Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 35, pp. 482-496, October 2019.

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

CHRISTOS KOUKOUVINOS[†], KHALIDE JBILOU[‡], MARILENA MITROULI[§], AND ONDŘEJ 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 $\boldsymbol{y} = X \cdot \boldsymbol{b}$, where $X = [\mathbf{1}, \boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_d}] \in \mathbb{R}^{n\times(d+1)}$ is the design matrix. The first column of the design matrix is $\mathbf{1}_n = [1, 1, \ldots, 1]^T$, i.e., the model includes the intercept, which corresponds to the mean effect. The j-th column of the design matrix is denoted by $\boldsymbol{x_j} = [x_{1j}, x_{2j}, \ldots, x_{nj}]^T$ and represents the factor j of the experiment. Vector $\boldsymbol{b} \in \mathbb{R}^{d+1}$ represents the regression parameter of the model and it is to be estimated. The vector $\boldsymbol{y} \in \mathbb{R}^n$ 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 $\Sigma = \sigma^2_{\mathrm{err}} I_n$.

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 $\boldsymbol{b}$ 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

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \{\|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2\}, \tag{1.1}$$

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

The specification of an appropriate value for λ 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 λ, and thereafter this method is followed in a vast majority of applications. However, this computation is very expensive, being of order $\mathcal{O}(n^3)$. Thus, with a view to overcome this disadvantage, there were proposed methods [1, 10] attaining the minimization of an estimate of the GCV

[*] Received by the editors on July 26, 2019. Accepted for publication on August 31, 2019. Handling Editor: Michael Tsatsomeros. Corresponding Author: Christos Koukouvinos.

[†] Department of Mathematics, National Technical University of Athens, Zografou 15773, Athens, Greece (ckoukouv@math.ntua.gr).

[‡] Department of Mathematics, Université du Littoral Côte d'Opale, Calais Cedex, 62228, France (jbilou@univ-littoral.fr).

[§] Department of Mathematics, University of Athens, Panepistemiopolis 15784, Athens, Greece (mmitroul@math.uoa.gr).

[¶] Department of Mathematics, University of Ostrava, 701 03 Ostrava, Czech Republic (ondrej.turek@osu.cz).

instead of its exact formula. In this way, the complexity is reduced to $\mathcal{O}(n^2)$. 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 $\boldsymbol{x} \in \mathbb{R}^n$ is denoted by $\|\boldsymbol{x}\|$, the i-th entry of the vector $\boldsymbol{x}$ is written as x_i. The trace of a matrix A is denoted $\mathrm{Tr}(A)$, the superscript ^{T} stands for the transpose of a matrix or a vector. The inner product of vectors $\boldsymbol{x}$, $\boldsymbol{y}$ is denoted $(\boldsymbol{x}, \boldsymbol{y})$.

2. Estimation of the generalised cross validation function via extrapolation.

The most common method for choosing the tuning parameter λ 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

$$V(\lambda) = \frac{\| (I_n - A_\lambda)\, \boldsymbol{y}\|^2}{(\mathrm{Tr}(I_n - A_\lambda))^2}, \tag{2.2}$$

where $A_\lambda = X(X^TX + \lambda I_d)^{-1}X^T$ is the $n \times n$ influence matrix. The value λ 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 $\mathcal{O}(n^3)$ 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

$$V(\lambda) = \frac{\boldsymbol{y}^T B^{-2} \boldsymbol{y}}{(\mathrm{Tr}(B^{-1}))^2}, \tag{2.3}$$

where $B = XX^T + \lambda I_n \in \mathbb{R}^{n\times n}$. The quadratic form $\boldsymbol{y}^T B^{-2}\boldsymbol{y}$ appearing in (2.3) is then extrapolated in the way described in Proposition 2.3 bellow. 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 $\boldsymbol{x} \in \mathbb{R}^n$, the moment $s_k(\boldsymbol{x})$ of the matrix XX^T with respect to the vector $\boldsymbol{x}$ is defined as

$$s_k(\boldsymbol{x}) = \big(\boldsymbol{x}, (XX^T)^k\boldsymbol{x}\big) = \boldsymbol{x}^T(XX^T)^k\boldsymbol{x}. \tag{2.4}$$

For a given $\boldsymbol{x} \in \mathbb{R}^n$, we also define quantities $c_0(\boldsymbol{x})$, $c_1(\boldsymbol{x})$ and $c_2(\boldsymbol{x})$ as follows:

$$\begin{aligned} c_0(\boldsymbol{x}) &= s_0(\boldsymbol{x}), \\ c_1(\boldsymbol{x}) &= s_1(\boldsymbol{x}) + \lambda s_0(\boldsymbol{x}), \\ c_2(\boldsymbol{x}) &= s_2(\boldsymbol{x}) + 2\lambda s_1(\boldsymbol{x}) + \lambda^2 s_0(\boldsymbol{x}). \end{aligned} \tag{2.5}$$

