MODEL SELECTION IN GAUSSIAN REGRESSION FOR HIGH-DIMENSIONAL DATA

Felix Abramovich and Vadim Grinshtein

Abstract We consider model selection in Gaussian regression, where the number of predictors might be even larger than the number of observations. The proposed procedure is based on penalized least square criteria with a complexity penalty on a model size. We discuss asymptotic properties of the resulting estimators corresponding to linear and so-called $2k\ln(p/k)$-type nonlinear penalties for nearly-orthogonal and multicollinear designs. We show that any linear penalty cannot be simultaneously adapted to both sparse and dense setups, while $2k\ln(p/k)$-type penalties achieve the wide adaptivity range. We also present Bayesian perspective on the procedure that provides an additional insight and can be used as a tool for obtaining a wide class of penalized estimators associated with various complexity penalties.

1 Introduction

Modern statistics encounters new challenges, where the problems have exploded both in size and complexity. Analysis of complex high-dimensional data sets of very large sizes requires a new look on traditional statistical methods.

Consider the standard Gaussian linear regression setup

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

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 p explanatory variables (predictors) $X_1,\ldots,X_p$, $\beta \in \mathbb{R}^p$ is a vector of un-

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

Vadim Grinshtein
Department of Mathematics, The Open University of Israel, Raanana 43107, Israel, e-mail: vadimg@openu.ac.il

known regression coefficients, $\varepsilon \sim N(\mathbf{0}, \sigma^2 I_n)$ and the noise variance σ^2 is assumed to be known.

The number of predictors p might be very large relatively even to the amount of available data n that causes a severe "curse of dimensionality" problem. However, it is usually believed that only a small fraction of them has a truly relevant impact on the response. Thus, the problem of model (or variable) selection for reduction dimensionality in (1) becomes of fundamental importance. Its main goal is to select the "best", parsimonious subset of predictors (model) among $X_1,...,X_p$. For a selected model M, the corresponding coefficients β_M are then typically estimated by least squares. The definition of the "best" subset however depends on the particular aim at hand. One should distinguish, for example, between estimating regression coefficients β, estimating the mean vector $X\beta$, identifying non-zero coefficients and predicting future observations. Different aims may lead to different optimal model selection procedures especially for the "p larger than n" setup. In this paper we focus on estimating the mean vector $X\beta$ and the goodness of a given model M is measured by the quadratic risk $E||X\hat{\beta}_M - X\beta||^2 = ||X\beta_M - X\beta||^2 + \sigma^2|M|$, where $X\beta_M$ is the projection of $X\beta$ on the span of M and $\hat{\beta}_M$ is the least square estimate of β_M. The bias term represents the approximation error of the projection, while the variance term is the price for estimating the projection coefficients β_M by $\hat{\beta}_M$ and is proportional to the model size. 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.

Since the above criterion involves the unknown β, the corresponding ideal minimal risk can be rather used as a benchmark for any available model selection procedure. Typical model selection criterion is based on the *empirical* quadratic risk $||\mathbf{y} - X\hat{\beta}_M||^2$, which is essentially the least squares. The empirical risk obviously decreases as the model size grows and to avoid overfitting, it is penalized by a complexity penalty $Pen(|M|)$ that increases with $|M|$. This leads to the *penalized* least square criterion of the form

$$||\mathbf{y} - X\hat{\beta}_M||^2 + Pen(|M|) \rightarrow \min_M \qquad (2)$$

The properties of the resulting estimator depends on the proper choice of the complexity penalty $Pen(\cdot)$ in (2). A large amount of works has studied various types of penalties. The most commonly used choice is a *linear* type penalty of the form $Pen(k) = 2\sigma^2\lambda k$ for some fixed $\lambda > 0$. The most known examples motivated by a wide variety of approaches include C_p (Mallows, 1973) and AIC (Akaike, 1974) for $\lambda = 1$, BIC (Schwarz, 1978) for $\lambda = (\ln n)/2$ and RIC (Foster & George, 1994) for $\lambda = \ln p$. On the other hand, a series of recent works suggested the so-called $2k\ln(p/k)$-type *nonlinear* complexity penalties of the form

$$Pen(k) = 2\sigma^2 ck(\ln(p/k) + \zeta_{p,k}), \qquad (3)$$

where $c > 1$ and $\zeta_{p,k}$ is some “negligible” term (see, e.g., Birgé & Massart, 2001, 2007; Johnstone, 2002; Abramovich *et al.*, 2006; Bunea, Tsybakov & Wegkamp, 2007; Abramovich & Grinshtein, 2010).

