Optimal directional statistic for general regression

Jonathan Gillard and Anatoly Zhigljavsky
Cardiff School of Mathematics
Cardiff University
{GillardJW,ZhigljavskyAA}@Cardiff.ac.uk

Abstract

For a general linear regression model we construct a directional statistic which maximizes the probability that the scalar product between the vector of unknown parameters and any linear estimator is positive. Special emphasis is given to comparison of this directional statistic with the BLUE and explaining why the BLUE could be relatively poor. We illustrate our results on analytical and numerical examples.

Keywords: Regression; Optimal direction; BLUE; Correlated errors.

1 Introduction and formulation of the main result

Consider the general linear regression model

$$y(x) = \theta^T f(x) + \varepsilon(x), \;\; x \in \mathcal{X}\,, \tag{1}$$

where $\mathcal{X}$ is a bounded Borel subset of $\mathbb{R}^d$ with $d \geq 1$, $\theta = (\theta_1, \ldots, \theta_m)^T$ is a vector of unknown parameters, $f = (f_1, \ldots, f_m)^T$ is a vector of base functions and $\varepsilon(x)$ is a Gaussian random noise process (or field) with zero mean and finite covariances $\mathbb{E}\varepsilon(x)\varepsilon(x') = \sigma^2 K(x, x')$ for $x, x' \in \mathcal{X}$, where $\sigma^2 > 0$ may be unknown and $K(\cdot, \cdot)$ is a known positive definite function (kernel) on $\mathcal{X} \times \mathcal{X}$. By Ξ_1 and Ξ_m we denote the linear spaces of all finite signed measures defined on $\mathcal{X}$ and all signed m-vector measures on $\mathcal{X}$, respectively.

We assume that the kernel $K(\cdot, \cdot)$ is integrally strictly positive definite (shortly, ISPD); that is, $\varphi(\nu) := \int\int K(x, x')\nu(dx)\nu(dx') > 0$ for any non-zero signed measure $\nu \in \Xi_1$; the integration domain is always assumed to be $\mathcal{X}$. ISPD kernels are studied in [8]; the majority of classical covariance kernels are ISPD. Note that since K is bounded and $\mathcal{X}$ is also bounded, $\varphi(\nu) < \infty$ for all $\nu \in \Xi_1$.

For any $\zeta \in \Xi_m$, there is an associated linear statistic

$$\widehat{\theta}_\zeta = \int y(x)\zeta(dx) \in \mathbb{R}^m\,. \tag{2}$$

The set of statistics (2) contains the set of linear estimators of θ but it is much broader than the latter.

We will be mostly interested in two particular linear statistics of the form (2). The first one is the BLUE (best linear unbiased estimator) of θ. Assume that the model (1) is such that the BLUE of θ of the form (2) exists; this is equivalent to the existence of the vector measure $\xi \in \Xi_m$ such that

$$\int K(x,x')\xi(dx') = f(x), \ \ \forall x \in \mathcal{X} \tag{3}$$

and the non-degeneracy of the matrix

$$C = \int\int K(x,x')\xi(dx)\xi^T(dx) = \int f(x)\xi^T(dx)\,. \tag{4}$$

In this case, the BLUE of θ is

$$\widehat{\theta}_{BLUE} = \int y(x)\zeta_{BLUE}(dx)\,; \tag{5}$$

with $\zeta_{BLUE}(dx) = C^{-1}\xi(dx)$; the covariance matrix of $\widehat{\theta}_{BLUE}$ is $\sigma^2 C^{-1}$, see [7, Theorem 2.3] for details. The second statistic of the form (2), which is of serious interest to us, is the statistic $\widehat{\theta}_\xi$ with ξ satisfying (3).

