Electronic Journal of Statistics
Vol. 7 (2013) 480–498
ISSN: 1935-7524
DOI: 10.1214/13-EJS780

Model selection in regression under structural constraints

Felix Abramovich*

Department of Statistics and Operations Research
Tel Aviv University
Tel Aviv 69978, Israel
e-mail: felix@post.tau.ac.il

and

Vadim Grinshtein*

Department of Mathematics and Computer Science
The Open University of Israel
Raanana 43537, Israel
e-mail: vadimg@openu.ac.il

Abstract: The paper considers model selection in regression under the additional structural constraints on admissible models where the number of potential predictors miht be even larger than the available sample size. We develop a Bayesian formalism which is used as a natural tool for generating a wide class of model selection criteria based on penalized least squares estimation with various complexity penalties associated with a prior on a model size. The resulting criteria are adaptive to structural constraints. We establish the upper bound for the quadratic risk of the resulting MAP estimator and the corresponding lower bound for the minimax risk over a set of admissible models of a given size. We then specify the class of priors (and, therefore, the class of complexity penalties) where for the "nearly-orthogonal" design the MAP estimator is asymptotically at least nearly-minimax (up to a log-factor) simultaneously over an entire range of sparse and dense setups. Moreover, when the numbers of admissible models are "small" (e.g., ordered variable selection) or, on the opposite, for the case of complete variable selection, the proposed estimator achieves the exact minimax rates.

AMS 2000 subject classifications: Primary 62C99; secondary 62C10, 62C20, 62J05.
Keywords and phrases: Adaptivity, complexity penalty, Gaussian linear regression, maximum *a posteriori* rule, minimaxity, model selection, sparsity, structural constraints.

Received June 2012.

Contents

*Supported by Israel Science Foundation (ISF).

1. Introduction

Consider the standard Gaussian linear regression model

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^n$ is a vector of the observed response variable Y, $X_{n\times p}$ is the design matrix of the p explanatory variables (predictors) $X_1, \ldots, X_p$, $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of unknown regression coefficients, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$ and the noise variance σ^2 is assumed to be known.

A variety of statistical applications of regression models in different fields nowadays involves a vast number of potential predictors. Moreover, p might be even large relative to the amount of available data n ($p \gg n$ setup) that raises a severe "curse of dimensionality" problem. However, typically only some of the predictors have a truly relevant impact on the response $\mathbf{y}$. Model (variable) selection by identifying the "best" sparse subset of these "significant" predictors becomes therefore crucial in the analysis of such large data sets. For a selected model (a subset of predictors) M, the corresponding coefficients $\boldsymbol{\beta}_M$ are then typically estimated by least squares.

The goodness of model selection depends on the particular goal at hand. One should distinguish, for example, between estimation of regression coefficients $\boldsymbol{\beta}$, estimation of the mean vector $X\boldsymbol{\beta}$, model identification and predicting future observations. Different goals may lead to different optimal model selection procedures especially when the number of potential predictors p might be much larger than the sample size n. In this paper we consider mainly the estimation of the mean vector $X\boldsymbol{\beta}$ and the goodness of a model M is measured by the quadratic risk $E||X\hat{\boldsymbol{\beta}}_M - X\boldsymbol{\beta}||^2$, where $\hat{\boldsymbol{\beta}}_M$ is the least squares estimate of $\boldsymbol{\beta}$ for M. The "best" model then is the one with the minimal quadratic risk. Note that the true underlying model in (1) is not necessarily the best in this sense since sometimes it is possible to reduce its risk by excluding predictors with small (but still nonzero!) coefficients.

Minimum quadratic risk criterion for model selection is evidently impossible to implement since it involves the unknown true $\boldsymbol{\beta}$ but the corresponding ideal minimal (oracle) risk can be used as a benchmark for any available model se-

lection procedure. Typical model selection criteria are based on minimizing the *empirical* quadratic risk $||\mathbf{y} - X\hat{\boldsymbol{\beta}}_M||^2$, which is the least squares, penalized by a complexity penalty $Pen(|M|)$ increasing with a model size $|M|$:

$$\min_M \left\{ ||\mathbf{y} - X\hat{\boldsymbol{\beta}}_M||^2 + Pen(|M|) \right\} \tag{2}$$

The properties of the resulting penalized least squares estimator depend obviously on the particular choice of the complexity penalty $Pen(\cdot)$ in (2). The most commonly used choice is a *linear* type penalty of the form $Pen(k) = 2\sigma^2\lambda k$ for some $\lambda > 0$. The most known examples include C_p (Mallows [20]) and AIC (Akaike [4]) for $\lambda = 1$, BIC (Schwarz [23]) for $\lambda = (\ln n)/2$ and RIC (Foster and George [13]) for $\lambda = \ln p$. A series of recent works proposed the so-called $2k\ln(p/k)$-type *nonlinear* complexity penalties of the form $Pen(k) = 2\sigma^2\lambda k(\ln(p/k) + 1)(1 + o(1))$ with $\lambda \geq 1$. (see, e.g., Birgé and Massart [6, 7], Abramovich *et al.* [1], Bunea *et al.* [8], Abramovich and Grinshtein [2], Rigollet and Tsybakov [22]).

As we have mentioned, in the analysis of large complex data sets it is typically believed that the underlying (unknown) model is sparse, where a natural measure of model's sparsity is its size p_0. Abramovich and Grinshtein [2] and Rigollet and Tsybakov [22] established the minimax rates for the quadratic risk of estimating the mean vector $X\boldsymbol{\beta}$ in (1) over the set of models of sizes at most p_0 (see also analogous results of Raskutti *et al.* [21] for the case when X is of a full rank, i.e. $rank(X) = \min(p, n)$). They also showed that for $2k\ln(p/k)$-type penalties, the resulting penalized estimators (2) are simultaneously asymptotically optimal (in the minimax sense) for the entire range of sparse and dense models, while linear penalties cannot achieve such a wide optimality range.

So far, the above minimax properties of various model selection procedures in (1) have been established for *complete variable selection*, where the set of admissible models contains all 2^p subsets of the predictors $X_1, \ldots, X_p$. However, in a variety of regression setups there exist additional *structural constraints* that restrict the set of admissible models. In some cases predictors have some natural order and X_j can enter the model only if X_{j-1} is already there (*ordered variable selection*). For example, in polynomial regression, higher order polynomials are usually considered only if polynomials of lower degrees are already in the model. In tree-based models, a predictor cannot enter a model unless all its ancestors are there (*hierarchical model selection*). Another important example of hierarchical model selection is models with interactions, where interactions are typically selected only with the corresponding main effects. For a model with factor predictors, each factor predictor with k levels is, in fact, associated with a group of $k-1$ indicator (dummy) predictors. In this case either none or all of this group can be selected. It is somewhat similar to group sparse models, where predictors are splitted into pre-defined groups (see, e.g. Lounici *et al.* [19]).

In this paper we investigate the minimaxity properties of model selection criteria in (2) over classes of sparse and dense models under additional general structural constraints extending the existing results for the complete variable selection. The key point is to adapt the choice of the complexity penalty in (2)

to the specific structural constraints. In fact, it turns out that it is only the number of admissible models of a given size that matters.

