AN EIGENVALUE APPROACH FOR ESTIMATING THE GENERALIZED CROSS VALIDATION FUNCTION FOR CORRELATED MATRICES

Christos Koukouvinos  
*Dept of Mathematics, National Technical University of Athens, Greece, ckoukouv@math.ntua.gr*

KHALIDE JBILOU  
*Universite du Littoral Cote d'Opale, Khalide.Jbilou@univ-littoral.fr*

Marilena Mitrouli  
*National and Kapodistrian University of Athens, mmitroul@math.uoa.gr*

Ondrej Turek  
*University of Ostrava, Czechia, ondrej.turek@osu.cz*

Follow this and additional works at: [https://repository.uwyo.edu/ela](https://repository.uwyo.edu/ela)

**Recommended Citation**  
DOI: [https://doi.org/10.13001/1081-3810.4090](https://doi.org/10.13001/1081-3810.4090)

This Article is brought to you for free and open access by Wyoming Scholars Repository. It has been accepted for inclusion in Electronic Journal of Linear Algebra by an authorized editor of Wyoming Scholars Repository. For more information, please contact [scholcom@uwyo.edu](mailto:scholcom@uwyo.edu).
AN EIGENVALUE APPROACH FOR ESTIMATING THE GENERALIZED CROSS VALIDATION FUNCTION FOR CORRELATED MATRICES∗

CHRISTOS KOUKOUVINOS†, KHALIDE JBILOU‡, MARILENA MITROULI§, AND ONĐREJ TUREK¶

Abstract. This works proposes a fast estimate for the generalized cross-validation function when the design matrix of an experiment has correlated columns. The eigenvalue structure of this matrix is used to derive probability bounds satisfied by an appropriate index of proximity, which provides a simple and accurate estimate for the numerator of the generalized cross-validation function. The denominator of the function is evaluated by an analytical formula. Several simulation tests performed in statistical models having correlated design matrix with intercept confirm the reliability of the proposed probabilistic bounds and indicate the applicability of the proposed estimate for these models.

Key words. Penalized least squares, Tuning parameter, Extrapolation, Generalized cross-validation.

AMS subject classifications. 15A18, 62J07, 62K15.

1. Introduction. In several applications arising from the field of Statistics, there appears the linear regression model

\[ y = X \cdot b, \]

where \( X = [1, x_1, x_2, \ldots, x_d] \in \mathbb{R}^{n \times (d+1)} \) is the design matrix. The first column of the design matrix is \( 1_n = [1, 1, \ldots, 1]^T \), i.e., the model includes the intercept, which corresponds to the mean effect. The \( j \)-th column of the design matrix is denoted by \( x_j = [x_{1j}, x_{2j}, \ldots, x_{nj}]^T \) and represents the factor \( j \) of the experiment. Vector \( b \in \mathbb{R}^{d+1} \) represents the regression parameter of the model and it is to be estimated. The vector \( y \in \mathbb{R}^n \) is the response vector. The error, which contaminates the model, is assumed to be independent identically distributed multivariate normal of dimension \( n \), with zero mean and with a variance matrix \( \Sigma = \sigma^2_{err} I_n \).

The complicated structure of the design matrix and the appearing correlations among its entries influence the solution of the model. The ordinary least square regression estimator of \( b \) is likely to produce an inaccurate solution. Therefore, penalization is employed in order to achieve better prediction; that is, one needs to obtain minimizers of the model

\[
\min_{\beta \in \mathbb{R}^d} \{ \|y - X\beta\|^2 + \lambda \|\beta\|^2 \},
\]

where \( \lambda \in \mathbb{R} \) is the tuning or regularization parameter.

The specification of an appropriate value for \( \lambda \) is an important issue. Several approaches were developed to handle it [6]. It was proved [3] that the minimizer of the generalized cross-validation (GCV) function provides a good value for \( \lambda \), and thereafter this method is followed in a vast majority of applications. However, this computation is very expensive, being of order \( O(n^3) \). Thus, with a view to overcome this disadvantage, there were proposed methods [1, 10] attaining the minimization of an estimate of the GCV.
Estimating the Generalized Cross Validation Function for Correlated Matrices

instead of its exact formula. In this way, the complexity is reduced to $O(n^2)$. In the present work, we improve a recently proposed extrapolation GCV estimate [8, 9] by analyzing the eigenvalue structure of the involved matrix. For a matrix with highly correlated columns, it is proved that the index of proximity, which arises in the extrapolation procedure, is close to one with high probability. This allows us to obtain an optimum extrapolation estimate for the numerator of the GCV function whereas the denominator can be expressed by an analytical formula.

The paper is organized as follows. Section 2 describes the procedure to derive an optimum extrapolation GCV estimate. In Section 3, we outline an efficient method of estimating the index of proximity [7, 8]. Section 4 is devoted to the study of the index of proximity for linear regression statistical models with intercept and highly correlated covariates. In particular, we estimate the probability that the index of proximity is close to one. Calculations in Section 5 lead to an analytical formula for the denominator of the GCV function, which is subsequently incorporated in an elegant and easily applicable formula estimating the whole GCV function. In order to verify the estimates numerically, we performed computer aided simulations; their results are given in Section 6 and further discussed in Section 7.

Throughout the paper, we use the symbol $I_n$ to represent the identity matrix of order $n$ and $J_n$ to represent the square matrix of order $n$ such that all of its entries are equal to 1. The Euclidean norm of a vector $x \in \mathbb{R}^n$ is denoted by $||x||$, the $i$-th entry of the vector $x$ is written as $x_i$. The trace of a matrix $A$ is denoted $\text{Tr}(A)$, the superscript $T$ stands for the transpose of a matrix or a vector. The inner product of vectors $x, y$ is denoted $(x, y)$.

2. Estimation of the generalised cross validation function via extrapolation. The most common method for choosing the tuning parameter $\lambda$ in the linear regression model is the generalized cross-validation (GCV), introduced by Craven and Wahba [3]. The GCV function $V(\lambda)$ is given as

\begin{equation}
V(\lambda) = \frac{\parallel (I_n - A_\lambda) y \parallel^2}{\text{Tr}(I_n - A_\lambda)^2},
\end{equation}

where $A_\lambda = X(X^TX + \lambda I_d)^{-1}X^T$ is the $n \times n$ influence matrix. The value $\lambda$ that minimizes the GCV function $V(\lambda)$ turns out to be a good approximation of the tuning parameter in (1.1) [3, 4].

