Direct methods for variable selection in statistical modeling

Marilena Mitrouli[a]

[a]Department of Mathematics,
National and Kapodistrian University of Athens

Joint work with: Prof. C. Koukouvinos,
National Technical University of Athens

Hellenic Open University

March 2026

Outline of the talk

Introduction

↪ Specify the output of a system for which there are no equal factors and observations.

↪Model describing any real data set is of the form:

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

$\mathbf{y} \in \mathbb{R}^{n\times 1}$ the vector of responses, $\mathbf{X} \in \mathbb{R}^{n\times p}$ the design matrix, $\boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$ the vector of unknown coefficients and $\boldsymbol{\varepsilon} \in \mathbb{R}^{n\times 1}$ the random error vector, where $\varepsilon_i \sim N(0, \sigma^2 I_n)$, for $i = 1, 2, \ldots, n$.
↪ This results in solving a least squares problem of the form:

$$\beta = argmin\|y - X\beta\|_2^2,\ X \in \mathbb{R}^{n\times p},\ n \geq p \text{ or } n < p$$

$$\beta = argmin\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2,\ \lambda\ \ regularization\ parameter$$

Main Task

↪ We consider the following cases of overdetermined matrices $n \geq p$:

1. Overdetermined design matrix $X = \begin{bmatrix} \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_d \end{bmatrix}$ or $X = \begin{bmatrix} \mathbf{1} & \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_d \end{bmatrix}$ with correlated covariates $\boldsymbol{x_i} \sim \boldsymbol{N(0_n, \sigma_i^2 I_n)}$, $i = 1, \ldots, d$, where σ_i^2 is the variance of $\boldsymbol{x_i}$.
2. Overdetermined design matrix $X = \begin{bmatrix} \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_d \end{bmatrix}$ or $X = \begin{bmatrix} \mathbf{1} & \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_d \end{bmatrix}$ with correlated covariates $\boldsymbol{x_i} \sim \boldsymbol{N(0_n, \Sigma)}$, $i = 1, \ldots, d$, where Σ an $n \times n$ square matrix, with $s_{ij} = r^{|i-j|}$.
3. Matrices from real experiments.

↪ Analyze their structure and derive the appropriate treatment.

↪ Direct vs Regularized solution

↪ To handle such models the following numerical methods are used:

- skinny-Golub-Kahan
- skinny-SVD
- $\ell_p - \ell_q$

Koukoudakis N., Koukouvinos C., Lappa A., Mitrouli M., Psitou A., Numerical Methods in Modelling with supersaturated designs, Appl. Numer. Math., 208 (2025) 271-283

The cast fatigue experiment

We have a real data set consisting of seven two-level factors. The design matrix of the main effects and the response data is

Table: The cast fatigue experiment with $n = 12$ Runs and $p = 7$ features

A	B	C	D	E	F	G	y
1	1	-1	1	1	1	-1	6.058
1	-1	1	1	1	-1	-1	4.733
-1	1	1	1	-1	-1	-1	4.625
1	1	1	-1	-1	-1	1	5.899
1	1	-1	-1	-1	1	-1	7.000
1	-1	-1	-1	1	-1	1	5.752
-1	-1	-1	1	-1	1	1	5.682
-1	-1	1	-1	1	1	-1	6.607
-1	1	-1	1	1	-1	1	5.818
1	-1	1	1	-1	1	1	5.917
-1	1	1	-1	1	1	1	5.863
-1	-1	-1	-1	-1	-1	-1	4.809

Variable selection

Which main effects are the most important ?

Table: Results only for main effects

Method	Active factors
SIS-CUSUM	D, F
skinny SVD or QR	F
$\ell_p - \ell_q$	A, B, D, F
Dantzig selector	D, F

Consider also the 7 main effects and the 21 two-factor interactions. In this matrix appears correlation $\pm\frac{1}{3}$ among the columns.

QUESTION:From the design matrix a direct or a regularized method is appropriate ? Can we learn from the data how to treat them ?

Correlated data sets

In Statistics the columns of a matrix can be correlated in a linear sense and linear independent. This sounds contradictory

$$X = \begin{bmatrix} c_1 & c_2 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 2 & 5 \\ 3 & 8 \end{bmatrix} \quad c_1, c_2 \quad \text{linearly independent}$$

$$c_2 = 3 * c_1 - 1 \text{ correlated}$$

The correlation coefficient is :

$$r = \frac{\sum_{i=1}^{3}(c_{i1} - \bar{c_1})(c_{i2} - \bar{c_2})}{\|c_1 - \bar{c_1}\| \|c_2 - \bar{c_2}\|} = 1, \ i = 1, 2, 3, \ \bar{c_1} = 2, \ \bar{c_2} = 5$$

$c_1, \ c_2$ are linearly independent and perfectly correlated, r=1

QUESTION: Given a correlated data set with correlation coefficient r, can we derive information about useful characteristics of the design matrix X (i.e. cond(X))?

Design matrix with correlated covariates

In all the applications the desing matrix $X = \begin{bmatrix} \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_d \end{bmatrix}$ or $X = \begin{bmatrix} \mathbf{1} & \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_d \end{bmatrix}$ has correlated covariates $\boldsymbol{x_i}$.

- In some models we select $\boldsymbol{x_i} \sim \boldsymbol{N(0_n, \sigma_i^2 I_n)}$, $i = 1, \ldots, d$, where σ_i^2 is the variance of $\boldsymbol{x_i}$.