We extend a Bayesian approach to model selection developed in Abramovich and Grinshtein [2] for the complete variable selection. The proposed Bayesian formalism is based on imposing a prior on a model size, where the penalty term in (2) is then naturally treated as proportional to its logarithm. From the Bayesian perspective, the model selection criterion (2) corresponds thus to the maximum *a posteriori* (MAP) Bayes rule. Depending on a specific choice of a prior, it implies a variety of penalized least squares estimators (or, equivalently, model selection criteria) with different complexity penalties. We show that under mild conditions on the prior, the resulting MAP estimator falls within a general class of penalized least squares estimators (2) considered in Birgé and Massart [6, 7] with complexity penalties $Pen(|M|)$ satisfying certain technical conditions depending on the number of admissible models of size $|M|$, that allows us to derive the upper bound for its quadratic risk. On the other hand, we establish the corresponding lower bound for the minimax quadratic risk over a set of models of a given size under structural constraints. We then specify the class of priors (and, therefore, the class of the corresponding complexity penalties), where for the "nearly-orthogonal" design, the resulting MAP estimators asymptotically are at least nearly-minimax (up to a log-factor) simultaneously over the wide range of sparse and dense models. In particular, when the numbers of admissible models are "small" (e.g., ordered variable selection) and for complete variable selection they lead to AIC-type and $2k\ln(p/k)$-type estimators respectively and achieve the exact minimax rate.

The paper is organized as follows. In Section 2 we develop a Bayesian formalism for model selection in regression under structural constraints and derive the MAP estimator. Its theoretical properties are presented in Section 3. In particular, we establish the upper bound for its quadratic risk, the corresponding lower bound for the minimax risk and discuss its asymptotic minimaxity in various setups. Section 4 presents the results of a simulation study. Concluding remarks are summarized in Section 5, while all the proofs are given in the Appendix.

2. MAP model selection procedure under structural constraints

We first extend the Bayesian formalism for the model selection in linear regression developed by Abramovich and Grinshtein [2] to structural constraints. We assume that the latter are known and the set of all admissible models is therefore fixed.

Consider the linear regression model (1), where the number of possible predictors p might be even larger then the number of observations n. Let $r = rank(X)(\leq \min(p,n))$ and assume that any r columns of X are linearly independent. For the "standard" linear regression setup, where all p predictors are linearly independent and there are at least p linearly independent design points, $r = p$.

Any model M is uniquely defined by the $p \times p$ diagonal indicator matrix $D_M = diag(\mathbf{d}_M)$, where $d_{Mj} = \mathbb{I}\{X_j \in M\}$ and, therefore, $|M| = tr(D_M)$. For

a given model M, we estimate its coefficients by least square estimator $\hat{\boldsymbol{\beta}}_M = (D_M X'XD_M)^+ D_M X'\mathbf{y}$, where "+" denotes the generalized inverse matrix.

Let $m(p_0)$ be the number of all admissible models of size p_0. The case $p_0 = 0$ corresponds to a null model with a single intercept and, therefore, $m(0) = 1$. In fact, we can consider only $p_0 \le r$ since otherwise, there necessarily exists another vector $\boldsymbol{\beta}^*$ with at most r nonzero entries such that $X\boldsymbol{\beta} = X\boldsymbol{\beta}^*$. Although for $p_0 = r$ there may be several different admissible models, all of them are evidently undistinguishable for estimating $X\boldsymbol{\beta}$ and can be associated with a single (saturated) model. Thus, without loss of generality, we can always assume that $m(r) = 1$. Obviously, for any p_0, $0 \le m(p_0) \le \binom{p}{p_0}$, where the two extreme cases $m(p_0) = 1$ and $m(p_0) = \binom{p}{p_0}$ for all $p_0 = 0, \ldots, r-1$, correspond respectively to the ordered and complete variable selection.

We start from imposing a prior on the model size $\pi(k) = P(|M| = k)$, $k = 0, \ldots, r$. Obviously, $\pi(k) = 0$ and $P(M\big|\, |M| = k) = 0$ iff there are no admissible models of a size k, i.e. $m(k) = 0$. For $m(k) > 0$, we assume all $m(k)$ admissible models of a given size k to be equally likely, that is, conditionally on the model size $|M| = k$,

$$P(M\big|\, |M| = k) = m(k)^{-1},$$

where recall that $m(r) = 1$ and hence $P(M\big|\, |M| = r) = 1$. To complete the prior, for any given model M we assume the normal prior on its unknown coefficients $\boldsymbol{\beta}_M$: $\boldsymbol{\beta}_M = \boldsymbol{\beta}|M \sim N_p(\mathbf{0}, \gamma\sigma^2(D_M X'XD_M)^+)$ which is the well-known conventional g-prior of Zellner [25].

For the proposed hierarchical prior, straightforward Bayesian calculus yields the posterior probability of a model M:

$$\begin{aligned} P(M|\mathbf{y}) \propto\ & \pi(|M|)m(|M|)^{-1}(1+\gamma)^{-\frac{|M|}{2}} \\ & \times \exp\left\{\frac{\gamma}{\gamma+1}\frac{\mathbf{y}'XD_M(D_M X'XD_M)^+ D_M X'\mathbf{y}}{2\sigma^2}\right\}, \end{aligned} \tag{3}$$

where we set $\pi(|M|)m(|M|)^{-1} = 0$ if $m(|M|) = 0$. Finding the most likely model leads therefore to the following maximum *a posteriori* (MAP) model selection criterion:

$$\begin{aligned} \max_M \Big\{ & \mathbf{y}'XD_M(D_M X'XD_M)^+ D_M X'\mathbf{y} \\ & + 2\sigma^2(1+1/\gamma)\ln\left\{m(|M|)^{-1}\pi(|M|)(1+\gamma)^{-\frac{|M|}{2}}\right\}\Big\} \end{aligned}$$

or, equivalently,

$$\min_M \left\{||\mathbf{y} - X\hat{\boldsymbol{\beta}}_M||^2 + 2\sigma^2(1+1/\gamma)\ln\left\{m(|M|)\pi(|M|)^{-1}(1+\gamma)^{\frac{|M|}{2}}\right\}\right\} \tag{4}$$

which is of the general type (2) with the complexity penalty

$$Pen(|M|) = 2\sigma^2(1+1/\gamma)\ln\left\{m(|M|)\pi(|M|)^{-1}(1+\gamma)^{\frac{|M|}{2}}\right\} \tag{5}$$

A specific form of the penalty (5) is defined by the choice of the prior $\pi(\cdot)$ on the model size.

3. Main results

3.1. Risk bounds

Denote the set of all admissible models by Ω. For a given model size $1 \le p_0 \le r$, define the set of all admissible models $\mathcal{M}_{p_0}$ of size p_0, that is, $\mathcal{M}_{p_0} = \{M \in \Omega : |M| = p_0\}$ and $card(\mathcal{M}_{p_0}) = m(p_0)$. Obviously, if a model $M \in \mathcal{M}_{p_0}$, the corresponding coefficients vector $\boldsymbol{\beta}_M \in \mathbb{R}^p$ has p_0 nonzero components, where $\beta_{Mj} \neq 0$ iff X_j is included in M. Let $\mathfrak{M}_{p_0} = \bigcup_{k=0}^{p_0} \mathcal{M}_k$ be the set of all admissible models with at most p_0 predictors. Following our arguments from Section 2, Ω can be essentially reduced to $\mathfrak{M}_r$.

In this section we derive the upper and lower bounds for the maximal risk of the proposed MAP model selector (4) over $\mathfrak{M}_{p_0}$.

Theorem 1 (upper bound)**.** *Let $\hat{M}$ be the solution of (4) and $\hat{\boldsymbol{\beta}}_{\hat{M}}$ be the corresponding least squares estimator of its coefficients. Define $c(\gamma) = 8(\gamma + 3/4)^2 > 9/2$ and assume that for some constant $c > 0$,*

$$\min\{m(k)^{-c}, e^{-ck}\} \le \pi(k) \le m(k)e^{-c(\gamma)k} \tag{6}$$

for all $k = 1, \ldots, r$ such that $m(k) > 0$.

