{I}(v_{1j} \neq v_{2j}) \geqslant \tilde{C}^2 C_0 d_0 = \frac{1}{16} \frac{\sigma^2}{N} C_0^2 d_0 \ln\left(\frac{d}{d_0} \right),$$

$$K(\mathbb{P}_{c^1}, \mathbb{P}_{c^2}) = \frac{\tilde{C}^2}{2\sigma^2/N} \sum_{j=1}^d \mathbb{I}(v_{1j} \neq v_{2j}) \leqslant \frac{\tilde{C}^2 d_0}{\sigma^2/N} \leqslant \frac{1}{16} \ln\{\mathrm{card}(\mathcal{C}_{d_0})\},$$

which completes the proof.

References

Abramovich, F. and Grinshtein, V. (2010) MAP model selection in Gaussian regression. *Electron. J. Statist.*, **4**, 932–949.

Abramovich, F. and Grinshtein, V. (2013) Estimation of a sparse group of sparse vectors. *Biometrika*, **100**, 335–370.

Abramovich, F., Grinshtein, V. and Pensky, M. (2007) On optimality of Bayesian testimation in the normal means problem. *Ann. Statist.*, **35**, 2261–2286.
Abramovich, F., Grinshtein, V., Petsa, A. and Sapatinas, T. (2010) On Bayesian testimation and its application to wavelet thresholding. *Biometrika*, **97**, 181–198.
Birgé, L. and Massart, P. (2001) Gaussian model selection. *J. Eur. Math. Soc.*, **3**, 203–268.
Bunea, F., Tsybakov, A. and Wegkamp, M. H. (2007) Aggregation for Gaussian regression. *Ann. Statist.*, **35**, 1674–1697.
Donoho, D. L. and Johnstone, I. M. (1994) Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.
Friedman, J. and Stuetzle, W. (1981) Projection pursuit regression. *J. Am. Statist. Ass.*, **76**, 817–823.
Guedj, B. and Alquier, P. (2013) PAC-Bayesian estimation and prediction in sparse additive models. *Electron. J. Statist.*, **7**, 264–291.
Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*. London: Chapman and Hall.
Johnstone, I. M. (2013) Gaussian estimation: sequence and multiresolution models. Stanford University, Stanford. (Available from `http://statweb.stanford.edu/imj/GE06-11-13.pdf`.)
Koltchinskii, V. and Yuan, M. (2010) Sparsity in multiple kernel learning. *Ann. Statist.*, **38**, 3660–3695.
Korostelev, A. and Korosteleva, O. (2011) *Mathematical Statistics: Asymptotic Minimax Theory*. Providence: American Mathematical Society.
Lin, Y. and Zhang, H. H. (2006) Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.*, **34**, 2272–2297.
Meier, L., van de Geer, S. and Buhlmann, P. (2009) High-dimensional additive modelling. *Ann. Statist.*, **37**, 3779–3821.
Raskutti, G., Wainwright, M. J. and Yu, B. (2011) Minimax rates of estimations for high-dimensional regression over l_q balls. *IEEE Trans. Inform. Theor.*, **57**, 6976–6994.
Raskutti, G., Wainwright, M. J. and Yu, B. (2012) Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *J. Mach. Learn. Res.*, **13**, 389–427.
Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2009) Sparse additive models. *J. R. Statist. Soc.* B, **71**, 1009–1030.
Rigollet, P. and Tsybakov, A. (2011) Exponential screening and optimal rates of sparse estimation. *Ann. Statist.*, **39**, 731–771.
Suzuki, T. (2012) PAC-Bayesian bound for Gaussian process regression and multiple kernel additive model. *J. Mach. Learn. Res. Wrkshp Conf. Proc.*, **23**, 8.1–8.20.
Suzuki, T. and Sugiyama, M. (2013) Fast learning rate of multiple kernel learning: trade-off between sparsity and smoothness. *Ann. Statist.*, **41**, 1381–1405.
Tsybakov, A. (2009) *Introduction to Nonparametric Estimation*. New York: Springer.
Wahba, G. (1990) *Spline Models for Observational Data*. Philadelphia: Society for Industrial and Applied Mathematics.
Yuan, M. and Lin, Y. (2006) Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc.* B, **68**, 49–67.