Optimal estimation of direction in regression models with large number of parameters

Jonathan Gillard[a], Anatoly Zhigljavsky[a,b]

[a] *School of Mathematics, Cardiff University, Cardiff, CF24 4AG, UK*
[b] *Lobachevsky Nizhny Novgorod State University, 603950, Nizhny Novgorod, Russia*

Abstract

We consider the problem of estimating the optimal direction in regression by maximizing the probability that the scalar product between the vector of unknown parameters and the chosen direction is positive. The estimator maximizing this probability is simple in form, and is especially useful for situations where the number of parameters is much larger than the number of observations. We provide examples which show that this estimator is superior to state-of-the-art methods such as the LASSO for estimating the optimal direction.

Keywords: Random balance, Screening experiments, Box–Wilson methodology, LASSO, Ridge regression

1. Introduction

In this paper, we are mainly interested in the problem of choosing the optimal direction in regression by maximizing the probability that the scalar product between the vector of unknown parameters and the chosen direction is positive. The results obtained are very general and could be applied to models where the number of parameters m exceeds the number of observations N. It turns out that the optimal directional vector has a very simple form, see (3), and can be easily computed even if the number of parameters m is extremely large. There are two very important practical areas where our directional statistic, denoted $\hat{\theta}_*$, can be used; see also Sections 5.1 and 5.2.

Email addresses: gillardjw@cardiff.ac.uk (Jonathan Gillard), zhigljavskyaa@cardiff.ac.uk (Anatoly Zhigljavsky)

- The Box–Wilson response surface methodology, see [1, 2] and [3, Ch.8A], where an unknown response function can be observed with random error and the aim of the experimentation is in reaching the experimental conditions where the response function achieves its maximum. The main step (applied many times) in this methodology is the construction of a local linear model of the response function and the estimation of the coefficients of this linear model for finding the direction of ascent. The standard advice is to use the LSE (least square estimator) for estimating the coefficients. As shown in this paper, this standard procedure can be much improved as the LSE does not provide the optimal direction. Also, the use of $\hat{\theta}_*$ in place of the LSE can expand the use of the Box–Wilson methodology to problems with very large number of input variables.

- The so-called 'sure independence screening' procedure for regression models with huge number of parameters, see [4] as a classical reference. This procedure consists of two stages. At the first stage, a computationally efficient method is used for screening out the most important variables quickly, thus reducing the dimensionality. At the second stage, a proper regression analysis is applied to the remaining variables. Our arguments show that $\hat{\theta}_*$ is not only computationally simple but also provides an optimal screening procedure to be applied at the first stage of the sure independence screening approach.

Assume we have N observations in the linear regression model

$$y_j = \theta_1 x_{j1} + \ldots + \theta_m x_{jm} + \varepsilon_j, \quad j = 1, \ldots, N\,. \tag{1}$$

In a standard way (see e.g. [5, Ch. 4]), we write the matrix version of this observation scheme as

$$Y = X\theta + \varepsilon \tag{2}$$

where $Y = (y_1, \ldots, y_N)^T$ is the observation vector (response variable), $X = (x_{ji})_{j,i=1}^{N,m}$ is the design matrix, $\theta = (\theta_1, \ldots, \theta_m)^T$ is the vector of unknown parameters and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_N)^T$ is a vector of noise. As usual in regression models we assume $\mathbb{E}\varepsilon = 0$ and the covariance matrix of errors is $D\varepsilon = \sigma^2 W$, where σ^2 is generally unknown and W is some positive definite $N \times N$ matrix. In Section 2 we assume that W is the identity $N \times N$ matrix (that is, $W = I_N$) and extend the main results to the general $W > 0$ in Section 2.2.

The main result of the paper is Theorem 2.1 which states that if $Y \sim N(0, \sigma^2 I_N)$ then the statistic

$$\hat{\theta}_* = X^T Y \tag{3}$$

maximizes the probability

$$\Pr\{v^T \theta_{true} > 0\} \tag{4}$$

over all vectors $v \in \mathbb{R}^m$, where θ_{true} is the true value of the unknown parameters θ.

Let us make two important remarks.

Remark 1. *For any vector v, the probability* (4) *is the same for all vectors γv with $\gamma > 0$. This means that our focus is solely on the directions generated by vectors $v \in \mathbb{R}^m$ rather than on the estimation of $\theta = \theta_{true}$ in the regression model* (2)*. Moreover, Theorem 2.1 implies that under appropriate assumptions all estimators of the form $\gamma X^T Y$ with $\gamma > 0$ are optimal with respect to the criterion* (4)*.*

Remark 2. *Careful examination of the proof of Theorem 2.1 shows that for given $\theta = \theta_{true}$ there could be other directions optimal for the criterion* (4)*. A remarkable property of the direction defined by $\hat{\theta}_*$ is the fact that this direction is optimal for any θ_{true}. We can state this property by saying that $\hat{\theta}_*$ is universally optimal with respect to the criterion* (4)*.*

The rest of the paper is organised as follows. In Section 2 we prove our main result, Theorem 2.1, and show how this result can be further generalized and used. In Section 3 we give two analytic examples which show that the direction created by $\hat{\theta}_*$ could be much superior to the direction generated by the BLUE and other linear estimators of θ. In Section 4 we provide results of several numerical studies which further confirm the superiority of $\hat{\theta}_*$. As a by-product of the numerical study of Section 4 we show that the celebrated LASSO can perform very poorly in terms of the criterion (4). We make further discussions in Section 5, where we also formulate conclusions.