Then, there exists a constant $C(\gamma) > 0$ depending only on γ such that

$$\sup_{\boldsymbol{\beta}_M : M \in \mathfrak{M}_{p_0}} E||X\hat{\boldsymbol{\beta}}_{\hat{M}} - X\boldsymbol{\beta}_M||^2 \le C(\gamma)\sigma^2 \min\{\max(p_0, \ln m(p_0)), r\} \tag{7}$$

simultaneously for all $1 \le p_0 \le r$.

Under the conditions (6) on the prior $\pi(k)$ in Theorem 1, the corresponding penalty (5) satisfies

$$C_1(\gamma)\sigma^2 k \;\le\; Pen(k) \;\le\; C_2(\gamma)\sigma^2 \max\{\ln m(k), k\}, \quad k = 1, \ldots, r \tag{8}$$

for some positive constants $C_1(\gamma)$ and $C_2(\gamma)$.

To assess the accuracy of the established upper bound for the quadratic risk of the proposed MAP estimator, we derive the lower bound for the minimax risk of estimating $X\boldsymbol{\beta}$ in (1).

The l_0 quasi-norm $||\boldsymbol{\beta}||_0$ of a vector $\boldsymbol{\beta}$ is defined as the number of its nonzero entries. For any given $k = 1, \ldots, r$, let $\phi_{min}[k]$ and $\phi_{max}[k]$ be the k-sparse minimal and maximal eigenvalues of the design defined as

$$\phi_{min}[k] = \min_{\boldsymbol{\beta}: 1 \le ||\boldsymbol{\beta}||_0 \le k} \frac{||X\boldsymbol{\beta}||^2}{||\boldsymbol{\beta}||^2},$$

$$\phi_{max}[k] = \max_{\boldsymbol{\beta}: 1 \le ||\boldsymbol{\beta}||_0 \le k} \frac{||X\boldsymbol{\beta}||^2}{||\boldsymbol{\beta}||^2}$$

In fact, $\phi_{min}[k]$ and $\phi_{max}[k]$ are respectively the minimal and maximal eigenvalues of all $k \times k$ submatrices of the matrix $X'X$ generated by any k columns

of X. Let $\tau[k] = \phi_{min}[k]/\phi_{max}[k],\ k = 1, \ldots, r$. By the definition, $\tau[k]$ is a non-increasing function of k. Obviously, $\tau[k] \le 1$ and for the orthogonal design the equality holds for all k.

Theorem 2 (minimax lower bound)**.** *Consider the model (1) and let* $1 \le p_0 \le r$. *There exists a universal constant* $C > 0$ *such that*

$$\inf_{\tilde{\mathbf{y}}} \sup_{\boldsymbol{\beta}_M : M \in \mathfrak{M}_{p_0}} \mathbb{E}||\tilde{\mathbf{y}} - X\boldsymbol{\beta}_M||^2 \ge \begin{cases} C\sigma^2 \max\left\{\tau[2p_0]\frac{\ln m(p_0)}{\max(1,\ln p_0)}, \tau[p_0]p_0\right\}, & 1 \le p_0 \le r/2 \\ C\sigma^2\tau[p_0]\, r, & r/2 \le p_0 \le r \end{cases} \tag{9}$$

where the infimum is taken over all estimates $\tilde{\mathbf{y}}$ *of the mean vector* $X\boldsymbol{\beta}$.

In some particular cases, e.g. for complete variable selection, the general minimax lower bounds established in Theorem 2 can be improved by removing the $\max(1, \ln p_0)$-term in (9) (Abramovich and Grinshtein [2], Rigollet and Tsybakov [22], Raskutti *et al.* [21]). Whether this additional log-term can be removed in the general case remains so far a conjecture. Note however that if $\ln m(p_0) = O(p_0)$ (in particular, for the ordered variable selection, where $m(p_0) = 1$), the dominating term in both bounds (7) and (9) is anyway p_0.

The upper bound (7) holds for any design matrix X of rank r, while the minimax lower bound (9) depends on X but only through the sparse eigenvalues ratios. Finally note that the structural constraints are manifested in the upper and lower bounds only through $m(p_0)$ – the number of admissible models of size p_0.

3.2. Asymptotic adaptive minimaxity of the MAP estimator

The established upper and lower risk bounds (7), (9) in the previous Section 3.1 is the key for investigating the asymptotic minimaxity of the proposed MAP estimator, where the number of possible predictors $p = p_n$ may increase with the sample size n. One can view such a setup as a series of projections of the vector $X\boldsymbol{\beta}$ on the expanding span of predictors. In particular, it may be $p_n > n$ or even $p_n \gg n$. Thus, formally, we consider now a *sequence* of design matrices $X_{p,n}$, where $r_n = rank(X_{n,p}) \to \infty$. For simplicity of exposition, hereafter, we omit the index n. Similarly, there are sequences of the coefficient vectors $\boldsymbol{\beta}_p$ and priors $\pi_p(\cdot)$. In these notations the original model (1) is transformed into a sequence of models

$$\mathbf{y} = X_p\boldsymbol{\beta}_p + \boldsymbol{\epsilon}, \tag{10}$$

where $rank(X_p) = r$ and any r columns of X_p are linearly independent (hence, $\tau_p[r] > 0$), $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$ and the noise variance σ^2 does not depend on n and p.

We consider the *nearly-orthogonal* design, where the sequence of sparse eigenvalues ratios $\tau_p[r]$ is bounded away from zero. Nearly-orthogonality means that there are no "too strong" linear relationships within any set of r columns of the design matrix X_p. Evidently, in this case p cannot be "too large" relative to r

and, therefore, to n. Indeed, Abramovich and Grinshtein [2] showed that nearly-orthogonality of a design necessarily implies $p = O(r)$ and, thus, $p = O(n)$. In this case,

$$\max\left(\ln m(p_0), p_0\right) \;\le\; \max\left(\ln \binom{p}{p_0}, p_0\right) \;\le\; p_0(\ln(p/p_0)+1) \;\le\; p = O(r)$$

and the following two corollaries are immediate consequences of Theorems 1 and 2:

Corollary 1 (bounds for the minimax risk)**.** *Let the design be nearly-orthogonal. There exist two constants* $0 < C_1 \le C_2 < \infty$ *such that for all sufficiently large* r*,*

$$C_1\sigma^2 \max\left\{\frac{\ln m(p_0)}{\max(1,\ln p_0)}, p_0\right\} \le \inf_{\tilde{\mathbf{y}}} \sup_{\boldsymbol{\beta}_M : M\in\mathfrak{M}_{p_0}} E||\tilde{\mathbf{y}} - X_p\boldsymbol{\beta}_M||^2 \le C_2\sigma^2 \max\left\{\ln m(p_0), p_0\right\}$$

for all $1 \le p_0 \le r$*.*

Corollary 2 (asymptotic adaptive minimaxity of the MAP estimator)**.** *Consider the nearly-orthogonal design and assume that for* $m(k) > 0$ *the prior* $\pi(k)$ *satisfies*

$$\min\{m(k)^{-c}, e^{-ck}\} \le \pi(k) \le m(k)e^{-c(\gamma)k}, \quad k = 1,\dots,r$$

for some $c > 0$ *and* $c(\gamma)$ *defined in Theorem 1. Then the corresponding MAP estimator (4) is asymptotically at least nearly-minimax (up to a* $\ln p_0$*-factor) simultaneously over all* $\mathfrak{M}_{p_0}$*,* $1 \le p_0 \le r$*.*

The above general results depend on the asymptotic behavior of $m(p_0)$ as a function of p_0. Similar to Birgé and Massart [7] we consider the following three typical cases:

1. Small numbers of admissible models: $\ln m(k) = O(k),\ k = 1,\dots,r-1$ (recall however that $m(k) \le \binom{p}{k}$).
In particular, for ordered variable selection, $m(k) = 1$ though "small" numbers allow also even exponential growth of $m(k)$ for small and moderate k.