In this paper we discuss the asymptotic properties of linear and $2k\ln(p/k)$-type penalized estimators (2) as both the sample size n and the number of predictors p increase. We distinguish between two different types of the design: *nearly-orthogonal*, where there is no strong collinearity between predictors, and *multicollinear*, that usually appears when $p \gg n$. Interesting, that the minimax rates for estimating the mean vector for multicollinear design are faster than those for nearly-orthogonal by a certain factor depending on the design properties. Such a phenomenon can be explained by a possibility of exploiting strong correlations between predictors to reduce the model size without paying much extra price in the bias.

We show that even for nearly-orthogonal design any linear penalty cannot be simultaneously optimal (in the minimax sense) for both sparse and dense cases. On the contrary, the $2k\ln(p/k)$-types penalties achieve the widest possible adaptivity range. Moreover, under some additional assumptions on the design and regression coefficients vector, they remain asymptotically optimal for the multicollinear design as well.

We also describe a Bayesian interpretation of penalized estimators (2) developed in Abramovich & Grinshtein (2010) for a general case and for the considered two types of penalties in particular. Bayesian approach provides an additional insight in these estimators and can be also used as a tool for obtaining a wide class of penalized estimators with various complexity penalties.

The paper is organized as follows. The notations, definitions and some preliminary results are given in Section 2, where, in particular, we present the (nonasymptotic) minimax lower bound for the risk of estimating the means vector $X\beta$ in (1). The asymptotic minimax properties of penalized estimators (2) for nearly-orthogonal and multicollinear designs are investigated respectively in Sections 3 and 4. Section 5 presents a Bayesian perspective on (2). Some concluding remarks are given in Section 6.

2 Preamble

Consider the general linear regression setup (1), where the number of possible predictors p may 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 $d_{jM} = \mathbb{I}\{X_j \in M\}$ and, therefore, $|M| = tr(D_M)$. For a given M, the least square estimate of its coefficients is $\hat{\beta}_M = (D_M X' X D_M)^+ D_M X' \mathbf{y}$, where “+” denotes the generalized inverse matrix.

For a fixed p_0, define the sets of models $\mathscr{M}_{p_0}$ that have at most p_0 predictors, that is, $\mathscr{M}_{p_0} = \{M : |M| \leq p_0\}$. Obviously, if a true model in (1) belongs to $\mathscr{M}_{p_0}$, then $||\beta||_0 \leq p_0$, where the l_0 quasi-norm of the coefficients vector β is defined as the number of its nonzero entries. We consider $p_0 \leq r$ since otherwise, there necessarily exists another vector β^* such that $||\beta^*||_0 \leq r$ and $X\beta = X\beta^*$.

Within the minimax framework, the performance of a penalized estimator $X\hat{\beta}_{\hat{M}}$ of the unknown mean vector $X\beta$ in (1) corresponding to the selected model $\hat{M}$ with respect to (2) over $\mathscr{M}_{p_0}$ is measured by its worst-case quadratic risk $\sup_{\beta:||\beta||_0 \leq p_0} E||X\hat{\beta}_{\hat{M}} - X\beta||^2$. It is then compared to the minimax risk – the best attainable worst-case risk among all possible estimators, $R(\mathscr{M}_{p_0}) = \inf_{\hat{\mathbf{y}}} \sup_{\beta:||\beta||_0 \leq p_0} E||\hat{\mathbf{y}} - X\beta||^2$.

We present first the following result of Abramovich & Grinshtein (2010) for the lower bound for the minimax risk $R(\mathscr{M}_{p_0})$.