2. Optimality of the directional statistic

2.1. The main result

In a general linear regression model (2), consider a family of linear statistics of the form

$$\hat{\theta}_C = CY\,, \tag{5}$$

where C is some $m \times N$ matrix. Define the scalar product in $\mathbb{R}^m$ by

$$\langle a, b\rangle = a^T S b\,, \quad a, b \in \mathbb{R}^m\,,$$

where S is an arbitrary positive definite $m \times m$ matrix. For given θ, define

$$\mathcal{C}_\theta = \mathrm{Argmax}_C \mathrm{Pr}\{\langle \hat{\theta}_C, \theta\rangle > 0\}\,; \tag{6}$$

that is, $\mathcal{C}_\theta = \{\mathrm{C}_\star\}$ is the set of $m \times N$ matrices $\mathrm{C}_\star$ such that

$$\mathrm{Pr}\{\langle \hat{\theta}_{\mathrm{C}_\star}, \theta\rangle > 0\} = \max_C \mathrm{Pr}\{\langle \hat{\theta}_C, \theta\rangle > 0\}\,. \tag{7}$$

For given θ, we say that a statistic $\hat{\theta}_{\mathrm{C}}$ is optimal if $\mathrm{C} \in \mathcal{C}_\theta$. The theorem below shows that if we assume normality of errors then the matrix $C_* = S^{-1}X^T \in \mathcal{C}_\theta$ for all θ. This matrix does not depend on θ and, if $S = I_m$, the corresponding optimal statistic $\hat{\theta}_{C_\star}$ coincides with $\hat{\theta}_\star$ defined in (3).

Theorem 2.1. *Consider the model* (2) *where* $\varepsilon \sim N(0, \sigma^2 I_N)$, $\sigma^2 > 0$, *and let* S *be any positive definite* $m \times m$ *matrix. Then for any* θ, $C_\star = S^{-1}X^T$ *belongs to the set* $\mathcal{C}_\theta$ *defined in* (6).

Proof If $\theta = 0$ then the statement of the theorem is trivial. Assume $\theta \neq 0$. For simplicity of notation, denote $t(C, \theta) = \langle \hat{\theta}_C, \theta\rangle = \theta^T SCY$. Straightforward calculations give

$$\mathbb{E}t(C, \theta) = \theta^T SCX\theta,\ \mathrm{var}[t(C, \theta)] = \sigma^2 \theta^T SCC^T S\theta\,.$$

Note that $\mathrm{var}[t(C, \theta)] = 0$ if and only if the vector $a = C^T S\theta \in \mathbb{R}^N$ is equal to 0. Assume $a = 0$. Then $Y^T a = 0$ and hence $t(C, \theta) = a^T Y = 0$. This yields that if $a = 0$ then $\mathrm{Pr}\{\langle \hat{\theta}_C, \theta\rangle > 0\} = 0$. Therefore if $a = 0$, then C cannot be optimal for (7). We can then assume that $a = C^T S\theta \neq 0$. This assumption implies $\mathrm{var}[t(C, \theta)] > 0$ and we can thus define the random variable

$$v(C, \theta) = \frac{\sigma[t(C, \theta) - \mathbb{E}t(C, \theta)]}{\sqrt{\mathrm{var}[t(C, \theta)]}} = \frac{t(C, \theta) - \theta^T SCX\theta}{\sqrt{\theta^T SCC^T S\theta}}\,, \tag{8}$$

which is normally distributed with mean 0 and variance σ^2.

For any C,

$$\mathrm{Pr}\{\langle \hat{\theta}_C, \theta\rangle > 0\} = \mathrm{Pr}\{v(C, \theta) > -\varphi(C, \theta)\}\,, \tag{9}$$

where

$$\varphi(C,\theta) = \frac{\theta^T SCX\theta}{\sqrt{\theta^T SCC^T S\theta}}\,. \tag{10}$$

Therefore, the probability $\Pr\{\langle\hat{\theta}_C,\theta\rangle > 0\}$ is large when $\varphi(C,\theta)$ is large. Hence any matrix $\mathrm{C}_\star$ defined by (7) is also

$$\mathrm{C}_\star = \arg\max_C \varphi(C,\theta) \tag{11}$$

where $\varphi(C,\theta)$ is defined in (10) and the maximum in (11) is taken over the set of $m \times N$ matrices C.

By the Cauchy-Schwartz inequality, recall that for two non-zero vectors a and b, $\sqrt{a^T a}\sqrt{b^T b} \geq a^T b$ with equality if and only if $a = \alpha b$ for some non-zero constant $\alpha \in \mathbb{R}$. Set $a = C^T S\theta$ and $b = X\theta$. Then

$$\sqrt{\theta^T WCC^T S\theta}\sqrt{\theta^T X^T X\theta} \geq \theta^T SCX\theta,$$

and it follows that

$$\varphi(S^{-1}X^T,\theta) = \frac{\theta^T X^T X\theta}{\sqrt{\theta^T X^T X\theta}} \geq \frac{\theta^T SCX\theta}{\sqrt{\theta^T WCC^T S\theta}} = \varphi(C,\theta)\,,$$

for all θ and any $m \times N$-matrix C. Thus it follows that $C_\star = S^{-1}X^T$ is one of the matrices $\mathrm{C}_\star$ defined by (11). □

2.2. Generalizations of the main result and some comments

Corollary 2.2. *Consider the model* (2) *where* $\varepsilon \sim N(0,\sigma^2 W)$ *for given positive definite* $N \times N$ *matrix* W*. Then*

$$\hat{\theta}_{C_*} = S^{-1}X^T W^{-1}Y, \tag{12}$$

is an optimal linear statistic in the sense of (7)*, for any* θ*.*

Proof Make the transformations $\tilde{X} = W^{-\frac{1}{2}}X$, $\tilde{Y} = W^{-\frac{1}{2}}Y$ and $\tilde{\varepsilon} = W^{-\frac{1}{2}}\varepsilon$ and apply Theorem 2.1 to $\tilde{X}$, $\tilde{Y}$ and $\tilde{\varepsilon}$. □

Remark 3. *Consider the model* (2)*, where* $\varepsilon = (\varepsilon_1,\ldots,\varepsilon_N)^T$ *is a vector of i.i.d. random variables, not necessarily normally distributed. Then the solution to the optimization problem* (11) *is given by* $C_* = S^{-1}X^T$*, for all* θ*.*

In view of the central limit theorem, for large enough N, the random variable $v(C,\theta)$ defined in (8) is approximately normal with mean 0 and variance σ^2. Therefore, the probability $\Pr\{\langle\theta,\hat{\theta}_C\rangle > 0\}$ is large when $\varphi(C,\theta)$ is large (approximately) and its accuracy depends on the value of N. Thus the solution to the main optimization problem (7) should be either exactly the same as or very close to the solution of the problem (11); this solution is provided in Theorem 2.1.