For this case, Corollary 1 and Theorem 1 imply:

Corollary 3. *Consider the nearly-orthogonal design and let* $\ln m(k) = O(k),\ k = 1,\dots,r-1$*. As* r *increases,*

1. $$\inf_{\tilde{\mathbf{y}}} \sup_{\boldsymbol{\beta}_M : M\in\mathfrak{M}_{p_0}} E||\tilde{\mathbf{y}} - X_p\boldsymbol{\beta}_M||^2 \asymp \sigma^2 p_0 \tag{11}$$
   for all $p_0 = 1,\dots,r$*.*
2. *For* $m(k) > 0$ *assume that* $e^{-ck} \le \pi(k) \le m(k)e^{-c(\gamma)k},\ k = 1,\dots,r$ *for some* $c > 0$*. The resulting MAP estimator (4) attains then the minimax rates simultaneously over all* $\mathfrak{M}_{p_0}, 1 \le p_0 \le r$*.*

Note that $\sigma^2 p_0$ is the risk of (unbiased) least squares estimation of $X\boldsymbol{\beta}_{M_0}$ for a true model M_0 of size p_0 in (1). Corollary 3 therefore verifies that when the number of admissible models is small, there is essentially no extra price for model selection.

It follows from (5) that priors satisfying the conditions of Corollary 3 lead to the AIC-type penalties of the form $Pen(k) \sim 2C(\gamma)\sigma^2 k$ for some $C(\gamma) > 1$.

2. Complete variable selection: $m(k) = \binom{p}{k}$, $k = 1, \ldots, r-1$.
As we have mentioned above, from the already known results of Abramovich and Grinshtein [2], Rigollet and Tsybakov [22] and Raskutti *et al.* [21] the $\max(1, \ln p_0)$-term in the lower minimax risk bound can be removed in this case:

Corollary 4. *Consider complete variable selection for the nearly-orthogonal design. As r increases,*

1. $$\inf_{\tilde{\mathbf{y}}} \sup_{\boldsymbol{\beta}_M : M \in \mathfrak{M}_{p_0}} E||\tilde{\mathbf{y}} - X_p \boldsymbol{\beta}_M||^2 \asymp \sigma^2 p_0 (\ln(p/p_0) + 1)$$
   for all $p_0 = 1, \ldots, r$.
2. *Assume that $\left(\frac{k}{pe}\right)^{ck} \le \pi(k) \le \left(\frac{p}{ke^{c(\gamma)}}\right)^k$, $k = 1, \ldots, r-1$ and $e^{-cr} \le \pi(r) \le e^{-c(\gamma)r}$ for some $c > c(\gamma)$. The resulting MAP estimator (4) attains then the minimax rates simultaneously over all $\mathfrak{M}_{p_0}, 1 \le p_0 \le r$.*

The upper bound on $\pi(k)$ in Corollary 4 trivially holds for all $k \le pe^{-c(\gamma)}$.

Corollary 4 shows that for complete variable selection, model selection yields an additional multiplicative factor of $\ln(p/p_0)$ to the risk $\sigma^2 p_0$ of estimating $X\boldsymbol{\beta}_{M_0}$ for a true model M_0 of size p_0 in (1).

The conditions on the prior of Corollary 4 hold, for example, for the truncated geometric prior $\pi(k) \propto q^k$, $k = 1, \ldots, r$ for some $0 < q < 1$, and the corresponding penalties in (5) are of the $2k\ln(p/k)$-type, where $Pen(k) = 2C(\gamma)\sigma^2 k(\ln(p/k) + 1)(1 + o(1))$ for some $C(\gamma) > 1$.

3. Intermediate case: $k = o(\ln m(k))$ and $m(k) < \binom{p}{k}$, $k = 1, \ldots, r-1$.
A practically interesting example of the intermediate case is hierarchical model selection with paired interactions mentioned in Section 1:

Example: hierarchical model selection with paired interactions.
Consider model selection in regression with K main predictors and their paired interactions. The overall number of possible predictors p in (1) is therefore $p = K + \binom{K}{2} = K(K+1)/2$. However, an interaction can be included in the model only together with the corresponding main effects. Obviously, $m(k) < \binom{p}{k}$, $k = 1, \ldots, r-1$. On the other hand, this is an example with "large" numbers of admissible models, where $\ln m(k) \ge ck\ln(p/k)$ for some $0 < c < 1$. Indeed, for models of sizes one and two only main effects can be selected, and the numbers of admissible models are $m(1) = K$ and $m(2) = \binom{K}{2}$ respectively.

One can trivially verify that in both cases $\ln m(k) \geq ck \ln(p/k),\ k = 1, 2$ for some positive constant $c < 1$. For $k \geq 3$, we have the following lemma:

Lemma 1. *For all* $3 \leq k \leq r - 1$,

$$\ln m(k) \geq \left\lfloor \frac{k}{3} \right\rfloor \ln(p/k)$$

For the intermediate case there is the $\ln p_0$-gap between the upper bound for the quadratic risk of the MAP estimator in Theorem 1 and the minimax lower bound in Theorem 2. So far, we can claim that if $m(k)^{-c} \leq \pi(k) \leq m(k)e^{-c(\gamma)k)},\ k = 1, \ldots, r$ and the design is nearly-orthogonal, the resulting MAP estimator is asymptotically at least nearly-minimax (up to $\ln p_0$-factor) over all $\mathfrak{M}_{p_0},\ 1 \leq p_0 \leq r$ and can only conjecture that, similar to the complete variable selection, this log-factor can be removed in this case as well.

Similar to complete variable selection, when the numbers of admissible models for the intermediate case are "large" (e.g., hierarchical model selection with main effects and paired interactions considered above), the conditions on the prior in Corollary 2 are satisfied for the truncated geometric prior $\pi(k) \propto q^k, k = 1, \ldots, r$ for some $0 < q < 1$ corresponding to complexity penalties of $2k \ln(p/k)$-type.

Note also that for the nearly-orthogonal design, $||X_p\hat{\boldsymbol{\beta}}_{p\hat{M}} - X_p\boldsymbol{\beta}_p|| \asymp ||\hat{\boldsymbol{\beta}}_{p\hat{M}} - \boldsymbol{\beta}_p||$ and all the results of this section for estimating the mean vector $X_p\boldsymbol{\beta}_M$ can be therefore straightforwardly applied to estimating the regression coefficients $\boldsymbol{\beta}_M$.

4. Simulation study

We conducted a short simulation study to demonstrate the performance of the proposed MAP model selection procedure. We considered polynomial regression which is an example of the ordered variable selection (see Section 1):

$$y_i = \beta_0 + \beta_1 x_i + \cdots + \beta_k x_i^k + \epsilon_i, \quad i = 1, \ldots, n,$$

where $0 \leq k \leq n - 1$ is the polynomial degree to be selected, and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ with the known variance σ^2 and independent. In this case obviously $m(k) = 1$ for all $k = 0, \ldots, n-1$. An example of a prior satisfying the conditions of Corollary 3 is the (truncated) geometric prior $Geom(1-q)$, where $\pi(k) = (1-q)q^k/(1-q^n) \propto q^k,\ k = 0, \ldots, n-1$ for some $0 < q < 1$. The corresponding complexity penalty (5) is the AIC-type linear penalty

$$Pen(k) = 2\sigma^2(1 + 1/\gamma) \ln\left(q^{-1}\sqrt{1+\gamma}\right) k \tag{12}$$

4.1. Estimation of parameters

To apply the developed MAP model selection procedure we need to specify the prior parameters γ and q in (12). They are rarely known *a priori* in practice and usually should be estimated from the data.