For the BLUE to exist the functions $f_1,\ldots,f_m$ should be linear independent on $\mathcal{X}$ and belong to $H(K)$, the reproducing kernel Hilbert space associated with the kernel K. In this case, if either $\mathcal{X}$ is discrete with $N \geq m$ distinct points or $\mathcal{X}$ is a regular closed subset of $\mathbb{R}^d$ ($\mathcal{X}$ is the closure of its interior) and $K(x,x')$ is continuous on $\mathcal{X}\times\mathcal{X}$ but the derivatives $\partial K(x,x')/\partial x$ are discontinuous on the diagonal $x = x'$, then the BLUE exists. Otherwise, the BLUE exists for only very specific functions $f_1,\ldots,f_m$ given the kernel K and set $\mathcal{X}$. For more information see [4, Sect.2.6].

For any $\theta \neq 0$ and any $\widehat{\theta}_\zeta \neq 0$, the cosine of the angle between $\widehat{\theta}_\zeta$ and θ is

$$c_{\zeta,\theta} = \frac{\widehat{\theta}_\zeta^T\theta}{\|\widehat{\theta}_\zeta\|\cdot\|\theta\|}\,. \tag{6}$$

We are interested in the probability

$$p_{\zeta,\theta} = \begin{cases} \Pr\{c_{\zeta,\theta} > 0\} & \text{if } \ \widehat{\theta}_\zeta \neq 0\,, \\ \frac{1}{2} & \text{if } \ \widehat{\theta}_\zeta = 0\,, \end{cases} \tag{7}$$

which is the probability that the angle between $\widehat{\theta}_\zeta$ and θ is acute; if $\widehat{\theta}_\zeta = 0$ and hence the direction is not defined then we assume that this direction is random and uniformly distributed. The probability (7) is only defined if $\theta \neq 0$. Note also that the value of the probability $p_{\zeta,\theta}$ is the same for all vectors of the form $c\widehat{\theta}_\zeta$ with any $c > 0$.

The main result of the paper is the following theorem.

Theorem 1.1 *Let ξ satisfy (3) and $\theta \neq 0$. Then we have: (a) $p_{\xi,\theta} \geq p_{\zeta,\theta}$ for any $\zeta \in \Xi_m$, and (b) if the signed measures $\theta^T\xi(\cdot)$ and $\theta^T\zeta(\cdot)$ are not positively proportional (so that there is no $c > 0$ such that $\theta^T\zeta(\cdot) = c\,\theta^T\xi(\cdot)$) then $p_{\xi,\theta} > p_{\zeta,\theta}$.*

Theorem 1.1 states that in the class of all linear statistics, the linear statistic $\widehat{\theta}_\xi$ with ξ satisfying (3) provides the best possible estimator of the direction of θ (whatever the value of θ), from the point of view of the criterion (7). We shall call $\widehat{\theta}_\xi$ 'the optimal directional statistic'.

We can see at least two important practical areas where the optimal directional statistic can be used. First, this is the Box–Wilson response surface methodology, see [2, 6], where an unknown response function is observed with random error and the aim of the experimentation is to determine the experimental conditions where the response function has its maximum. The main step in this methodology (this step is applied many times) is estimation of coefficients of a local linear model for finding the direction of ascent. The standard advice is to use the OLSE for estimating these coefficients. As shown in this paper, this standard procedure can be much improved as the OLSE is not a good estimator of the direction of ascent. Second, it is the so-called 'sure independence screening' procedure for regression models with many parameters, see e.g. [5]. This procedure consists of two stages. At the first stage, a computationally efficient method is used for screening out the most important variables thus reducing the dimensionality. At the second stage, a proper regression analysis is applied to the remaining variables. Our arguments show that the optimal directional statistic 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.

The rest of the paper is organized as follows. Section 2 contains some discussions, extensions and analytic expressions for the probability (3) in important special cases. Section 3 proves Theorem 1.1. Analytic and numerical examples are discussed in Section 4 and the paper is concluded in Section 5.