  ↪ For this design matrix the correlation coefficient is :

$$r_{ij} = cor(\boldsymbol{x_i}, \boldsymbol{x_j}) = \frac{\boldsymbol{x_i^T x_j}}{\|\boldsymbol{x_i}\| \|\boldsymbol{x_j}\|}, \; i, j = 1, \ldots, d, \; i \neq j,$$

  with $-1 \leq r_{ij} \leq 1$.

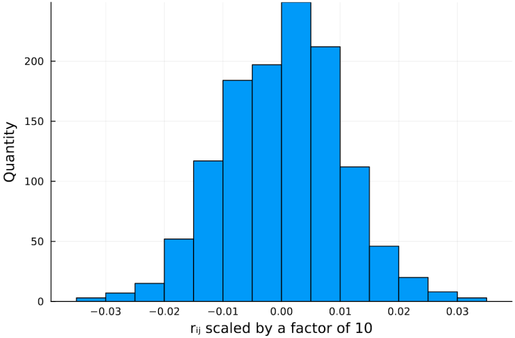
- For models with $\boldsymbol{x_i} \in \boldsymbol{N(0_n, \Sigma)}$, $i = 1, \ldots, d$, where Σ the $n \times n$ covariance matrix, with $s_{ij} = r^{|i-j|}$, r the correlation.

For the above models the correlation affects the characteristics of the design matrix?

1. Histograms

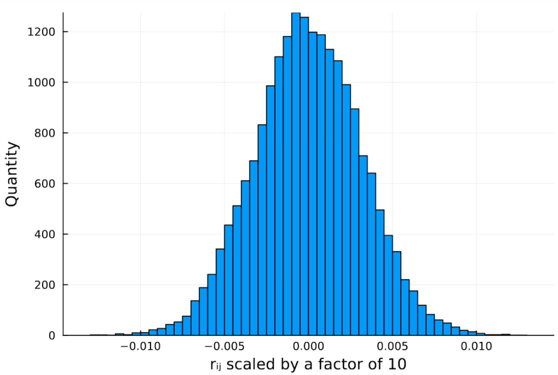
↪ We consider the covariates of the design matrix X to be chosen as $\boldsymbol{x_i \sim N(0, 1I_n)}$.
↪ We tested 100 random matrices and in the following histograms we demonstrate the **mean correlation** $\boldsymbol{r_{ij}}$ that appears between the columns (x_i, x_j).

Figure: Correlations of 100 matrices $x \in \mathbb{R}^{100 \times 50}$



1. Histograms

Figure: Correlations of 100 matrices $x \in \mathbb{R}^{1000\times 200}$



Remark

We notice that as the size of the matrix increases the correlations are approaching zero i.e. **the covariates tend to be independent**.

2. Constant correlation model

Design matrices with **same correlation** $\boldsymbol{r}$ correspond to a constant correlation model which often appears in heritability.

In the following theorems, we specify the singular values of the design matrix X.

The cases of the design matrix under study are:

1. Without intercept

$$X = \begin{bmatrix} \boldsymbol{x_1} & \boldsymbol{x_2} & \cdots & \boldsymbol{x_d} \end{bmatrix} \in \mathbb{R}^{n \times d}$$

2. With intercept

$$\tilde{X} = \begin{bmatrix} \mathbf{1} & X \end{bmatrix}$$

where $X = \begin{bmatrix} \boldsymbol{x_1} & \boldsymbol{x_2} & \cdots & \boldsymbol{x_d} \end{bmatrix} \in \mathbb{R}^{n \times d}$ whose columns are $\boldsymbol{x}_i \sim N(\mathbf{0}_n, \sigma^2 I_n)$, $i = 1, 2, \ldots, d$, with same correlation r.

Case 1: Correlated covariates with same variance and correlation, without intercept

Theorem 1

Let $X = \begin{bmatrix} \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_d \end{bmatrix} \in \mathbb{R}^{n \times d}$, $n \geq d$ *be a design matrix of full rank whose columns* $\boldsymbol{x}_i \sim N(\mathbf{0}_n, \sigma^2 I_n)$, $i = 1, 2, \ldots, d$, *with correlation structure* r.

The generalized condition number of X *is given by*

$$\kappa(X) = \sqrt{\frac{(d-1)r + 1}{1 - r}}.$$

Koukouvinos C., Mitrouli M., Turek O.
Using Direct Versus Regularized Solvers for Realistic Statistical Models.
Springer Proceedings in Mathematics and Statistics, 247, 251–266, 2025.

Case 1: Correlated covariates with same variance and correlation, without intercept

Proof. In [6], there exists a brief description of the eigenvalues of the matrix $X^T X$. It follows from the assumptions on the covariates $\boldsymbol{x_i}$ that

$$X^T X = \begin{bmatrix} (n-1)\sigma^2 & r(n-1)\sigma^2 & \cdots & r(n-1)\sigma^2 \\ r(n-1)\sigma^2 & (n-1)\sigma^2 & \cdots & r(n-1)\sigma^2 \\ \vdots & \vdots & \ddots & \vdots \\ r(n-1)\sigma^2 & r(n-1)\sigma^2 & \cdots & (n-1)\sigma^2 \end{bmatrix}$$

$$= (n-1)\sigma^2 \left[(1-r)I + rJ\right],$$

where J denotes the $d \times d$ matrix with all entries equal to 1.

Koukouvinos C., Lappa A., Mitrouli M., Roupa P., Turek O.
Numerical methods for estimating the tuning parameter in penalized least squares problems.
Communications in Statistics - Simulation and Computation, 2019, 1–22.