For any given $k = 1,...,r$, let $\phi_{min}[k]$ and $\phi_{max}[k]$ be the k-sparse minimal and maximal eigenvalues of the design defined as

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

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

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

Theorem 1. *(Abramovich & Grinshtein, 2010). Consider the model (1) and let* $1 \leq p_0 \leq r$*. There exists a universal constant* $C > 0$ *such that*

$$R(\mathscr{M}_{p_0}) \geq \begin{cases} C_2\sigma^2\tau[2p_0]\; p_0(\ln(p/p_0)+1) & , 1 \leq p_0 \leq r/2 \\ C_2\sigma^2\tau[p_0]\; r & , r/2 \leq p_0 \leq r \end{cases} \quad (4)$$

Note that the minimax lower bound (4) depends on a design matrix X only through its sparse eigenvalues ratios. Computationally simpler but less accurate minimax lower bound can be obtained by replacing $\tau[2p_0]$ and $\tau[p_0]$ in (4) by $\tau[r]$, that for the case $r = p \leq n$ is just the ratio of the minimal and maximal eigenvalues of $X'X$.

Consider now the asymptotics as the sample size n increases. We allow $p = p_n$ to increase with n as well in such a way that r tends to infinity and look for a projection of the unknown mean vector on an expanding span of predictors. In the “classical” regression setup, $p_n = o(n)$, while in the “modern” one, p_n may be larger than n or even $p_n \gg n$.

In such asymptotic setting one should essentially consider a *sequence* of design matrices X_{n,p_n}, where $r_n \longrightarrow \infty$. For simplicity of exposition, in what follows the index n is omitted and X_{n,p_n} will be denoted by X_p emphasizing the dependence on the number of predictors p when r tend to infinity. Similarly, we consider now sequences of corresponding coefficients vectors β_p. In these notations, the original model (1) is transformed into a sequence of models

$$\mathbf{y} = X_p \boldsymbol{\beta}_p + \varepsilon, \tag{5}$$

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

The minimax lower bound (4) indicates that depending on the asymptotic behavior of the sparse eigenvalues ratios, one should distinguish between nearly-orthogonal and multicollinear designs:

Definition 1. Consider the sequence of design matrices X_p. The design is called nearly-orthogonal if the corresponding sequence of sparse eigenvalues ratios $\tau_p[r]$ is bounded away from zero by some constant $c > 0$. Otherwise, the design is called multicollinear.

Nearly-orthogonality assumption essentially means that there is no collinearity in the design in the sense that there are no "too strong" linear relationships within any set of r columns of X_p. It is intuitively clear that it can happen only when p is not "too large" relative to r (and hence to n), while for the $p_n \gg n$ setup, multicollinearity between predictors is inherent. Indeed, Abramovich & Grinshtein (2010) showed that for nearly-orthogonal design necessarily $p = O(r)$ and, thefore, $p = O(n)$.

In what follows we consider separately the two types of the design and investigate the asymptotic optimality (in the minimax sense) of linear and $2k\ln(p/k)$-type penalties.

3 Nearly-orthogonal design

From the definition of nearly-orthogonal design it follows that there exists a constant $c > 0$ such that $c \leq \tau_p[r] \leq ... \leq \tau_p[1] = 1$. In addition, as we have mentioned in the previous Section 2, for this type of design $p = O(r)$ and, therefore, the minimax lower bound (4) over $\mathscr{M}_{p_0}$ in this case is essentially of the order $p_0(\ln(p/p_0)+1)$ for all $p_0 = 1, ..., r$.