A drawback of this approach consists in the computational cost: Exact evaluation of the GCV function needs $O(n^3)$ operations. It is therefore desirable to search for efficient estimates of the GCV function that can be used instead of computing it exactly. Mitrouli and Roupa [9] proposed a method that allows to estimate the GCV function with quadratic complexity. The approach takes advantage of a reformulation of (2.2) by Reichel et al. [10], which reads

\begin{equation}
V(\lambda) = \frac{y^TB^{-2}y}{(\text{Tr}(B^{-1}))^2},
\end{equation}

where $B = XX^T + \lambda I_n \in \mathbb{R}^{n \times n}$. The quadratic form $y^TB^{-2}y$ appearing in (2.3) is then extrapolated in the way described in Proposition 2.3 below. To formulate the result, we need two definitions.

**Definition 2.1.** Let $X$ be a design matrix having $n$ rows. For any integer $k$ and a vector $x \in \mathbb{R}^n$, the moment $s_k(x)$ of the matrix $XX^T$ with respect to the vector $x$ is defined as

\begin{equation}
s_k(x) = (x, (XX^T)^k x) = x^T(XX^T)^k x.
\end{equation}
For a given $x \in \mathbb{R}^n$, we also define quantities $c_0(x)$, $c_1(x)$ and $c_2(x)$ as follows:

$$
\begin{align*}
    c_0(x) &= s_0(x), \\
    c_1(x) &= s_1(x) + \lambda s_0(x), \\
    c_2(x) &= s_2(x) + 2\lambda s_1(x) + \lambda^2 s_0(x).
\end{align*}
$$

**Definition 2.2.** The index of proximity with respect to a given vector $x \in \mathbb{R}^n$ is defined as

$$\rho(x) = \frac{c_0(x)c_2(x)}{(c_1(x))^2},$$

where $c_0(x)$, $c_1(x)$ and $c_2(x)$ are introduced in Definition 2.1.

The index of proximity is a crucial notion. It is closely related to the applicability of extrapolation estimates.

**Proposition 2.3.** (Optimum Extrapolation Estimate (OEE) [9]) Let $y \in \mathbb{R}^n$ be a given vector. If $\rho(y)$ is close to one, then an optimum extrapolation estimate for the bilinear form $y^T B^{-2} y$ is given by

$$OEE = \frac{(\rho(y)c_0(y))^3}{(c_1(y))^2}. $$

Although the assumption that the index of proximity is close to one may seem restrictive, it turns out to be satisfied very often. For example, it was shown heuristically that the index of proximity is around $O(\epsilon)$ when the covariates have the same variance and correlation [7, 8]. In this paper, we will analyze the index of proximity for the linear regression model with intercept and highly correlated covariates.

**3. The index of proximity.** In this section, we will find a bound on the index of proximity for a given $X \in \mathbb{R}^{n \times d}$ with respect to a general vector $y \in \mathbb{R}^n$. Then we will specify the result for a linear regression model, where $y = X\beta + \epsilon$.

**Proposition 3.1.** We have

$$\rho(y) = \frac{\|XX^T y\|^2\|y\|^2 + f(\lambda, y)}{\|X^2 y\|^4 + f(\lambda, y)},$$

where $f(\lambda, y) = 2\lambda \|XX^T y\|^2\|y\|^2 + \lambda^2 \|y\|^4$.

**Proof.** The statement follows directly from Definition 2.2 and relations (2.4) and (2.5):

$$
\rho(y) = \frac{s_0(y)s_2(y) + 2\lambda s_1(y) + \lambda^2 s_0(y)}{(s_1(y) + \lambda s_0(y))^2} = \frac{s_2(y)s_0(y) + 2\lambda s_1(y)s_0(y) + \lambda^2 s_0^2(y)}{s_1^2(y) + 2\lambda s_1(y)s_0(y) + \lambda^2 s_0^2(y)} = \frac{\|XX^T y\|^2\|y\|^2 + 2\lambda \|XX^T y\|^2\|y\|^2 + \lambda^2 \|y\|^4}{\|X^2 y\|^4 + 2\lambda \|XX^T y\|^2\|y\|^2 + \lambda^2 \|y\|^4}. \quad \square
$$

Removing $f(\lambda, y)$ from (3.8), we obtain bounds on $\rho(y)$ that are independent of $\lambda$:...
3.2. Proposition. It holds

\begin{equation}
1 \leq \rho(y) \leq \frac{\|XX^T y\|_2^2 \|y\|^2}{\|X^T y\|^4}.
\end{equation}

Proof. The Cauchy-Schwarz inequality gives

\[ \|XX^T y\|_2^2 \|y\|^2 \geq \|(XX^T y, y)\|^2 = (X^T y, X^T y)^2 = \|X^T y\|^4. \]

Since moreover \( f(\lambda, y) \geq 0 \), we have \( \|XX^T y\|_2^2 f(\lambda, y) \equiv \|X^T y\|^4 + f(\lambda, y) \); hence, \( \rho \geq 1 \) by (3.8).

Regarding the upper bound, we have

\[ \rho(y) = \frac{\|XX^T y\|_2^2 \|y\|^2}{\|X^T y\|^4} - \frac{\|(XX^T y\|_2^2 \|y\|^2 - \|X^T y\|^4) f(\lambda, y)}{\|X^T y\|^4 \cdot (\|X^T y\|^4 + f(\lambda, y))} \leq \frac{\|XX^T y\|_2^2 \|y\|^2}{\|X^T y\|^4}. \]

Let us now present a method to analyze the upper bound in (3.9) for the linear regression model, where \( y = X\beta + \epsilon \). The method \([7, 8]\) is based on the singular value decomposition \([5]\) of \( X \),

\begin{equation}
X = USV^T,
\end{equation}

where \( U \in \mathbb{R}^{n \times n} \) and \( V \in \mathbb{R}^{d \times d} \) are orthogonal matrices and \( S \in \mathbb{R}^{n \times d} \) is a rectangular diagonal matrix having the singular values of \( X \) on its main diagonal in (3.11). Substituting (3.10) into the estimate

\begin{equation}
\rho(X\beta + \epsilon) \leq \frac{\|XX^T (X\beta + \epsilon)\|_2^2 \cdot \|X\beta + \epsilon\|^2}{\|X^T (X\beta + \epsilon)\|^4},
\end{equation}

we get

\begin{equation}
\rho \leq \frac{\|USST (SV^T \beta + UT\epsilon)\|_2^2 \cdot \|USV^T \beta + UT\epsilon\|^2}{\|VST (SV^T \beta + UT\epsilon)\|^4} = \frac{\|SS^T (SV^T \beta + UT\epsilon)\|^2 \cdot \|SV^T \beta + UT\epsilon\|^2}{\|ST SV^T \beta + ST UT\epsilon\|^4}.
\end{equation}