Remark 4. *If $m = 1$ then the statistic* (12) *is proportional to the BLUE and therefore BLUE provides the optimal direction in the sense of* (7)*.*

Remark 5. *As follows from the Gauss-Markov theorem, for any $m \geq 1$ the BLUE provides the optimal direction in the sense of* (7) *if we only consider the directions made by linear unbiased estimators of θ.*

As shown in the analytical example of Section 3.1 and numerical examples of Section 4, the BLUE, despite providing the best direction compared to all linear unbiased estimators, can be rather poor relative to the direction of statistic $\hat{\theta}_{C_*}$. To measure the quality of a direction v we introduce its P-efficiency as

$$\operatorname{eff}_P(v;\theta) = \frac{\Pr\{\theta^T\hat{\theta}_{C_*} \leq 0\}}{\Pr\{\theta^T v \leq 0\}} \tag{13}$$

which depends on the unknown parameter θ. If the errors in the regression model are normal, then Theorem 2.1 and Corollary 2.2 imply that for any θ, we have $0 \leq \operatorname{eff}_P(v;\theta) \leq 1$ for any vector $v \in \mathbb{R}^m$. If the efficiency $\operatorname{eff}_P(v;\theta)$ is small then the vector v gives a poor direction, in terms of the criterion defined in (7).

3. Analytical examples

3.1. Efficiency of the BLUE direction

Theorem 2.1 implies that the P-efficiency (13) of any linear statistic $\hat{\theta} = CY$ never exceeds 1. If the errors are Gaussian then as stated in Remark 5 the BLUE direction has the highest P-efficiency among all directions computed from unbiased estimators of θ but its P-efficiency still cannot exceed 1. Let us show that P-efficiency of the BLUE direction can be very low.

Assume $N = 2$, $m = 2$ and a family of regression models $y_j = \theta_1 x_j + \theta_2 x_j^2 + \varepsilon_j$ $(j = 1, 2)$ where $x_1 = 0.5$, $x_2 = 1$, ε_1 and ε_2 are independent Gaussian random variables with zero mean and variances $\mathrm{var}(\varepsilon_1) = \alpha$ and $\mathrm{var}(\varepsilon_2) = 1 - \alpha$, where $\alpha \in (0, 1)$. Assume that $\theta = (\theta_1, \theta_2)^T \neq 0$, $\theta_2 \neq -2\theta_1$ and $\theta_2 \neq -\theta_1$.

For all $\alpha \in (0, 1)$, we have

$$\hat{\theta}_{C_*} = X^T W^{-1} Y = \begin{pmatrix} y_1/(2\alpha) + y_2/(1-\alpha) \\ y_1/(4\alpha) + y_2/(1-\alpha) \end{pmatrix},$$

$$\hat{\theta}_{BLUE} = (X^T W^{-1} X)^{-1} X^T W^{-1} Y = \begin{pmatrix} 4y_1 - y_2 \\ -4y_1 + 2y_2 \end{pmatrix}.$$

For the probability $\Pr\{\theta^T \hat{\theta}_{C_*} > 0\}$ we have

$$\Pr\{\theta^T \hat{\theta}_{C_*} > 0\} = \Phi\left(\frac{(2\theta_1 + \theta_2)^2}{16\alpha(1-\alpha)} + \frac{(2\theta_1 + 3\theta_2)(6\theta_1 + 5\theta_2)}{16(1-\alpha)} \right),$$

where $\Phi(\cdot)$ is the c.d.f. of the standard normal distribution. This probability tends to 1 if $\alpha \to 0$ or $\alpha \to 1$.

On the other hand,

$$\Pr\{\theta^T \hat{\theta}_{BLUE} > 0\} = \Phi\left(\frac{(\theta_1^2 + \theta_2^2)^2}{\alpha(15\theta_1^2 + 12\theta_2^2 - 28\theta_1\theta_2) + (\theta_1 - 2\theta_2)^2} \right).$$

This probability does not come close to 0 for any α. This means that if $\alpha \to 0$ or $\alpha \to 1$ then the P-efficiency of the BLUE tends to 0.

Assume now the true values of parameters $\theta = (\theta_1, \theta_2)^T$ are $\theta = (1, -1)^T$; note that the formulas above do not cover this particular case. Then

$$\Pr\{\theta^T \hat{\theta}_{C_*} > 0\} = \Phi\left(\frac{1}{16\alpha} \right), \quad \Pr\{\theta^T \hat{\theta}_{BLUE} > 0\} = \Phi\left(\frac{4}{9 + 55\alpha} \right)$$

and therefore

$$\mathrm{eff}_P(\hat{\theta}_{BLUE}; (1, -1)) = \frac{\Phi\left(-1/16\alpha\right)}{\Phi\left(-4/(9 + 55\alpha)\right)}$$

This yields the following expression showing the rate of decrease of the P-efficiency of the BLUE as $\alpha \to 0$:

$$\mathrm{eff}_P(\hat{\theta}_{BLUE}; (1, -1)) = \frac{8\sqrt{2}\,\alpha}{\sqrt{\pi}\Phi\left(-4/9\right)} \exp\left\{-1/(512\alpha^2)\right\} \left(1 + O(\alpha^2)\right) \quad (14)$$

If $\alpha \to 0$ then $\Pr\{\theta^T \hat{\theta}_{C_*} > 0\} \to 1$ but $\Pr\{\theta^T \hat{\theta}_{BLUE} > 0\} \to \Phi\left(4/9\right) \simeq .823 < 1$.