We start from linear penalties, where $Pen(k) = 2\sigma^2\lambda_p k$. Foster & George (1994) and Birgé & Massart (2001, Section 5.2) showed that the best possible risk of corresponding penalized estimators over $\mathscr{M}_{p_0}$ is of the order $\sigma^2 p_0 \ln p$ achieved for $\lambda_p = (1+\delta)\ln p,\ \delta > 0$ corresponding to the RIC criterion. This risk is of the same order as $p_0(\ln(p/p_0)+1)$ in the minimax lower bound (4) when $p_0 = O(r^\alpha)$ for some $0 < \alpha < 1$ (sparse cases), but higher than the latter for the dense cases, where $p_0 \sim r$. On the other hand, it is the AIC estimator ($\lambda_p = 1$) with the risk of the order $\sigma^2 p$, that is asymptotically similar to (4) for dense but much higher for sparse cases. In other words, no penalized estimator (2) with a linear penalty can be simultaneously rate-optimal for both sparse and dense cases. Note that a linear penalty $Pen(k) = 2\sigma^2\lambda_p k$ yields the *constant* per predictor price $2\sigma^2\lambda_p$ that cannot be adapted to both cases.

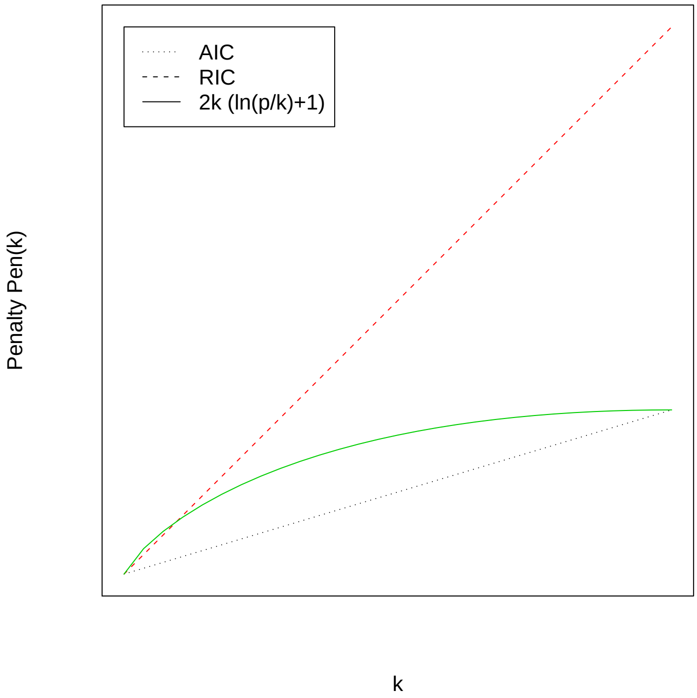

Fig. 1 Various penalties: AIC (dotted line), RIC (dashed line) and $2k(\ln(p/k)+1)$ (solid line).

On the other hand, the nonlinear penalties of the $2k\ln(p/k)$-type imply *different* per predictor price: higher for small models but decreasing as the model size grows. In fact, such type of penalty behaves like RIC for sparse and AIC for dense cases (see Figure 1). As we shall show, it allows the corresponding estimators to achieve the widest adaptivity range.

Consider a general $2k\ln(p/k)$-type penalty (3), where $c > 1$. From the results of Birgé & Massart (2001, 2007) and Abramovich & Grinshtein (2010) for the corresponding penalized estimators (2) it follows that for any $1 \leq p_0 \leq r$,

$$\sup_{\beta:||\beta||_0 \leq p_0} E||X\hat{\beta}_{\hat{M}} - X\beta||^2 \leq C\sigma^2 p_0(\ln(p/p_0)+1) \tag{6}$$

for some $C > 0$. Comparing the risk upper bound (6) with the minimax lower bound implies the following Corollary:

Corollary 1. *Let the design be nearly-orthogonal. Consider the penalized least square estimation (2) with a $2k\ln(p/k)$-type complexity penalty (3), where $c > 1$. Then, as $r \to \infty$, the corresponding penalized estimator attains the minimax convergence rates simultaneously over all $\mathscr{M}_{p_0},\ p_0 = 1,\ldots,r$.*

Furthermore, for sparse cases Birgé & Massart (2007) showed that for $c < 1$ in (3), the risk of the corresponding penalized estimator is much larger than that in (6). The value $c = 1$ for the $2k\ln(p/k)$-type penalty (3) is, therefore, a borderline. For the *orthogonal* design and various sparse settings, Abramovich *et al.* (2006) and Wu & Zhou (2009) proved that $c = 1$ yields even *sharp* (with an exact constant) asymptotic minimaxity. However, to the best of our knowledge, it is not clear what happens for the choice $c = 1$ in a general case.