Let $X_{n\times n}$ be the design matrix of the saturated model, that is, $X_{ij} = x_i^{j-1},\ i = 1,\ldots,n; j = 0,\ldots,n-1$. For a given k, define the corresponding diagonal indicator matrix $D_k = diag(\mathbf{d}_k)$, where $d_{kj} = 1,\ j = 1,\ldots,k$ and zero otherwise (see Section 2). Based on the proposed Bayesian model from Section 2,

$$\mathbf{y}|k \sim \mathcal{N}(\mathbf{0}, \sigma^2(I + \gamma H_k)),$$

where $H_k = XD_k(D_kX'XD_k)^+D_kX'$, and straightforward calculus yields the following marginal likelihood of the observed data $\mathbf{y}$:

$$\begin{aligned} L(\mathbf{y};\gamma,q) &= \sum_{k=0}^{n-1} \frac{1}{\sqrt{|I+\gamma H_k|}} \exp\left\{-\frac{\mathbf{y}'(I+\gamma H_k)^{-1}\mathbf{y}}{2\sigma^2}\right\}\pi(k) \\ &\propto \sum_{k=0}^{n-1} \frac{1}{(1+\gamma)^{k/2}} \exp\left\{\frac{\gamma}{\gamma+1}\frac{\mathbf{y}'H_k\mathbf{y}}{2\sigma^2}\right\}\frac{(1-q)q^k}{1-q^n} \end{aligned}$$

The MLEs for γ and q can be obtained (numerically) by the EM algorithm. Regard k as a "missing" data and define the corresponding latent indicator variables $u_j = \delta_{jk},\ j = 0,\ldots,n-1$. The complete log-likelihood for the "augmented" data $(\mathbf{y},\mathbf{u})$, up to an additive constant, is then

$$\begin{aligned} l(\mathbf{y},\mathbf{u};\gamma,q) = \frac{\gamma}{\gamma+1}\sum_{k=0}^{n-1} u_k\frac{\mathbf{y}'H_k\mathbf{y}}{2\sigma^2} - \frac{\ln(1+\gamma)}{2}\sum_{k=0}^{n-1}u_k k \\ + \ln q\sum_{k=0}^{n-1}u_k k + \ln(1-q) - \ln(1-q^n) \end{aligned}$$

On the E-step at the h-th iteration we compute the conditional expectation

$$\begin{aligned} \hat{l}^{[h]} &= E\left(l(\mathbf{y},\mathbf{u};\gamma,q)|\mathbf{y},\gamma^{[h]},q^{[h]}\right) \\ &= \frac{\gamma}{\gamma^+1}\sum_{k=0}^{n-1}\hat{u}_k^{[h]}\frac{\mathbf{y}'H_k\mathbf{y}}{2\sigma^2} - \frac{\ln(1+\gamma)}{2}\sum_{k=0}^{n-1}\hat{u}_k^{[h]}k \\ &+ \ln q\sum_{k=0}^{n-1}\hat{u}_k^{[h]}k + \ln(1-q) - \ln(1-q^n), \end{aligned} \tag{13}$$

where

$$\hat{u}_k^{[h]} = E(u_k|\mathbf{y},\gamma^{[h]},q^{[h]}) = \frac{\exp\left\{\frac{\gamma^{[h]}}{\gamma^{[h]}+1}\frac{\mathbf{y}'H_k\mathbf{y}}{2\sigma^2}\right\}q^{[h]k}}{\sum_{j=0}^{n-1}\exp\left\{\frac{\gamma^{[h]}}{\gamma^{[h]}+1}\frac{\mathbf{y}'H_j\mathbf{y}}{2\sigma^2}\right\}q^{[h]j}}$$

At the M-step we maximize $\hat{l}^{[h]}$ w.r.t. γ and q to get

$$\hat{\gamma}^{[h+1]} = \left(\frac{\sum_{k=0}^{n-1}\hat{u}_k^{[h]}\mathbf{y}'H_k\mathbf{y}}{\sigma^2\sum_{k=0}^{n-1}\hat{u}_k^{[h]}k} - 1\right)_+$$

There is no closed form solution for $\hat{q}^{[h+1]}$. However, replacing the truncated geometric distribution by a usual one and ignoring thus the last term $\ln(1-q^n)$ in the RHS of (13), implies a good approximation of $\hat{q}^{[h+1]}$ for large n:

$$\hat{q}^{[h+1]} = \frac{\sum_{k=0}^{n-1} \hat{u}_k^{[h]} k}{1 + \sum_{k=0}^{n-1} \hat{u}_k^{[h]} k}$$

4.2. The results

We used two test functions: a fifth degree polynomial $g_1(x) = (x+0.1)(x-0.2)(x-0.4)(x-0.8)(x-1.1)$ and the Doppler function $g_2(x) = \sqrt{x(1-x)}\sin(2\pi \cdot 1.05/(x+0.05))$ (see Donoho and Johnstone [11]) which does not have a sparse polynomial approximation. Both functions were then normalized to have unit $L_2[0,1]$-norms. The data were generated according to the model

$$y_i = g_{1,2}(i/n) + \epsilon_i, \quad i = 1, \ldots, n$$

for $n = 100$, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ and independent. The noise variance σ^2 was chosen to ensure the signal-to-noise ratio SNR at levels $3, 5$ and 7 and was assumed to be known. The number of replications was 100.

The proposed MAP model selector results in a model selection procedure with a linear type penalty of the form $Pen(k) = 2\sigma^2 \lambda k$ with $\lambda_{MAP} = (1 + 1/\gamma)\ln(q^{-1}\sqrt{1+\gamma})$ (see (12)). We compared it with two other well-known model selection procedures with linear penalties, namely, AIC ($\lambda_{AIC} = 1$) and RIC ($\lambda_{RIC} = \ln p$) (see Section 1). In our case, $\lambda_{RIC} = \ln 100 = 4.6$.

Table 1 summarizes mean squared errors averaged over 100 replications (AMSE). We also present the average polynomial degrees selected by the three methods for approximating the true response functions.

As expected, a more conservative RIC tends to include less terms in the model and outperforms AIC for a polynomial g_1, while the latter is superior for g_2, where a high order polynomial approximation is required. The MAP estimator with estimated γ and q yields a data-driven λ and is adaptive to the unknown polynomial degree – it behaves very similar to RIC when it is low and to AIC when it is high.

TABLE 1
AMSEs and polynomial degrees (in parentheses) averaged over 100 replications for three estimators

	g_1			g_2		
SNR	MAP	AIC	RIC	MAP	AIC	RIC
3	0.661	0.813	0.652	28.844	27.818	35.188
	(5.01)	(5.50)	(5.00)	(29.07)	(33.66)	(18.11)
5	0.238	0.293	0.235	24.677	23.450	27.056
	(5.01)	(5.50)	(5.00)	(44.13)	(53.48)	(29.40)
7	0.121	0.149	0.120	22.995	22.698	26.036
	(5.01)	(5.50)	(5.00)	(52.72)	(58.57)	(32.21)

5. Concluding remarks

In this paper we considered model selection in linear regression under general structural constraints and extended the existing results for the complete variable selection. In particular, we utilized a Bayesian MAP model selection procedure of Abramovich and Grinshtein [2] and modified it correspondingly. From a frequentist view, the resulting MAP model selector is a penalized least squares estimator with a complexity penalty associated with a prior on the model size which is adaptive to the structural constraints. In fact, the proposed Bayesian approach can be used as a tool for generating a wide class of penalized least squares estimators with various complexity penalties.

We established the general upper bound for the quadratic risk of the MAP estimator over a set of admissible models of a given size and the lower bound for the corresponding minimax risk. Based on these results, we showed that for the nearly-orthogonal design, the MAP estimator is asymptotically at least nearly minimax (up to a log-factor) for the entire range of sparse and dense models. Moreover, when the numbers of admissible models are "small" or, on the opposite, for the case of complete variable selection, it achieves the exact minimax rates. The corresponding MAP model selection procedures lead respectively to AIC-type and $2k\ln(p/k)$-type criteria. Whether these results on asymptotic minimaxity are true for the intermediate case remains so far a conjecture.