Case 1: Correlated covariates with same variance and correlation, without intercept

Matrix J has eigenvalue d with the normalized eigenvector $\boldsymbol{v_1} = \frac{1}{\sqrt{d}}[1, 1, \ldots, 1]^T = \frac{1}{\sqrt{d}}\mathbf{1}$. Moreover, since J is of rank 1, it has eigenvalue 0 of multiplicity $d - 1$.

The eigenvalues of $X^T X$ are

$$\begin{aligned} \lambda_1 &= (n-1)\sigma^2 \left[(1-r) + rd\right] \\ \lambda_2 = \cdots = \lambda_d &= (n-1)\sigma^2(1-r) \end{aligned} \qquad (1)$$

In the sequel, we express the singular value decomposition of the matrix X. Let

$$X = USV^T \qquad (2)$$

be the singular value decomposition of the matrix $X \in \mathbb{R}^{n\times d}$.

Case 1: Correlated covariates with same variance and correlation, without intercept

From relation (1) it holds that the singular values of the matrix X are

$$s_1 = \sigma\sqrt{(n-1)[(d-1)r+1]},$$
$$s_2 = \cdots = s_d = \sigma\sqrt{(n-1)(1-r)}$$

The generalized condition number can be expressed as

$$\kappa(X) = \frac{s_{max}}{s_{min}}$$

From the relations of the singular values we notice that it holds $s_1 > s_i, \ i = 2, \ldots, d.$ Therefore,

$$\kappa(X) = \sqrt{\frac{(d-1)r+1}{1-r}}.$$

Datta, B.N. Numerical Linear Algebra and Applications, 2nd ed, SIAM: Philadelphia, PA, USA, 2010

Case 2: Correlated covariates with same variance and correlation, with intercept

Theorem 2

Let $X = \begin{bmatrix} \mathbf{1} & \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_d \end{bmatrix} \in \mathbb{R}^{n\times(d+1)}$ *be a design matrix of full rank such as* $X \sim N(\mathbf{0}_n, \Sigma)$, $s_{ii} = 1, \quad s_{ij} = r$

The generalized condition number of X *is given by*

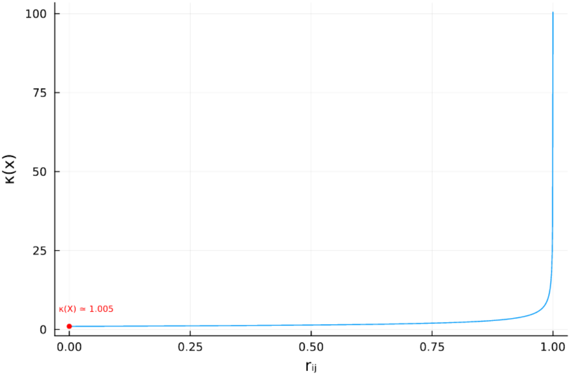
1. $\kappa(X) = \sqrt{1 + \dfrac{dr}{1-r}}, \qquad$ *if* $\quad r \geq 0,$

2. $\kappa(X) = \sqrt{\dfrac{1-r}{(d-1)r+1}}, \qquad$ *if* $\quad r < 0$

Example: Plots of the generalized condition number of X as a function of the correlation r

$\hookrightarrow$ We consider design matrix with $n = 100$, $d = 50$ and $\sigma^2 = 1$.

Figure: Condition number for Case I of Theorem 2



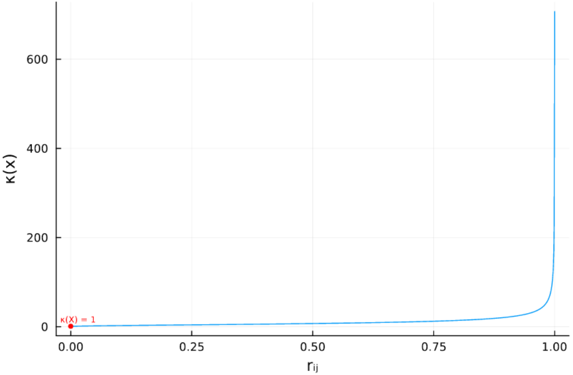
Note:

- $r = 0$, $\kappa(X) = \sqrt{\frac{n}{n-1}} = \sqrt{\frac{100}{99}} \simeq 1.005$
- $r = 0.9999$, $\kappa(X) \simeq 101,50$

Example: Plots of the generalized condition number of X as a function of the correlation r

$\hookrightarrow$ We consider design matrix with $n = 100$, $d = 50$ and $\sigma^2 = 1$.

Figure: Condition number for Case II of Theorem 2



Note:

- $r = 0$, $\kappa(X) = 1$
- $r = 0.9999$, $\kappa(X) \simeq 707,07$

Example: Plots of the generalized condition number of X as a function of the correlation r

Remark

We notice that as the digits increase the condition number increases enormously, if for example:

1. For $r = 0.999999$
   - Case I: $\kappa(X) \simeq 1005,04$
   - Case II: $\kappa(X) \simeq 7071,06$
2. For $r = 0.9999999$
   - Case I: $\kappa(X) \simeq 3178.21$
   - Case II: $\kappa(X) \simeq 22360.68$

meaning that the matrix is ill-conditioned.