Finally, note that for the nearly-orthogonal design, $||X_p\hat{\beta}_{p\hat{M}} - X_p\beta_p|| \asymp ||\hat{\beta}_{p\hat{M}} - \beta_p||$, where "$\asymp$" means that their ratio is bounded from below and above. Therefore, all the results of this section for estimating the mean vector $X_p\beta_p$ in (5) can be straightforwardly applied for estimating the regression coefficients β_p. This equivalence, however, does not hold for the multicollinear design considered below.

4 Multicollinear design

Recall that nearly-orthogonality assumption necessarily implies $p = O(n)$. Thus, it may be reasonable in the "classical" setup, where p is not too large relatively to n but is questionable for the analysis of high-dimensional data, where $p \gg n$. In this section we investigate the performance of $2k\ln(p/k)$-type penalties for the multicollinear design.

When nearly-orthogonality does not hold, the sparse eigenvalues ratios in (4) may tend to zero as p increases and, thus, decrease the minimax lower bound rate relatively to the nearly-orthogonal design. In this case there is a gap between the rates in the lower and upper bounds (4) and (6). Intuitively, strong correlations between predictors can be exploited to diminish the size of a model (hence, to decrease the variance) without paying much extra price in the bias, and, therefore, to reduce the overall risk. It turns out that under certain additional assumptions on the design and the regression coefficients vector in (5) given below, the upper risk bound (6) of the $2k\ln(p/k)$-type estimator can be indeed reduced to the minimax lower bound rate (4). Under these conditions, $2k\ln(p/k)$-type penalized estimator, therefore, remains asymptotically rate-optimal even for the multicollinear design.

For simplicity of exposition we consider $p_0 \leq r/2$ although the results for $r/2 \leq p_0 \leq r$ can be obtained in a similar way with necessary changes. In particular, for the latter case one should slightly modify the $2k\ln(p/k)$-type penalty for $k = r$ to be of the form $Pen(r) \sim 2\sigma^2 cr$ for some $c > 0$ (Abramovich & Grinshtein, 2010). Note that for the nearly-orthogonal design, where $p = O(r)$, $2k\ln(p/k)$-type penalties automatically imply this condition on $Pen(r)$.

For any model M of size $k \leq r/2$ let X_M be the $n \times k$ submatrix X_M containing the corresponding k columns of X_p. Consider the matrix $(X_M'X_M)^{-1}$ and the maximum of minimal eigenvalues $\phi_{min}[(X_M'X_M)^{-1}]_{M'}$ of all its symmetric $\lfloor k(1-\tau_p[2k])\rfloor \times \lfloor k(1-\tau_p[2k])\rfloor$ submatrices corresponding to various submodels $M' \subset M$ of size $\lfloor k(1-\tau_p[2k])\rfloor$. Define $\tilde{\phi}_p[k] = \min_M \phi_{min}[(X_M'X_M)^{-1}]_{M'}$, that is,

$$\tilde{\phi}_p[k] = \min_{M:|M|=k} \max_{\substack{M' \subset M \\ |M'| = \lfloor k(1-w_p[k]) \rfloor}} \phi_{min}[(X'_M X_M)^{-1}]_{M'}$$

It can be shown (Abramovich & Grinshtein, 2010) that $\tilde{\phi}_p^{-1}[k]$ measures an error of approximating mean vectors $X_p\beta_p$, where $||\beta_p||_0 = k$, by their projections on lower dimensional subspans of predictors. The stronger is multicollinearity, the better is the approximation and the larger is $\tilde{\phi}_p[k]$.