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

$$\rho(\boldsymbol{x}) = \frac{c_0(\boldsymbol{x})c_2(\boldsymbol{x})}{(c_1(\boldsymbol{x}))^2}, \tag{2.6}$$

where $c_0(\boldsymbol{x})$, $c_1(\boldsymbol{x})$ and $c_2(\boldsymbol{x})$ 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 $\boldsymbol{y} \in \mathbb{R}^n$ be a given vector. If $\rho(\boldsymbol{y})$ is close to one, then an optimum extrapolation estimate for the bilinear form $\boldsymbol{y}^T B^{-2}\boldsymbol{y}$ is given by*

$$OEE = \frac{(\rho(\boldsymbol{y})c_0(\boldsymbol{y}))^3}{(c_1(\boldsymbol{y}))^2}. \tag{2.7}$$

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 one for matrices arising from discrete ill-posed problems with error contaminated data [2]. Analytical proofs were given for linear regression models in case 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 $\boldsymbol{y} \in \mathbb{R}^n$. Then we will specify the result for a linear regression model, where $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

Proposition 3.1. *We have*

$$\rho(\boldsymbol{y}) = \frac{\|XX^T\boldsymbol{y}\|^2\|\boldsymbol{y}\|^2 + f(\lambda,\boldsymbol{y})}{\|X^T\boldsymbol{y}\|^4 + f(\lambda,\boldsymbol{y})}, \tag{3.8}$$

where $f(\lambda,\boldsymbol{y}) = 2\lambda\|X^T\boldsymbol{y}\|^2\|\boldsymbol{y}\|^2 + \lambda^2\|\boldsymbol{y}\|^4$.

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

$$\begin{aligned} \rho(\boldsymbol{y}) &= \frac{s_0(\boldsymbol{y})(s_2(\boldsymbol{y}) + 2\lambda s_1(\boldsymbol{y}) + \lambda^2 s_0(\boldsymbol{y}))}{(s_1(\boldsymbol{y}) + \lambda s_0(\boldsymbol{y}))^2} = \frac{s_2(\boldsymbol{y})s_0(\boldsymbol{y}) + 2\lambda s_1(\boldsymbol{y})s_0(\boldsymbol{y}) + \lambda^2 s_0^2(\boldsymbol{y})}{s_1^2(\boldsymbol{y}) + 2\lambda s_1(\boldsymbol{y})s_0(\boldsymbol{y}) + \lambda^2 s_0^2(\boldsymbol{y})} \\ &= \frac{\|XX^T\boldsymbol{y}\|^2\|\boldsymbol{y}\|^2 + 2\lambda\|X^T\boldsymbol{y}\|^2\|\boldsymbol{y}\|^2 + \lambda^2\|\boldsymbol{y}\|^4}{\|X^T\boldsymbol{y}\|^4 + 2\lambda\|X^T\boldsymbol{y}\|^2\|\boldsymbol{y}\|^2 + \lambda^2\|\boldsymbol{y}\|^4}. \qquad \square \end{aligned}$$

Removing $f(\lambda,\boldsymbol{y})$ from (3.8), we obtain bounds on $\rho(\boldsymbol{y})$ that are independent of λ:

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 35, pp. 482-496, October 2019.

Proposition 3.2. *It holds*

$$1 \leq \rho(\boldsymbol{y}) \leq \frac{\|XX^T\boldsymbol{y}\|^2\|\boldsymbol{y}\|^2}{\|X^T\boldsymbol{y}\|^4}. \tag{3.9}$$

Proof. The Cauchy-Schwarz inequality gives

$$\|XX^T\boldsymbol{y}\|^2\|\boldsymbol{y}\|^2 \geq |(XX^T\boldsymbol{y},\boldsymbol{y})|^2 = (X^T\boldsymbol{y},X^T\boldsymbol{y})^2 = \|X^T\boldsymbol{y}\|^4.$$

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

Regarding the upper bound, we have

$$\rho(\boldsymbol{y}) = \frac{\|XX^T\boldsymbol{y}\|^2\|\boldsymbol{y}\|^2}{\|X^T\boldsymbol{y}\|^4} - \frac{\left(\|XX^T\boldsymbol{y}\|^2\|\boldsymbol{y}\|^2 - \|X^T\boldsymbol{y}\|^4\right) f(\lambda,\boldsymbol{y})}{\|X^T\boldsymbol{y}\|^4\cdot\left(\|X^T\boldsymbol{y}\|^4 + f(\lambda,\boldsymbol{y})\right)} \leq \frac{\|XX^T\boldsymbol{y}\|^2\|\boldsymbol{y}\|^2}{\|X^T\boldsymbol{y}\|^4}. \qquad \square$$

