VARIABLE SELECTION IN HIGH-DIMENSIONAL ADDITIVE MODELS BASED ON NORMS OF PROJECTIONS

MARTIN WAHL

Abstract. We consider the problem of variable selection in high-dimensional sparse additive models. The proposed method is motivated by geometric considerations in Hilbert spaces, and consists in comparing the norms of the projections of the data on various additive subspaces. Our main results are concentration inequalities which lead to conditions making variable selection possible. In special cases these conditions are known to be optimal. As an application we consider the problem of estimating single components. We show that, up to first order, one can estimate a single component as well as if the other components were known.

1. Introduction

In this paper, we consider the two related problems of variable selection and component estimation in high-dimensional additive random regression models when the number of covariates can be much larger than the number of observations. We study these models under the assumption that most components are equal to zero.

High-dimensional linear models have been investigated intensively in the literature. A great deal of attention has been given to the Lasso (see, e.g., the book by Bühlmann and van de Geer [3] and the references therein). The Lasso is based on $l_1$-penalization, and can be used for both estimation and variable selection. There is also a huge literature on estimation and variable selection via $l_0$-penalization. These procedures can be found, e.g., in the book by Massart [19], where a general approach to model selection via penalization is developed (see also the work by Barron, Birgé, and Massart [2] and the references therein).

More recently, high-dimensional additive models have been studied, e.g., in the work by Meier, van de Geer, and Bühlmann [20],

2010 Mathematics Subject Classification. 62G05, 62G08, 94A12.

Key words and phrases. Variable selection, sparse additive models, model selection, geometry of Hilbert spaces, compressive sensing, structured random matrices, dimension reduction, component estimation.
Huang, Horowitz, and Wei [13], Koltchinskii and Yuan [17], Raskutti, Wainwright, and Yu [21], Gayraud and Ingster [12], and Dalalyan, Ingster, and Tsybakov [10]. One approach generalizes the (group) Lasso, and combines sparsity penalties with smoothness penalties or constraints (see [20, 13, 17, 21]). As in the case of the Lasso, these procedures can be used for both estimation and variable selection (see [20, 13]). Another approach focuses on the problem of estimation, and is based on exponentially weighted aggregation (see [10] and the references therein). In fact, Dalalyan, Ingster, and Tsybakov [10] considered a more general model which they called the compound model and which includes the additive model as a special case. In a Gaussian white noise setting, they showed that their estimator achieves non-asymptotic minimax rates of convergence.

Comminges and Dalalyan [9] considered the problem of variable selection in high-dimensional Gaussian white noise models, and established tight conditions which make the estimation of the relevant variables possible. Similar results were obtained earlier by Wainwright [26] for high-dimensional linear models with Gaussian measurement matrices. Comminges and Dalalyan [9] also extended their method to a high-dimensional random regression model, but they assumed that the joint density of all covariates is known.

Several results in the theory of high-dimensional statistical inference are initiated by achievements in the theory of compressive sensing (see, e.g., the introductory book chapters by Fornasier and Rauhut [11] and Rauhut [22] and the references therein). A popular method is the $l_1$-minimization which enables sparse recovery if the measurement matrix satisfies, for instance, a restricted isometry property (RIP). It is known that several random matrices satisfy the RIP with probability close to one, important examples being the Gaussian random matrices and the so-called structured random matrices (see, e.g., the work by Candès and Tao [8], Baraniuk, Davenport, DeVore, and Walkin [1], and Rauhut [22]). These results were generalized to high-dimensional linear models by Candès and Tao [7] (see also the work by Bickel, Ritov, and Tsybakov [4] and the book by Koltchinskii [16] Chapters 7 and 8).

In this paper, we study a method for variable selection which consists in comparing the norms of the projections of the data on various additive subspaces. Given a positive integer $q^*$ which has to be an upper bound for the number of nonzero components, the procedure selects a subset of cardinality smaller than or equal to $q^*$ which contains the non-zero components with probability close to one. The basis of this procedure is a selection criterion in the population setting which works well under the essential assumption that the minimal angles between
various disjoint additive subspaces are bounded away from zero. Applying this assumption and recent results in random matrix theory (see, e.g., [22]), we obtain that, with probability close to one, a generalized restricted isometry property holds in the finite sample setting. This property enables us to carry over the geometry from the population to the finite sample setting, and thus leads to an analysis of our procedure. Our results are of theoretical interest. We prove upper bounds for the probability that our procedure misses relevant variables. These concentration inequalities lead to sufficient conditions making the estimation of the relevant variables possible. In special cases, these conditions are known to be optimal. Moreover, as an application of our variable selection procedure, we consider the problem of estimating single components in a high-dimensional additive random regression model. We find conditions under which a single component can be estimated with the same non-asymptotic optimal rate of convergence as in the case where the other components are known.

The paper is organized as follows. In Section 2, we present the main assumption and discuss a variable selection criterion in the population setting. From this, we derive our variable selection criterion in the finite sample setting and state our main results in Theorems 1 and 2. The proofs of Theorems 1 and 2 are contained in Section 3, the main part being the analysis of the finite sample geometry. We also discuss a restricted isometry property which shows up if the covariates are independent. Section 4 is devoted to the application of our variable selection procedure to the problem of estimating single components. Finally, some technical parts of the proofs are given in the Appendix.

2. Main result

2.1. The variable selection problem. Let \((Y, X)\) be a pair of random variables such that \(X = (X_1, \ldots, X_q)^T\) and

\[
Y = \sum_{j=1}^{q} f_j(X_j) + \epsilon, \tag{2.1}
\]

where the \(X_j\) are real-valued random variables, the \(f_j\) are unknown functions which are contained in \(L^2(\mathbb{P}^{X_j})\), and \(\epsilon\) is a Gaussian random variable with expectation 0 and variance \(\sigma^2\) which is independent of \(X\). Moreover, we suppose that \(f_j\) satisfies \(\mathbb{E}[f_j(X_j)] = 0\) for \(j = 1, \ldots, q - 1\). We denote by \(f\) the whole regression function given by \(f(x) = \sum_{j=1}^{q} f_j(x_j)\). We assume that we observe \(n\) independent copies
The number of covariates $q$ can be much larger than the number of observations $n$, but we assume that most of the components are equal to 0. Thus we consider a high-dimensional sparse additive model. If we define $J_0 = \{j \in \{1, \ldots, q\} | f_j \text{ is non-zero}\}$, then we have $f(x) = \sum_{j \in J_0} f_j(x_j)$. The set $J_0$ is supposed to be unknown, but we assume that we are given an integer $q^*$ such that $|J_0| \leq q^*$. We aim at selecting a subset of cardinality smaller than or equal to $q^*$ which contains $J_0$.

2.2. The main assumption. Without any further assumption, the components are not necessarily uniquely determined. In this subsection we give an assumption which implies uniqueness and furthermore makes the variable selection task accessible. We define $H_{q^*} = L^2(\mathbb{P}^{X_q})$ and

$$H_j = \{h_j \in L^2(\mathbb{P}^{X_j}) | \mathbb{E}[h_j(X_j)] = 0\}$$

for $j = 1, \ldots, q - 1$. Note that $f_j \in H_j$. The spaces $H_j$ are all canonically contained in $L^2(\mathbb{P}^X)$ which is a Hilbert space with the inner product $\langle g, h \rangle = \mathbb{E}[g(X)h(X)]$ and the corresponding norm $\|g\| = \sqrt{\langle g, g \rangle}$. Moreover, for $J \subseteq \{1, \ldots, q\}$ we define

$$H_J = \sum_{j \in J} H_j$$

(with the convention that $H_J = 0$ if $J = \emptyset$).

**Assumption 1.** There exists a constant $0 \leq \rho < 1$ such that for all subsets $J_1, J_2 \subseteq \{1, \ldots, q\}$ satisfying $J_1 \cap J_2 = \emptyset$ and $|J_1|, |J_2| \leq q^*$, we have

$$\langle h_{J_1}, h_{J_2} \rangle \leq \rho \|h_{J_1}\| \|h_{J_2}\|$$

for all $h_{J_1} \in H_{J_1}, h_{J_2} \in H_{J_2}$.

It follows from the fact that the spaces $H_J$ are closed combined with Assumption 1 and [15, Theorem 1a] (applied inductively) that all spaces $H_J$ with $J \subseteq \{1, \ldots, q\}$ and $|J| \leq 2q^*$ are closed. The real number

$$\rho_0(H_{J_1}, H_{J_2}) = \sup \left\{ \frac{\langle h_{J_1}, h_{J_2} \rangle}{\|h_{J_1}\| \|h_{J_2}\|} \right\}$$

$0 \neq h_{J_1} \in H_{J_1}, 0 \neq h_{J_2} \in H_{J_2}$

is the cosine of the minimal angle between $H_{J_1}$ and $H_{J_2}$ (see, e.g., [14, Definition 1]). Letting $\rho_{q^*} = \max \rho_0(H_{J_1}, H_{J_2})$, where the maximum is taken over all subsets $J_1, J_2 \subseteq \{1, \ldots, q\}$ satisfying $J_1 \cap J_2 = \emptyset$.
and $|J_1|, |J_2| \leq q^*$, then Assumption \(\Pi\) says that $\rho_{q^*} < 1$. By a simple argument which is given in Appendix A, one can show that Assumption \(\Pi\) can be written as follows:

**Remark 1** (Equivalent form of Assumption \(\Pi\)). For all subsets $J_1, J_2 \subseteq \{1, \ldots, q\}$ satisfying $J_1 \cap J_2 = \emptyset$ and $|J_1|, |J_2| \leq q^*$, we have

$$
\|h_{J_1} + h_{J_2}\|^2 \geq (1 - \rho_{q^*}^2) \|h_{J_1}\|^2
$$

(2.4)

for all $h_{J_1} \in H_{J_1}$, $h_{J_2} \in H_{J_2}$.