There are also other challenges for future research. The assumption of nearly-orthogonal design used in investigating asymptotic minimaxity of MAP estimators in Section 3.2 typically does not hold for $p \gg n$ setup due to the multicollinearity phenomenon. The analysis of multicollinear design, where the sequence of sparse eigenvalues ratios $\tau_p[r]$ may tend to zero as p increases is much more delicate. In this case there is a gap (in addition to a log-factor) between the rates in the upper and lower bounds (7) and (9). Unlike model identification or coefficients estimation, where multicollinearity is a "curse", it may essentially become a "blessing" for estimating the mean vector allowing one to exploit correlations between predictors to reduce the size of a model (hence, to decrease the variance) without paying much extra price in the bias term. Interestingly, a similar phenomenon also occurs in a testing setup (e.g., Hall and Jin [17]). Abramovich and Grinshtein [2] investigated the complete variable selection for multicollinear design and showed that under certain additional assumptions on the design and the regression coefficients, the MAP estimator corresponding to the $2k\ln(p/k)$-type complexity penalty remains asymptotically rate-optimal (in the minimax sense) even for this case. Whether it is true and what are the additional conditions in the presence of structural constraints is a challenging topic for future research.

Computational issues are another important problem. When the numbers of admissible models are large (e.g., complete variable selection or hierarchical model selection with interactions), minimizing (2) (and (4) in particular) requires generally an NP-hard combinatorial search. During the last decade there have been substantial efforts to develop various approximated algorithms for solving (2) that are computationally feasible for high-dimensional data (see,

e.g. Tropp and Wright [24] for a survey and references therein). The common remedies involve either greedy algorithms (e.g., forward selection, matching pursuit) approximating the global solution by a stepwise sequence of local ones, or convex relaxation methods replacing the original combinatorial problem by a related convex program (e.g., Lasso and Dantzig selector for linear penalties). Abramovich and Grinshtein [2] proposed to utilize the developed Bayesian formalism for solving (4) by using a stochastic search variable selection (SSVS) techniques originated in George and McCullogh [14, 15]. The underlying idea of SSVS is based on generating a sequence of models from the posterior distribution $P(M|\mathbf{y})$ in (3). The key point is that the relevant models with the highest posterior probabilities will appear most frequently and can be identified even for a generated sample of a relatively small size avoiding computations of the entire posterior distribution. However, most of the above approaches have been developed and studied for complete variable selection. Their adaptation to minimization of (4) subject to the additional structural constraints while remaining computationally feasible should depend on the specific type of constraints at hand. In particular, for somewhat different priors, Chipman [9] and Farcomeni [12] considered SVSS for hierarchical model selection in regression with paired interactions (see the example in Section 3) and for model selection with factor predictors, where the corresponding dummy variables are all included or excluded. Bien *et al.* [5] modified Lasso for hierarchical paired interactions (see also references therein). However, to the best of our knowledge, there are no theoretical results on the optimality of the resulting estimators.

Finally, we should note that the obtained theoretical results assume that the noise variance σ^2 is known which is rarely the case in practical applications. One can estimate σ^2 from the data with the additional tuning parameters of the MAP procedure (see Section 4.1). Alternatively, he can follow the fully Bayesian approach and impose some prior distribution on it (see, e.g., Chipman [9], Chipman *et al.* [10] and Farcomeni [12]).

6. Appendix

Throughout the proofs we use C to denote a generic positive constant, not necessarily the same each time it is used, even within a single equation.

6.1. Proof of Theorem 1

We first show that under the conditions on a prior in Theorem 1, the corresponding penalty $Pen(k)$ in (5) belongs to the class of penalties considered in Birgé and Massart [6] and then use their Theorem 2 to establish the general upper bound for the quadratic risk of the MAP estimator (4).

Define

$$L_k = \frac{1}{k} \ln \left(m(k) \pi^{-1}(k) \right), \quad k = 1, \ldots, r, \tag{14}$$

where under the conditions on the prior $\pi(\cdot)$ in Theorem 1, $L_k \geq c(\gamma)$. In terms of L_k the complexity penalty (5) is $Pen(k) = \sigma^2(1 + 1/\gamma)k(2L_k + \ln(1 + \gamma))$.

In our notations the conditions (3.3) and (3.4) on L_k in Theorem 2 of Birgé and Massart [6] correspond respectively to

$$\sum_{k=1}^{r} m(k)e^{-kL_k} < c \tag{15}$$

and

$$(1+1/\gamma)(2L_k+\ln(1+\gamma)) \geq C(1+\sqrt{2L_k})^2, \quad k=1,\ldots,r \tag{16}$$

for some $C > 1$.

The condition (15) follows immediately from the definition of L_k:

$$\sum_{k=1}^{r} m(k)e^{-kL_k} = \sum_{k=1}^{r} \pi(k) = 1-\pi(0) < 1$$

Consider now (16) which is equivalent to the inequality

$$2(1+1/\gamma-C)L_k - 2C\sqrt{2L_k} + (1+1/\gamma)\ln(1+\gamma) - C \geq 0 \tag{17}$$

Repeating the calculus in the proof of Theorem 1 of Abramovich *et al.* [3], one verifies that with the upper bound on the prior $\pi(k)$ in (6), for $C = 1+1/(2\gamma)$, (17) and therefore (16) are satisfied for all $L_k,\ k=1,\ldots,r$.

Given (15)-(16), Theorem 2 of Birgé and Massart [6] yields the following upper bound for the quadratic risk of the MAP estimator (4) of the mean vector $X\boldsymbol{\beta}$ in (1):

$$E||X\hat{\boldsymbol{\beta}}_{\hat{M}} - X\boldsymbol{\beta}||^2 \leq c_0(\gamma) \inf_{M\in\mathfrak{M}_r} \left\{||X\boldsymbol{\beta}_M - X\boldsymbol{\beta}||^2 + Pen(|M|)\right\} + c_1(\gamma)\sigma^2 \tag{18}$$

for some $c_0(\gamma)$ and $c_1(\gamma)$ depending only on γ, where $Pen(|M|)$ is given in (5).

Recall that $m(r) = 1$. Then, using the lower bound on $\pi(r)$, (18) and (8) imply

$$\begin{aligned} \sup_{\boldsymbol{\beta}_M : M\in\mathfrak{M}_{p_0}} E||X\hat{\boldsymbol{\beta}}_{\hat{M}} - X\boldsymbol{\beta}_M||^2 &\leq \sup_{\boldsymbol{\beta}_M : M\in\mathfrak{M}_r} E||X\hat{\boldsymbol{\beta}}_{\hat{M}} - X\boldsymbol{\beta}_M||^2 \\ &\leq c_0(\gamma)Pen(r) + c_1(\gamma)\sigma^2 \ \leq\ C(\gamma)\sigma^2 r \end{aligned} \tag{19}$$

for all $1 \leq p_0 \leq r$.

On the other hand, for $p_0 < r$ from (18) and (8) we have

$$\begin{aligned} \sup_{\boldsymbol{\beta}_M : M\in\mathfrak{M}_{p_0}} E||X\hat{\boldsymbol{\beta}}_{\hat{M}} - X\boldsymbol{\beta}_M||^2 &\leq c_0(\gamma)Pen(p_0) + c_1(\gamma)\sigma^2 \\ &\leq C(\gamma)\sigma^2 \max\{\ln m(p_0), p_0\} \end{aligned} \tag{20}$$

Combining (19) and (20) completes the proof. □

6.2. Proof of Theorem 2

We first show that

$$\inf_{\tilde{\mathbf{y}}} \sup_{\boldsymbol{\beta}_M : M \in \mathfrak{M}_{p_0}} \mathbb{E}||\tilde{\mathbf{y}} - X\boldsymbol{\beta}_M||^2 \geq \tau[p_0]\sigma^2 p_0 \tag{21}$$

for all $1 \leq p_0 \leq r$.