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

$$X = USV^T, \tag{3.10}$$

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

$$\rho(X\boldsymbol{\beta}+\boldsymbol{\epsilon}) \leq \frac{\|XX^T(X\boldsymbol{\beta}+\boldsymbol{\epsilon})\|^2\cdot\|X\boldsymbol{\beta}+\boldsymbol{\epsilon}\|^2}{\|X^T(X\boldsymbol{\beta}+\boldsymbol{\epsilon})\|^4}, \tag{3.11}$$

we get

$$\rho \leq \frac{\|USS^T(SV^T\boldsymbol{\beta}+U^T\boldsymbol{\epsilon})\|^2\cdot\|USV^T\boldsymbol{\beta}+\boldsymbol{\epsilon}\|^2}{\|VS^T(SV^T\boldsymbol{\beta}+U^T\boldsymbol{\epsilon})\|^4} = \frac{\|SS^T(SV^T\boldsymbol{\beta}+U^T\boldsymbol{\epsilon})\|^2\cdot\|SV^T\boldsymbol{\beta}+U^T\boldsymbol{\epsilon}\|^2}{\|S^TSV^T\boldsymbol{\beta}+S^TU^T\boldsymbol{\epsilon}\|^4}. \tag{3.12}$$

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

$$\rho \leq \frac{\|SS^T(SV^T\boldsymbol{\beta}+\boldsymbol{e})\|^2\cdot\|SV^T\boldsymbol{\beta}+\boldsymbol{e}\|^2}{\|S^TSV^T\boldsymbol{\beta}+S^T\boldsymbol{e}\|^4}.$$

Furthermore, let us denote the vector $V^T\boldsymbol{\beta}$ by $\boldsymbol{z}$. Hence,

$$\rho \leq \frac{\|SS^T(S\boldsymbol{z}+\boldsymbol{e})\|^2\cdot\|S\boldsymbol{z}+\boldsymbol{e}\|^2}{\|S^TS\boldsymbol{z}+S^T\boldsymbol{e}\|^4}. \tag{3.13}$$

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

$$\rho \leq \frac{\left(\sum\limits_{j=1}^{d}\lambda_j^2(\sqrt{\lambda_j}z_j+e_j)^2\right)\cdot\left(\sum\limits_{j=1}^{d}(\sqrt{\lambda_j}z_j+e_j)^2+\sum\limits_{j=d+1}^{n}e_j^2\right)}{\left(\sum\limits_{j=1}^{d}\lambda_j(\sqrt{\lambda_j}z_j+e_j)^2\right)^2}, \tag{3.14}$$

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 35, pp. 482-496, October 2019.

where λ_j $(j = 1, \ldots, d)$ are the eigenvalues of X^TX.

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

$$\rho \le 1 + \frac{\sum_{j=1}^{d}\sum_{k=j+1}^{d}(\lambda_j-\lambda_k)^2(\sqrt{\lambda_j}z_j+e_j)^2(\sqrt{\lambda_k}z_k+e_k)^2}{\left(\sum_{j=1}^{d}\lambda_j(\sqrt{\lambda_j}z_j+e_j)^2\right)^2} + \frac{\sum_{j=1}^{d}\lambda_j^2(\sqrt{\lambda_j}z_j+e_j)^2}{\left(\sum_{j=1}^{d}\lambda_j(\sqrt{\lambda_j}z_j+e_j)^2\right)^2}\cdot\sum_{j=d+1}^{n}e_j^2. \tag{3.15}$$

4. Linear regression model with intercept.

Consider the model

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{4.16}$$

where $X = \begin{bmatrix} \mathbf{1} & \boldsymbol{x_1} & \boldsymbol{x_2} & \cdots & \boldsymbol{x_d} \end{bmatrix}_{n\times d+1}$, the vector $\boldsymbol{\epsilon}_{n\times 1}$ is normally distributed as $\mathcal{N}(\mathbf{0}, I_n)$ and $\boldsymbol{\beta}$ is a randomly chosen $(d+1)\times 1$ vector on a sphere of radius R. Let the covariates $\boldsymbol{x_i}$, $i = 1, \ldots, d$, be jointly Gaussian marginally distributed as $\mathcal{N}(0, \sigma_i^2)$ random variables with correlation structure $r = cor(\boldsymbol{x_i}, \boldsymbol{x_j})$, 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 $\sum_{j=1}^{d}\sigma_j^2 \ge \frac{n}{2(n-1)}$, then the index of proximity satisfies*

$$\rho \le 1 + a + \mathcal{O}(\delta)$$

with probability at least

$$\frac{2}{\pi}\operatorname{arccot}\left[\left(1-\frac{n}{(n-1)\sum_{j=1}^{d}\sigma_j^2}\right)\sqrt{\frac{2n}{a(n-1)\sum_{j=1}^{d}\sigma_j^2}}\right]\cdot\left(1-\sqrt{\frac{4d}{\pi aR^2\sum_{j=1}^{d}\sigma_j^2}}\right). \tag{4.17}$$

Proof. The information matrix is

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

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

$$X^TX = \begin{bmatrix} n & 0 & 0 & \cdots & 0 \\ 0 & (n-1)\sigma_1^2 & (n-1)\sigma_1\sigma_2 & \cdots & (n-1)\sigma_1\sigma_d \\ 0 & (n-1)\sigma_2\sigma_1 & (n-1)\sigma_2^2 & \cdots & (n-1)\sigma_2\sigma_d \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & (n-1)\sigma_d\sigma_1 & (n-1)\sigma_d\sigma_2 & \cdots & (n-1)\sigma_d^2 \end{bmatrix} - \mathcal{O}(\delta) = A - \mathcal{O}(\delta).$$

Matrix A has a simple eigenvalue n (with associated eigenvector $[1,0,0,\dots,0]^T$), a simple eigenvalue $\lambda_1 = (n-1)\sum\limits_{j=1}^{d}\sigma_j^2$ (with associated eigenvector $[0,\sigma_1,\sigma_2,\dots,\sigma_d]^T$), and an eigenvalue 0 of multiplicity $d+1-\operatorname{rank}(A) = d+1-2 = d-1$.