Remark \(\Pi\) shows that Assumption \(\Pi\) is essential for variable selection: if (2.4) does not hold, then it is possible that $f$ is arbitrary close to a sparse additive function which is based on a completely different set of variables. From (2.4) and the definition of $J_0$, we obtain:

**Lemma 1.** Let Assumption \(\Pi\) be satisfied. Then

$$
\kappa := \min_{\emptyset \neq J \subseteq J_0} \left\| \sum_{j \in J} f_j \right\|^2 > 0.
$$

For $J \subseteq \{1, \ldots, q\}$ let $\Pi_{H_J}$ be the orthogonal projection from $L^2(\mathbb{P}^X)$ to $H_J$. In the following we abbreviate $\Pi_{H_J}$ as $\Pi_J$. Since projections lower the norm, the set $J_0$ maximizes the quantity $\|\Pi_J f\|^2$. If Assumption \(\Pi\) holds, the following Lemma shows that $\|\Pi_{J_0} f\|^2 - \|\Pi_J f\|^2$ is strictly positive for all subsets $J \subseteq \{1, \ldots, q\}$ with $|J| \leq q^*$ and $J_0 \setminus J \neq \emptyset$. This means that a subset $J \subseteq \{1, \ldots, q\}$ with $|J| \leq q^*$ which maximizes $\|\Pi_J f\|^2$ always contains $J_0$ (and is equal to $J_0$ in the special case when $|J_0| = q^*$). These observations will be the theoretical basis for our selection criterion in the finite sample setting.

**Proposition 1.** Let Assumption \(\Pi\) be satisfied. Let $J \subseteq \{1, \ldots, q\}$ be a subset such that $|J| \leq q^*$ and $J_0 \setminus J \neq \emptyset$. Then

$$
\|\Pi_{J_0} f\|^2 - \|\Pi_J f\|^2 \geq (1 - \rho_{q^*}^2) \kappa_l,
$$

where $l = |J_0 \setminus J|$ and

$$
\kappa_l := \min_{J' \subseteq J_0 \setminus J} \left\| \sum_{j \in J'} f_j \right\|^2.
$$

**Proof.** We have $f = \sum_{j \in J_0 \cap J} f_j + \sum_{j \in J_0 \setminus J} f_j =: f_{J_0 \cap J} + f_{J_0 \setminus J}$. Hence

$$
\Pi_J f = f_{J_0 \cap J} + \Pi_J f_{J_0 \setminus J}
$$

and

$$
\Pi_{J_0} f = f = f_{J_0 \cap J} + f_{J_0 \setminus J}
$$

$$
= (f_{J_0 \cap J} + \Pi_J f_{J_0 \setminus J}) + (f_{J_0 \setminus J} - \Pi_J f_{J_0 \setminus J}) \in H_J + H_J^\perp
$$
From this we conclude that
\[ \| \Pi_{J_0} f \|_2^2 = \| \Pi_J f \|_2^2 + \| f_{J_0 \setminus J} - \Pi_J f_{J_0 \setminus J} \|_2^2 \]
We have \( f_{J_0 \setminus J} \in H_{J_0 \setminus J}, \) \( \Pi_J f_{J_0 \setminus J} \in H_J, \) and \( l = |J_0 \setminus J| \geq 1. \) Thus (2.4) and the definition of \( \kappa_l \) yield
\[ \| f_{J_0 \setminus J} - \Pi_J f_{J_0 \setminus J} \|_2^2 \geq (1 - \rho_{q^*}^2) \| f_{J_0 \setminus J} \|_2^2 \geq (1 - \rho_{q^*}^2) \kappa_l. \]
This completes the proof. \( \square \)

Finally, we show that \( \rho_{q^*} \) can be related to a quantity which is known in the literature on sparse additive models (see, e.g., [17]).

**Lemma 2.** Let \( \epsilon_{2q^*} \) be the smallest numbers such that
\[ \left\| \sum_{j \in J} f_j \right\|_2^2 \geq (1 - \epsilon_{2q^*}) \left( \sum_{j \in J} |f_j| \right)^2 \] (2.5)
for all \( J \subseteq \{1, \ldots, q\} \) with \( |J| \leq 2q^* \) and all \( \sum_{j \in J} f_j \in H_J. \) Then we have \( \rho_{q^*} < 1 \) if and only if \( \epsilon_{2q^*} < 1. \)

A proof of this lemma is given in Appendix [B].

2.3. The selection criterion. Now we construct the selection criterion. For \( j = 1, \ldots, q \) let \( V_j \subseteq H_j \) be finite-dimensional linear subspaces. For \( J \subseteq \{1, \ldots, q\} \) we define
\[ V_J = \sum_{j \in J} V_j \]
and \( d_J = \dim V_J. \) In order to proceed, we need some further notation. Let \( \| \cdot \|_n \) be the empirical norm which is defined by
\[ \| h \|_n^2 = \frac{1}{n} \sum_{i=1}^n h^2(X_i) \]
for \( h \in L^2(\mathbb{P}^X), \) and which is defined by \( \| \cdot \|_n^2 = (1/n) \| \cdot \|_2^2 \) if applied to vectors in \( \mathbb{R}^n. \) (\( \| \cdot \|_2 \) denotes the usual Euclidean norm). Moreover, let \( \hat{\Pi}_J \) be the orthogonal projection from \( \mathbb{R}^n \) to the subspace \( \{ (g_J(X^1), \ldots, g_J(X^n))^T | g_J \in V_J \} \). If \( h \in L^2(\mathbb{P}^X), \) then we abbreviate \( \hat{\Pi}_J(h(X^1), \ldots, h(X^n))^T \) as \( \hat{\Pi}_J h. \) Finally, let \( Y = (Y^1, \ldots, Y^n)^T \) and \( \epsilon = (\epsilon^1, \ldots, \epsilon^n)^T. \) Motivated by Proposition [B] we define an estimator \( \hat{J}_0 \) of \( J_0 \) as follows:
\[ \hat{J}_0 = \arg \max_{J \subseteq \{1, \ldots, q\}, |J| \leq q^*} \| \hat{\Pi}_J Y \|_n^2 - \sigma^2 d_J/n. \] (2.6)
Conditioning on \( X^1, \ldots, X^n, \) the random variable \( (n/\sigma^2)\| \hat{\Pi}_J \epsilon \|_n^2 \) has a chi-square distribution with rank(\( \hat{\Pi}_J \)) \( \leq d_J \) degrees of freedom and the
last term is supposed to cancel its expectation. The last term can also
be seen as a penalty term. In fact, the criterion in (2.6) can be written
as a penalized least squares criterion (see, e.g., [19]).

The success of the criterion depends on a suitable choice of the \( V_j \)
which in turn depends on the regularity assumptions on the \( f_j \). For
instance, if the \( f_j \) belong to some known finite-dimensional linear sub-
spaces of \( H_j \), then we let the \( V_j \) be equal to these spaces. In the
following we consider the nonparametric case. Without loss of gener-
ality we shall restrict our attention to Hölder smoothness and spaces
of piecewise polynomials. A similar treatment is possible, e.g., for (pe-
riodic) Sobolev smoothness and spaces of trigonometric polynomials,
with slightly modified results.

**Assumption 2.** For \( j = 1, \ldots, q \) suppose that \( X_j \) takes values in \([0, 1]\),
and has a density \( p_j \) with respect to the Lebesgue measure on \([0, 1]\).
Suppose that these densities satisfy \( c \leq p_j \leq 1/c \) for some constant \( c > 0 \).

Moreover, for \( j = 1, \ldots, q \) suppose that \( f_j \) belongs to the Hölder class
\( \mathcal{H}_j(\alpha_j, K_j) \) on \([0, 1]\), where \( \alpha_j, K_j \) are positive real numbers.

For \( j = 1, \ldots, q \) we let \( V_j \) be the intersection of \( H_j \) with the space of
regular piecewise polynomials on \([0, 1]\) with integer-valued parameters
\( r_j = \lfloor \alpha_j \rfloor \) and \( m_j \), where \( r_j \) is the maximal degree of the polynomials
and \( \{ 0 < 1/m_j < 2/m_j < \cdots < 1 \} \) generates the partition of \([0, 1]\)
into \( m_j \) intervals (see [5]). This implies that the \( V_j \) and thus also
the procedure in (2.6) depend on the conditions \( \mathbb{E}[g_j(X_j)] = 0 \) for
\( j = 1, \ldots, q - 1 \). In Remark 5 we show how one can eliminate this
dependence. We have \( d_j = (r_j + 1)m_j - 1 \) for \( j = 1, \ldots, q - 1 \) and
\( d_q = (r_q + 1)m_q \). Moreover, the \( V_j \) have good approximation properties
with respect to Hölder classes. In fact one can show that there exists a
constant \( c_{\alpha_j} \) depending only on \( \alpha_j \) such that
\[ \| f_j - \Pi_{V_j} f_j \| \leq c_{\alpha_j} K_j d_j^{-\alpha_j} \]
for all \( h_j \in \mathcal{H}_j(\alpha_j, K_j) \cap H_j \), where \( \Pi_{V_j} \) is the orthogonal projection
from \( L^2(\mathbb{P}^X) \) to \( V_j \). The next lemma shows that up to a constant this
bound also holds when the \( L^2(\mathbb{P}^X) \)-norm is replaced by the \( L^\infty \)-norm.