For a constant correlation model regularization will be applied only if the correlation is very high (approaching 1)

The skinny Golub-Kahan regression

Theorem 3 (Bidiagonalization Golub-Kahan)

Let $A \in \mathbb{R}^{m \times n}$, $m \geq n$. Then there exists orthogonal $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$, so that:

$$U^T AV = B = \left[\begin{array}{ccccc} a_1 & \beta_1 & \dots & \dots & 0 \\ 0 & a_2 & \beta_2 & \dots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \dots & 0 & a_{n-1} & \beta_{n-1} \\ 0 & \dots & \dots & 0 & a_n \\ \hline & & \mathit{0} & & \end{array}\right] \tag{3.1}$$

Analysis of Bidiagonalization Golub-Kahan

Since A and B are related orthogonally, they have the same eigenvalues.

Through the equations $AV = UB$ and $A^T U = V B^T$, we are going to develop formulas for the vectors a and β.

Column partitions of U and V:

$$U = [u_1|\ldots|u_m] \quad V = [v_1|\ldots|v_n],$$

so we have

$$Av_k = a_k u_k + \beta_{k-1} u_{k-1}, \tag{3.2}$$

$$A^T u_k = a_k v_k + \beta_k v_{k+1} \tag{3.3}$$

Analysis of Bidiagonalization Golub-Kahan

For $k = 1 : n$, with the convention that $\beta_0 u_0 \equiv 0$ and $\beta_n u_{n+1} \equiv 0$, we define vectors:

$$r_k = Av_k + \beta_{k-1} u_{k-1}, \tag{3.4}$$

$$p_k = A^T u_k - a_k v_k \tag{3.5}$$

From (3.2), (3.4) and orthonormality of vectors u,we get:

$$a_k = \pm ||r_k||_2,$$
$$u_k = r_k / a_k, \quad (a_k \neq 0)$$

Note that if $a_k = 0$, then from (3.1) $A(:, 1 : k)$ is rank deficient. Likewise from (3.3) and (3.5) we get:

$$\beta_k = \pm ||p_k||_2,$$
$$v_{k+1} = p_k / \beta_k, \quad (\beta_k \neq 0)$$

Analysis of Bidiagonalization Golub-Kahan

If $\beta_k = 0$ then from the equations $AV = UB$ and $A^T U = VB^T$ we get:

$$AV(:,1:k) = U(:,1:k)B(1:k,1:k), \tag{3.6}$$

$$A^T U(:,1:k) = V(:,1:k)B(1:k,1:k)^T, \tag{3.7}$$

This calculation was first described by Golub and Kahan in 1965.

If $V_k = [v_1 | \dots | v_k]$, $U_k = [u_1 | \dots | u_k]$ and

$$B_k = \begin{bmatrix} a_1 & \beta_1 & \dots & \dots & 0 \\ 0 & a_2 & \beta_2 & \dots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \dots & 0 & a_{k-1} & \beta_{k-1} \\ 0 & \dots & \dots & 0 & a_k \end{bmatrix} \tag{3.8}$$

then after the k-th iteration we have:

$$AV_k = U_k B_k, \tag{3.9}$$

$$A^T U_k = V_k B_k^T + p_k e_k^T, \tag{3.10}$$

Algorithm of Bidiagonalization Golub-Kahan

Algorithm 1: Golub-Kahn Bidiagonalization

For $A \in \mathbb{R}^{m \times n}$ with $rank(A) = n$ and unit vector $v_c \in \mathbb{R}^n$ the following algorithm calculates the factorizations (3.6) and (3.7) for some k, with $1 \leq k \leq n$. The first column of V is v_c.

$k = 0,\ p_0 = v_c,\ \beta_0 = 1,\ u_0 = 0.$

while $b_k \neq 0$

$$v_{k+1} = p_k/\beta_k$$

$$k = k + 1$$

$$r_k = Av_k - \beta_{k-1}u_{k-1} \quad \longrightarrow 1\ mvp$$

$$a_k = \|r_k\|_2 \quad \longrightarrow 1\ ip$$

$$u_k = r_k/a_k$$

$$p_k = A^T u_k - a_k v_k \quad \longrightarrow 1\ mvp$$

$$\beta_k = \|p_k\|_2 \quad \longrightarrow 1\ ip$$

end

Skinny Golub-Kahan Regression

Skinny GK Regression

Let us suppose that we have the linear model

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

where $X \in \mathbb{R}^{m \times n}$, $m \geq n$, $rank(X) = n$. Then using Golub-Kahan Bidiagonalization X can be factored as follows:

$$X = UBV^T = [U_1 \ U_2] \begin{bmatrix} B_1 \\ 0 \end{bmatrix} V^T = U_1 B_1 V^T$$

where $U_1 \in \mathbb{R}^{m \times n}$, $V^T \in \mathbb{R}^{n \times n}$ with orthonormal columns and $B_1 \in \mathbb{R}^{n \times n}$ is bidiagonal. The columns of U_1 form an orthocanonical base for the range of X $(R(X))$, i.e.

$$R(U_1) = R(X)$$

Skinny GK Regression

The matrix X can be factored as:

$$X = U_1 B_1 V^T,\ U_1 \in \mathbb{R}^{m\times n},\ B_1 \in \mathbb{R}^{n\times n},\ V^T \in \mathbb{R}^{n\times n}$$

The solution of model (1) following the skinny GK regression is performed as follows:

$$\left.\begin{array}{l} y = X\beta = U_1 B_1 V^T \beta \\ \text{Set } \ V^T\beta = w \end{array}\right\} \Rightarrow \ y = U_1 B_1 w \Rightarrow \ U_1^T y = B_1 w \Rightarrow$$