The eigenvalue $\lambda_0 = n$ is common for both A and X^TX; the other eigenvalues of X^TX are given as perturbations of the eigenvalues of A, i.e.,

$$\lambda_1 = (n-1)\sum_{j=1}^{d}\sigma_j^2 + \mathcal{O}(\delta), \quad \lambda_2 = \mathcal{O}(\delta), \quad \dots, \quad \lambda_d = \mathcal{O}(\delta). \tag{4.18}$$

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

$$\begin{aligned}
\rho \le\ & 1 + \frac{(\lambda_0-\lambda_1)^2\left(\sqrt{\lambda_0}z_0+e_0\right)^2\left(\sqrt{\lambda_1}z_1+e_1\right)^2}{\left[\lambda_0\left(\sqrt{\lambda_0}z_0+e_0\right)^2+\lambda_1(\sqrt{\lambda_1}z_1+e_1)^2\right]^2} + \frac{\lambda_0^2\left(\sqrt{\lambda_0}z_0+e_0\right)^2\sum\limits_{k=2}^{d}e_k^2+\lambda_1^2(\sqrt{\lambda_1}z_1+e_1)^2\sum\limits_{k=2}^{d}e_k^2}{\left[\lambda_0\left(\sqrt{\lambda_0}z_0+e_0\right)^2+\lambda_1(\sqrt{\lambda_1}z_1+e_1)^2\right]^2} \\
& + \frac{\lambda_0^2\left(\sqrt{\lambda_0}z_0+e_0\right)^2+\lambda_1^2(\sqrt{\lambda_1}z_1+e_1)^2}{\left[\lambda_0\left(\sqrt{\lambda_0}z_0+e_0\right)^2+\lambda_1(\sqrt{\lambda_1}z_1+e_1)^2\right]^2}\sum_{k=d+1}^{n}e_k^2 + \mathcal{O}(\delta) \\
=\ & 1 + \frac{(\lambda_0-\lambda_1)^2\left(\sqrt{\lambda_0}z_0+e_0\right)^2(\sqrt{\lambda_1}z_1+e_1)^2}{\left[\lambda_0\left(\sqrt{\lambda_0}z_0+e_0\right)^2+\lambda_1(\sqrt{\lambda_1}z_1+e_1)^2\right]^2} + \frac{\lambda_0^2\left(\sqrt{\lambda_0}z_0+e_0\right)^2+\lambda_1^2(\sqrt{\lambda_1}z_1+e_1)^2}{\left[\lambda_0\left(\sqrt{\lambda_0}z_0+e_0\right)^2+\lambda_1(\sqrt{\lambda_1}z_1+e_1)^2\right]^2}\sum_{k=2}^{n}e_k^2 + \mathcal{O}(\delta) \\
=\ & 1 + K_1 + K_2 + \mathcal{O}(\delta).
\end{aligned}$$

At first we examine K_2.

$$\begin{aligned}
K_2 &= \frac{\lambda_0^2\left(\sqrt{\lambda_0}z_0+e_0\right)^2+\lambda_1^2(\sqrt{\lambda_1}z_1+e_1)^2}{\lambda_0^2\left(\sqrt{\lambda_0}z_0+e_0\right)^4+2\lambda_0\lambda_1\left(\sqrt{\lambda_0}z_0+e_0\right)^2(\sqrt{\lambda_1}z_1+e_1)^2+\lambda_1^2(\sqrt{\lambda_1}z_1+e_1)^4}\sum_{k=2}^{n}e_k^2 \\
&\le \frac{\lambda_0^2\left(\sqrt{\lambda_0}z_0+e_0\right)^2+\lambda_1^2(\sqrt{\lambda_1}z_1+e_1)^2}{2\lambda_0\lambda_1\left(\sqrt{\lambda_0}z_0+e_0\right)^2(\sqrt{\lambda_1}z_1+e_1)^2+\lambda_1^2(\sqrt{\lambda_1}z_1+e_1)^4}\sum_{k=2}^{n}e_k^2.
\end{aligned}$$