**Lemma 3.** Let Assumption 2 be satisfied. Then there exists a constant \( c_j \)
depending only on \( l_j \) and \( c \) (given explicitly in the proof) such that
\[ \| f_j - \Pi_{V_j} f_j \|_\infty \leq c_j K_j m_j^{-\alpha_j} \]
for all \( f_j \in \mathcal{H}_j(\alpha_j, K_j) \cap H_j \).

For completeness, a proof of this lemma is given in Appendix C. In
order to state our main result, we define the events
\[ \mathcal{E}_{\delta, J} = \{ (1 - \delta)\| g_J \|_2^2 \leq \| g_J \|_n^2 \leq (1 + \delta)\| g_J \|_2^2 \text{ for all } g_J \in V_J \} \quad (2.7) \]
for $J \subseteq \{1, \ldots, q\}$ and $0 < \delta < 1$, and

$$E_{\delta,q^*} = \bigcap_{J \subseteq \{1, \ldots, q\}, |J| \leq q^*} \mathcal{E}_{\delta,J \cup J_0}.$$  \hfill (2.8)

Now we suppose that $m_j$ satisfies the lower bound

$$m_j \geq \left( \frac{c_j K_j q^* (\|f\| \vee 1)}{c'(1 - \rho_2^q) (1 \wedge \kappa)} \right)^{1/\alpha_j} \hfill (2.9)$$

for $j = 1, \ldots, q$, where $c'$ is a small constant satisfying (3.3).

**Theorem 1.** Let Assumptions 1 and 2 be satisfied. Let $0 < \delta < 1$. Suppose that (2.9) is satisfied for $j = 1, \ldots, q$ and assume that $d_{q^*} = \max_{|J| \leq q^*} d_{J \cup J_0} \leq n$. Then

$$\mathbb{P} \left( J_0 \subseteq \hat{J}_0 \right) \geq 1 - \mathbb{P} \left( E_{\delta,q^*} \right)$$

$$- \sum_{l=1}^{|J_0|} \sum_{m=0}^{q^* - (|J_0| - l)} \binom{|J_0|}{l} (q - |J_0|) \frac{4 \exp \left( - \frac{c_1}{\kappa^2} \sigma^2 (1 - \rho_2^q)^2 \kappa_l \right)}{\sigma^2 (1 - \rho_2^q)^2 \kappa_l} \hfill (2.10)$$

where $c_1 = (1 - \delta)^4 / (2^{11} (1 + \delta)^3)$.

In Remarks 2 and 3, we discuss conditions under which the right-hand side is close to one. In special cases these conditions are also known to be optimal. But first of all, we want to state a more concrete version of Theorem 1 in the general case. For $J \subseteq \{1, \ldots, q\}$ let

$$\varphi_J = 1 \sqrt{d_J} \sup_{0 \neq g \in V_J} \frac{\|g\|_\infty}{\|g\|}.$$ 

Under Assumptions 1 and 2 and if $J$ satisfies $|J| \leq q^*$, we have that $\varphi_{J \cup J_0}$ satisfies

$$\varphi_{J, J_0}^2 \leq \frac{2(r + 1)}{c(1 - \epsilon_{2q^*})}, \hfill (2.11)$$

where $r = \max_{j=1,\ldots,q} r_j$. A proof of (2.11) is given in Appendix D.

**Theorem 2.** Let Assumptions 1 and 2 be satisfied. Let $0 < \delta < 1$. Suppose that (2.9) is satisfied for $j = 1, \ldots, q$ and assume that $d_{q^*} = \max_{|J| \leq q^*} d_{J \cup J_0} \leq n$. Moreover, let $\varphi_{q^*} = \max_{|J| \leq q^*} \varphi_{J \cup J_0}$. Then there
exists a constant $c_2 > 0$ such that

$$\mathbb{P}(J_0 \subseteq \hat{J}_0) \geq 1 - \left(\frac{eq}{q^*}\right)^{q^*} 2^{3/4} d_{q^*} \exp\left(-c_2 \frac{n\delta^2}{\varphi_{q^*}^2 d_{q^*}}\right)$$

$$- \sum_{l=1}^{\lfloor |J_0| / (|J_0| - l) \rfloor} \sum_{m=0}^{l} \left(\frac{|J_0|}{l}\right) \left(\frac{q - |J_0|}{m}\right) 4 \exp\left(-\frac{c_1}{\sigma^2 \|f\|^2 \vee \sigma^2 \vee (1 - \rho_{q^*}^2)\kappa_l}\right),$$

(2.12)

where $c_1 = (1 - \delta)^4 / (2^{11} (1 + \delta)^3)$.

Remark 2. If $q^* = |J_0|$, then we have $J_0 \subseteq \hat{J}_0$ if and only if $J_0 = \hat{J}_0$. Thus, in this case we can write (2.12) as follows:

$$\mathbb{P}(J_0 \neq \hat{J}_0) \leq \left(\frac{eq}{q^*}\right)^{q^*} 2^{3/4} d_{q^*} \exp\left(-c_2 \frac{n\delta^2}{\varphi_{q^*}^2 d_{q^*}}\right)$$

$$+ \sum_{l=1}^{q^*} \sum_{m=0}^{l} \left(\frac{q^*}{l}\right) \left(\frac{q - q^*}{m}\right) 4 \exp\left(-\frac{c_1}{\sigma^2 \|f\|^2 \vee \sigma^2 \vee (1 - \rho_{q^*}^2)\kappa_l}\right),$$

(2.13)

If all quantities except $n$ and $q$ are bounded from above (resp. form below) by a constant independent of $n$, then one can see that there are constants $c_3, c_4 > 0$ such that $\mathbb{P}(J_0 \neq \hat{J}_0) \leq \exp(-c_4 n)$, provided

$$\log q \leq c_3 n.$$

Note that this condition is known to be optimal (see, e.g., [22], [9]).

Remark 3. We continue the discussion of Remark 2. We want to see which conditions on $n, q, q^*$ are sufficient such that the right-hand side of (2.13) is exponentially small. We again suppose that $q^* = |J_0|$. To simplify the exposition, we suppose that $1/(1 - \rho_{q^*}^2), 1/(1 - \epsilon_{2q^*}), 1/\kappa,$ and $\max_j \|f_j\|$ are all bounded by a constant independent of $n$.

First, we consider the linear model $Y = \sum_{j=1}^{q} X_j \beta_j + \epsilon$, where the $\beta_j$ are real numbers (and we have $\mathbb{E}[X_j] = 0$ for $j = 1, \ldots, q - 1$). In Theorems 1 and 2 the linear model corresponds to the case $c_j = 0, m_j = 0,$ and $r_j = 1$. Moreover, analogous results also hold for unbounded covariates, see below. By (3.4), the second term on the right-hand side of (2.13) can be bounded by

$$\sum_{l=1}^{q^*} \left(\frac{eq^*}{l}\right) \left(\frac{e(q - q^*)}{l}\right) 4 \exp\left(-\frac{c_1}{\sigma^2 \|f\|^2 \vee \sigma^2 \vee (1 - \rho_{q^*}^2)\kappa_l}\right).$$

(2.14)
Applying the bounds $\kappa_l, \|f\|^2 \leq (1 + \epsilon' q^*) q^* \max_j \|f_j\|^2$ (see (B.1), note that always $1 + \epsilon' q^* \leq q^*$ and that $\epsilon' q^* = 0$ if the covariates are independent) and $\kappa_l \geq (1 - \epsilon q^*) l \kappa_1$, one can show that there are constants $c_3, c_4 > 0$ such that (2.14) is smaller than $\exp(-c_4 n/(1 + \epsilon' q^*) q^*)$, provided

$$(1 + \epsilon' q^*) q^* \log(q) \leq c_3 n.$$ 

There are also constants $c_3, c_4 > 0$ such that the first term on the right-hand side of (2.13) is smaller than $\exp(-c_4 n/q^*)$, provided

$q^* d^* \log(q/q^*) \leq c_3 n.$

This gives a stronger condition which in special cases can be weakened by using other concentration inequalities for $\mathbb{P} (E_{\delta, q^*})$. For instance, in Subsection 3.2, we also discuss the model $Y = \sum_{j=1}^q X_j \beta_j + \epsilon$, where the $X_j$ are independent Gaussian random variables and the $\beta_j$ are real numbers. We show that there are constants $c_3, c_4 > 0$ such that $\mathbb{P} (E_{\delta, q^*}) \leq \exp(-c_4 n)$, provided

$q^* \log(q/q^*) \leq c_3 n.$

Thus in this particular example, we obtain conditions which are (up to a small logarithmic change) also known to be necessary (see, e.g., [22, Section 2.6] for the setting without noise and [27, Theorem 2] for the noisy setting).