(For the sake of brevity, from now on we will not write the argument of \( \rho \) explicitly.) Since \( \epsilon \sim \mathcal{N}(0, \sigma I_n) \), its unitary transformation satisfies \( UT\epsilon \sim \mathcal{N}(0, \sigma I_n) \). This allows us to denote \( \epsilon = UT\epsilon \sim \mathcal{N}(0, \sigma I_n) \) and rewrite (3.12) in the following form, which is independent of \( U \):

\[ \rho \leq \frac{\|SS^T (SV^T \beta + \epsilon)\|^2 \cdot \|SV^T \beta + \epsilon\|^2}{\|ST SV^T \beta + ST UT\epsilon\|^4}. \]

Furthermore, let us denote the vector \( V^T \beta \) by \( z \). Hence,

\begin{equation}
\rho \leq \frac{\|SS^T (Sz + \epsilon)\|^2 \cdot \|Sz + \epsilon\|^2}{\|ST Sz + ST UT\epsilon\|^4}.
\end{equation}

It follows from the singular value decomposition of \( X \) that the diagonal terms of \( S \) coincide with the square roots of the eigenvalues of \( X^TX \), while the off-diagonal terms of \( S \) vanish. This allows us to rewrite (3.13) as

\begin{equation}
\rho \leq \left( \sum_{j=1}^d \lambda_j^2 (\sqrt{x_j z_j + e_j})^2 \right) \cdot \left( \sum_{j=1}^d (\sqrt{x_j z_j + e_j})^2 + \sum_{j=d+1}^n e_j^2 \right) \left( \sum_{j=1}^d \lambda_j (\sqrt{x_j z_j + e_j})^2 \right)^2.
\end{equation}
where \( \lambda_j \ (j = 1, \ldots, d) \) are the eigenvalues of \( X^T X \).

Numerical experiments show that the index of proximity usually attains values close to 1. In view of this fact, it is convenient to rewrite the right hand side of formula (3.14) in the following manner [8].

**Theorem 3.3.** Let \( X \in \mathbb{R}^{n \times d}, \beta \in \mathbb{R}^d \) and \( y = X \beta + \epsilon \in \mathbb{R}^n \), where \( \epsilon \sim N(0,1) \). Let \( \lambda_1, \ldots, \lambda_d \) be the eigenvalues of \( X^T X \), \( (v_1, \ldots, v_d) \) be an orthonormal basis of \( \mathbb{R}^d \) given by the associated eigenvectors of \( X^T X \), and \( z \in \mathbb{R}^d \) be the coordinates of the vector \( \beta \) in this basis. Then the index of proximity satisfies

\[
(3.15) \quad \rho \leq 1 + \frac{\sum_{j=1}^{d} \sum_{k=j+1}^{d} (\lambda_j - \lambda_k)^2 (\sqrt{\lambda_j z_j} + e_j)^2 (\sqrt{\lambda_k z_k} + e_k)^2}{\left( \sum_{j=1}^{d} \lambda_j (\sqrt{\lambda_j z_j} + e_j)^2 \right)^2} + \frac{\sum_{j=1}^{d} \lambda_j^2 (\sqrt{\lambda_j z_j} + e_j)^2}{\left( \sum_{j=d+1}^{d} \lambda_j (\sqrt{\lambda_j z_j} + e_j)^2 \right)^2} \cdot \sum_{j=d+1}^{d} e_j^2.
\]

**4. Linear regression model with intercept.** Consider the model

\[
(4.16) \quad y = X \beta + \epsilon,
\]

where \( X = \begin{bmatrix} 1 & x_1 & x_2 & \cdots & x_d \end{bmatrix} \in \mathbb{R}^{n \times d+1} \), the vector \( \epsilon_{n \times 1} \) is normally distributed as \( \mathcal{N}(0, I_n) \) and \( \beta \) is a randomly chosen \((d+1) \times 1\) vector on a sphere of radius \( R \). Let the covariates \( x_i, i = 1, \ldots, d \), be jointly Gaussian marginally distributed as \( \mathcal{N}(0, \sigma_i^2) \) random variables with correlation structure \( r = \text{cor}(x_i, x_j) \), if \( i \neq j \), for a given \( r \). We assume that the covariates are highly correlated, i.e., \( r = 1 - \delta \) for \( \delta \ll 1 \).

**Theorem 4.1.** Let \( a > 0 \). If \( \sum_{j=1}^{d} \sigma_j^2 \geq \frac{n}{\pi(n-1)} \), then the index of proximity satisfies

\[
(4.17) \quad \rho \leq 1 + a + O(\delta)
\]

with probability at least

\[
2 \pi \arccot \left[ \left( 1 - \frac{n}{(n-1) \sum_{j=1}^{d} \sigma_j^2} \right) \sqrt{\frac{2n}{a(n-1) \sum_{j=1}^{d} \sigma_j^2}} \right] \cdot \left( 1 - \sqrt{\frac{4d}{\pi a R^2 \sum_{j=1}^{d} \sigma_j^2}} \right).
\]

**Proof.** The information matrix is

\[
X^T X = \begin{bmatrix}
  n & 0 & 0 & \cdots & 0 \\
  0 & (n-1)\sigma_1^2 & r_{12}(n-1)\sigma_1\sigma_2 & \cdots & r_{1d}(n-1)\sigma_1\sigma_d \\
  0 & r_{21}(n-1)\sigma_2\sigma_1 & (n-1)\sigma_2^2 & \cdots & r_{2d}(n-1)\sigma_2\sigma_d \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  0 & r_{d1}(n-1)\sigma_d\sigma_1 & r_{d2}(n-1)\sigma_d\sigma_2 & \cdots & (n-1)\sigma_d^2 \\
\end{bmatrix}
\]