3.2. Comparison of different estimators in a one-parameter model

Consider the linear regression model (2) with $\sigma^2 = 1$ and some matrix $W > 0$ (which can be unknown) and a class of estimators $\hat{\theta}_A = C_A Y$ with $C_A = \left(X^T A^{-1} X\right)^{-1} X^T A^{-1}$. If A is proportional to the true covariance matrix W then $\hat{\theta}_A$ becomes the BLUE $\hat{\theta}_W$. The estimator $\hat{\theta}_A$ minimizes the weighted sum of squares

$$SS_A(\theta) = \hat{\varepsilon}(\theta)^T A^{-1} \hat{\varepsilon}(\theta)\,, \tag{15}$$

where $\hat{\varepsilon}(\theta) = (Y - X\theta)$. Define

$$SS_{error,A} = SS_A(\hat{\theta}_A) = \arg\min_{\theta} SS_A(\theta)\,.$$

If the covariance matrix W is unknown but a parametric form of it is known (so that $W = \sigma^2 W(\kappa)$ with some parameters κ and unknown multiplier σ^2), then it is customary (see, for example, [6]) to minimize the weighted sum of squares of residuals $SS_{W(\kappa)}(\theta)$ with respect to both θ and κ in the belief that for κ_{true}, the true parameters κ, we have $SS_{error,W(\kappa_{true})} \leq SS_{error,W(\kappa)}$ for all κ. In the example below we show that this could be completely wrong.

Note, however, that the danger of the simultaneous minimization of $SS_{W(\kappa)}(\theta)$ with respect to θ and κ is not unknown to statisticians, as they sometimes advocate more conservative approach to estimating θ and κ using adaptive procedures like the one studied in [7].

Set $\hat{\varepsilon}_A = \hat{\varepsilon}(\theta_A)$. We have $\hat{\varepsilon}_A = G_A \varepsilon$, where

$$G_A = I_N - X\, C_A = I_N - X\left(X^T A^{-1} X\right)^{-1} X^T A^{-1}\,.$$

Note $G_A^2 = G_A$, $\mathrm{tr} G_A = N - m$ and $G_A^T = A^{-1} G_A\, A$.

This gives

$$SS_{error,A} = \varepsilon^T G_A^T A^{-1} \varepsilon = \varepsilon^T A^{-1} G_A \varepsilon\,. \tag{16}$$

Assume that the vector of errors ε is normally distributed $\varepsilon \sim N(0, W)$. Define $\eta = W^{-\frac{1}{2}} \varepsilon$ and $B_A = W^{\frac{1}{2}} G_A^T A^{-1} W^{\frac{1}{2}}$. Then (16) can be rewritten as

$$SS_{error,A} = \eta^T B_A \eta \tag{17}$$

where $\eta \sim N(0, I_N)$.

We now consider an example. Assume $N = 2$, $m = 1$ and a family of regression models $y_j = \theta x_j + \varepsilon_j$ $(j = 1, 2)$ where $x_1 = x$, $x_2 = 1$, ε_1 and ε_2 are independent Gaussian random variables with zero mean and variances $\mathrm{var}(\varepsilon_1) = \alpha$ and $\mathrm{var}(\varepsilon_2) = 1-\alpha$. Here $x \in (0,1)$ and $\alpha \in (0, \frac{1}{2})$ are arbitrary.

BLUE of θ is $\hat{\theta}_{BLUE} = (X^T W^{-1} X)^{-1} X^T W^{-1} Y$, where in our example $X^T = (x, 1)$ and

$$(X^T W^{-1} X)^{-1} = \frac{\alpha(1-\alpha)}{\alpha + x^2(1-\alpha)} = \mathrm{var}(\hat{\theta}_{BLUE}) \, .$$

For the BLUE, $SS_{error,W} = \eta^T B_W \eta$, where B_W is a symmetric 2×2 matrix with eigenvalues 0 and 1. Hence $SS_{error,W}$ has chi-square distribution with 1 degree of freedom (d.f.) and density

$$p_W(x) = \frac{1}{\sqrt{2\pi x}} \exp\{-x/2\}, \quad x > 0 \, . \tag{18}$$

In particular, $ESS_{error,W} = 1$.

Define 2×2 matrix A as a diagonal matrix with diagonal elements $1-\alpha$ and α (that is, A is a flipped W). The variance of the estimator $\hat{\theta}_A = C_A Y$ with $C_A = (X^T A^{-1} X)^{-1} X^T A^{-1}$ is

$$\mathrm{var}(\hat{\theta}_A) = C_A W C_A^T = \frac{x^2\alpha^3 + (1-\alpha)^3}{(x^2\alpha + (1-\alpha))^2} \, . \tag{19}$$

By the Gauss-Markov theorem, $\mathrm{var}(\hat{\theta}_A) \geq \mathrm{var}(\hat{\theta}_{BLUE})$. The efficiency of the estimator $\hat{\theta}_A$ is

$$\mathrm{eff}(\hat{\theta}_A) = \frac{\mathrm{var}(\hat{\theta}_{BLUE})}{\mathrm{var}(\hat{\theta}_A)} = \frac{\alpha \, (1-\alpha) \, (x^2\alpha - \alpha + 1)^2}{(x^2 \, (1-\alpha) + \alpha) \left(x^2\alpha^3 + (1-\alpha)^3\right)} \leq 1 \, .$$

For any fixed $x \in (0,1)$, the efficiency of $\hat{\theta}_A$ can be arbitrary small if α is small enough. Indeed, if $x \in (0,1)$ is fixed and $\alpha \to 0$ then $\mathrm{var}(\hat{\theta}_{BLUE}) \to 0$ and $\mathrm{var}(\hat{\theta}_A) \to 1$ implying $\mathrm{eff}(\hat{\theta}_A) \to 0$. More precisely,

$$\mathrm{eff}(\hat{\theta}_A) = \frac{\alpha}{x^2} + \frac{(2\,x^2 - 1) \, (x^2 + 1)}{x^4} \alpha^2 + O\left(\alpha^3\right) \, .$$

Consider now $SS_{error,A}$. In view of (17) it can be written as $SS_{error,A} = \eta^T B_A \eta$, where B_A is symmetric with eigenvalues 0 and

$$\lambda_{x,\alpha} = (x^2\alpha + 1 - \alpha)/(x^2(1-\alpha) + \alpha) > 1 \, .$$

Therefore $SS_{error,A}$ is a random variable (r.v.) which is $\sqrt{\lambda_{x,\alpha}}$ times a r.v. with chi-square distribution with 1 d.f. The density of $SS_{error,A}$ is

$$p_A(t) = \frac{1}{\sqrt{2\pi\lambda_{x,\alpha}t}} \exp\{-\sqrt{\lambda_{x,\alpha}}t/2\}, \quad t > 0\,.$$