The following theorem is a consequence of Theorem 5 of Abramovich & Grinshtein (2010):

Theorem 2. *Let $\tau_p[r] \to 0$ as $r \to \infty$ (multicollinear design). Assume the following additional assumptions on the design matrix X_p and the (unknown) vector of coefficients β_p in (5):*

(D) for all p there exist $1 \le \kappa_{p1} \le \kappa_{p2} \le r/2$ such that

1. $\tilde{c}_1 \le \tau_p[2k] \cdot k \le k-1,\ k = \kappa_{p1},...,\kappa_{p2}$
2. $\tau_p[2\kappa_{p2}] \ge (\kappa_{p2}/(pe))^{\tilde{c}_2}$
3. $\phi_{p,min}[2k] \cdot \tilde{\phi}_p[k] \ge \tilde{c}_3,\ k = \kappa_{p1},...,\kappa_{p2}$

(B) $||\beta_p||_\infty^2 \le \tilde{c}_4 \tau_p[2p_0] \cdot \tilde{\phi}_p[p_0] \cdot (\ln(p/p_0)+1)$*, where* $p_0 = ||\beta_p||_0$

for some positive constants $\tilde{c}_1,\ \tilde{c}_2,\ \tilde{c}_3$ and $\tilde{c}_4$.

Then, the penalized least square estimator (2) with a $2k\ln(p/k)$-type complexity penalty (3), where $c > 1$, is asymptotically simultaneously minimax (up to a constant multiplier) over all $\mathscr{M}_{p_0},\ \ \kappa_{p1} \le p_0 \le \kappa_{p2}$.

Generally, Assumptions (D.1, D.2) and Assumption (B) allow one to reduce the upper bound (6) for the risk of the $2k\ln(p/k)$-type estimator by the factor $\tau_p[2p_0]$, while Assumption (D.3) is required to guarantee that the additional constraint on β_p in Assumption (B) does not affect the lower bound (4). We have mentioned that multicollinearity typically arises when $p \gg n$. One can easily verify that for $n = O(p^\alpha),\ 0 \le \alpha < 1$, Assumption (D.2) always follows from Assumption (D.1) and, therefore, can be omitted in this case.

5 Bayesian perspective

In this section we discuss the Bayesian approach to model selection in the Gaussian regression model (1) proposed by Abramovich & Grinshtein (2010). Bayesian framework naturally interpretates the penalized least square estimation (2) by treating the penalty term as proportional to the logarithm of a prior distribution on the model size. Minimization of (2) corresponds then to the maximum *a posteriori* (MAP) rule. Choosing different types of a prior, the resulting Bayesian MAP estimator can imply various complexity penalties, linear and $2k\ln(p/k)$-type penalties in particular, that gives an additional insight in motivation behind such types of penalties.

Consider the model (1), where the number of possible predictors p may be larger then the number of observations n. Recall that $r = rank(X)$ and we assume that any r columns of X are linearly independent.

Assume some prior on the model size $\pi(k) = P(|M| = k)$, where $\pi(k) > 0,\ k = 0,...,r$ and $\pi(k) = 0$ for $k > r$ since the model becomes nonidentifiable when the number of its parameters is larger than the number of observations (see Section 2).

For any $k = 0,...,r-1$, assume all $\binom{p}{k}$ various models of size k to be equally likely, that is, conditionally on $|M| = k$,

$$P(M \mid |M| = k) = \binom{p}{k}^{-1}$$

The case $k = r = rank(X)$ is slightly different. Although there are $\binom{p}{r}$ various sets of predictors of size r, all of them evidently result in the same estimator for the mean vector and, in this sense, are essentially undistinguishable and associated with a *single* (saturated) model. Hence, in this case, we set

$$P(M \mid |M| = r) = 1 \tag{7}$$

Finally, assume the normal prior on the unknown vector of k coefficients of the model M: $\beta_M \sim N_p(\mathbf{0}, \gamma\sigma^2(D_M X'XD_M)^+)$, where $\gamma > 0$ and the diagonal indicator matrix D_M was defined in Section 2. This is a well-known conventional g-prior of Zellner (1986).