We assume \( 1 - r_{ij} = O(\delta) \) as \( \delta \to 0 \); hence,

\[
X^T X = \begin{bmatrix}
  n & 0 & 0 & \cdots & 0 \\
  0 & (n-1)\sigma_1^2 & (n-1)\sigma_1\sigma_2 & \cdots & (n-1)\sigma_1\sigma_d \\
  0 & (n-1)\sigma_2\sigma_1 & (n-1)\sigma_2^2 & \cdots & (n-1)\sigma_2\sigma_d \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  0 & (n-1)\sigma_d\sigma_1 & (n-1)\sigma_d\sigma_2 & \cdots & (n-1)\sigma_d^2 \\
\end{bmatrix} - O(\delta) = A - O(\delta).
\]
At first we examine Matrix Electronic Journal of Linear Algebra, ISSN 1081-3810 (4.19) 

\[ \lambda_0 = n \] is common for both A and \( X^T X \); the other eigenvalues of \( X^T X \) are given as perturbations of the eigenvalues of A, i.e.,

\[ \lambda_1 = (n - 1) \sum_{j=1}^{d} \sigma_j^2 + \mathcal{O}(\delta), \quad \lambda_2 = \mathcal{O}(\delta), \quad \ldots, \quad \lambda_d = \mathcal{O}(\delta). \]  

Plugging the eigenvalues (4.18) of \( X^T X \) into formula (3.15), we get an upper bound on \( \rho \) as follows:

\[ \rho \leq 1 + \frac{(\lambda_0 - \lambda_1)^2 (\sqrt{\lambda_0} z_0 + e_0)^2 (\sqrt{\lambda_1} z_1 + e_1)^2}{\lambda_0 (\sqrt{\lambda_0} z_0 + e_0)^2 + \lambda_1 (\sqrt{\lambda_1} z_1 + e_1)^2} + \frac{\lambda_0^2 (\sqrt{\lambda_0} z_0 + e_0)^2 \sum_{k=2}^{d} \sigma_k^2 + \lambda_1^2 (\sqrt{\lambda_1} z_1 + e_1)^2 \sum_{k=2}^{d} \sigma_k^2}{\lambda_0 (\sqrt{\lambda_0} z_0 + e_0)^2 + \lambda_1 (\sqrt{\lambda_1} z_1 + e_1)^2} \]

\[ + \frac{\lambda_0^2 (\sqrt{\lambda_0} z_0 + e_0)^2 + \lambda_1^2 (\sqrt{\lambda_1} z_1 + e_1)^2}{\lambda_0 (\sqrt{\lambda_0} z_0 + e_0)^2 + \lambda_1 (\sqrt{\lambda_1} z_1 + e_1)^2} \sum_{k=2}^{n} \sigma_k^2 + \mathcal{O}(\delta) \]

\[ = 1 + K_1 + K_2 + \mathcal{O}(\delta). \]

At first we examine \( K_2 \).

\[ K_2 = \frac{\lambda_0^2 (\sqrt{\lambda_0} z_0 + e_0)^2 + \lambda_1^2 (\sqrt{\lambda_1} z_1 + e_1)^2}{\lambda_0 (\sqrt{\lambda_0} z_0 + e_0)^2 + \lambda_1 (\sqrt{\lambda_1} z_1 + e_1)^2} \sum_{k=2}^{n} \sigma_k^2 \]

\[ = \frac{\lambda_0^2 (\sqrt{\lambda_0} z_0 + e_0)^2 + \lambda_1^2 (\sqrt{\lambda_1} z_1 + e_1)^2}{2 \lambda_0 \lambda_1 (\sqrt{\lambda_0} z_0 + e_0)^2 (\sqrt{\lambda_1} z_1 + e_1)^2 + \lambda_1^2 (\sqrt{\lambda_1} z_1 + e_1)^4} \sum_{k=2}^{n} \sigma_k^2. \]

The assumption \( \sum_{j=1}^{d} \sigma_j^2 \geq \frac{n}{2(n-1)} \) implies \( 2 \lambda_1 \geq \lambda_0 \), which allows to estimate \( 2 \lambda_0 \lambda_1 \) in the denominator by \( \lambda_0^2 \); i.e.,

\[ K_2 \leq \frac{\lambda_0^2 (\sqrt{\lambda_0} z_0 + e_0)^2 + \lambda_1^2 (\sqrt{\lambda_1} z_1 + e_1)^2}{\lambda_0 (\sqrt{\lambda_0} z_0 + e_0)^2 + \lambda_1 (\sqrt{\lambda_1} z_1 + e_1)^2} \sum_{k=2}^{n} \sigma_k^2 = \frac{\sum_{k=2}^{n} \sigma_k^2}{(\sqrt{\lambda_1} z_1 + e_1)^2}. \]

Let us proceed to \( K_1 \).

\[ K_1 \leq \frac{(\lambda_0 - \lambda_1)^2 (\sqrt{\lambda_0} z_0 + e_0)^2 (\sqrt{\lambda_1} z_1 + e_1)^2}{\lambda_1^2 (\sqrt{\lambda_1} z_1 + e_1)^4} = \frac{(\lambda_0 - \lambda_1)^2 (\sqrt{\lambda_0} z_0 + e_0)^2}{\lambda_1^2 (\sqrt{\lambda_1} z_1 + e_1)^2}. \]

\[ \leq \frac{(\lambda_0 - \lambda_1)^2 \lambda_0 e_0^2}{\lambda_1^2 (\sqrt{\lambda_1} z_1 + e_1)^2} + \frac{(\lambda_0 - \lambda_1)^2 2 \sqrt{\lambda_0} z_0 e_0}{\lambda_1^2 (\sqrt{\lambda_1} z_1 + e_1)^2} + \frac{(\lambda_0 - \lambda_1)^2 e_0^2}{\lambda_1^2 (\sqrt{\lambda_1} z_1 + e_1)^2}. \]
We use once again the assumption \( \sum_{j=1}^{d} \sigma_{j}^{2} \geq \frac{n}{\pi a \lambda_{1}} \) (i.e., \( 2\lambda_{1} \geq \lambda_{0} \)) to estimate \( |\lambda_{0} - \lambda_{1}| \leq \lambda_{1} \) in the last summand of (4.20). We neglect the middle summand in (4.20), which is small – \( O(\frac{1}{n}) \) – and, besides, vanishes on average due to \( e_{0} \sim N(0,1) \). So,

\[
K_{1} \lesssim \frac{(\lambda_{0} - \lambda_{1})^{2} \lambda_{0} z_{0}^{2}}{\lambda_{1}^{2} (\sqrt{\lambda_{1} z_{1}} + e_{1})^{2}} + \frac{e_{0}^{2}}{(\sqrt{\lambda_{1} z_{1}} + e_{1})^{2}}.
\]

In view of (4.19) and (4.21), we obtain

\[
\rho \lesssim 1 + \frac{(\lambda_{0} - \lambda_{1})^{2} \lambda_{0} z_{0}^{2}}{\lambda_{1}^{2} (\sqrt{\lambda_{1} z_{1}} + e_{1})^{2}} + \frac{e_{0}^{2} + \sum_{k=2}^{n} e_{k}^{2}}{(\sqrt{\lambda_{1} z_{1}} + e_{1})^{2}}.
\]