In particular,

$$\mathbb{E}SS_{error,A} = \frac{1}{\lambda_{x,\alpha}} = \frac{x^2(1-\alpha)+\alpha}{x^2\alpha-\alpha+1}\,.$$

This value is always smaller than 1 as long as $x \in (0,1)$ and $\alpha \in (0, \frac{1}{2})$. For small x and α the value of $\mathbb{E}SS_{error,A}$ is close to 0. Indeed, if we assume that $x = \alpha$ then

$$\mathbb{E}SS_{error,A} = \alpha\,\frac{1+\alpha-\alpha^2}{1-\alpha+\alpha^3} = \alpha + 2\,\alpha^2 + \alpha^3 + O\left(\alpha^5\right) \quad \text{as } \alpha \to 0\,.$$

Let us assume now that $\theta > 0$ and consider the probability

$$\Pr\{\langle\hat{\theta},\theta\rangle > 0\} = \Pr\{\hat{\theta} > 0\}$$

for the following three estimators: $\hat{\theta}_* = X^T W^{-1} Y = \tilde{X}^T\tilde{Y}$, $\hat{\theta}_{BLUE} = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\tilde{Y}$ and $\hat{\theta}_A = (X^T A^{-1} X)^{-1} X^T A^{-1} Y$; here $\tilde{X} = W^{-1/2}X$ and $\tilde{Y} = W^{-1/2}Y$. In view of (9) with $m = 1$, for any $\hat{\theta} = C^T\tilde{Y}$, where C is a vector of size N, we have

$$\Pr\{C^T\tilde{Y} > 0\} = 1 - \Phi\left(-\theta C^T\tilde{X}/\sqrt{C^TC}\right)\,.$$

This gives

$$\Pr\{\hat{\theta}_* > 0\} = \Pr\{\hat{\theta}_{BLUE} > 0\} = 1 - \Phi\left(-\theta\sqrt{\frac{x^2}{\alpha} + \frac{1}{1-\alpha}}\right)\,.$$

These probabilities approach 1 exponentially fast as $\alpha \to 0$. For example, for $\theta = 1$ and $x = 0.5$, $\Pr\{\hat{\theta}_* > 0\} < 1.7 \cdot 10^{-5}$ for all $\alpha \le 0.1$. On the other hand,

$$\Pr\{\hat{\theta}_A > 0\} = 1 - \Phi\left(-\theta/\sqrt{\mathrm{var}(\hat{\theta}_A)}\right)\,,$$

where $\mathrm{var}(\hat{\theta}_A)$ is given in (19). For small α, $\mathrm{var}(\hat{\theta}_A) \simeq 1$ and hence the probability $\Pr\{\hat{\theta}_A > 0\}$ is close to $1 - \Phi(-\theta)$. For $\theta = 1$ this is $1 - \Phi(-1) \simeq$

0.841345, which is much smaller than 1. For any θ, the P-efficiency of the direction created by $\hat{\theta}_A$ tends to 0 as $\alpha \to 0$. This is counter-intuitive to the fact established above which says that for small x and α the value of $SS_{error,A}$ is probabilistically close to 0 whereas the distribution of $SS_{error,W}$ is not (it has the density (18)).

4. Numerical examples

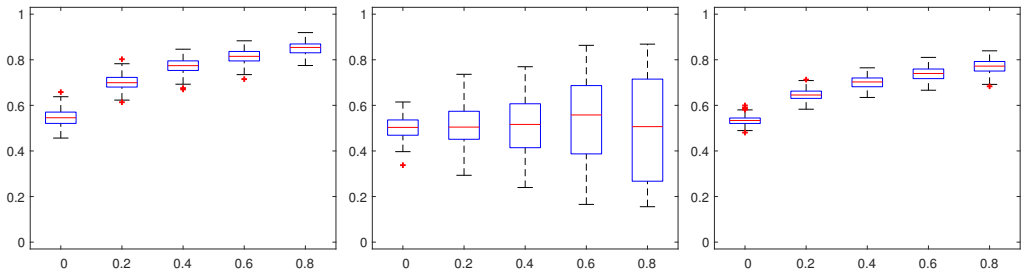
Suppose we have data consisting of N observations taken on m variables $X_1, X_2, \ldots, X_m$ drawn from a multivariate normal distribution with zero mean and $m \times m$ covariance matrix $\Sigma = (\Sigma_{i,j})_{m\times m}$. The entries of Σ are given by $\Sigma_{i,j} = 1, i = j$ and $\Sigma_{i,j} = \rho, i \neq j$. Suppose that the model is of the form

$$Y = \beta X_1 + \beta X_2 + \beta X_3 + \varepsilon,$$

so $\theta = (\beta, \beta, \beta, 0, \ldots, 0)^T$ and $\varepsilon \sim N(0, I_N)$. This is one of the standard models used in studying variable selection in problems with large number of parameters, see [4]. We use $S = I_m$ and are hence interested in the event $\langle \hat{\theta}, \theta \rangle = \hat{\theta}^T \theta > 0$.

Example 1. In this example we take $m = 100$ and evaluate (9) for three cases: (i) $C = C_1 = X^T$ in accordance with the statement of Theorem 2.1, (ii) $C = C_2 = (X^T X)^- X^T$ and (iii) $C = C_3 = X^T (XX^T)^-$. Here $(XX^T)^-$ is the Moore-Penrose pseudoinverse of the matrix XX^T. Note that the choice $C = C_2$ makes the estimator (5) the standard ordinary least squares estimator and the matrix $(X^T X)^- X^T$ is a pseudoinverse for X commonly used when $m < N$. The matrix $X^T (XX^T)^-$ is a pseudoinverse for X commonly used when $N > m$.