Second, we consider the nonparametric case. The second term on the right-hand side of (2.13) does not change. Therefore it remains to consider the first term. Again, it is easy to see that there are constants $c_3, c_4 > 0$ such that the first term is smaller than $\exp(-c_4 n/d q^*)$, provided $q^* d q^* \log(q/q^*) \leq c_3 n$. The latter condition can be simplified to

$q^* (2\alpha + 1)/\alpha \log(q/q^*) \leq c_3 n,$

where $\alpha = \min_j \alpha_j$ and $c_3 > 0$ is some constant. Again, in special cases of independent covariates, a factor 2 can be removed from the exponent of $q^*$ (see (3.6)), which leads to a condition which is (up to some additional logarithmic terms) also necessary for consistent estimation of the regression function $f$ (see [21] and [10]).

Remark 4. The proposed selection procedure has good theoretical properties, as discussed in the previous remarks. However, the procedure has practical drawbacks. First, it depends on $q^*$ which might be not known in practice. Second, except for a few circumstances, the selection procedure is computationally expensive: finding the maximum of (2.6) by looking at all subsets of $\{1, \ldots, q\}$ of cardinality less than or
equal to \( q^* \) has large complexity, since there are at least \((q/q^*)q^*\) such subsets.

**Remark 5.** If \( j' \in \{1, \ldots, q\} \) is a fixed element, then one can modify the procedure in (2.6) such that the argmax is over all subsets \( J \subseteq \{1, \ldots, q\} \) satisfying \(|J| \leq q^* + 1\) and \( j' \in J \). Theorem 2 remains valid if Assumption 1 is satisfied with \( q^* \) replaced by \( q^* + 1 \). Moreover, in the special case \( j' = q \), the modified procedure does no longer depend on the conditions \( \mathbb{E}[g_j(X_j)] = 0 \) for \( j = 1, \ldots, q - 1 \), since the spaces \( V_J \) with \( q \in J \) contain all constant functions.

### 3. Outline of the proof of Theorem 2

#### 3.1. The finite sample geometry.

In this subsection we present empirical versions of Assumption 1 and Proposition 1. Throughout this section \( 0 < \delta < 1 \) is considered as fixed. Written in the equivalent form of Remark 1, we have:

**Lemma 4.** Let Assumption 1 be satisfied. Let \( J_1, J_2 \subseteq \{1, \ldots, q\} \) be two subsets such that \( J_1 \cap J_2 = \emptyset \) and \(|J_1|, |J_2| \leq q^*\). If \( \mathcal{E}_{\delta,J_1 \cup J_2} \) holds, then we have

\[
\|g_{J_1} + g_{J_2}\|_n^2 \geq \frac{(1 - \delta)}{(1 + \delta)} (1 - \rho^2_{q^*}) \|g_{J_1}\|_n^2
\]  

for all \( g_{J_1} \in V_{J_1}, g_{J_2} \in V_{J_2} \).

**Proof.** Under the assumptions of Lemma 4 we have

\[
\|g_{J_1} + g_{J_2}\|_n^2 \geq (1 - \delta)\|g_{J_1} + g_{J_2}\|^2 \\
\geq (1 - \delta)(1 - \rho^2_{q^*})\|g_{J_1}\|^2 \\
\geq \frac{(1 - \delta)}{(1 + \delta)} (1 - \rho^2_{q^*}) \|g_{J_1}\|_n^2.
\]

This completes the proof. \( \square \)

Applying (3.1) as in the proof of Proposition 1 we obtain:

**Proposition 2.** Let Assumption 1 be satisfied. Let \( J \subseteq \{1, \ldots, q\} \) be a subset such that \(|J| \leq q^* \) and \( J_0 \setminus J \neq \emptyset \). Let \( l = |J_0 \setminus J| \). Let \( v = \sum_{j \in J_0} v_j \) with \( v_j \in V_j \) for \( j \in J_0 \). If \( \mathcal{E}_{\delta,J \cup J_0} \) holds, then we have

\[
\left\| \hat{\Pi}_{J_0} v \right\|_n^2 - \left\| \hat{\Pi}_J v \right\|_n^2 \geq \frac{(1 - \delta)}{(1 + \delta)} (1 - \rho^2_{q^*}) \min_{J' \subseteq J_0, |J'| = l} \left\| \sum_{j \in J'} v_j \right\|_n^2.
\]

By appropriately decomposing \( f \) as \( v + f - v \) with \( v \in V_{J_0} \), we can apply Proposition 2 to \( v \) and (2.9) and Lemma 3 and to \( f - v \). The result is the following empirical version of Proposition 1.
Proposition 3. Let Assumption $\mathcal{A}$ and Assumption $\mathcal{B}$ be satisfied. Let $J \subseteq \{1, \ldots, q\}$ be a subset such that $|J| \leq q^*$ and $J_0 \setminus J \neq \emptyset$. Let $l = |J_0 \setminus J|$. If $E_{\delta,J_0}$ holds, then we have

$$
\| \hat{\Pi}_{J_0} f \|^2_n - \| \hat{\Pi}_J f \|^2_n \geq \frac{1}{2} \left( 1 - \delta \right)^2 (1 - \rho_{q^*}) \kappa_l,
$$

provided that

$$
(1 - c')^2 - \frac{(1 + \delta)^2}{(1 - \delta)^2} (4c' + 5c'^2) \geq 1/2.
$$

A proof of Proposition 3 is given in Appendix E. In the absence of noise, Propositions 2 and 3 already prove Theorem 1. In fact, if the event $E_{\delta,J}$ holds, then (2.6) selects a subset $\hat{J}_0 \subseteq \{1, \ldots, q\}$ with $|\hat{J}_0| \leq q^*$ and $J_0 \subseteq \hat{J}_0$. Proposition 3 applies to the general nonparametric setting, while Proposition 2 applies if the components $f_j$ belong to some known finite-dimensional linear subspaces of $H_j$, the latter being a commonly used setting in the theory of compressive sensing (see, e.g., [11] and the references therein). Finally, we derive a concentration inequality for the event $E_{\delta,q^*}$.

Theorem 3. Let Assumptions $\mathcal{A}$ and $\mathcal{B}$ be satisfied. Let $J \subseteq \{1, \ldots, q\}$ be a subset such that $|J| \leq q^*$. Then we have

$$
\mathbb{P}(E_{\delta,J}^c) \leq 2^{3/4} d_{J \cup J_0} \exp \left( -c_2 \frac{n \delta^2}{\varphi_{q^*}^2 d_{J \cup J_0}} \right),
$$

where $c_2 > 0$ is a universal constant.

Theorem 3 is a consequence of [22, Theorem 7.3], the details are given in Appendix E. A similar result can be obtained by Talagrand’s inequality combined with Rudelson’s lemma (see [23, Theorem 1] and [24, Theorem 3.1]). Applying Theorem 3, the union bound, and the following combinatorial result

$$
\sum_{j=0}^{q^*} \binom{q}{j} \leq \left( \frac{eq}{q^*} \right)^{q^*}
$$

(for a proof see, e.g., [19, Proposition 2.5]), we obtain the following concentration inequality

$$
\mathbb{P}(E_{\delta,q^*}^c) \leq \left( \frac{eq}{q^*} \right)^{q^*} 2^{3/4} d_{q^*} \exp \left( -c_2 \frac{n \delta^2}{\varphi_{q^*}^2 d_{q^*}} \right),
$$

where $\varphi_{q^*}$ and $d_{q^*}$ are given in Theorem 2.
3.2. Independent covariates and the RIP. In this subsection, we suppose that the covariates are independent which implies that the spaces $V_1, \ldots, V_q$ are orthogonal in $L^2(\mathbb{P}^X)$. In this particular case we rewrite the event $\mathcal{E}_{\delta,q^*}$ as a restricted isometry property for block-matrices. This allows us to apply concentration inequalities leading to improvements and extensions of (3.5).