Since \( e_{j} \sim N(0,1) \), we have \( e_{0}^{2} + \sum_{k=2}^{n} e_{k}^{2} \approx n \); hence,

\[
(4.22)
\rho \lesssim 1 + \frac{(\lambda_{0} - \lambda_{1})^{2} \lambda_{0} z_{0}^{2}}{\lambda_{1}^{2} (\sqrt{\lambda_{1} z_{1}} + e_{1})^{2}} + \frac{n}{(\sqrt{\lambda_{1} z_{1}} + e_{1})^{2}}.
\]

In order to have \( \rho \) small, the quantity \( (\sqrt{\lambda_{1} z_{1}} + e_{1})^{-2} \) must be greater than \( \frac{1}{n} \). This is possible only for \( \sqrt{\lambda_{1} z_{1}} \) being greater than roughly \( \sqrt{n} \); i.e., \( \sqrt{\lambda_{1} z_{1}} \) must largely outweigh the error term \( e_{1} \). Consequently, it is necessary that \( (\sqrt{\lambda_{1} z_{1}} + e_{1})^{-1} \approx \lambda_{1} z_{1}^{-1} \), which allows us to simplify the estimate (4.22) to

\[
(4.23)
\rho \lesssim 1 + \frac{(\lambda_{0} - \lambda_{1})^{2} \lambda_{0} z_{0}^{2}}{\lambda_{1}^{2} \lambda_{1} z_{1}^{2}} + \frac{n}{\lambda_{1} z_{1}^{2}} + O(\delta) = 1 + \left( 1 - \frac{\lambda_{0}}{\lambda_{1}} \right)^{2} \frac{\lambda_{0} z_{0}^{2}}{\lambda_{1} z_{1}^{2}} + \frac{n}{\lambda_{1} z_{1}^{2}} + O(\delta).
\]

Now let us estimate the probability of \( \rho \leq 1 + a + O(\delta) \). A sufficient condition is

\[
(4.24)
\left| \frac{z_{1}}{z_{0}} \right| \geq 1 - \frac{\lambda_{0}}{\lambda_{1}} \sqrt{\frac{2 \lambda_{0}}{a \lambda_{1}}}
\]

which can be rewritten as

\[
(4.25)
\frac{|z_{1}|}{R} \geq \sqrt{\frac{2n}{a R^{2} \lambda_{1}}}.
\]

Since \( \frac{z_{1}}{R} \) is one coordinate of a unit vector in \( \mathbb{R}^{d+1} \), we can infer\(^3\) that \( \frac{|z_{1}|}{R} < \sqrt{\frac{2n}{a R^{2} \lambda_{1}}} \) occurs with probability at most \( \sqrt{\frac{2d}{\pi}} \) for \( c = \sqrt{\frac{2n}{a R^{2} \lambda_{1}}} \). In other words, (4.25) is violated with probability at most \( \sqrt{\frac{4nd}{\pi a R^{2} \lambda_{1}}} \). Thus, (4.25) is satisfied with probability at least

\[
(4.26)
1 - \sqrt{\frac{4nd}{\pi a R^{2} \lambda_{1}}}.
\]

\(^3\text{Lemma. (8)}\) Let \( c \in (0,1) \) and \( d \geq 3 \) be an integer. The probability that a given coordinate of a random unit vector in \( \mathbb{R}^{d} \) satisfies \( |x| < c \) is bounded from above by the value \( \sqrt{\frac{2d-1}{\pi}} c \).
The probability of satisfying (4.24) is equal to the probability that a randomly chosen vector \( [z_0, z_1] = [Z \sin \zeta, Z \cos \zeta] \) satisfies

\[
| \cot \zeta | \geq \left| 1 - \frac{\lambda_0}{\lambda_1} \right| \sqrt{\frac{2\lambda_0}{a\lambda_1}},
\]

which is equal to

\[
(4.27) \quad \frac{2}{\pi} \arccot \left( \left| 1 - \frac{\lambda_0}{\lambda_1} \right| \sqrt{\frac{2\lambda_0}{a\lambda_1}} \right).
\]

Now we observe that the condition (4.24) requires that \( z_1 \) is (in some sense) large compared to \( z_0 \). Since (4.25) required \( z_1 \) to be large as well, we conclude that \( \{ (4.24) \text{ subject to (4.25)} \} \) is satisfied with higher probability than (4.24). Therefore, the probability that both (4.24) and (4.25) are satisfied at the same time is greater than the product of probabilities of both inequalities being satisfied independently. Consequently, \( \rho \leq 1 + a + O(\delta) \) occurs with probability greater or equal to the product of expressions (4.26) and (4.27).

Using explicit values \( \lambda_0 = n, \lambda_1 = (n - 1) \sum_{j=1}^{d} \sigma_j^2 \), we arrive at the sought lower bound on the probability,

\[
\frac{2}{\pi} \arccot \left( \left| 1 - \frac{n}{(n - 1) \sum_{j=1}^{d} \sigma_j^2} \right| \sqrt{\frac{2n}{a(n - 1) \sum_{j=1}^{d} \sigma_j^2}} \right) \cdot \left( 1 - \sqrt{\frac{4d}{\pi aR^2 \sum_{j=1}^{d} \sigma_j^2}} \right). \]

**Remark 4.2.** A similar result can be derived for a model with error \( \epsilon_{n \times 1} \) being normally distributed as \( N(0, \sigma_{err} I_n) \) for any given \( \sigma_{err} > 0, \) instead of \( N(0, I_n) \). In this generalized case, formula (4.17) acquires the form

\[
\frac{2}{\pi} \arccot \left( \left| 1 - \frac{n}{(n - 1) \sum_{j=1}^{d} \sigma_j^2} \right| \sqrt{\frac{2n}{a(n - 1) \sum_{j=1}^{d} \sigma_j^2}} \right) \cdot \left( 1 - \sqrt{\frac{4d\sigma_{err}^2}{\pi aR^2 \sum_{j=1}^{d} \sigma_j^2}} \right).
\]

**5. An estimate for the GCV function.** Based on the results of the previous sections, we can formulate an estimate for the GCV function of formula (2.3) for regression models \( y = X\beta + \epsilon \), with a highly correlated design matrix \( X \).

**5.1. Estimation of the numerator.** Since the numerator is a bilinear form for which \( \rho(y) \) is usually close to one, the formula (2.7) from Proposition 2.3 can be applied for its estimation.