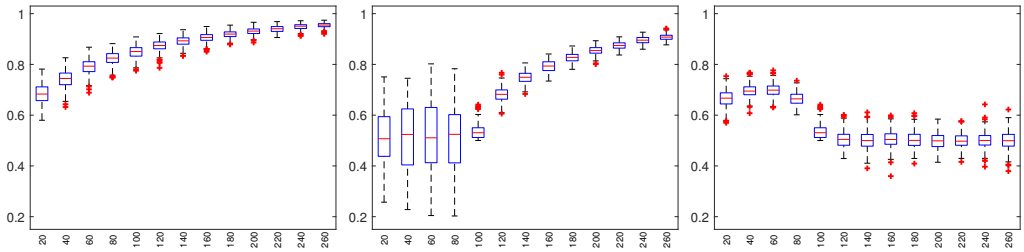
Figure 1 contains boxplots of the probabilities computed from (9), for different values of ρ, taken over 250 simulated data sets with $N = 20$ and $\beta = 0.05$. We make the following remarks. The choice $C = C_1$ yields larger probabilities across all values of ρ, and the probabilities increase as a function of ρ. Such a trend is not apparent with the choice $C = C_2$, where the distribution of the probabilities becomes more variable as ρ increases. The larger value of β gives larger probabilities for the choice $C = C_1$, whilst it makes the distribution of the probabilities more variable for the choice $C = C_2$. The choice $C = C_3$ gives larger probabilities with increasing ρ, but they are not as large as those from $C = C_1$.



(a) $C = C_1 = X^T$ (b) $C = C_2 = (X^T X)^- X^T$ (c) $C = C_3 = X^T (X X^T)^-$

Figure 1: Boxplots of the probabilities computed from (9), for different values of ρ, taken over 250 simulated data sets with $N = 20$.

Figure 2 contains boxplots of the probabilities computed from (9), for different values of N, taken over 250 simulated data sets with $\rho = 0.4$ and $\beta = 0.05$. Again the choice $C = C_1$ yields larger probabilities across all values of N, and the probabilities increase when N grows. The probabilities when $C = C_2$ and $C = C_3$ relate to the discussion given earlier on the correct choice of pseudoinverse for X depending on the dimension N. Consider first the choice $C = C_2$. The probabilities from (9) are smaller than for any other choice for C when $N < 100$. This is because $(X^T X)^- X^T$ is the wrong psuedoinverse for X when $N < 100$. The discussion is similar for the choice $C = C_3$ when $N > 100$. However the choice $C = C_3$ has a smaller variation of probabilities when $N > 100$ than $C = C_2$ when $N < 100$.

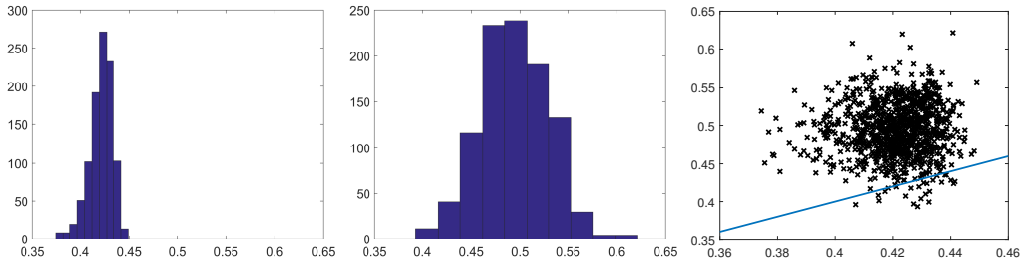


(a) $C = C_1 = X^T$ (b) $C = C_2 = (X^T X)^- X^T$ (c) $C = C_3 = X^T (X X^T)^-$

Figure 2: Boxplots of the probabilities computed from (9), for different values of N, taken over 250 simulated data sets with $\rho = 0.4$.

We now consider the angles between the two vectors θ and $\hat{\theta}$, where $\hat{\theta}$ is an estimate of θ. We estimate θ by using the estimator (5) with $C = C_1 = X^T$ and $C = C_2 = (X^T X)^- X^T$. Figure 3 contains plots of the angles found from

1000 simulated data sets. The choice $C = C_1$ yields an estimate with smaller angle between the estimate $\hat{\theta}$ and θ, than the choice $C = C_2$.



(a) $C = C_1 = X^T$

(b) $C = C_2 = (X^T X)^- X^T$

(c) Scatterplot of angles from using $C = C_2$ against $C = C_1$ with '$y = x$' line

Figure 3: Plots of angles between the two vectors θ and $\hat{\theta}$. Parameters are $m = 100$, $N = 20$, $\rho = 0.4$ and $\beta = 0.05$. Plots standardised in the interval $[0, 1]$ such that 1 represents π radians.

Example 2. We simulate 250 data sets, and for each data set estimate θ by using the estimator (5) with $C = C_1 = X^T$ in accordance with the statement of Theorem 2.1 with $\gamma = 1$ and by using the LASSO. For the 250 simulated data sets we count the frequencies of the event $\langle \hat{\theta}, \theta \rangle > 0$, where $\hat{\theta}$ is either $\hat{\theta}_{RB} = X^T Y$ or LASSO. Note that because the LASSO estimate of θ cannot, in general, be written in the form (5) we are unable to use (9). The LASSO estimate of θ is given by the solution to the following optimization problem:

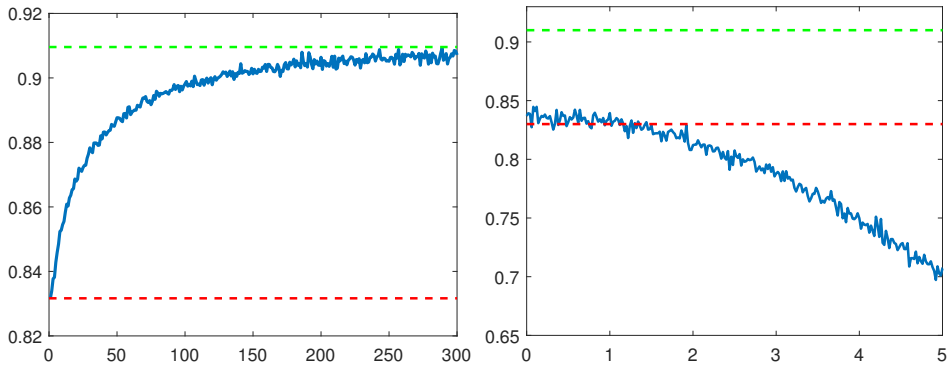
$$\hat{\theta}_\lambda = \arg\min_{\theta \in \mathbb{R}^m} ||Y - X\theta||_2^2 + \lambda ||\theta||_1 \,. \tag{20}$$