For $j = 1, \ldots, q$ let $\phi_{j,1}, \ldots, \phi_{j,d_j}$ be an orthonormal basis of $V_j$. Then we define the $n \times d_j$-matrix

$$A_j = \frac{1}{\sqrt{n}} \left( \phi_{j,k}(X^j_\ell) \right)_{1 \leq i \leq n, 1 \leq k \leq d_j}$$

and for $J \subseteq \{1, \ldots, q\}$ we define the $n \times d_J$-matrix $A_J = (A_j)_{j \in J}$ (we abbreviate $A_{\{1,\ldots,q\}}$ as $A$). With these definitions, it is easy to see that $\mathcal{E}_{\delta,J}$ is the event such that

$$(1 - \delta) \|z_J\|_2^2 \leq \|A_J z_J\|_2^2 \leq (1 + \delta) \|z_J\|_2^2$$

for all $z_J \in \mathbb{R}^{d_J}$, where we have used that the spaces $V_1, \ldots, V_q$ are orthogonal. Thus, if we define

$$\delta_{q^*} = \max_{J \subseteq \{1,\ldots,q\}, |J| \leq q^*} \|A_J^T A_{J \cup J_0} - I\|_{op},$$

then we have

$$\mathcal{E}_{\delta,q^*} = \{\delta_{q^*} \leq \delta\}.$$

The constant $\delta_{q^*}$ is bounded by the restricted isometry constant of order $d_{q^*}$ of the matrix $A$ (see [22, Definition 2.4]). Note that the restricted isometry constant plays a prominent role in the theory of sparse recovery, and that there exist many concentration inequalities for the restricted isometry constant in many ensembles of random matrices. For instance, in the model $Y = \sum_{j=1}^q X_j \beta_j + \epsilon$, where the $X_j$ are independent Gaussian random variables and the $\beta_j$ are real numbers, $A$ is a Gaussian random matrix (the entries are independent Gaussian random variables, each with expectation zero and variance $1/n$), and [1] Theorem 5.2] implies that there exist constants $c_3, c_4 > 0$ depending only on $\delta$ such that $\mathbb{P}(\delta_{q^*} \leq \delta) \geq 1 - 2 \exp(c_4 n)$, provided that $q^* \log(q^*/q) \leq c_3 n$. Finally, in the nonparametric case, one can also apply [22, Theorem 8.4] which does not lead to improvements for spaces of piecewise polynomials, but leads to improvements for spaces which have an orthonormal basis satisfying a boundedness condition.

We briefly discuss these improvements in the case where $X_1, \ldots, X_n$ are independent and uniformly distributed on $[0, 1]$ and the $f_j$ belong to periodic Sobolev classes with parameters $\alpha_j$ and $K_j$ (see [25, Chapter 1]). In this case, we replace the spaces of piecewise polynomials by spaces of trigonometric polynomials. We return to the discussion of
Remark 3. Applying [22, Theorem 8.4], we can replace the condition $q^*d_q^* \log(q/q^*) \leq c_3n$ in Remark 3 by the condition

$$d_q^* \log^4 \left( n \vee \sum_{j=1}^{q} d_j \right) \leq c_3n.$$ 

Moreover, using the independence and Bernstein’s inequality, one can improve the bounds in (E.3) leading to a weaker lower bound in (2.9) with respect to $q^*$. To summarize, we end up with the condition

$$q^{*(2\alpha+1)/(2\alpha)} \log^4 \left( n \vee \sum_{j=1}^{q} d_j \right) \leq c_3n,$$

(3.6)

where $\alpha = \min_j \alpha_j$. The condition that $q^{*(2\alpha+1)/(2\alpha)}/n$ is small is also necessary for estimation, since the term $q^*/n^{2\alpha/(2\alpha+1)}$ is part of the minimax rate of convergence (see [10]).

3.3. End of the proofs of Theorems 1 and 2. We have

$$\mathbb{P} \left( J_0 \notin \hat{J}_0 \right) = \mathbb{P} \left( J_0 \setminus \hat{J}_0 \neq \emptyset \right)$$

$$= \mathbb{P} \left( \exists J \subseteq \{1, \ldots, q\}, |J| \leq q^* \text{ with } J_0 \setminus J \neq \emptyset \right.$$

and $\|\hat{\Pi}_J Y\|_n^2 - d_J/n > \|\hat{\Pi}_{J_0} Y\|_n^2 - d_{J_0}/n \right).$

Applying the union bound, we obtain

$$\mathbb{P} \left( J_0 \notin \hat{J}_0 \right) \leq \mathbb{P} \left( E_{\delta,J_0}^c \right)$$

$$+ \sum_{J \subseteq \{1, \ldots, q\}, |J| \leq q^*, d_J \setminus J \neq \emptyset} \mathbb{P} \left( E_{\delta,J \cup J_0} \cap \|\hat{\Pi}_J Y\|_n^2 - d_J/n > \|\hat{\Pi}_{J_0} Y\|_n^2 - d_{J_0}/n \right).$$

We have:

Lemma 5. Let Assumptions [2] and [2] be satisfied. Let $J \subseteq \{1, \ldots, q\}$ be a subset such that $|J| \leq q^*$ and $J_0 \setminus J \neq \emptyset$. Let $l = |J_0 \setminus J|$. Then

$$\mathbb{P} \left( E_{\delta,J \cup J_0} \cap \|\hat{\Pi}_J Y\|_n^2 - \sigma^2 d_J/n > \|\hat{\Pi}_{J_0} Y\|_n^2 - \sigma^2 d_{J_0}/n \right)$$

$$\leq 4 \exp \left( -c_1 \frac{n(1 - \rho_{q^*}^2)^2 \kappa_l^2}{\sigma^2 \|f\|^2 \vee \sigma^2 \vee (1 - \rho_{q^*}^2) \kappa_l} \right),$$

where $c_1 = (1 - \delta)^4/(2^{11} (1 + \delta)^3).$
A proof of Lemma is given in Appendix G. We conclude that

\[ P \left( J_0 \not\subseteq \hat{J}_0 \right) \leq P \left( E^{c}_{\delta,q^*} \right) + \sum_{l=1}^{|J_0|} \sum_{m=0}^{q^* - |J_0| - l} \left( \frac{|J_0|}{l} \right) \left( q - |J_0| \right) 4 \exp \left( - \frac{c_1}{2} \frac{n(1 - \rho_{q^*}^2)^2 \kappa_l^2}{\sigma^2 \| f \|^2 \vee \sigma^2 \vee (1 - \rho_{q^*}^2) \kappa_l} \right), \]

which gives Theorem 1. Applying (3.5), we obtain (2.12). This completes the proof. □

4. Estimation of single components

4.1. The dimension reduction step. The proposed variable selection method can be seen as a method to reduce the dimension of the model. We start with \( n \) independent observations of a sparse additive model with \( q \) covariates and an unknown subset \( J_0 \subseteq \{1, \ldots, q\} \) of non-zero components, and we end up with a subset \( \hat{J}_0 \subseteq \{1, \ldots, q\} \) such that \( |\hat{J}_0| \leq q^* \) and \( J_0 \subseteq \hat{J}_0 \) with probability close to one. More precisely, if \( \{J_0 \subseteq \hat{J}_0\} \) holds, we have successfully reduced the model (2.1) to

\[ Y = \sum_{j \in J_0} f_j(X_j) + \epsilon. \] (4.1)

4.2. The estimation method. In this subsection we consider the problem of estimating a single component \( f_j \) of the model (2.1) with \( j \in J_0 \). We may assume without loss of generality that \( j = 1 \). To proceed we split the sample into two parts. We use the first part to perform a variable selection step and we use the selected variables and the second part to construct an estimator of \( f_1 \). More precisely, we assume that we observe an even number of independent copies \((Y_1, X_1), \ldots, (Y_{2n}, X_{2n})\) of \((Y, X)\). The estimator \( \hat{J}_0 \) of \( J_0 \) is constructed as in Subsection 2.3 using the sample \((Y^1, X^1), \ldots, (Y^n, X^n)\). The estimator \( \hat{f}_1 \) of \( f_1 \) is constructed using \( \hat{J}_0 \) and the sample \((Y^{n+1}, X^{n+1}), \ldots, (Y_{2n}, X_{2n})\).

For \( j = 1, \ldots, q \) let \( V'_j \) be the intersection of \( H_j \) with the space of regular piecewise polynomials on \( [0, 1] \) with integer-valued parameters \( r_j = \lfloor \alpha_j \rfloor \) and \( m'_j \). Let \( d'_j = \dim V'_j \). Moreover, for \( J \subseteq \{1, \ldots, q\} \) let

\[ V'_j = \sum_{j \in J} V'_j. \]
and \( d'_f = \text{dim} \, V'_f \). Then the least squares estimator on the (random) model \( V'_{J_0} \) based on the second sample is given (not uniquely) by

\[
\hat{f} = \arg \min_{g \in V'_{J_0}} \frac{1}{n} \sum_{i=n+1}^{2n} (Y^i - g(X^i))^2.
\]

By Assumption 1, \( V'_1 \) and \( V'_{J_0 \setminus \{1\}} \) have intersection equal to 0. Therefore, we have \( \hat{f} = \check{f}_1 + \check{f}_{-1} \) with \( \check{f}_1 \in V'_1 \) and \( \check{f}_{-1} \in V'_{J_0 \setminus \{1\}} \) uniquely determined. We now define

\[
\tilde{f}_1 = \check{f}_1 \text{ if } \|\check{f}_1\|_\infty \leq k_n \text{ and } \tilde{f}_1 = 0 \text{ otherwise}, \quad (4.2)
\]

where \( k_n \) is a real number to be chosen appropriately.