Set $y' = U_1^T y$

Solve the bidiagonal system: $B_1 w = y'$

Specify the regression parametres $\beta :\ \beta = Vw$

Skinny GK Regression

Algorithm 2: Skinny GK Regression

1. Perform Golub-Kahan Bidiagonalization to X, i.e.:

$$X = U_1 \cdot B_1 \cdot V^T$$

   where U_1 is a $m \times n$ orthogonal matrix, B_1 is a $n \times n$ bidiagonal and V is a $n \times n$ orthogonal matrix.
2. Set $y' = U_1^T y$.
3. Solve the bidiagonal system: $B_1 w = y'$
4. The solution is given by:

$$\hat{\beta} = V \cdot w$$

An Illustrative Example

Let $X \in \mathbb{R}^{6\times 5}$ where $x_j \sim N_6(0_6, I_6)$, for $j = 1, \ldots, 4$

$$X = [x_0, x_1, x_2, x_3, x_4]$$
$$= \begin{bmatrix} 1 & 0.575413 & -1.38779 & -1.88444 & -0.442166 \\ 1 & 0.282931 & 0.46637 & 0.817164 & -0.00944475 \\ 1 & -0.105546 & 1.26251 & -0.266505 & -0.10252 \\ 1 & -0.907191 & 0.778388 & 0.109954 & 0.0591161 \\ 1 & 0.281636 & -1.53836 & -0.813759 & -1.05514 \\ 1 & -0.0720256 & 0.469853 & 2.07275 & 1.57046 \end{bmatrix}$$

We consider the linear model

$$y = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 = 1x_1 + 2x_3 + \varepsilon$$

We must find a sparse solution which leads to variable selection i.e. to identify as important factors x_1, x_3

An Illustrative Example: Choice of the noise

We set $S = \{i : \beta_i \neq 0\}$ the sparsity of the solution, ie. S will express the number of the non-zero entries.

↪ We choose $\varepsilon \sim N_6(0_6, \sigma^2 I_6)$, with $\sigma = \frac{1}{3}\sqrt{\frac{S}{6}}$.

Candes, E. , Tao, T. The Dantzig selector: Statistical estimation when p is much larger than n, The Annals of Statistics, 2007, 35, (6), 2313-51.

↪ For our example $S = 2$ and $\sigma \simeq 0.1925$, so the noise ε is

$$\varepsilon \simeq [0.0211, -0.0084, -0.0207, 0.0895, 0.0722, 0.0244]^T,$$

with $||\varepsilon||_2 \simeq 0.1214$.

↪ A response y, obtained by using the above designed model is

$$y^T \simeq [-3.1724, 1.9089, -0.6592, -0.5978, -1.2737, 4.0978]$$

An Illustrative Example

Next we perform Golub-Kahan Bidiagonalization on $X = U_1 \cdot B_1 \cdot V^T$:

$$U_1 = \begin{bmatrix} -0.398689 & -0.668795 & -0.236615 & 0.044511 & -0.517814 \\ 0.324705 & -0.358785 & -0.187102 & -0.210248 & 0.578726 \\ 0.223194 & -0.321493 & -0.0224815 & -0.75505 & -0.0727194 \\ 0.0487052 & 0.538088 & -0.600939 & -0.374722 & -0.336822 \\ -0.444753 & -0.0354572 & -0.624113 & 0.131525 & 0.483682 \\ 0.696865 & -0.172722 & -0.397311 & 0.475393 & -0.210381 \end{bmatrix}$$

$$B_1 = \begin{bmatrix} 2.77754 & 2.71892 & 0 & 0 & 0 \\ 0 & 1.40148 & 1.0689 & 0 & 0 \\ 0 & 0 & 2.10175 & 0.439786 & 0 \\ 0 & 0 & 0 & 1.80529 & 0.556666 \\ 0 & 0 & 0 & 0 & 0.957317 \end{bmatrix}$$

$$V = \begin{bmatrix} 0.194875 & -0.0335597 & -0.909471 & -0.357167 & -0.0786871 \\ 0.576222 & -0.728676 & 0.0777627 & 0.108817 & 0.345126 \\ 0.504567 & 0.233396 & 0.392991 & -0.699364 & -0.217687 \\ 0.383248 & 0.626858 & -0.0895532 & 0.239843 & 0.628193 \\ 0.478044 & 0.14311 & -0.0659851 & 0.560318 & -0.657786 \end{bmatrix}$$

where U_1 is a 6×5 and V is 5×5 orthogonal matrices and B_1 is a 5×5 bidiagonal matrix.

An Illustrative Example: The regression parameter

Next we calculate the estimation $\hat{\beta}$ from Step 4 of algorithm 2.

$$\hat{\beta} = V \cdot w \simeq \begin{bmatrix} 0.0310 \\ 0.8978 \\ -0.0463 \\ 2.0107 \\ -0.0086 \end{bmatrix}$$

which provides a very good approximation of the original designed model, with a relative error

$$Rel = \frac{||\hat{\beta} - \beta||_2}{||\beta||_2} \simeq 0.0524$$

In most applications an accuracy of $O(10^{-2})$ is satisfactory.