The assumption $\sum\limits_{j=1}^{d}\sigma_j^2 \ge \frac{n}{2(n-1)}$ implies $2\lambda_1 \ge \lambda_0$, which allows to estimate $2\lambda_0\lambda_1$ in the denominator by λ_0^2; i.e.,

$$K_2 \le \frac{\lambda_0^2\left(\sqrt{\lambda_0}z_0+e_0\right)^2+\lambda_1^2(\sqrt{\lambda_1}z_1+e_1)^2}{(\sqrt{\lambda_1}z_1+e_1)^2\left(\lambda_0^2\left(\sqrt{\lambda_0}z_0+e_0\right)^2+\lambda_1^2(\sqrt{\lambda_1}z_1+e_1)^2\right)}\sum_{k=2}^{n}e_k^2 = \frac{\sum\limits_{k=2}^{n}e_k^2}{(\sqrt{\lambda_1}z_1+e_1)^2}. \tag{4.19}$$

Let us proceed to K_1.

$$\begin{aligned}
K_1 &\le \frac{(\lambda_0-\lambda_1)^2\left(\sqrt{\lambda_0}z_0+e_0\right)^2(\sqrt{\lambda_1}z_1+e_1)^2}{\lambda_1^2(\sqrt{\lambda_1}z_1+e_1)^4} = \frac{(\lambda_0-\lambda_1)^2\left(\sqrt{\lambda_0}z_0+e_0\right)^2}{\lambda_1^2(\sqrt{\lambda_1}z_1+e_1)^2} \\
&\le \frac{(\lambda_0-\lambda_1)^2\lambda_0z_0^2}{\lambda_1^2(\sqrt{\lambda_1}z_1+e_1)^2} + \frac{(\lambda_0-\lambda_1)^2 2\sqrt{\lambda_0}z_0e_0}{\lambda_1^2(\sqrt{\lambda_1}z_1+e_1)^2} + \frac{(\lambda_0-\lambda_1)^2e_0^2}{\lambda_1^2(\sqrt{\lambda_1}z_1+e_1)^2}.
\end{aligned} \tag{4.20}$$

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 35, pp. 482-496, October 2019.

We use once again the assumption $\sum_{j=1}^{d}\sigma_j^2 \geq \frac{n}{2(n-1)}$ (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 – $\mathcal{O}(\frac{1}{n})$ – and, besides, vanishes on average due to $e_0 \sim \mathcal{N}(0,1)$. So,

$$K_1 \lesssim \frac{(\lambda_0-\lambda_1)^2\lambda_0 z_0^2}{\lambda_1^2(\sqrt{\lambda_1}z_1+e_1)^2} + \frac{e_0^2}{(\sqrt{\lambda_1}z_1+e_1)^2}. \tag{4.21}$$

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

$$\rho \lesssim 1 + \frac{(\lambda_0-\lambda_1)^2\lambda_0 z_0^2}{\lambda_1^2(\sqrt{\lambda_1}z_1+e_1)^2} + \frac{e_0^2+\sum_{k=2}^{n}e_k^2}{\left(\sqrt{\lambda_1}z_1+e_1\right)^2}.$$

Since $e_j \sim \mathcal{N}(0,1)$, we have $e_0^2+\sum_{k=2}^{n}e_k^2 \approx n$; hence,

$$\rho \lesssim 1 + \frac{(\lambda_0-\lambda_1)^2\lambda_0 z_0^2}{\lambda_1^2(\sqrt{\lambda_1}z_1+e_1)^2} + \frac{n}{(\sqrt{\lambda_1}z_1+e_1)^2}. \tag{4.22}$$

In order to have ρ small, the quantity $\left(\sqrt{\lambda_1}z_1+e_1\right)^2$ must be greater than 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 outweight the error term e_1. Consequently, it is necessary that $\left(\sqrt{\lambda_1}z_1+e_1\right)^2 \approx \lambda_1 z_1^2$, which allows us to simplify the estimate (4.22) to

$$\rho \lesssim 1 + \frac{(\lambda_0-\lambda_1)^2\lambda_0 z_0^2}{\lambda_1^2\lambda_1 z_1^2} + \frac{n}{\lambda_1 z_1^2} + \mathcal{O}(\delta) = 1 + \left(1-\frac{\lambda_0}{\lambda_1}\right)^2\frac{\lambda_0 z_0^2}{\lambda_1 z_1^2} + \frac{n}{\lambda_1 z_1^2} + \mathcal{O}(\delta). \tag{4.23}$$

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

$$\left(1-\frac{\lambda_0}{\lambda_1}\right)^2\frac{\lambda_0 z_0^2}{\lambda_1 z_1^2} \leq \frac{a}{2} \quad \text{and} \quad \frac{n}{\lambda_1 z_1^2} \leq \frac{a}{2},$$

which can be rewritten as

$$\left|\frac{z_1}{z_0}\right| \geq \left|1-\frac{\lambda_0}{\lambda_1}\right|\sqrt{\frac{2\lambda_0}{a\lambda_1}} \tag{4.24}$$

and

$$\frac{|z_1|}{R} \geq \sqrt{\frac{2n}{aR^2\lambda_1}}. \tag{4.25}$$

Since $\frac{z_1}{R}$ is one coordinate of a unit vector in $\mathbb{R}^{d+1}$, we can infer[1] that $\frac{|z_1|}{R} < \sqrt{\frac{2n}{aR^2\lambda_1}}$ occurs with probability at most $\sqrt{\frac{2d}{\pi}}c$ for $c = \sqrt{\frac{2n}{aR^2\lambda_1}}$. In other words, (4.25) is violated with probability at most $\sqrt{\frac{4nd}{\pi aR^2\lambda_1}}$. Thus, (4.25) is satisfied with probability at least

$$1 - \sqrt{\frac{4nd}{\pi aR^2\lambda_1}}. \tag{4.26}$$

[1] 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 $\sqrt{\frac{2(d-1)}{\pi}}c$.