**5.2. Estimation of the denominator.** The value of \( \text{Tr}(B^{-1}) \) is equal to the sum of the eigenvalues of \( B^{-1} \), where \( B = XX^T + \lambda I_n \). Let us start from finding the eigenvalues of \( XX^T \). The singular value decomposition \( X = USV^T \) gives

\[
XX^T = USS^T U^T \quad \text{and} \quad X^T X = V S^T S V^T.
\]

\[
(5.28)
\]
The first equation in (5.28) implies that the eigenvalues of $XX^T$ coincide with the eigenvalues of $SS^T$. Moreover, since $S$ is a diagonal rectangular matrix of order $n \times (d+1)$ for $n \geq d+1$, the first $d+1$ eigenvalues of $SS^T$ coincide with the eigenvalues of $S^TS$, while the remaining $n-d$ eigenvalues of $SS^T$ are equal to 0. Now we use the second equation in (5.28) to infer that the eigenvalues of $S^TS$ coincide with the eigenvalues of $X^TX$. These eigenvalues were found in the proof of Theorem 4.1; see (4.18). To sum up, the eigenvalues of $XX^T$ are

$$
\lambda_0 = n, \quad \lambda_1 = (n - 1) \sum_{j=1}^{d} \sigma_j^2 + O(\delta), \quad \lambda_2 = O(\delta), \ldots, \lambda_d = O(\delta) \quad \text{as} \quad \delta \to 0, \quad \lambda_{d+1} = \cdots = \lambda_n = 0.
$$

Since $B = XX^T + \lambda I_n$, the eigenvalues of $B$ take the form $\lambda_j + \lambda$, where $\lambda_j$ ($j = 0, 1, \ldots, n$) are given by (5.29). The eigenvalues of $B^{-1}$ are reciprocals of the eigenvalues of $B$. Therefore, the trace of $B^{-1}$ is

$$
\text{Tr}(B^{-1}) = \frac{1}{\lambda_0 + \lambda} + \frac{1}{\lambda_1 + \lambda} + \frac{1}{\lambda_2 + \lambda} + \cdots + \frac{1}{\lambda_d + \lambda} + \frac{n-d}{\lambda},
$$

Combining (5.30) with (5.29), we obtain

$$
\text{Tr}(B^{-1}) = \frac{1}{n+\lambda} + \frac{1}{(n-1) \sum_{j=1}^{d} \sigma_j^2 + O(\delta) + \lambda} + \frac{1}{O(\delta) + \lambda} + \cdots + \frac{1}{O(\delta) + \lambda} + \frac{n-d}{\lambda}
$$

$$
= \frac{n-1}{\lambda} + \frac{1}{(n-1) \sum_{j=1}^{d} \sigma_j^2 + \lambda} + \frac{1}{n+\lambda} + O(\delta) \quad \text{as} \quad \delta \to 0.
$$

Thus, in view of (2.3), Proposition 2.3 and (5.31), an estimate for the GCV function is given by

$$
\tilde{V}(\lambda) = \left( \frac{\rho(y)c_0(y)}{c_1(y)} \right)^3 \cdot \frac{1}{\left( \frac{n-1}{\lambda} + \frac{1}{(n-1) \sum_{j=1}^{d} \sigma_j^2 + \lambda} + \frac{1}{n+\lambda} \right)^2}.
$$
6. Simulations. In order to verify the estimates derived in Section 4, numerical simulations were carried out. We considered several examples of the model

\[ y = X\beta + \epsilon, \]

where the covariates \( x_i \) (\( i = 1, \ldots, d \)) in the design matrix \( X = \begin{bmatrix} 1 & x_1 & x_2 & \cdots & x_d \end{bmatrix}_{n \times (d+1)} \) are jointly Gaussian marginally distributed random variables satisfying \( x_i \sim N(0, \sigma_i^2) \) with correlation \( r = \text{cor}(x_i, x_j) = \frac{x_i^T x_j}{\|x_i\|\|x_j\|} \approx 1 \) for all \( i, j = 1, \ldots, d, i \neq j \).

For each considered example, we generated 10000 datasets, each of them representing \( n \) observations, i.e., we took \( \beta \) as a randomly chosen \( d \times 1 \) vector having norm \( R \) and \( \epsilon \) as a random \( n \times 1 \) vector being normally distributed as \( N(0, 1) \). Since the parameter \( R \) stands for the norm of a \( d \)-dimensional vector with random entries, for our simulation study we examined the case \( R = d \) and a few other values close to \( d \) in order to illustrate the trend.

The results of the simulation are presented in the tables below. Each table corresponds to some design matrix \( X \) of given parameters. The tables show the probability of \( \rho \leq 1 + a \) for various values of \( a \) and \( R = \|\beta\| \). For each combination of \( a \) and \( R \), the theoretical approximate lower bound on the probability, as given by formula (4.17), is written in the upper row, while the lower row shows how many times the bound \( \rho \leq 1 + a \) was satisfied during the simulation.

\[
\begin{array}{|c|c|c|c|c|c|}
\hline
R & a & 0.1 & 0.2 & 0.3 & 0.4 & 0.5 \\
10 & & 0.3117 & 0.4252 & 0.4941 & 0.5423 & 0.5786 \\
& & 4173/10000 & 5489/10000 & 6338/10000 & 6825/10000 & 7291/10000 \\
11 & & 0.3178 & 0.4308 & 0.4993 & 0.5471 & 0.5831 \\
& & 4308/10000 & 5651/10000 & 6353/10000 & 7029/10000 & 7372/10000 \\
12 & & 0.3229 & 0.4340 & 0.5035 & 0.5511 & 0.5869 \\
& & 4327/10000 & 5766/10000 & 6384/10000 & 7243/10000 & 7610/10000 \\
\hline
\end{array}
\]

\[
\begin{array}{|c|c|c|c|c|c|}
\hline
R & a & 0.1 & 0.2 & 0.3 & 0.4 & 0.5 \\
10 & & 0.3764 & 0.4914 & 0.5583 & 0.6037 & 0.6373 \\
& & 4755/10000 & 5947/10000 & 6699/10000 & 7155/10000 & 7511/10000 \\
11 & & 0.3890 & 0.4954 & 0.5619 & 0.6071 & 0.6405 \\
& & 4827/10000 & 6087/10000 & 6713/10000 & 7194/10000 & 7614/10000 \\
12 & & 0.3846 & 0.4987 & 0.5649 & 0.6099 & 0.6431 \\
& & 4839/10000 & 6153/10000 & 6793/10000 & 7243/10000 & 7610/10000 \\
\hline
\end{array}
\]

Table 1: Probability of \( \rho \leq 1 + a \) in the model with parameters \( d = 10, n = 100, r = 0.999 \) and \( [\sigma_1, \ldots, \sigma_{10}] = [0.5, 0.5, 0.5, 0.8, 0.8, 0.8, 1.0, 1.0, 1.0, 1.0] \) (top) and \( [\sigma_1, \ldots, \sigma_{10}] = [0.8, 0.8, 0.8, 1.0, 1.0, 1.0, 1.0, 1.0, 1.2, 1.2] \) (bottom).
Comparing various $r$. Tables 2 and 3 illustrate in more detail the behaviour of the index of proximity in dependence on the correlation $r$.