Table 1 contains these proportions for different estimators of θ, and different m when $N = 20$, $\beta = 0.05$. We compute the LASSO estimator (20) of θ for each simulated data set by taking 100 equally spaced values of λ, in the interval $[0, \lambda_*]$ where $\lambda_* > 0$ is the largest value of λ which gives a non-null solution of (20). In Table 1 the column headed $0.1\lambda_{\max}$ refers to the proportion of solutions of (20) with $\lambda = 0.1\lambda_{\max}$ where $\langle \hat{\theta}, \theta \rangle > 0$. Other columns are described similarly. We make the following remarks. The proportion of the simulations which yield positive values of $\langle \hat{\theta}, \theta \rangle$ is greater for the estimator (5) with $C = C_1 = X^T$ than for the LASSO, and it is less affected by m. The proportion increases with larger ρ for this estimator, whilst it decreases for the LASSO. As λ in (20) increases, the proportion decreases.

Table 1: Proportion of 250 simulations which give $\langle\hat{\theta},\theta\rangle > 0$ for different m and ρ.

	$m = 100$				$m = 1000$			
ρ	$C = C_1$	$0.1\lambda_*$	$0.5\lambda_*$	$0.9\lambda_*$	$C = C_1$	$0.1\lambda_*$	$0.5\lambda_*$	$0.9\lambda_*$
0	0.648	0.344	0.304	0.252	0.604	0.048	0.040	0.032
0.2	0.684	0.308	0.264	0.208	0.668	0.064	0.064	0.048
0.4	0.692	0.236	0.216	0.176	0.700	0.036	0.020	0.024
0.6	0.704	0.248	0.232	0.144	0.756	0.044	0.032	0.016
0.8	0.720	0.244	0.224	0.116	0.728	0.048	0.028	0.016

Example 3. In this example we take $m = 100$, $N = 200$, $\rho = 0.4$ and evaluate (9) with $C = C_a = (X^TX + aI_m)^{-1}X^T$, where $a > 0$ is the so-called ridge parameter. We set $\beta = 0.05$. Figure 4(a) contains a plot of the average probability (evaluated using (9)) over 250 simulated data sets, against a. Figure 4(b) contains a plot of the average probability obtained from the LASSO (20) against λ. The lower horizontal line is the average probability when $a = 0$. The upper horizonal line is the average probability when $C = C_1 = X^T$. Note that as $a \to \infty$ then $a(X^TX+aI_m)^{-1}X^T \to X^T$. This is why the probability as a gets larger becomes the same as for when $C = C_1 = X^T$. The LASSO yields smaller probabilities than when (9) is evaluated when $C = C_a = (X^TX + aI_m)^{-1}X^T$. As the penalty coefficient λ in (20) is taken larger then the average probability declines.



(a) $C = C_a = (X^TX + aI_m)^{-1}X^T$ (b) LASSO with penalty coefficient λ

Figure 4: $\Pr\{\langle\hat{\theta},\theta\rangle > 0\}$ for different estimators $\hat{\theta}$ of θ averaged over 250 simulations, and for different a and λ respectively.

5. Further discussions and conclusions

5.1. Random balance estimator

Consider the regression model (1) where errors are i.i.d. normal. By Theorem 2.1 and Remark 1, all estimators of the form $\gamma X^T Y$ with $\gamma > 0$ are optimal with respect to the criterion (4). Let us use $\gamma = 1/N$.

Assume that $x_{ji} = \pm 1$ for all i, j. In this case, the statistic $\hat{\theta}_{RB} = \frac{1}{N} X^T Y$ is known as the Random Balance (RB) estimator of θ, see [8] and also [9, 10, 11]. The RB estimator $\hat{\theta}_{RB,i}$ of an individual parameter θ_i is the LSE (least-squares estimator) in the one-parameter regression model $y_j = \theta_i x_{ji} + \varepsilon_j^{(i)}$, where $\varepsilon_j^{(i)} = \varepsilon_j + \sum_{k \neq i} \theta_k x_{jk}$. In this approach, all θ_i's are estimated one-by-one by merging the input of other terms in the original model with noise. Obviously, this estimation method produces the same estimator if the assumption $x_{ji} = \pm 1$ for all i, j is replaced with a more general assumption $\sum_{j=1}^N x_{ji}^2 = N$. Note that if $\sum_{j=1}^N x_{ji}^2$ has different values as i varies then the statistic (3) is not associated with any classical method of estimating the parameters θ.

5.2. The sure independence screening

Let us give more details about the sure independence screening. Assume the inputs x_{ji} are normalised so that $\sum_{j=1}^N x_{ji}^2 = 1$ for all $i = 1, \ldots, m$. Then the statistic (3) coincides with the estimator used in the sure independence screening. In this approach, the statistic (3) simply provides a 'cheap' estimator of θ. It can be proved that under certain conditions the use of (3) can significantly reduce the dimensionality while preserving the true model with overwhelming probability. In problems of ultrahigh dimensions there are three problems when trying to identify variables that contribute most to the response: computational cost, statistical accuracy and model interpretability [12]. Existing variable selection methods (such as the LASSO [13, 14]) can become computationally burdensome in high dimensions. The LASSO can give non-consistent models if certain conditions are not met [14]. Some methods have been developed to circumnavigate this problem, but they are computationally intensive [15, 16].

Work which explains why $\hat{\theta}_*$ is a good estimator for screening out the most important variables is based around the following ideas:

- Assume that Y and the columns of X are all standardized so that $\sum_{j=1}^N y_j^2 = 1$ and $\sum_{j=1}^N x_{ji}^2 = 1$ for all i. Then $\hat{\theta}_*$ is the vector of

marginal correlations of the variables in the design matrix X with the response variable Y. The justification of this estimator is that if $\theta_j \neq 0$ then with high probability $X_j^T Y \neq 0$, where X_j is the j-th column of the design matrix X.