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 35, pp. 482-496, October 2019.

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

$$\frac{2}{\pi}\operatorname{arccot}\left(\left|1 - \frac{\lambda_0}{\lambda_1}\right| \sqrt{\frac{2\lambda_0}{a\lambda_1}}\right). \tag{4.27}$$

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) 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, $\rho \leq 1 + a + \mathcal{O}(\delta)$ occurs with probability greater or equal to the product of expressions (4.26) and (4.27). Using explicit values $\lambda_0 = n$, $\lambda_1 = (n-1)\sum\limits_{j=1}^{d}\sigma_j^2$, we arrive at the sought lower bound on the probability,

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

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

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

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 $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with a highly correlated design matrix X.

5.1. Estimation of the numerator.

Since the numerator is a bilinear form for which $\rho(\boldsymbol{y})$ 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 $\operatorname{Tr}(B^{-1})$ 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

$$XX^T = USS^TU^T \quad \text{and} \quad X^TX = VS^TSV^T. \tag{5.28}$$

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 35, pp. 482-496, October 2019.

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

$$\lambda_0 = n, \quad \lambda_1 = (n-1)\sum_{j=1}^{d} \sigma_j^2 + \mathcal{O}(\delta), \quad \lambda_2 = \mathcal{O}(\delta), \;\ldots,\; \lambda_d = \mathcal{O}(\delta) \quad \text{as} \quad \delta \to 0, \quad \lambda_{d+1} = \cdots = \lambda_n = 0. \tag{5.29}$$

Since $B = XX^T + \lambda I_n$, the eigenvalues of B take the form $\lambda_j + \lambda$, where λ_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

$$\operatorname{Tr}(B^{-1}) = \frac{1}{\lambda_0 + \lambda} + \frac{1}{\lambda_1 + \lambda} + \frac{1}{\lambda_2 + \lambda} + \cdots + \frac{1}{\lambda_d + \lambda} + \frac{n-d}{\lambda}, \tag{5.30}$$

Combining (5.30) with (5.29), we obtain

$$\begin{aligned}\operatorname{Tr}(B^{-1}) &= \frac{1}{n+\lambda} + \frac{1}{(n-1)\sum\limits_{j=1}^{d} \sigma_j^2 + \mathcal{O}(\delta) + \lambda} + \frac{1}{\mathcal{O}(\delta) + \lambda} + \cdots + \frac{1}{\mathcal{O}(\delta) + \lambda} + \frac{n-d}{\lambda} \\ &= \frac{n-1}{\lambda} + \frac{1}{(n-1)\sum\limits_{j=1}^{d} \sigma_j^2 + \lambda} + \frac{1}{n+\lambda} + \mathcal{O}(\delta) \quad \text{as} \quad \delta \to 0.\end{aligned} \tag{5.31}$$

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

$$\tilde{V}(\lambda) = \frac{(\rho(\boldsymbol{y})c_0(\boldsymbol{y}))^3}{(c_1(\boldsymbol{y}))^2} \cdot \frac{1}{\left(\frac{n-1}{\lambda} + \frac{1}{(n-1)\sum_{j=1}^{d} \sigma_j^2 + \lambda} + \frac{1}{n+\lambda}\right)^2}. \tag{5.32}$$

6. Simulations. In order to verify the estimates derived in Section 4, numerical simulations were carried out. We considered several examples of the model

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

where the covariates $\boldsymbol{x_i}$ $(i = 1, \dots, d)$ in the design matrix $X = \begin{bmatrix} \mathbf{1} & \boldsymbol{x_1} & \boldsymbol{x_2} & \cdots & \boldsymbol{x_d} \end{bmatrix}_{n\times(d+1)}$ are jointly Gaussian marginally distributed random variables satisfying $\boldsymbol{x_i} \sim \mathcal{N}(0, \sigma_i^2)$ with correlation

$$r = cor(\boldsymbol{x_i}, \boldsymbol{x_j}) = \frac{\boldsymbol{x_i}^T \boldsymbol{x_j}}{\|\boldsymbol{x_i}\|\|\boldsymbol{x_j}\|} \approx 1 \quad \text{for all } \ i, j = 1, \dots, d, \ i \neq j.$$

For each considered example, we generated 10000 datasets, each of them representing n observations, i.e., we took $\boldsymbol{\beta}$ as a randomly chosen $d \times 1$ vector having norm R and $\boldsymbol{\epsilon}$ as a random $n \times 1$ vector being normally distributed as $\mathcal{N}(0, 1)$. 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 $R = \|\boldsymbol{\beta}\|$. 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.