Consider the original regression model (1) and assume that the true coefficients vector $\boldsymbol{\beta}_M \in \mathfrak{M}_{p_0}$. Define $X_M = XD_M$, where the $p \times p$ diagonal indicator matrix D_M was defined in Section 2. In the coefficients domain one then has

$$\mathbf{w} = \boldsymbol{\beta}_M + \boldsymbol{\epsilon}', \tag{22}$$

where $\mathbf{w} = (X_M'X_M)^+X_M'\mathbf{y}$ and $\boldsymbol{\epsilon}' \sim \mathcal{N}(\mathbf{0}, \sigma^2(X_M'X_M)^+)$.

Let $R(\mathfrak{M}_{p_0}, \sigma^2(X_M'X_M)^+)$ be the minimax risk of estimating $\boldsymbol{\beta}_M$ in (22) over $\mathfrak{M}_{p_0}$, that is,

$$R(\mathfrak{M}_{p_0}, \sigma^2(X_M'X_M)^+) = \inf_{\tilde{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}_M : M \in \mathfrak{M}_{p_0}} \mathbb{E}||\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_M||^2$$

Evidently,

$$\inf_{\tilde{\mathbf{y}}} \sup_{\boldsymbol{\beta}_M : M \in \mathfrak{M}_{p_0}} \mathbb{E}||\tilde{\mathbf{y}} - X\boldsymbol{\beta}_M||^2 \geq \phi_{min}[p_0]\ R(\mathfrak{M}_{p_0}, \sigma^2(X_M'X_M)^+) \tag{23}$$

Consider also the model (22) but with the uncorrelated noise $\boldsymbol{\epsilon}'' \sim \mathcal{N}(\mathbf{0}, \sigma^2\phi_{max}^{-1}[p_0]D_M)$:

$$\mathbf{w} = \boldsymbol{\beta}_M + \boldsymbol{\epsilon}'' \tag{24}$$

and the corresponding minimax risk $R(\mathfrak{M}_{p_0}, \sigma^2\phi_{max}^{-1}[p_0]D_M)$.

Since $(X_M'X_M)^+ \geq \phi_{max}^{-1}[p_0]D_M$ in the usual sense that $(X_M'X_M)^+ - \phi_{max}^{-1}[p_0]D_M$ is positive semi-definite,

$$R(\mathfrak{M}_{p_0}, \sigma^2(X_M'X_M)^+) \geq R(\mathfrak{M}_{p_0}, \sigma^2\phi_{max}^{-1}[p_0]D_M) \tag{25}$$

(see, e.g., Lemma 4.27 of Johnstone [18]).

No estimator of $\boldsymbol{\beta}_M$ in (24) obviously cannot outperform the oracle estimator that knows the true $\boldsymbol{\beta}_M \in \mathfrak{M}_{p_0}$ whose ideal minimal quadratic risk is $\sum_{j=1}^p \min(\beta_{M,j}^2, \sigma^2\phi_{max}^{-1}[p_0])$ (e.g., Donoho and Johnstone [11]). Hence,

$$\begin{aligned} R(\mathfrak{M}_{p_0}, \sigma^2\phi_{max}^{-1}[p_0]D_M) &\geq \sup_{\boldsymbol{\beta}_M : M \in \mathfrak{M}_{p_0}} \sum_{j=1}^p \min(\beta_{M,j}^2, \sigma^2\phi_{max}^{-1}[p_0]) \\ &\geq \sigma^2\phi_{max}^{-1}[p_0]\ p_0 \end{aligned} \tag{26}$$

Combining (23), (25) and (26) implies (21).

Consider now $1 \leq p_0 \leq r/2$ and show that in this case, in addition to the lower bound (21),

$$\inf_{\tilde{\mathbf{y}}} \sup_{\boldsymbol{\beta}_M : M \in \mathfrak{M}_{p_0}} \mathbb{E}||\tilde{\mathbf{y}} - X\boldsymbol{\beta}_M||^2 \geq C\sigma^2\tau[2p_0]\frac{\ln m(p_0)}{\max(1, \ln p_0)}$$

Let

$$s^2(p_0) = C\sigma^2\tau[2p_0]\frac{\ln m(p_0)}{\max(1,\ln p_0)} \tag{27}$$

The core of the proof is to find a subset $\mathcal{B}_{p_0}$ of vectors $\boldsymbol{\beta}_M, M \in \mathfrak{M}_{p_0}$ and the corresponding subset of mean vectors $\mathcal{G}_{p_0} = \{\mathbf{g} \in \mathbb{R}^n : \mathbf{g} = X\boldsymbol{\beta}_M,\ \boldsymbol{\beta}_M \in \mathcal{B}_{p_0}\}$ such that for any $\mathbf{g}_1,\ \mathbf{g}_2 \in \mathcal{G}_{p_0}$, $||\mathbf{g}_1 - \mathbf{g}_2||^2 \geq 4s^2(p_0)$ and the Kullback-Leibler divergence $K(\mathbb{P}_{\mathbf{g}_1}, \mathbb{P}_{\mathbf{g}_2}) = \frac{||\mathbf{g}_1-\mathbf{g}_2||^2}{2\sigma^2} \leq (1/16)\ln \mathrm{card}(\mathcal{G}_{p_0})$. Lemma A.1 of Bunea *et al.* [8] will imply then that $s^2(p_0)$ is the minimax lower bound over $\mathfrak{M}_{p_0}$.

The standard techniques for constructing such sets of vectors for the complete variable selection setup is based on generalizations of Varshamov-Gilbert bound (e.g, Abramovich and Grinshtein [2], Rigollet and Tsybakov [22], Raskutti *et al.* [21]). Unfortunately, it cannot be applied when there are additional structural constraints on the set of admissible models. We utilize instead the recent combinatorial results of Gutin and Jones [16].

Define the subset $\tilde{\mathcal{B}}_{p_0}$ of all vectors $\boldsymbol{\beta}_M, M \in \mathcal{M}_{p_0} \subset \mathfrak{M}_{p_0}$ that have p_0 entries equal to C_{p_0} defined later, while the remaining entries are zeros: $\tilde{\mathcal{B}}_{p_0} = \{\boldsymbol{\beta}_M : \boldsymbol{\beta}_M \in \{\{0, C_{p_0}\}^p\},\ M \in \mathcal{M}_{p_0}\}$. Let $\rho(\boldsymbol{\beta}_{1M}, \boldsymbol{\beta}_{2M}) = \sum_{j=1}^{p} \mathbb{I}\{\beta_{1M,j} \neq \beta_{2M,j}\}$ be the Hamming distance between $\boldsymbol{\beta}_{1M},\ \boldsymbol{\beta}_{2M} \in \tilde{\mathcal{B}}_{p_0}$ and define

$$\rho_{max} = \max_{\boldsymbol{\beta}_{1M},\, \boldsymbol{\beta}_{2M} \in \tilde{\mathcal{B}}_{p_0}} \rho(\boldsymbol{\beta}_{1M}, \boldsymbol{\beta}_{2M}) \quad \text{and} \quad \rho_{min} = \min_{(\boldsymbol{\beta}_{1M} \neq \boldsymbol{\beta}_{2M}) \in \tilde{\mathcal{B}}_{p_0}} \rho(\boldsymbol{\beta}_{1M}, \boldsymbol{\beta}_{2M})$$