Sensitivity to noise

Theorem 4

Let $\Delta\beta$ and Δy perturbations in the β and y of the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$. Then it holds

$$\frac{||\Delta\beta||_2}{||\beta||_2} \le \kappa(X) \cdot \frac{||\Delta y_r||_2}{||y_r||_2}$$

where $\kappa(X)$ is the generalized condition number of X and y_r, Δy_r are the projections of vectors y and Δy onto the range of X (R(X)). It holds that $y_r = P_X y$ and $\Delta y_r = P_X \Delta y$, where $P_X = U_1 U_1^T$.

$k(X)$ serves as the condition number in the sensitivity analysis. If this number is large, then even with a small relative error in the projection of y onto $R(X)$, we can have a drastic change in the solution.

An Illustrative Example: Sensitivity analysis

We test the bound of Theorem 4 for our data set. The condition number $\kappa(X) \simeq 4.8858$. The projection y_r of y onto the range of X is

$$y_r = P_X \cdot y = U_1 U_1^T \cdot y \simeq \begin{bmatrix} -3.1734 \\ 1.9066 \\ -0.6572 \\ -0.5990 \\ -1.2721 \\ 4.0988 \end{bmatrix}, \quad \|y_r\|_2 \simeq 5.7371.$$

Thus, an upper bound for the relative error is

$$\kappa(X) \cdot \frac{\|\Delta y_r\|_2}{\|y_r\|_2} \simeq 0.1034$$

So the solution will not be perturbed much. Indeed

$$\frac{\|\Delta \beta\|_2}{\|\beta\|_2} \simeq 0.0524.$$

An Illustrative Example: The Ftest

Identification of the important factors

From the solution it is important for a fixed significance level a, to evaluate the importance of each factor.

We test the null hypothesis

$H_0 : \beta_i = 0$ against the alternatives $H_1 : \beta_i \neq 0$, $i = 1, 2, \ldots, t$.

We use the test statistic:

$$F = \frac{\hat{\beta}_i^{\,2}}{S^2 g_{ii}}$$

where $S^2 = \frac{(y - X\hat{\beta})^T (y - X\hat{\beta})}{n-t} = \frac{1}{n-t} \left\| y - X\hat{\beta} \right\|_2^2$, $n - t$ are the degrees of freedom for estimating the mean square error.

g_{ii} is the i-th diagonal element of $\left(X^T X\right)^{-1}$.

Estimating the diagonal of matrix inverses

↪ The exact computation of A^{-1} results to $\mathcal{O}(n^3)$ floating point operations. Very expensive for matrices of large size.

↪ In practice, this exact computation can be replaced by an estimate that is cheaper and faster to evaluate.

↪ We estimate the quadratic form $x^T A^{-1} x$. For $x = e_i$ estimates for the diagonal elements of A^{-1} are produced.

$$(e_i, A^{-1} e_i) = A^{-1}_{ii}$$

Bizas, E. , Mitrouli M, Turek, O. Efficient estimates for matrix inverse quadratic forms, Appl. Numer. Math., 208 (2025) 76-91

Formulae of estimates

$$(x, A^{-1}x) \approx \frac{\|x\|^4}{x^T Ax}$$

$$(x, A^{-1}x) \approx \frac{(x,Ax)^3}{\|Ax\|^4}$$

$$(x, A^{-1}x) \approx \frac{\|x\|^2 \cdot \|A^2x\| \cdot (x,Ax)}{\|Ax\| \cdot (x,A^3x)}$$

$$(x, A^{-1}x) \approx \frac{\|x\|^2 \cdot (x,Ax) \cdot (x,A^3x)}{\|A^2x\| \cdot \|Ax\|^3}$$

$$(x, A^{-1}x) \approx \frac{\|x\|^6 \cdot \|Ax\| \cdot (x,A^3x)}{\|A^2x\| \cdot (x,Ax)^3}$$

Computational complexity

- multiptications, divisions, square roots $\rightarrow \mathcal{O}(1)$
- vector x non zero of order n

mvp's	ip's	COMPLEXITY
2	5	$2n^2 + 5n + \mathcal{O}(1)$

Model Covariance Matrix

A Model Covariance Matrix A of order n is symmetric and positive definite with elements $A_{ij} = 1 + i^a, \; i = j$ and

$$A_{ij} = \frac{1}{|i-j|^b}, \; i \neq j, \quad i = 1, 2, \ldots, n \quad \text{where } a, b \in R \text{ and } b \geq 1.$$

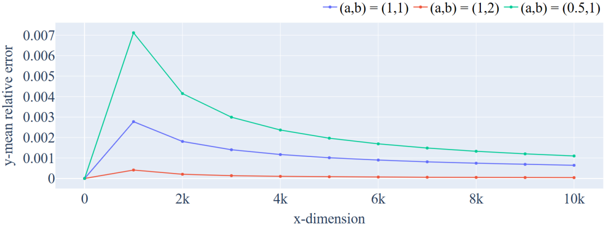


Figure: Estimation of the whole diagonal of 33 matrices up to order 10k

An Illustrative Example

In order to reduce the complexity of calculating the inverse of X^TX, we can estimate its diagonal.
Since X^TX is symmetric and positive definite, we adopt the heuristic estimation formula

$$(x, A^{-1}x) \approx \frac{\|x\|^6 \cdot \|Ax\| \cdot (x, A^3x)}{\|A^2x\| \cdot (x, Ax)^3}$$

and we apply it for $A = X^TX$ and $x = e_i$.
Direct computation of the diagonal of $(X^TX)^{-1}$ equals to $[0.1669, 1.3117, 0.3562, 0.4102, 0.9061]$. Complexity of this computation is $\mathcal{O}(n^3)$ flops.