### Table 2
Probability of $\rho \leq 1 + a$ in the model with parameters $d = 10$, $n = 200$ and $[\sigma_1, \ldots, \sigma_{10}] = [0.5, 0.5, 0.5, 0.8, 0.8, 1.0, 1.0, 1.0, 1.0, 1.0]$. The theoretical upper bound, given by formula (4.17), is compared with the results of numerical simulations carried out for the correlation $r = 0.9703$, $r = 0.9765$ and $r = 0.999$.

<table>
<thead>
<tr>
<th>$a$</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>$r = 0.9703$</th>
<th>$r = 0.9765$</th>
<th>$r = 0.999$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.3121</td>
<td>0.4257</td>
<td>0.4946</td>
<td>0.5428</td>
<td>0.5790</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3227/10000</td>
<td>4786/10000</td>
<td>5567/10000</td>
<td>6281/10000</td>
<td>6704/10000</td>
<td>6076/10000</td>
<td>6826/10000</td>
<td>7312/10000</td>
</tr>
<tr>
<td></td>
<td>3380/10000</td>
<td>4903/10000</td>
<td>5847/10000</td>
<td>6431/10000</td>
<td>6826/10000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4018/10000</td>
<td>5538/10000</td>
<td>6314/10000</td>
<td>6860/10000</td>
<td>7312/10000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0.3182</td>
<td>0.4312</td>
<td>0.4997</td>
<td>0.5476</td>
<td>0.5835</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3251/10000</td>
<td>4848/10000</td>
<td>5634/10000</td>
<td>6268/10000</td>
<td>6847/10000</td>
<td>6076/10000</td>
<td>6893/10000</td>
<td>7355/10000</td>
</tr>
<tr>
<td></td>
<td>3434/10000</td>
<td>4967/10000</td>
<td>5826/10000</td>
<td>6385/10000</td>
<td>7093/10000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4225/10000</td>
<td>5668/10000</td>
<td>6354/10000</td>
<td>6946/10000</td>
<td>7355/10000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.3233</td>
<td>0.4359</td>
<td>0.5040</td>
<td>0.5516</td>
<td>0.5873</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3298/10000</td>
<td>4848/10000</td>
<td>5803/10000</td>
<td>6470/10000</td>
<td>6826/10000</td>
<td>6076/10000</td>
<td>6893/10000</td>
<td>7355/10000</td>
</tr>
<tr>
<td></td>
<td>3928/10000</td>
<td>5060/10000</td>
<td>5846/10000</td>
<td>6391/10000</td>
<td>6937/10000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4337/10000</td>
<td>5726/10000</td>
<td>6506/10000</td>
<td>6999/10000</td>
<td>7469/10000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Comparing various $\sigma$. The following tables illustrate the behaviour of the index of proximity in dependence on the variance $\sigma$. Table 4 is devoted to a model with 10 to 12 parameters and 200 observations, Table 5 to a model with 6 to 8 parameters and 100 observations.
Table 4

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_{10}$</td>
<td>0.3733</td>
<td>0.4883</td>
<td>0.5552</td>
<td>0.6008</td>
<td>0.6346</td>
</tr>
<tr>
<td></td>
<td>4567/10000</td>
<td>5982/10000</td>
<td>6648/10000</td>
<td>7110/10000</td>
<td>7555/10000</td>
</tr>
<tr>
<td>$R_{11}$</td>
<td>0.3778</td>
<td>0.4923</td>
<td>0.5589</td>
<td>0.6043</td>
<td>0.6378</td>
</tr>
<tr>
<td></td>
<td>4713/10000</td>
<td>6106/10000</td>
<td>6802/10000</td>
<td>7141/10000</td>
<td>7477/10000</td>
</tr>
<tr>
<td>$R_{12}$</td>
<td>0.3816</td>
<td>0.4957</td>
<td>0.5620</td>
<td>0.6071</td>
<td>0.6405</td>
</tr>
<tr>
<td></td>
<td>4842/10000</td>
<td>6088/10000</td>
<td>6761/10000</td>
<td>7276/10000</td>
<td>7498/10000</td>
</tr>
</tbody>
</table>

Table 5

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_{6}$</td>
<td>0.4314</td>
<td>0.5458</td>
<td>0.6097</td>
<td>0.6522</td>
<td>0.6831</td>
</tr>
<tr>
<td></td>
<td>5185/10000</td>
<td>6420/10000</td>
<td>6970/10000</td>
<td>7398/10000</td>
<td>7730/10000</td>
</tr>
<tr>
<td>$R_{7}$</td>
<td>0.4349</td>
<td>0.5488</td>
<td>0.6124</td>
<td>0.6547</td>
<td>0.6855</td>
</tr>
<tr>
<td></td>
<td>5230/10000</td>
<td>6574/10000</td>
<td>7086/10000</td>
<td>7511/10000</td>
<td>7772/10000</td>
</tr>
<tr>
<td>$R_{8}$</td>
<td>0.4378</td>
<td>0.5514</td>
<td>0.6147</td>
<td>0.6568</td>
<td>0.6875</td>
</tr>
<tr>
<td></td>
<td>5428/10000</td>
<td>6516/10000</td>
<td>7132/10000</td>
<td>7527/10000</td>
<td>7767/10000</td>
</tr>
</tbody>
</table>

Table 4

Probability of $\rho \leq 1 + \alpha$ in the model with parameters $d = 10$, $n = 200$, $r = 0.999$ and $\sigma_1 = \cdots = \sigma_{10} = 1.0$ (top) and $\sigma_1 = \cdots = \sigma_6 = 1.2$ (bottom).