**Theorem 4.** Let the assumptions of Theorem 1 be satisfied. Let

\[
\varphi_{q^*} = \max_{|J| \leq q^*} \sup_{0 \neq g \in V'_J} \frac{\|g\|_\infty}{\sqrt{d'_q \|g\|}}
\]

and \( d'_{q^*} = \max_{|J| \leq q^*} d'_J \). Then there is a constant \( c_2 > 0 \) such that

\[
\mathbb{E} \left[ \|f_1 - \tilde{f}_1\|^2 \right] \leq \frac{(1 + \delta)^2}{(1 - \delta)^2} \frac{1}{1 - \rho^2_{q^*}} \left( 2c_2^2 K_1^2 d'_1 - 2\alpha_1 + \frac{\sigma^2 d'_1}{n} \right)
\]

\[
+ \frac{2(1 + \delta)^2 (1 + \epsilon'_{q^*})}{(1 - \delta)^2} \frac{1}{1 - \rho^2_{q^*}} \sum_{j \in J_0 \setminus \{1\}} c_2^2 K_j^2 d'_j
\]

\[
+ R_n + (\|f_1\| + k_n)^2 \mathbb{P} \left( J_0 \not\subseteq \hat{J}_0 \right)
\]

with

\[
R_n = \frac{2c_2^2 d'_{q^*} \|f_1\|^2 (\|f\|^2 + \sigma^2)}{(1 - \delta)(1 - \rho^2_{q^*}) k_n^2} + (\|f_1\| + k_n)^2 2^{3/4} d'_{q^*} \exp \left( -c_2 \frac{n \delta^2}{c_2^2 k_n^2 d'_{q^*}} \right).
\]

**Remark 6.** Theorem 4 shows that the estimator \( f_1 \) can attain the non-asymptotic optimal rate of convergence. One simple setting is the following: All involved quantities except \( n \) and \( q \) are bounded from above (resp. from below) by a constant independent of \( n \) (see Remark 2), and we have \( \alpha_1 > 0 \) and \( \alpha = \min_j \alpha_j > \alpha_1/(2\alpha_1 + 1) \).

**Proof.** Theorem 4 is a consequence of [26, Theorem 1]. We have

\[
\mathbb{E} \left[ \|f_1 - \tilde{f}_1\|^2 \right] \leq \mathbb{E} \left[ 1_{\{J_0 \subseteq J_0\}} \|f_1 - \tilde{f}_1\|^2 \right] + (\|f_1\| + k_n)^2 \mathbb{P} \left( J_0 \not\subseteq \hat{J}_0 \right).
\]

The first term considers the case described in (4.1). Moreover, by conditioning on the first sample, we may assume that \( \hat{J}_0 \) is fixed. In order to
apply [26, Theorem 1], we need that [26, Assumptions 1 and 2] are satisfied, i.e. that $H_1$ and $H_{V_{J_0 \setminus \{1\}}}$ are closed and that $\rho_0(H_1, H_{V_{J_0 \setminus \{1\}}}) < 1$, which is satisfied by Assumption 1.

\section*{Appendix A. Proof of Remark 1}

Suppose that (2.3) holds, and let $h_{J_1} \in H_{J_1}$ and $h_{J_2} \in H_{J_2}$. Then we have $\|h_{J_1} + h_{J_2}\|^2 \geq \|h_{J_1}\|^2 - 2\rho_q^\ast \|h_{J_1}\|\|h_{J_2}\| + \|h_{J_2}\|^2$ and (2.4) follows from the inequality $2\rho_q^\ast \|h_{J_1}\|\|h_{J_2}\| \leq \rho_q^2 \|h_{J_1}\|^2 + \|h_{J_2}\|^2$.

Conversely, suppose that (2.4) holds, and let $h_{J_1} \in H_{J_1}$ and $h_{J_2} \in H_{J_2}$. We may assume without loss of generality that $h_{J_2} \neq 0$ and that $\|h_{J_2}\| = 1$. Then $\|h_{J_1}\|^2 - \langle h_{J_1}, h_{J_2} \rangle^2 = \|h_{J_1} - \langle h_{J_1}, h_{J_2} \rangle h_{J_2} \|^2 \geq (1 - \rho_q^2) \|h_{J_1}\|^2$ which gives (2.3). This completes the proof.

\section*{Appendix B. Proof of Lemma 2}

Let $\epsilon_k'$ be the smallest numbers such that

$$\left\| \sum_{j \in J} f_j \right\|^2 \leq (1 + \epsilon_k') \left( \sum_{j \in J} \|f_j\|^2 \right) \quad (B.1)$$

for all $J \subseteq \{1, \ldots, q\}$ with $|J| \leq k$ and all $f_j = \sum_{j \in J} f_j \in H_J$. Note that we always have $1 + \epsilon_k' \leq k$. Let $J_1, J_2 \subseteq \{1, \ldots, q\}$ be two subsets satisfying $J_1 \cap J_2 = \emptyset$ and $|J_1|, |J_2| \leq q^\ast$. Applying (2.5) and (B.1), we see that

$$\|f_{J_1} + f_{J_2}\|^2 \geq \frac{1 - \epsilon_{2q^\ast}}{1 + \epsilon_{q^\ast}} \left( \|f_{J_1}\|^2 + \|f_{J_2}\|^2 \right)$$

for all $f_{J_1} \in H_{J_1}$, $f_{J_2} \in H_{J_2}$. Thus Remark 1 gives the “if” part. Conversely, applying (2.3) iteratively, one gets for instance

$$1 - \epsilon_{2q^\ast} \geq (1 - \rho_q^2)^{\log_2 q^\ast + 1}$$

which gives the “only if” part.

\section*{Appendix C. Proof of Lemma 3}

We may assume without loss of generality that $j = 1$. Let $U_1 \subseteq L^2(\mathbb{P}^{X_1})$ be the subspace of piecewise polynomials with parameters $r_1$ and $m_1$ (thus $V_1 = \{g_1 \in U_1 | \mathbb{E}[g_1(X_1)] = 0\}$). First, we show that

$$\|\Pi_{U_1} f_1\|_\infty \leq \frac{(r_1 + 1)^2}{c} \|f_1\|_\infty. \quad (C.1)$$

The space $U_1$ is the orthogonal sum of $m_1$ $(r_1 + 1)$-dimensional subspaces $U_1^{(k)}$, where $g_1^{(k)} \in U_1^{(k)}$ is a polynomial of degree equal to or smaller than $r_1$ if restricted to the interval $[k/m_1, (k + 1)/m_1]$ (resp. $[1 - 1/m_1, 1]$ if $k = m_1 - 1$) and is zero elsewhere. Since $\Pi_{U_1} = \sum_{k=1}^{m_1} \Pi_{U_1^{(k)}}$, we have $\|\Pi_{U_1} f_1\|_\infty \leq \max_{k=1}^{m_1} \|\Pi_{U_1^{(k)}} f_1\|_\infty$. Applying (C.1) to each $U_1^{(k)}$, we obtain

$$\|\Pi_{U_1^{(k)}} f_1\|_\infty \leq \frac{(r_1 + 1)^2}{c} \|f_1\|_\infty.$$
\[ \sum_{k=0}^{m-1} \Pi_{U_1^{(k)}} \text{ and } \| \Pi_{U_1} f_1 \|_\infty = \max_{0 \leq k \leq m-1} \| \Pi_{U_1^{(k)}} f_1 \|_\infty, \text{ it suffices to show that} \]
\[ \| \Pi_{U_1^{(k)}} f_1 \|_\infty \leq \frac{(r_1 + 1)^2}{c} \sup_{x_1 \in [k/m_1, (k+1)/m_1]} |f_1(x_1)| \]  
(C.2)

(resp. \( x_1 \in [1 - 1/m_1, 1] \) if \( k = m_1 - 1 \)). For this let \( \phi_1^{(k)}, \ldots, \phi_{r_1+1}^{(k)} \) be an orthonormal basis of \( U_1^{(k)} \). In Appendix D, it is shown that
\[ \| \phi_l^{(k)} \|_\infty \leq \frac{(r_1 + 1)\sqrt{m_1}}{\sqrt{c}} \]  
(C.3)

Moreover, by the Cauchy-Schwarz inequality and Assumption 2, we have
\[ \int_0^1 \phi_l^{(k)}(x_1)^t f_1(x_1)p_1(x_1)dx_1 \]
\[ = \int_0^1 \phi_l^{(k)}(x_1)1_{[k/m_1,(k+1)/m_1]}(x_1)f_1(x_1)p_1(x_1)dx_1 \]
\[ \leq \frac{\sup_{x_1 \in [k/m_1, (k+1)/m_1]} |f_1(x_1)|}{\sqrt{cm_1}} \]  
(C.4)

for \( l = 1, \ldots, r_1 + 1 \). Since
\[ \Pi_{U_1^{(k)}} f_1 = \sum_{l=1}^{r_1+1} \left( \int_0^1 \phi_l^{(k)}(x_1)f_1(x_1)p_1(x_1)dx_1 \right) \phi_l^{(k)}, \]
(C.2) follows from (C.3) and (C.4). This completes the proof of (C.1).