Theorem 2 of Gutin and Jones [16] ensures that for any constant $C > 2$ there exists a subset $\mathcal{B}_{p_0} \subset \tilde{\mathcal{B}}_{p_0}$ such that

$$\frac{\rho_{max}}{\rho_{min}} \leq C \quad \text{and} \quad \ln \mathrm{card}(\mathcal{B}_{p_0}) \geq \alpha \ln m(p_0),$$

where

$$\alpha = \left\lceil \frac{\ln(p_0/2)}{\ln(C/2)} \right\rceil^{-1}$$

Consider the corresponding subset of mean vectors $\mathcal{G}_{p_0}$, where $\mathrm{card}(\mathcal{G}_{p_0}) = \mathrm{card}(\mathcal{B}_{p_0})$. For any $\mathbf{g}_1,\ \mathbf{g}_2 \in \mathcal{G}_{p_0}$ and the associated with them $\boldsymbol{\beta}_{1M},\ \boldsymbol{\beta}_{2M} \in \mathcal{B}_{p_0}$ we then have

$$||\mathbf{g}_1 - \mathbf{g}_2||^2 = ||X(\boldsymbol{\beta}_{1M} - \boldsymbol{\beta}_{2M})||^2 \geq \phi_{min}[2p_0]\ ||\boldsymbol{\beta}_{1M} - \boldsymbol{\beta}_{2M}||^2 \geq \phi_{min}[2p_0]C_{p_0}^2\rho_{min} \tag{28}$$

On the other hand, by similar arguments, the Kullback-Leibler divergence satisfies

$$K(\mathbb{P}_{\mathbf{g}_1}, \mathbb{P}_{\mathbf{g}_2}) \leq \frac{\phi_{max}[2p_0]C_{p_0}^2\rho(\boldsymbol{\beta}_{1M}, \boldsymbol{\beta}_{2M})}{2\sigma^2} \leq \frac{\phi_{max}[2p_0]C_{p_0}^2\rho_{max}}{2\sigma^2} \tag{29}$$

Set now

$$C_{p_0}^2 = \frac{\sigma^2\alpha \ln m(p_0)}{8\rho_{\max}\phi_{max}[2p_0]}$$

Then, (28) and (29) yield $||\mathbf{g}_1 - \mathbf{g}_2||^2 \geq \tau[2p_0]\sigma^2\alpha \ln m(p_0)/(8C)$, $K(\mathbb{P}_{\mathbf{g}_1}, \mathbb{P}_{\mathbf{g}_2}) \leq (1/16)\ln \mathrm{card}(\mathcal{G}_{p_0})$, and Lemma A.1 of Bunea *et al.* [8] with $s^2(p_0)$ from (27) completes the proof. □

6.3. Proof of Lemma 1

1. $3 \le k \le \frac{3}{2}K$
Consider first all models that include any $\lfloor \frac{k}{3} \rfloor$ paired interactions and the corresponding main effects. Evidently, the size of any such a model is at most k. If it is less than k (it happens when the same main effect appears in several interactions), we complete it to k by adding other main effects from the remaining ones (one can easily verify that there are enough remaining main effects since $k - \lfloor \frac{k}{3} \rfloor \le K$). We then have

$$m(k) \ge \binom{\frac{K(K-1)}{2}}{\lfloor \frac{k}{3} \rfloor}$$

and, therefore, by straightforward calculus,

$$\ln m(k) \ge \left\lfloor \frac{k}{3} \right\rfloor \ln \left(\frac{K(K-1)}{2\lfloor \frac{k}{3} \rfloor} \right) \ge \left\lfloor \frac{k}{3} \right\rfloor \ln \left(\frac{K(K+1)}{2k} \right) = \left\lfloor \frac{k}{3} \right\rfloor \ln(p/k)$$

2. $\frac{3}{2}K < k \le r - 1$
For this case we consider models of size k that include all K main effects and any $k - K$ paired interactions. Thus,

$$m(k) \ge \binom{\frac{K(K-1)}{2}}{k-K}$$

and

$$\ln m(k) \ge (k-K) \ln \left(\frac{K(K-1)}{2(k-K)} \right) \ge \frac{k}{3} \ln \left(\frac{K(K-1)}{2(k-K)} \right)$$

To complete the proof one can easily verify that $\frac{K(K-1)}{2(k-K)} \ge \frac{K(K+1)}{2k}$. □

References

[1] Abramovich, F., Benjamini, Y., Donoho, D.L. and Johnstone, I.M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.* **34**, 584–653. MR2281879

[2] Abramovich, F. and Grinshtein, V. (2010). MAP model selection in Gaussian regression. *Electr. J. Statist.* **4**, 932–949. MR2721039

[3] Abramovich, F., Grinshtein, V. and Pensky, M. (2007). On optimality of Bayesian testimation in the normal means problem. *Ann. Statist.* **35**, 2261–2286. MR2363971

[4] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. in *Second International Symposium on Information Theory.* (eds. B.N. Petrov and F. Czáki). Akademiai Kiadó, Budapest, 267–281. MR0483125

[5] Bien, J., Taylor, J. and Tibshirani, R. (2013). A Lasso for hierarchical interactions. *Ann, Statist., to appear.*

[6] Birgé, L. and Massart, P. (2001). Gaussian model selection. *J. Eur. Math. Soc.* **3**, 203–268. MR1848946
[7] Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probab. Theory Relat. Fields* **138**, 33–73. MR2288064
[8] Bunea, F., Tsybakov, A. and Wegkamp, M.H. (2007). Aggregation for Gaussian regression. *Ann. Statist.* **35**, 1674–1697. MR2351101
[9] Chipman, H. (1996). Bayesian variable selection with related predictors. *Canad. J. Statist.* **24**, 17–36. MR1394738
[10] Chipman, H., George, E.I. and McCullogh, R.E. (2001). *The Practical Implementation of Bayesian Model Selection. IMS Lecture Notes – Monograph Series* **38**. MR2000752
[11] Donoho, D. and Johnstone, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–455. MR1311089
[12] Farcomeni, A. (2010). Bayesian constrained variable selection. *Statistica Sinica* **20**, 1043–1062. MR2729852
[13] Foster, D.P. and George, E.I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22**, 1947–1975. MR1329177
[14] George, E.I. and McCullogh, R.E. (1993). Variable selection via Gibbs sampling. *J. Am. Statist. Assoc.* **88**, 881–889.
[15] George, E.I. and McCullogh, R.E. (1997). Approaches to Bayesian variable selection. *Statistica Sinica* **7**, 339–373.
[16] Gutin, G. and Jones, M. (2012). Note on large subsets of binary vecrtors with similar distances. *SIAM J. Discrete Math.* **26**, 1108-1111.
[17] Hall, P. and Jin, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *Ann. Statist.* **38**, 1681–1732. MR2662357
[18] Johnstone, I.M. (2011). *Function Estimation and Gaussian Sequence Models*, *unpublished manuscript.*
[19] Lounici, K., Pontil, M., Tsybakov, A. and van de Geer, S. (2011). Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.* **39**, 2164–2204. MR2893865
[20] Mallows, C.L. (1973). Some comments on C_p. *Technometrics* **15**, 661–675.
[21] Raskutti, G., Wainwright, M.J. and Yu, B. (2011). Minimax rates of estimation for high-dimensional regression over l_q balls. *IEEE Trans. Inform. Theory* **57**, 6976–6994. MR2882274
[22] Rigollet, P. and Tsybakov, A. (2011). Exponential screening and optimal rates of sparse estimation. *Ann. Statist.* **39**, 731–771. MR2816337
[23] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464. MR0468014
[24] Tropp, J.A. and Wright, S.J. (2010). Computational methods for sparse solution of linear inverse problems. *Proc. IEEE, special issue "Applications of sparse representation and compressive sensing".*
[25] Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* (eds. Goel, P.K. and Zellner, A.), North-Holland, Amsterdam, 233–243. MR0881437