For the proposed hierarchical prior, a straightforward calculus yields the posterior probability of a model M of size $|M| = 0,...,r-1$:

$$P(M|\mathbf{y}) \propto \pi(|M|)\binom{p}{|M|}^{-1}(1+\gamma)^{-\frac{|M|}{2}} \exp\left\{\frac{\gamma}{\gamma+1}\frac{\mathbf{y}'XD_M(D_MX'XD_M)^+D_MX'\mathbf{y}}{2\sigma^2}\right\} \tag{8}$$

Finding the most likely model leads then to the following MAP model selection criterion:

$$\mathbf{y}'XD_M(D_MX'XD_M)^+D_MX'\mathbf{y} + 2\sigma^2(1+1/\gamma)\ln\left\{\binom{p}{|M|}^{-1}\pi(|M|)(1+\gamma)^{-\frac{|M|}{2}}\right\} \to \max_M$$

or, equivalently,

$$||\mathbf{y} - X\hat{\beta}_M||^2 + 2\sigma^2(1+1/\gamma)\ln\left\{\binom{p}{|M|}\pi(|M|)^{-1}(1+\gamma)^{\frac{|M|}{2}}\right\} \to \min_M, \tag{9}$$

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

$$Pen(k) = 2\sigma^2(1+1/\gamma)\ln\left\{\binom{p}{k}\pi(k)^{-1}(1+\gamma)^{\frac{k}{2}}\right\}, \quad k = 0,...,r-1 \tag{10}$$

Similarly, for $|M| = r$ from (7) one has

$$Pen(r) = 2\sigma^2(1+1/\gamma)\ln\left\{\pi(r)^{-1}(1+\gamma)^{\frac{r}{2}}\right\} \qquad (11)$$

In particular, the (truncated if $p > r$) binomial prior $B(p,\xi)$ corresponds to the prior assumption that the indicators d_{jM} are independent. The binomial prior yields the linear penalty $Pen(k) = 2\sigma^2(1+1/\gamma)k\ln(\sqrt{1+\gamma}(1-\xi)/\xi) \sim 2\sigma^2 k\ln(\sqrt{\gamma}(1-\xi)/\xi),\ k = 1,\ldots,r-1$ for sufficiently large variance ratio γ. The AIC criterion corresponds then to $\xi \sim \sqrt{\gamma}/(e+\sqrt{\gamma})$, while $\xi \sim \sqrt{\gamma}/(p+\sqrt{\gamma})$ leads to the RIC criterion. These relations again confirm our previous arguments in Section 3 that RIC should be appropriate for sparse cases, where the size of the true (unknown) model is believed to be much less than the number of possible predictors, while AIC is suitable for dense cases, where they are of the same order. Any binomial prior or, equivalently, any linear penalty cannot “kill two birds with one stone”.

On the other hand, there is a class of priors associated with the $2k\ln(p/k)$-type penalties. In particular, the (truncated) geometric distribution $\pi(k) \propto q^k,\ k = 1,\ldots,r$ yields $Pen(k) \sim 2\sigma^2(1+1/\gamma)k(\ln(p/k)+\zeta(\gamma,q)),\ k = 1,\ldots,r-1$, where we used that $k\ln(p/k) \leq \ln(\binom{p}{k}) < k(\ln(p/k)+1)$ (see Lemma 1 of Abramovich *et al.*, 2010). In addition, (11) implies $Pen(r) = 2\sigma^2 c(q,\gamma)r$ for some constant $c(q,\gamma) > 1$ that goes along the lines with the remark on the requirement on $Pen(r)$ for $2k\ln(p/k)$-type penalties in Section 4.

The Bayesian interpretation of the complexity penalized estimators can be also exploited for their computations. Generally, minimizing (2) requires an NP-hard combinatorial search over all possible models. To make computations for high-dimensional data feasible in practice, one typically applies either various greedy algorithms (e.g., forward selection) approximating the global solution in (2) by a stepwise sequence of local ones, or convex relaxation methods (e.g., Lasso (Tibshirani, 1996) and Dantzig selector (Candés & Tao, 2007) for linear penalties) replacing the original combinatorial problem by a related convex program. The proposed Bayesian approach allows one instead to use the Gibbs sampler to efficiently generate a sequence of models from the posterior distribution $P(M|\mathbf{y})$ in (8) (see, e.g. George & McCullooch, 1993 for more detail). The key point is that the relevant models with highest posterior probabilities will appear most frequently and can be easily identified even for a generated sample of a relatively small length.