By Taylor’s theorem and the definition of the Hölder class, we can find an element \( g_1 \in U_1 \) such that
\[ \| f_1 - g_1 \|_\infty \leq \frac{K_1 m_1^{-\alpha_1}}{r_1!}. \]

Applying this and the fact that \( \Pi_{U_1} g_1 = g_1 \), we obtain
\[ \| f_1 - \Pi_{U_1} f_1 \|_\infty \leq \| f_1 - g_1 \|_\infty + \| \Pi_{U_1} (g_1 - f_1) \|_\infty \]
\[ \leq \left( 1 + \frac{(r_1 + 1)^2}{c} \right) \frac{K_1 m_1^{-\alpha_1}}{r_1!} \]

for all \( f_1 \in \mathcal{H}_1(\alpha_1, K_1) \). Finally, we have \( V_1 + \mathbb{R} = U_1 \). Moreover, \( V_1 \) and \( \mathbb{R} \) are orthogonal. This implies that \( \Pi_{U_1} f_1 = \Pi_{V_1} f_1 + \mathbb{E}[f_1(X_1)] \) for all \( f_1 \in \mathcal{H}_1(\alpha_1, K_1) \). In particular, if \( f_1 \in \mathcal{H}_1(\alpha_1, K_1) \cap H_1 \), then we have \( \Pi_{U_1} f_1 = \Pi_{V_1} f_1 \). This completes the proof. \( \square \)
Appendix D. Proof of (2.11)

In [5] it is shown that
\[ \|g_j\|_\infty^2 \leq (r_j + 1)^2m_j \int_0^1 g_j^2(x_j)dx_j \]
for all \( g_j \in V_j \). This implies that
\[ \|g_j\|_\infty^2 \leq \varphi_j^2 \| v\| \]
with \( \varphi_j^2 = 2(r_j + 1)/c \). Now let \( J \subseteq \{1, \ldots, q\} \) be a subset with \( |J| \leq q^* \). Applying (D.1), the Cauchy-Schwarz inequality, and Lemma 2, we obtain
\[ \|g_{J \cup J_0}\|_\infty \leq \sum_{j \in J \cup J_0} \|g_j\|_\infty \leq \max_j \varphi_j \sqrt{\sum_{j \in J \cup J_0} d_j} \sqrt{\sum_{j \in J \cup J_0} \|g_j\|^2} \leq \frac{\max_j \varphi_j}{\sqrt{1 - \epsilon_{2q^*}}} \sqrt{d_{J \cup J_0}} \|g_{J \cup J_0}\| \]
for all \( g_{J \cup J_0} = \sum_{j \in J \cup J_0} g_j \in V_{J \cup J_0} \). This completes the proof. \( \square \)

Appendix E. Proof of Proposition 3

Let \( v = \sum_{j \in J_0} v_j \) with \( v_j = \Pi_{V_j} f_j \) for \( j \in J_0 \). Then
\[ \|\hat{\Pi}_{J_0} f \|_n^2 - \|\hat{\Pi} f \|_n^2 \]
\[ = \|\hat{\Pi}_{J_0} v \|_n^2 + 2\langle \hat{\Pi}_{J_0} v, \hat{\Pi}_{J_0} (f - v) \rangle_n + \|\hat{\Pi}_{J_0} (f - v) \|_n^2 \]
\[ - \|\hat{\Pi} v \|_n^2 - 2\langle \hat{\Pi} v, \hat{\Pi} (f - v) \rangle_n - \|\hat{\Pi} (f - v) \|_n^2 \]
\[ \geq \|\hat{\Pi}_{J_0} v \|_n^2 - \|\hat{\Pi} v \|_n^2 - 4\|v\|_n \|f - v\|_n - \|f - v\|_n^2, \quad (E.1) \]
where the inequality holds since projections lower the norm. By (2.9) and Lemma 3, we have
\[ \|f_j - v_j\|_n \leq \frac{c'(1 - \rho_{q^*}^2)(1 \land \kappa)}{q^*(1 \lor \|f\|)} \quad \text{and} \quad \|f_j - v_j\| \leq \frac{c'(1 - \rho_{q^*}^2)(1 \land \kappa)}{q^*(1 \lor \|f\|)} \quad (E.2) \]
and thus by the triangle inequality, we have
\[ \|f - v\|_n \leq \frac{c'(1 - \rho_{q^*}^2)(1 \land \kappa)}{1 \lor \|f\|} \quad \text{and} \quad \|f - v\| \leq \frac{c'(1 - \rho_{q^*}^2)(1 \land \kappa)}{1 \lor \|f\|}. \quad (E.3) \]
For \( J' \subseteq J_0 \) with \( |J'| = l \) let \( f_{J'} = \sum_{j \in J'} f_j \) and \( v_{J'} = \sum_{j \in J'} v_j \). By (E.2), we have \( \|f_{J'} - v_{J'}\| \leq c'(1 \land \kappa) \leq c'\sqrt{\kappa} \leq c'\sqrt{\kappa_l} \). If \( \mathcal{E}_{\delta, J_0} \) holds,
then we obtain
\[ \|v_n\|^2 \geq (1 - \delta)\|v_J\|^2 \]
\[ \geq (1 - \delta) (\|f_J\| - \|f_J - v_J\|)^2 \]
\[ \geq (1 - \delta)(1 - c')^2 \kappa_l. \]

Thus, if \( E_{\delta,J,J_0} \) holds, then Proposition 2 yields
\[ \|\hat{\Pi}_{J_0}v\|^2 - \|\hat{\Pi}_Jv\|^2 \geq \frac{(1 - \delta)^2}{(1 + \delta)} (1 - \rho_{q^*}^2)(1 - c')^2 \kappa_l. \]  

(E.4)

On the other hand, if \( E_{\delta,J_0} \) holds, then we have
\[ 4\|v\|_n \|f - v\|_n \|f - v\|_n \]
\[ \leq 4\sqrt{1 + \delta}\|v\|_n \|f - v\|_n \|f - v\|_n \]
\[ \leq 4\sqrt{1 + \delta}\|f\|_n \|f - v\|_n \|f - v\|_n + 4\sqrt{1 + \delta}\|f - v\|_n \|f - v\|_n \]
which by (E.3) is bounded by
\[ \sqrt{1 + \delta}(1 - \rho_{q^*}^2)\kappa (4c' + 5c^2) \leq (1 + \delta)(1 - \rho_{q^*}^2)\kappa (4c' + 5c^2). \]  

(E.5)

Combining (E.1), (E.4), and (E.5), we obtain (3.2). This completes the proof. \( \square \)

Appendix F. Proof of Theorem 3

Let \( \phi_1, \ldots, \phi_{d_{J\cup J_0}} \) be an orthonormal basis of \( V_{J\cup J_0} \). By [6, Lemma 1], we have
\[ \left\| \sum_{j=1}^{d_{J\cup J_0}} \phi_j \right\|_\infty^2 = \varphi^2_{d_{J\cup J_0}} d_{J\cup J_0}. \]  

(F.1)

Now let
\[ B_n = (\langle \phi_j, \phi_k \rangle_n)_{1 \leq j, k \leq d_{J\cup J_0}}. \]

Then [22, Theorem 7.3] (in the case \( s = N = d_{J\cup J_0} \)) yields
\[ \mathbb{P}(\|B_n - I\|_{op} \leq \delta) \geq 1 - 2^{3/4}d_{J\cup J_0} \exp \left(-c_2\frac{n\delta^2}{\varphi_{d_{J\cup J_0}}^2 d_{J\cup J_0}} \right) \]  

(F.2)

for \( 0 < \delta < 1 \) and a universal constant \( c_2 > 0 \). Note that we can apply [22, Theorem 7.3], since in the proof the condition [22, (4.2)] is only used in the form [22, (7.5)], which is satisfied by (F.1). A similar result follows from [24, Theorem 3.1]. A function \( g \in V_{J\cup J_0} \) with \( \|g\| = 1 \) can
be written uniquely as $g = \sum_{j=1}^{d_{J\cup J^0}} x_j \phi_j$ with $x \in \mathbb{R}^{d_{J\cup J^0}}$ and $\|x\|_2 = 1$. Then we have $\|g\|_n^2 = x^T B_n x$ and thus
\[
\sup_{g \in V_{J\cup J^0}, \|g\| = 1} ||g||_n^2 - ||g||^2 = \sup_{x \in \mathbb{R}^{d_{J\cup J^0}}, \|x\|_2 = 1} |x^T (B_n - I) x| = \|B_n - I\|_{op},
\]
where the latter equality follows from the spectral theorem. (F.2) and (F.3) yield that $\|g\|_n^2 - ||g||^2 \leq \delta \|g\|^2$ for all $g \in V_{J\cup J^0}$ with probability greater or equal to $1 - 2^{-\frac{3}{4}} d_{J\cup J^0} \exp\left(-cn^2/(\phi_d^2 n^2)\right)$. Applying the triangle inequality completes the proof.