Estimation of the diagonal of $(X^TX)^{-1}$ equals to $[0.1667, 3.100, 0.2484, 0.1712, 0.8701]$. The complexity is of order $\mathcal{O}(2n^2)$

Identification of the

— — Important Factors

The value of S^2 equals to 1.525×10^{-5} and the value of the F distribution with $a = 0.01$ with 1 degree of freedom for the numerator and with $n - t = 1$ degrees of freedom for the denominator, is about 4052.18.

The values of the test statistics for $\hat{\beta}_i$, $i = 0, 1, 2, 3, 4$ are about

$$[377.69, 40296.85, 394.71, 646343.61, 5.35]$$

The estimated important coefficients at significance level $a = 0.01$ are

$$\hat{\beta}_1, \quad \hat{\beta}_3$$

The final estimated model is

$$\hat{y} = 0.8978x_1 + 2.0107x_3 + \varepsilon$$

which provides a very good approximation of the original model.

Simulation scheme

The implementation of the simulation study has been done by using the Julia Programming Language.

Algorithm 3: Simulation scheme

1. Create the design matrix X.
2. Construct the sparse vector β.
3. Create noise vector $\varepsilon \sim N(0, \sigma^2 I)$, with $\sigma = \frac{1}{3}\sqrt{\frac{S}{n}}$.
4. Construct the response vector $y = X\beta + \varepsilon$.
5. Find $\hat{\beta}$ using skinny-GK bidiag, skinny-SVD, $\ell_p - \ell_q$ regression.

The quality of the generated approximation solution $\hat{\boldsymbol{\beta}}$ is assessed by the relative mean square error (RMSE) between $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}$ given by the formula

$$RMSE(\hat{\beta}) = E\left[\frac{\|\hat{\beta} - \beta\|^2}{\|\beta\|^2}\right]$$

Simulations

Experiment 1:

Applying skinny-GK bidiag, skinny-SVD and $\ell_p - \ell_q$ regularization on 1000 overdetermined matrices with $\boldsymbol{x_i} \in \boldsymbol{N(0_n, I_n)}$.

Buccini A., De la Cruz Cabrera O., Koukouvinos C., Mitrouli M., Reichel L.,
Variable selection in saturated and supersaturated designs via $l_p - l_q$ minimization.
Communications in Statistics - Simulation and Computation, 2023, 4326–4347.

- Result for matrices $X \in \mathbb{R}^{500\times15}$, with $x_i \in N(0_{15}, I_{15})$.

Table: Results for β vector sparsity equal to 7.

Method	RMSE	Time
skinny GK	$3.194470294566166e-10$	170.952 μs
skinny SVD	$3.193602115930959e-10$	147.311 μs
$\ell_p - \ell_q$	$6.415944573055048e-10$	330.493 μs

- Result for matrices $X \in \mathbb{R}^{1000\times20}$, with $x_i \in N(0_{20}, I_{20})$

Table: Results for β vector sparsity equal to 8.

Method	RMSE	Time
skinny GK	$9.838952325729841e-10$	449.403 μs
skinny SVD	$5.922014376447575e-11$	436.313 μs
$\ell_p - \ell_q$	$8.543321332684556e-11$	558.754 μs

Simulations

Experiment 2: Toeplitz model

Applying skinny-GK bidiag, skinny-SVD and $\ell_p - \ell_q$ regularization on 1000 overdetermined matrices with $\boldsymbol{x_i} \in \boldsymbol{N(0_n, \Sigma)}$, where Σ an $n \times n$ square matrix, with $s_{ij} = r^{|i-j|}$.

- Result for matrices $X \in \mathbb{R}^{100\times 10}$, with $x_i \in N(0_{10}, \Sigma)$ and $r = 0.2$.

Table: Results for β vector sparsity equal to 5.

Method	RMSE	Time
skinny GK	$2.2573506159651647e-8$	20.630 μs
skinny SVD	$2.2573513939068584e-8$	32.470 μs
$\ell_p - \ell_q$	$1.2951717334361495e-8$	192.552 μs

- Result for matrices $X \in \mathbb{R}^{500\times 15}$, with $x_i \in N(0_{15}, \Sigma)$ and $r = 0.15$

Table: Results for β vector sparsity equal to 7

Method	RMSE	Time
skinny GK	$3.154491834280147e-10$	174.882 μs
skinny SVD	$3.154413255261233e-10$	141.731 μs
$\ell_p - \ell_q$	$7.409345843230896e-10$	322.872 μs

- Result for matrices $X \in \mathbb{R}^{1000\times 20}$, with $x_i \in N(0_{20}, \Sigma)$ and $r = 0.1$.

Table: Results for β vector sparsity equal to 8

Method	RMSE	Time
skinny GK	$7.783456363159013e-11$	443.533 μs
skinny SVD	$5.913669150385043e-11$	444.183 μs
$\ell_p - \ell_q$	$1.0339405465156034e-10$	551.424 μs

Real datasets

Skinny GK regression is tested over real datasets.
In all experiments the significan features were correctly identified.

Dataset	Records	Features	Output	Active
Glass Ident.	214	10	Silicon	8
Turkish Music	400	51	Mean-1	15
Breast Cancer	569	31	Smoothness	13
Wine Quality	6497	13	Density	12
Gamma Telesc.	19020	11	Linearity	8

Iordanis I. Koukouvinos C., Mitrouli M., Psitou A., Golub-Kahan Regression for Variable Selection in Statistical Modeling, 14th IEEE International Conference on Dependable Systems, Services and Technologies, (DESSERT), Athens, Greece, 2024.