- For high-dimensional problems the matrix $X^T X$ is likely to be singular or nearly singular. This causes problems both in theory and in numerical computations. Let $\hat{\theta}_a$ denote the so-called ridge estimator of θ, given by $\hat{\theta}_a = (X^T X + aI_m)^{-1} X^T Y$, for $a > 0$. It is clear that $\hat{\theta}_a \to (X^T X)^{-1} X^T Y$ as $a \to 0$ and similarly $a\hat{\theta}_a \to \hat{\theta}_*$ as $a \to \infty$. The estimator $a\hat{\theta}_a$ becomes less dependent on the degeneracy of $X^T X$ for larger a. Since ranking of the absolute values of components of $\hat{\theta}_a$ is the same as $a\hat{\theta}_a$, $\hat{\theta}_*$ can be viewed as a special case of the ridge estimator $\hat{\theta}_a$ with $a = \infty$.

- It has been shown that variable selection based on $\hat{\theta}_*$ does preserve the true model with high probability (this is known as the sure screening property). This property is considered pivotal for the success of the approach. See for example [4].

5.3. Computability and presentation of asymptotic results using the concept of grossone

In Section 3, we have based some of our conclusions on certain asymptotic expansions and limiting relations. All these relations can be easily rewritten in the language of 'grossone' developed by Ya.Sergeyev, see for example [17, 18, 19] and also [20], where some logical arguments related to the grossone are discussed. For example, the relation (14) written in the language of grossone has the following form:

$$\mathrm{eff}_P(\hat{\theta}_{BLUE}; (1, -1)) = \frac{8\sqrt{2}}{\sqrt{\pi}\Phi\left(-4/9\right)\ ①} \exp\left(-\ ①^2/512\right) \left(1 + O\left(\frac{1}{①^2}\right)\right),$$

where the grossone ① can be thought of as 'numerical infinity'. The advantage of writing asymptotic expressions in the form involving the grossone is two-fold: first, one can easily evaluate the limiting values even if these limiting values cannot be directly computed and may have little sense (in our example, $\alpha = 0$ cannot be used as the matrix W^{-1} does not exist when $\alpha = 0$); second, all the calculations can be made on the so-called 'infinity computer' (see e.g. [19]). Calculations on the infinity computer (under the

assumption of its existence) give us a possibility of operating with infinite and infinitesimals as with numbers and not symbols (like in MAPLE) which makes computations much faster.

5.4. Conclusions

We have considered the problem of estimating the optimal direction in regression by maximizing the probability that the scalar product between the vector of unknown parameters and the chosen direction is positive. For the case when the errors are normal we have derived the explicit form for the universally optimal estimator. It appears that this estimator is simple in form, does nor require matrix inversion and hence is especially useful for situations where the number of parameters is larger than the number of observations. We have shown that in particular cases our universally optimal estimator coincides with the random balance estimator and the estimator used in the sure independence screening approach. We have provided examples which demonstrate that our estimator is superior to the BLUE and the state-of-the-art methods such as the LASSO.

Acknowledgement

The work of the second author was supported by the Russian Science Foundation, project No. 15-11-30022 'Global optimization, supercomputing computations, and applications'.

References

[1] G. E. Box, K. Wilson, On the experimental attainment of optimum conditions, in: Breakthroughs in Statistics, Springer, 1992, pp. 270–310.

[2] W. J. Hill, W. G. Hunter, A review of response surface methodology: a literature survey, Technometrics 8 (4) (1966) 571–590.

[3] W. Cochran, G. Cox, Experimental designs, 2nd Edition, Wiley, 1957.

[4] J. Fan, J. Lv, Sure independence screening for ultrahigh dimensional feature space, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70 (5) (2008) 849–911.

[5] C. R. Rao, H. Toutenburg, H. C. Shalabh, M. Schomaker, Linear models and generalizations, Least Squares and Alternatives (3rd edition) Springer, Berlin Heidelberg New York.

[6] A. P. Dempster, Covariance selection, Biometrics (1972) 157–175.

[7] V. Fedorov, Regression problems with controllable variables subject to error, Biometrika (1974) 49–56.

[8] F. Satterthwaite, Random balance experimentation, Technometrics 1 (2) (1959) 111–137.

[9] L. Meshalkin, On the substantiation of the random balance method, Zavodskaya Laboratory 36 (3) (1970) 316–318.

[10] M. B. Maljutov, H. P. Wynn, Sequential screening of significant variables of an additive model, in: The Dynkin Festschrift, Springer, 1994, pp. 253–265.

[11] A. Slinko, A generalization of Komlos theorem on random matrices, New Zealand Journal of Mathematics 30 (2001) 81–86.

[12] J. Fan, J. Lv, A selective overview of variable selection in high dimensional feature space, Statistica Sinica 20 (1) (2010) 101.

[13] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society. Series B (Statistical Methodology) 58 (1) (1996) 267–288.

[14] P. Zhao, B. Yu, On model selection consistency of lasso, Journal of Machine Learning Research 7 (Nov) (2006) 2541–2563.

[15] N. Meinshausen, P. Bühlmann, Stability selection, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72 (4) (2010) 417–473.

[16] R. D. Shah, R. J. Samworth, Variable selection with error control: another look at stability selection, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75 (1) (2013) 55–80.

[17] Y. D. Sergeyev, Numerical computations and mathematical modelling with infinite and infinitesimal numbers, Journal of Applied Mathematics and Computing 29 (1) (2009) 177–195.

[18] Y. D. Sergeyev, Numerical point of view on calculus for functions assuming finite, infinite, and infinitesimal values over finite, infinite, and infinitesimal domains, Nonlinear Analysis: Theory, Methods & Applications 71 (12) (2009) e1688–e1707.

[19] Y. D. Sergeyev, Higher order numerical differentiation on the infinity computer, Optimization Letters 5 (4) (2011) 575–585.

[20] G. Lolli, Metamathematical investigations on the theory of grossone, Applied Mathematics and Computation 255 (2015) 3–14.