### Appendix G. Proof of Lemma 5

We have
\[
\|\hat{\Pi}_J Y\|_n^2 - \sigma^2 d_J/n > \|\hat{\Pi}_{J^0} Y\|_n^2 - \sigma^2 d_{J^0}/n
\]
if and only if
\[
\|\hat{\Pi}_J f\|_n^2 + 2\langle \hat{\Pi}_J f, \epsilon \rangle_n + \|\hat{\Pi}_J \epsilon\|_n^2 - \sigma^2 d_J/n
\]
\[
> \|\hat{\Pi}_{J^0} f\|_n^2 + 2\langle \hat{\Pi}_{J^0} f, \epsilon \rangle_n + \|\hat{\Pi}_{J^0} \epsilon\|_n^2 - \sigma^2 d_{J^0}/n.
\]
The random variables $\epsilon^1, \ldots, \epsilon^n$ are independent and Gaussian, each with expectation 0 and variance $\sigma^2$. Moreover, they are independent of $X^1, \ldots, X^n$. Thus conditioned on $X^1, \ldots, X^n$ and if $E_{\delta,J\cup J^0}$ holds, then we have $\|\hat{\Pi}_J \epsilon\|_n^2 \sim (\sigma^2/n) \chi^2(d_J)$ and $\|\hat{\Pi}_{J^0} \epsilon\|_n^2 \sim (\sigma^2/n) \chi^2(d_{J^0})$, where $\chi^2(d_J)$ and $\chi^2(d_{J^0})$ denote chi-square distributions with $d_J$ and $d_{J^0}$ degrees of freedom, respectively. By (3.2), if $E_{\delta,J\cup J^0}$ holds, then we have
\[
\|\hat{\Pi}_{J^0} f\|_n^2 - \|\hat{\Pi}_J f\|_n^2 \geq \frac{1}{2} \left(1 - \frac{\delta}{1 + \delta}\right) \left(1 - \rho^2_n\right) \kappa_l.
\]
We conclude that
\[
P\left(E_{\delta,J\cup J^0} \cap \|\hat{\Pi}_J Y\|_n^2 - \sigma^2 d_J/n > \|\hat{\Pi}_{J^0} Y\|_n^2 - \sigma^2 d_{J^0}/n\right)
\leq P\left(\frac{\sigma^2}{n} \left(\chi^2(d_J) - d_J\right) \geq \frac{1}{8} \left(1 - \frac{\delta}{1 + \delta}\right) \left(1 - \rho^2_n\right) \kappa_l\right)
\]  
\[
+ P\left(\frac{\sigma^2}{n} \left(\chi^2(d_{J^0}) - d_{J^0}\right) \leq -\frac{1}{8} \left(1 - \frac{\delta}{1 + \delta}\right) \left(1 - \rho^2_n\right) \kappa_l\right)
\]  
\[
+ P\left(E_{\delta,J\cup J^0} \cap 2\langle \hat{\Pi}_J f, \epsilon \rangle_n - 2\langle \hat{\Pi}_{J^0} f, \epsilon \rangle_n \geq \frac{1}{4} \left(1 - \frac{\delta}{1 + \delta}\right) \left(1 - \rho^2_n\right) \kappa_l\right).
\]
The first and the second term can be bounded by standard concentration inequalities for chi-square distributions.
Lemma 6. Let $d$ be a positive integer. Then, for all $x \geq 0$, we have

$$\mathbb{P}(\chi^2(d) - d \geq x) \leq \exp \left( -\frac{x^2}{2(2d + 2x)} \right)$$

(G.1)

and

$$\mathbb{P}(\chi^2(d) - d \leq -x) \leq \exp \left( -\frac{x^2}{4d} \right).$$

(G.2)

For a proof of this lemma see [18, Lemma 1] and [6, Lemma 8]. Using the bounds $d_J \leq n$ and $d_{J_0} \leq n$, we obtain

$$\mathbb{P} \left( \frac{\sigma^2}{n} (\chi^2(d_J) - d_J) \geq \frac{1}{16} \frac{(1 - \delta)^2}{(1 + \delta)} (1 - \rho_{q^*}^2) \kappa_l \right)
+ \mathbb{P} \left( \frac{\sigma^2}{n} (\chi^2(d_{J_0}) - d_{J_0}) \leq -\frac{1}{16} \frac{(1 - \delta)^2}{(1 + \delta)} (1 - \rho_{q^*}^2) \kappa_l \right)
\leq 2 \exp \left( -\frac{1}{29\sigma^2 (1 + \delta)^2 \sigma^2 \vee (1 - \rho_{q^*}^2) \kappa_l} \right).$$

(G.3)

Thus it remains the third term. It can be bounded by

$$\mathbb{P} \left( \mathcal{E}_{\delta,J\cup J_0} \cap \langle \hat{\Pi}_J f, \epsilon \rangle_n \geq \frac{1}{16} \frac{(1 - \delta)^2}{(1 + \delta)} (1 - \rho_{q^*}^2) \kappa_l \right)
+ \mathbb{P} \left( \mathcal{E}_{\delta,J\cup J_0} \cap -\langle \hat{\Pi}_{J_0} f, \epsilon \rangle_n \geq \frac{1}{16} \frac{(1 - \delta)^2}{(1 + \delta)} (1 - \rho_{q^*}^2) \kappa_l \right).$$

(G.4)

If $\mathcal{E}_{\delta,J\cup J_0}$ holds, then we have

$$\|\hat{\Pi}_{J_0} f\|_n \leq \|f\|_n
\leq \|v\|_n + \|f - v\|_n
\leq \sqrt{1 + \delta} \|v\| + \|f - v\|_n
\leq \sqrt{1 + \delta} (\|f\| + \|f - v\| + \|f - v\|_n)
$$

We have $\kappa \leq \|f\|^2$. Thus, by (E.3) and since $c' < 1/2$, we have $\|f - v\| \leq \|f\|/2$ and $\|f - v\|_n \leq \|f\|/2$. If $\mathcal{E}_{\delta,J\cup J_0}$ holds, then we conclude that

$$\|\hat{\Pi}_{J_0} f\|_n^2 \leq 4(1 + \delta) \|f\|^2$$

Applying a standard concentration inequality for Gaussian random variables, we obtain

$$\mathbb{P} \left( \mathcal{E}_{\delta,J\cup J_0} \cap \langle \hat{\Pi}_J f, \epsilon \rangle_n \geq \frac{1}{16} \frac{(1 - \delta)^2}{(1 + \delta)} (1 - \rho_{q^*}^2) \kappa_l \right)
\leq \exp \left( -\frac{1}{29\sigma^2 (1 + \delta)^2 \sigma^2 \vee (1 - \rho_{q^*}^2) \kappa_l} \right).$$
The second term in (G.4) can be bounded analogously. This completes the proof. □

Acknowledgements

Finally, I sincerely would like to thank Prof. Enno Mammen for his support during the preparation of this paper.

References

[1] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. Constr. Approx., 28:253–263, 2008.
[2] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. Probab. Theory Related Fields, 113:301–413, 1999.
[3] P. Bühlmann and S. van de Geer. Statistics for High-Dimensional Data. Methods, Theory and Applications. Springer, Heidelberg, 2011.
[4] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. Ann. Statist., 37:1705–1732, 2009.
[5] L. Birgé and P. Massart. From model selection to adaptive estimation. In Festschrift for Lucien Le Cam, pages 55–87. Springer, New York, 1997.
[6] L. Birgé and P. Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. Bernoulli, 4:329–375, 1998.
[7] E. Candes and T. Tao. The dantzig selector: Statistical estimation when $p$ is much larger than $n$. Ann. Statist., 35:2313–2351, 2007.
[8] E. J. Candes and T. Tao. Near-optimal signal recovery from random projections: universal encoding strategies? IEEE Trans. Inform. Theory, 52:5406–5425, 2006.
[9] L. Comminges and A. S. Dalalyan. Tight conditions for consistency of variable selection in the context of high dimensionality. Ann. Statist., 40:2667–2696, 2012.
[10] A. Dalalyan, Y. Ingster, and A. B. Tsybakov. Statistical inference in compound functional models. Probab. Theory Related Fields, 158:513–532, 2014.
[11] M. Fornasier and H. Rauhut. Compressive sensing. In Handbook of Mathematical Methods in Imaging, pages 187–228. Springer, 2011.
[12] G. Gayraud and Y. Ingster. Detection of sparse additive functions. Electron. J. Stat., 6:1409–1448, 2012.
[13] J. Huang, J. L. Horowitz, and F. Wei. Variable selection in nonparametric additive models. Ann. Statist., 38:2282–2313, 2010.
[14] S. Kayalar and H. L. Weinert. Error bounds for the method of alternating projections. Math. Control Signals Systems, 1:43–59, 1988.
[15] H. Kober. A theorem on banach spaces. Compositio Math., 7:135–140, 1939.
[16] V. Koltchinskii. Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems. Springer, Heidelberg, 2011.
[17] V. Koltchinskii and M. Yuan. Sparsity in multiple kernel learning. Ann. Statist., 38:3660–3695, 2010.
[18] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. Ann. Statist., 28:1302–1338, 2000.
[19] P. Massart. *Concentration Inequalities and Model Selection*. Springer, Berlin, 2007.

[20] L. Meier, S. van de Geer, and P. Bühlmann. High-dimensional additive modeling. *Ann. Statist.*, 37:3779–3821, 2009.

[21] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *J. Mach. Learn. Res.*, 13:389–427, 2012.

[22] H. Rauhut. Compressive sensing and structured random matrices. In *Theoretical Foundations and Numerical Methods for Sparse Recovery*, Radon Ser. Comput. Appl. Math., 9, pages 1–92. Walter de Gruyter, Berlin, 2010.

[23] M. Rudelson. Random vectors in the isotropic position. *J. Funct. Anal.*, 164:60–72, 1999.

[24] M. Rudelson and R. Vershynin. Sampling from large matrices: an approach through geometric functional analysis. *J. ACM*, 54:19 pp, 2007.

[25] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, 2009.

[26] M. Wahl. Optimal estimation of components in structured nonparametric models. preprint. available at http://arxiv.org/abs/1403.1088, 2014.

[27] M. J. Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inform. Theory*, 55:5728–5741, 2009.

DEPARTMENT OF ECONOMICS, UNIVERSITY OF MANNHEIM, L7, 3-5, 68131, GERMANY

E-mail address: mawahl@mail.uni-mannheim.de