Real dataset: Beer Consumption - Sao Paulo

The data were collected in Sao Paulo - Brazil, in a university area with period of one year, where there are some parties with groups of students from 18 to 28 years of age.

The dataset has 5 factors:

1. Temperatura Media (C) ~ Average temperature
2. Temperatura Minima (C) ~ Minimum temperature
3. Temperatura Maxima (C) ~ Maximum temperature
4. Precipitacao (mm) ~ Precipitation
5. Final de Semana ~ Weekend

The target is:

- Consumo de cerveja (litros) ~ Beer consumption (liters)

Real dataset: Beer Consumption - Sao Paulo

We are working on a matrix with dimensions $\mathbf{365 \times 5}$ and response $\mathbf{365 \times 1}$.

Sample of the first 10 rows:

$$\left(\begin{array}{ccccc|c}
27.3 & 23.9 & 32.5 & 0.0 & 0.0 & 25.461 \\
27.02 & 24.5 & 33.5 & 0.0 & 0.0 & 28.972 \\
24.82 & 22.4 & 29.9 & 0.0 & 1.0 & 30.814 \\
23.98 & 21.5 & 28.6 & 1.2 & 1.0 & 29.799 \\
23.82 & 21.0 & 28.3 & 0.0 & 0.0 & 28.9 \\
23.78 & 20.1 & 30.5 & 12.2 & 0.0 & 28.218 \\
24.0 & 19.5 & 33.7 & 0.0 & 0.0 & 29.732 \\
24.9 & 19.5 & 32.8 & 48.6 & 0.0 & 28.397 \\
28.2 & 21.9 & 34.0 & 4.4 & 0.0 & 24.886 \\
26.76 & 22.1 & 34.2 & 0.0 & 1.0 & 37.937
\end{array}\right)$$

Real dataset: Beer Consumption - Sao Paulo

- **Results**

$$\text{skinny-GK} \quad \begin{bmatrix} 0.119197 \\ 0.114593 \\ 0.731302 \\ -0.055200 \\ 5.480156 \end{bmatrix}, \quad \text{skinny-SVD} \quad \begin{bmatrix} 0.119201 \\ 0.114589 \\ 0.731290 \\ -0.055201 \\ 5.481556 \end{bmatrix}, \quad \ell_p - \ell_q \quad \begin{bmatrix} 0.119179 \\ 0.114595 \\ 0.731303 \\ -0.055200 \\ 5.481555 \end{bmatrix}$$

Method	Time	Important factors
skinny GK	27.700 μs	3, 5
skinny SVD	25.701 μs	3, 5
$\ell_p - \ell_q$	246.692 μs	3, 5

- **Conclusion**

As the temperature increases and during weekends the beer consumption is increased!!!!

Concluding remarks

$\hookrightarrow$ For models with $\boldsymbol{x_i} \sim N(\mathbf{0}_n, \sigma_i^2 I_n)$, as the dimension increases the correlations approach zero.

$\hookrightarrow$ Matrices with small correlation tend to have a small condition number, leading to a well-conditioned system.

$\hookrightarrow$ A well-conditioned system implies that direct methods are effective and there is no need for regularization.

$\hookrightarrow$ For tall - skinny matrices, where $n \gg p$, skinny-GK regression is reliable and fast. Improvements are under study.

- **LEARNING FROM DATA HOW TO TREAT THEM**

$\hookrightarrow$ Each data set may require a different handling.

$\hookrightarrow$ Study the properties of the design matrix.

References

Buccini, A., De la Cruz Cabrera, O., Koukouvinos, C., Mitrouli, M., Reichel, L. Variable selection in saturated and supersaturated designs via-minimization. Communications in Statistics-Simulation and Computation, 2023, 4326-4347.

Bizas, E. , Mitrouli M, Turek, O. Efficient estimates for matrix inverse quadratic forms, Applied Numerical Math., 208 (2025) 76-91.

Candes, E. , Tao, T. The Dantzig selector: Statistical estimation when p is much larger than n, The Annals of Statistics, 2007, 35, (6), 2313-51.

Datta, B.N. Numerical Linear Algebra and Applications, 2nd ed.; SIAM: Philadelphia, PA, USA, 2010

Gene H. Golub, Charles F. Van Loan, D.C. Matrix Computations, (4th edition), USA, 2013.

Koukouvinos C., Lappa A., Mitrouli M., Roupa P., Turek O., Numerical methods for estimating the tuning parameter in penalized least squares problems, Communications in Statistics - Simulation and Computation, 2019, 1–22.

Koukouvinos, C., Jbilou, K., Mitrouli M., Turek, O. An eigenvalue approach for estimating the generalized cross validation function for correlated matrices, Electron. J. Linear Algebra, 2019, 35 (1), 482–496.

Koukouvinos, C., Mitrouli M., Turek, O. Using Direct Versus Regularized Solvers for Realistic Statistical Models, Springer Proceedings in Mathematics and Statistics, Vol. 497, 251–266, 2025.

Koukoudakis N., Koukouvinos C., Lappa A., Mitrouli M., Psitou A., Numerical Methods in Modelling with supersaturated designs, Applied Numerical Mathematics, 208 (2025) 271-283.

Koukoudakis N., Mitrouli M., Learning from the data how to treat correlated data sets, 2023 13th IEEE International Conference on Dependable Systems, Services and Technologies (DESSERT)

Thank you!
for your attention