		a				
		0.1	0.2	0.3	0.4	0.5
R	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

		a				
		0.1	0.2	0.3	0.4	0.5
R	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

Table 1

Probability of $\rho \leq 1 + a$ in the model with parameters $d = 10$, $n = 100$, $r = 0.999$ and $[\sigma_1, \dots, \sigma_{10}] = [0.5,\ 0.5,\ 0.5,\ 0.8,\ 0.8,\ 0.8,\ 1.0,\ 1.0,\ 1.0,\ 1.0]$ (top) and $[\sigma_1, \dots, \sigma_{10}] = [0.8,\ 0.8,\ 0.8,\ 1.0,\ 1.0,\ 1.0,\ 1.0,\ 1.2,\ 1.2,\ 1.2]$ (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.

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

Table 2

Probability of $\rho \leq 1 + a$ in the model with parameters $d = 10$, $n = 200$ and $[\sigma_1, \ldots, \sigma_{10}] = [0.5,\ 0.5,\ 0.5,\ 0.8,\ 0.8,\ 0.8,\ 1.0,\ 1.0,\ 1.0,\ 1.0]$. 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 $r = 0.999$.

		a					
		0.1	0.2	0.3	0.4	0.5	
R	10	0.3770	0.4920	0.5588	0.6042	0.6378	
		3643/10000	5083/10000	5891/10000	6448/10000	6808/10000	$r = 0.9703$
		3774/10000	5209/10000	6077/10000	6590/10000	6921/10000	$r = 0.9765$
		4746/10000	5951/10000	6700/10000	7119/10000	7489/10000	$r = 0.999$
	11	0.3814	0.4960	0.5625	0.6076	0.6410	
		3613/10000	5168/10000	5978/10000	6455/10000	6890/10000	$r = 0.9703$
		3852/10000	5286/10000	6116/10000	6578/10000	6940/10000	$r = 0.9765$
		4811/10000	6137/10000	6743/10000	7206/10000	7528/10000	$r = 0.999$
	12	0.3852	0.4993	0.5655	0.6104	0.6436	
		3817/10000	5139/10000	5957/10000	6496/10000	6811/10000	$r = 0.9703$
		3975/10000	5273/10000	6159/10000	6653/10000	7029/10000	$r = 0.9765$
		4938/10000	6151/10000	6840/10000	7234/10000	7585/10000	$r = 0.999$

Table 3

Probability of $\rho \leq 1 + a$ in the model with parameters $d = 10$, $n = 200$, $[\sigma_1, \ldots, \sigma_{10}] = [0.8,\ 0.8,\ 0.8,\ 1.0,\ 1.0,\ 1.0,\ 1.0,\ 1.2,\ 1.2,\ 1.2]$ and $r = 0.9703$, $r = 0.9765$ and $r = 0.999$. The table is similar to Table 2, the only difference consists in larger values of σ_j.

Comparing various σ. The following tables illustrate the behaviour of the index of proximity in dependence on the variance σ. 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.

		a				
		0.1	0.2	0.3	0.4	0.5
R	10	0.3733 4567/10000	0.4883 5982/10000	0.5552 6648/10000	0.6008 7110/10000	0.6346 7555/10000
	11	0.3778 4713/10000	0.4923 6106/10000	0.5589 6802/10000	0.6043 7141/10000	0.6378 7477/10000
	12	0.3816 4842/10000	0.4957 6088/10000	0.5620 6761/10000	0.6071 7276/10000	0.6405 7498/10000

		a				
		0.1	0.2	0.3	0.4	0.5
R	10	0.4314 5185/10000	0.5458 6420/10000	0.6097 6970/10000	0.6522 7398/10000	0.6831 7730/10000
	11	0.4349 5230/10000	0.5488 6574/10000	0.6124 7086/10000	0.6547 7511/10000	0.6855 7772/10000
	12	0.4378 5428/10000	0.5514 6516/10000	0.6147 7132/10000	0.6568 7527/10000	0.6875 7767/10000

TABLE 4

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

		a				
		0.1	0.2	0.3	0.4	0.5
R	6	0.2725 3594/10000	0.3876 5154/10000	0.4584 5954/10000	0.5084 6612/10000	0.5462 6999/10000
	7	0.2863 3866/10000	0.4002 5433/10000	0.4701 6231/10000	0.5193 6859/10000	0.5566 7246/10000
	8	0.2967 4111/10000	0.4096 5456/10000	0.4789 6347/10000	0.5275 6910/10000	0.5643 7374/10000

		a				
		0.1	0.2	0.3	0.4	0.5
R	6	0.3319 4160/10000	0.4484 5656/10000	0.5176 6422/10000	0.5653 6820/10000	0.6009 7273/10000
	7	0.3425 4431/10000	0.4579 5668/10000	0.5263 6484/10000	0.5734 7021/10000	0.6085 7327/10000
	8	0.3504 4498/10000	0.4650 5890/10000	0.5328 6577/10000	0.5795 7104/10000	0.6142 7429/10000

TABLE 5

Probability of $\rho \leq 1 + a$ *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).