Table 5

Probability of $\rho \leq 1 + \alpha$ in the model with parameters $d = 6$, $n = 100$, $r = 0.999$ and $\sigma_1 = \cdots = \sigma_6 = 1.0$ (top) and $\sigma_1 = \cdots = \sigma_6 = 1.2$ (bottom).
Christos Koukouvinos, Khalide Jbilou, Marilena Mitrouli, and Ondřej Turek

**Large** $d$. The remaining tables are devoted to models with a large number of parameters. We considered models with 50 parameters (Table 6) and 100 parameters (Table 7).

<table>
<thead>
<tr>
<th>$R$</th>
<th>$a$</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.4334</td>
<td>0.5454</td>
<td>0.6086</td>
<td>0.6509</td>
<td>0.6817</td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>0.4337</td>
<td>0.5457</td>
<td>0.6089</td>
<td>0.6511</td>
<td>0.6820</td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>0.4341</td>
<td>0.5460</td>
<td>0.6092</td>
<td>0.6514</td>
<td>0.6822</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$R$</th>
<th>$a$</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.6400</td>
<td>0.7316</td>
<td>0.7765</td>
<td>0.8045</td>
<td>0.8240</td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>0.6402</td>
<td>0.7317</td>
<td>0.7766</td>
<td>0.8046</td>
<td>0.8241</td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>0.6403</td>
<td>0.7318</td>
<td>0.7767</td>
<td>0.8046</td>
<td>0.8242</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$R$</th>
<th>$a$</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.6898</td>
<td>0.7718</td>
<td>0.8110</td>
<td>0.8351</td>
<td>0.8518</td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>0.6899</td>
<td>0.7719</td>
<td>0.8111</td>
<td>0.8352</td>
<td>0.8519</td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>0.6900</td>
<td>0.7702</td>
<td>0.8111</td>
<td>0.8352</td>
<td>0.8519</td>
<td></td>
</tr>
</tbody>
</table>

**Table 6**

Probability of $\rho \leq 1 + a$ in the model with parameters $d = 50$, $n = 1000$, $r = 0.999$. The standard deviation is supposed to have the same value for all the covariates, namely $\sigma = 0.5$ (top), $\sigma = 1.0$ (middle) and $\sigma = 1.2$ (bottom).
Estimating the Generalized Cross Validation Function for Correlated Matrices

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$R = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.5404/10000</td>
</tr>
<tr>
<td>0.2</td>
<td>0.6459/10000</td>
</tr>
<tr>
<td>0.3</td>
<td>0.7011/10000</td>
</tr>
<tr>
<td>0.4</td>
<td>0.7365/10000</td>
</tr>
<tr>
<td>0.5</td>
<td>0.7616/10000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$R = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.7320/10000</td>
</tr>
<tr>
<td>0.2</td>
<td>0.8045/10000</td>
</tr>
<tr>
<td>0.3</td>
<td>0.8389/10000</td>
</tr>
<tr>
<td>0.4</td>
<td>0.8597/10000</td>
</tr>
<tr>
<td>0.5</td>
<td>0.8741/10000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$R = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.7724/10000</td>
</tr>
<tr>
<td>0.2</td>
<td>0.8355/10000</td>
</tr>
<tr>
<td>0.3</td>
<td>0.8647/10000</td>
</tr>
<tr>
<td>0.4</td>
<td>0.8824/10000</td>
</tr>
<tr>
<td>0.5</td>
<td>0.8946/10000</td>
</tr>
</tbody>
</table>

Table 7

Probability of $\rho \leq 1 + \alpha$ in the model with parameters $d = 100$, $n = 2000$, $r = 0.999$. The standard deviation is the same for all covariates, having value $\sigma = 0.5$ (top), $\sigma = 1.0$ (middle) and $\sigma = 1.2$ (bottom).

7. Conclusions. In the present paper, we proposed an elegant formula for a GCV estimate for models with highly correlated design matrices. Theoretical probability bounds concerning the numerator of the estimate were proved and tested through various simulation experiments. The denominator of the estimate was expressed explicitly by means of an analytical formula, which was derived by exploiting the eigenvalue structure of the matrix. The theoretical results together with the numerical simulations allow us to draw several conclusions, which are summarized below.

- As a first rough observation, we can say that the index of proximity indeed tends to be close to 1. We considered various experimental settings, and in all of them we obtained $\rho \leq 1.5$ with probability exceeding 50%.
- The situation is particularly favourable when $d$ (the number of parameters in the model) is high. Numerical results in Tables 6 and 7 show that the probability of $\rho \leq 1.3$ exceeds 70% – 80%. This behaviour is obvious also from the analytical estimate (4.17).
- The estimate (4.17) depends not only on $d$, but also on $R$. Note, however, that $R = \|\beta\|$ itself is related to $d$, because $\beta$ is a (generally arbitrary) vector having $d$ entries. Therefore, even though $\|\beta\|$ can take in general any value, it is natural to expect it to depend roughly linearly on $d$. As a consequence, if $d$ grows to infinity, we have $R \to \infty$ as well, and – by the estimate (4.17) – the probability of $\rho \leq 1 + \alpha$ approaches 100% for any $\alpha > 0$.
- An important role is played by the variance. Roughly speaking, the larger $\sigma_j$, the larger probability of $\rho$ being close to 1. This trend, which is indicated by formula (4.17), is visible in each of Tables 1–7. Note also that if the sum of variances grows to infinity, then (4.17) implies that the probability of $\rho \leq 1 + \alpha$ tends to 100% for any $\alpha > 0$.
- On the other hand, the number of observations has only a little effect. To see it numerically, compare Table 1 with Tables 2 and 3 (see the rows corresponding to $r = 0.999$). The models considered in the tables differ only in $n$; the other parameters of the models (i.e., $d$, $\sigma_j$, $r$) take the same values. In spite of a big difference in $n$ (Table 1 concerns $n = 100$, while Tables 2 and 3 concern $n = 200$)
the probability of $\rho \leq 1 + a$ remains nearly unchanged. This is not surprising; such a behaviour is obviously expectable in view of formula (4.17).

We wish to emphasize mainly the fact that the probability of $\rho \leq 1 + a$ grows with growing $d$. When the number of parameters ($d$) in the model is expected to be large, the number of observations ($n$) needs to be large as well (due to the natural assumption $n \geq d$), which makes the search for the exact solution computationally hard. In such a situation, however, we have $\rho \approx 1$ with high probability, and therefore the minimization of the inexpensive GCV estimate (5.32) can be applied for the selection of the tuning parameter $\lambda$ in penalized regression models.

Acknowledgment. O.T. appreciates a support from the Czech Science Foundation (GAČR) within the project 17-01706S.

REFERENCES