6 Concluding remarks

In this paper we considered model selection in Gaussian linear regression for high-dimensional data, where the number of possible predictors may be even larger than the number of available observations. The procedure is based on minimizing penalized least squares with a penalty on a model size. We discussed asymptotic properties of the resulting estimators corresponding to different types of penalties. Bayesian interpretation allows one to better understand the intuition behind various penalties and provides a natural tool for obtaining a wide class of estimators of this type.

We showed that any linear penalty, including widely used AIC, C_p, BIC and RIC, cannot be simultaneously minimax for both sparse and dense cases. Moreover, the same conclusions are valid for the well-known Lasso (Tibshirani, 1996) and Dantzig (Candés & Tao, 2007) estimators that for the optimally chosen tuning parameter, under nearly-orthogonality conditions similar to those considered in this paper, can achieve only the same sub-optimal rate $p_0 \ln p$ as RIC (Bickel, Ritov & Tsybakov, 2009). These results are, in fact, should not be surprising since both Lasso and Dantzig estimators are essentially based on convex relaxations of $|M| = ||\beta_M||_0$ in the linear complexity penalty in (2) in order to replace the original combinatorial problem by a convex program (see also remarks in the conclusion of Section 5). Thus, Lasso approximates the l_0-norm $||\beta_M||_0$ by the the corresponding l_1-norm $||\beta_M||_1$. On the other hand, the nonlinear $2k\ln(p/k)$-type penalty adapts to both sparse and dense cases.

It is also interesting to note that, unlike model identification or coefficients estimation problems, where multicollinearity is a “curse”, it may become a “blessing” for estimating the mean vector. One can exploit strong correlations between predictors to reduce the size of a model (hence, to decrease the variance) without paying much extra price in the bias.

Acknowledgements The work was supported by Israel Science Foundation (ISF), grant ISF-248/08. Valuable remarks of the two referees are gratefully acknowledged.

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.
2. Abramovich, F. and Grinshtein, V. (2010). MAP model selection in Gaussian regression. Electr. J. Statist. **4**, 932–949.
3. Abramovich, F., Grinshtein, V., Petsa, A. and Sapatinas, T. (2010). On Bayesian testimation and its application to wavelet thresholding. Biometrika **97**, 181–198.
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.
5. Bickel, P., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. Ann. Statist. **35**, 1705–1732.
6. Birgé, L. and Massart, P. (2001). Gaussian model selection. J. Eur. Math. Soc. **3**, 203–268.
7. Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. Probab. Theory Relat. Fields **138**, 33–73.
8. Bunea, F., Tsybakov, A. and Wegkamp, M.H. (2007). Aggregation for Gaussian regression. Ann. Statist. **35**, 1674–1697.
9. Candés, E.J. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. Ann. Statist. **35**, 2313–2351.
10. Foster, D.P. and George, E.I. (1994). The risk inflation criterion for multiple regression. Ann. Statist. **22**, 1947–1975.
11. George, E.I. and McCuloch, R.E. (1993). Variable selection via Gibbls sampling. J. Amer. Statist. Assoc. **88**, 881–889.
12. Mallows, C.L. (1973). Some comments on C_p. Technometrics **15**, 661–675.

13. Schwarz, G. (1978). Estimating the dimension of a model. Ann. Statist. **6**, 461–464.
14. Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. J. Roy. Statist. Soc. Ser. B **58**, 267–288.
15. Wu, Z. and Zhou, H.H. (2010). Model selection and sharp asymptotic minimaxity. Technical Report, Statistics Department, Yale University.
16. 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 Finietti (eds. Goel, P.K. and Zellner, A.), North-Holland, Amsterdam, 233–243.