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).

		a				
		0.1	0.2	0.3	0.4	0.5
R	50	0.4334	0.5454	0.6086	0.6509	0.6817
		5453/10000	6521/10000	7143/10000	7500/10000	7801/10000
	51	0.4337	0.5457	0.6089	0.6511	0.6820
		5400/10000	6560/10000	7177/10000	7491/10000	7797/10000
	52	0.4341	0.5460	0.6092	0.6514	0.6822
		5406/10000	6525/10000	7204/10000	7515/10000	7802/10000

		a				
		0.1	0.2	0.3	0.4	0.5
R	50	0.6400	0.7316	0.7765	0.8045	0.8240
		7148/10000	7892/10000	8287/10000	8521/10000	8646/10000
	51	0.6402	0.7317	0.7766	0.8046	0.8241
		7214/10000	7942/10000	8364/10000	8491/10000	8680/10000
	52	0.6403	0.7318	0.7767	0.8046	0.8242
		7120/10000	7941/10000	8294/10000	8559/10000	8663/10000

		a				
		0.1	0.2	0.3	0.4	0.5
R	50	0.6898	0.7718	0.8110	0.8351	0.8518
		7483/10000	8198/10000	8506/10000	8752/10000	8894/10000
	51	0.6899	0.7719	0.8111	0.8352	0.8519
		7500/10000	8163/10000	8494/10000	8711/10000	8846/10000
	52	0.6900	0.7702	0.8111	0.8352	0.8519
		7536/10000	8177/10000	8523/10000	8721/10000	8869/10000

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).

	a				
	0.1	0.2	0.3	0.4	0.5
$R = 100$	0.5404 6263/10000	0.6459 7248/10000	0.7011 7738/10000	0.7365 7973/10000	0.7616 8209/10000

	a				
	0.1	0.2	0.3	0.4	0.5
$R = 100$	0.7320 7768/10000	0.8045 8436/10000	0.8389 8756/10000	0.8597 8889/10000	0.8741 9053/10000

	a				
	0.1	0.2	0.3	0.4	0.5
$R = 100$	0.7724 8089/10000	0.8355 8622/10000	0.8647 8857/10000	0.8824 8995/10000	0.8946 9168/10000

Table 7

Probability of $\rho \le 1 + a$ 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 $\rho \le 1.5$ 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 $\rho \le 1.3$ 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 $R = \|\boldsymbol{\beta}\|$ itself is related to d, because $\boldsymbol{\beta}$ is a (generally arbitrary) vector having d entries. Therefore, even though $\|\boldsymbol{\beta}\|$ 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 $\rho \le 1 + a$ approaches 100% for any $a > 0$.
- An important role is played by the variance. Roughly speaking, the larger σ_j, the larger probability of ρ 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 $\rho \le 1 + a$ tends to 100% for any $a > 0$.
- 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, σ_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 λ in penalized regression models.

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

REFERENCES

[1] C. Fenu, L. Reichel, and G. Rodriguez. GCV for Tikhonov regularization via global Golub–Khan decomposition. *Numerical Linear Algebra with Applications*, 23:467–484, 1992.

[2] P. Fika, M. Mitrouli, P. Roupa, and D. Triantafyllou. The e-MoM approach for approximating matrix functionals. *Journal of Computational and Applied Mathematics*, to appear, 2019. Available at `https://doi.org/10.1016/j.cam.2019.04.023`.

[3] G.H. Golub, M. Heath, and G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21:215–223, 1979.

[4] G.H. Golub and G. Meurant. *Matrices, Moments and Quadrature with Applications.* Princeton University Press, Princeton, 2010.

[5] G.H. Golub and C.F. Van Loan. *Matrix Computations.* The John Hopkins University Press, Baltimore, 1989.

[6] P.C. Hansen. Regularization Tools Version 4.0 for MATLAB 7.3. *Numerical Algorithms*, 46:189–194, 2007.

[7] C. Koukouvinos, A. Lappa, M. Mitrouli, P. Roupa, and O. Turek. Numerical methods for estimating the tuning parameter in penalized least squares problems. *Communications in Statistics - Simulation and Computation*, to appear, 2020. Available at `https://doi.org/10.1080/03610918.2019.1676436`.

[8] C. Koukouvinos, M. Mitrouli, and O. Turek. Efficient estimates in regression models with highly correlated covariates. *Journal of Computational and Applied Mathematics*, to appear, 2019. Available at `https://doi.org/10.1016/j.cam.2019.112416`.

[9] M. Mitrouli and P. Roupa. Estimates for the generalized cross-validation function via an extrapolation and statistical approach. *Calcolo*, 55:1–25, 2018.

[10] L. Reichel, G. Rodriguez, and S. Seatzu. Error estimates for large-scale ill-posed problems. *Numerical Algorithms*, 51:341–361, 2009.