2 Explicit formulae and extensions

2.1 Expressing the probability (7) via the c.d.f. of the standard normal distribution

Let $\zeta \in \Xi_m$ be an arbitrary signed vector-measure and $\widehat{\theta}_\zeta$ be the associated linear statistic (2). For any $\theta \in \mathbb{R}^m$, the random variable $\theta^T\widehat{\theta}_\zeta = \theta^T \int y(x)\zeta(dx)$ has normal distribution with mean and variance

$$\mathbb{E}\theta^T\widehat{\theta}_\zeta = \theta^T \left[\int f(x)\zeta^T(dx)\right]\theta\,, \quad \mathrm{var}(\theta^T\widehat{\theta}_\zeta) = \sigma^2\,\theta^T \left[\int\int K(x,x')\zeta(dx)\zeta^T(dx')\right]\theta\,. \tag{8}$$

In the following lemma we express the probability (7) through the c.d.f. of the standard normal distribution.

Lemma 2.1 *Let $\zeta \in \Xi_m$ be any signed vector-measure. For any $\theta \in \mathbb{R}^m \setminus \{0\}$, the probability (7) is equal to $p_{\zeta,\theta} = \Phi(t_{\zeta,\theta}/\sigma)$, where*

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-u^2/2} du$$

is the c.d.f. of the standard normal distribution and

$$t_{\zeta,\theta} = \begin{cases} \frac{\theta^T\left[\int f(x)\zeta^T(dx)\right]\theta}{\sqrt{\theta^T\left[\int\int K(x,x')\zeta(dx)\zeta^T(dx')\right]\theta}} & \text{if } \operatorname{var}(\theta^T\widehat{\theta}_\zeta) \neq 0\,, \\ 0 & \text{if } \operatorname{var}(\theta^T\widehat{\theta}_\zeta) = 0\,. \end{cases} \tag{9}$$

Proof of Lemma 2.1. If $\operatorname{var}(\theta^T\widehat{\theta}_\zeta) = 0$ the statement follows from the definition of $p_{\zeta,\theta}$. Assuming $\operatorname{var}(\theta^T\widehat{\theta}_\zeta) \neq 0$, the standardized random variable $\kappa = [\mathbb{E}\theta^T\widehat{\theta}_\zeta - \theta^T\widehat{\theta}_\zeta]/\sqrt{\operatorname{var}(\theta^T\widehat{\theta}_\zeta)}$ has normal distribution with zero mean and variance 1 and hence

$$p_{\zeta,\theta} = \Pr\{c_{\zeta,\theta} > 0\} = \Pr\{\theta^T\widehat{\theta}_\zeta > 0\} = \Pr\left\{\kappa < \mathbb{E}\theta^T\widehat{\theta}_\zeta/\sqrt{\operatorname{var}(\theta^T\widehat{\theta}_\zeta)}\right\} = \Phi(t_{\zeta,\theta}/\sigma)\,.$$

□

2.2 Probability (7) in important special cases

Case 1: $\zeta = \xi$ satisfy (3). From (4) we obtain

$$t_{\xi,\theta} = \sqrt{\theta^T C\theta} \tag{10}$$

so that $p_{\xi,\theta} = \Phi(\sqrt{\theta^T C\theta}/\sigma)$. Since matrix C is positive definite and $\theta \neq 0$, $t_{\xi,\theta} > 0$ and $p_{\xi,\theta} > \frac{1}{2}$.

Case 2: $\zeta = \zeta_{BLUE} = C^{-1}\xi$. From (4) and (9) we get $p_{\zeta_{BLUE},\theta} = \Phi\left(\theta^T\theta/\sqrt{\sigma^2\theta^T C^{-1}\theta}\right)$.

Case 3: Any unbiased estimator $\widehat{\theta}_\zeta = \int y(x)\zeta(dx)$ of θ. Unbiasedness condition for $\widehat{\theta}_\zeta$ is $\int \zeta(dx)f^T(x) = I_m$, where I_m is the identity matrix of size $m \times m$. The covariance matrix of $\widehat{\theta}_\zeta$ is $\operatorname{Cov}(\widehat{\theta}_\zeta) = \sigma^2\int\int K(x,x')\zeta(dx)\zeta^T(dx')$. From (9) we get $p_{\zeta,\theta} = \Phi(t_{\zeta,\theta}/\sigma)$ with $t_{\zeta,\theta} = \theta^T\theta/\sqrt{\theta^T\int\int K(x,x')\zeta(dx)\zeta^T(dx')\theta}\,.$

Two particular instances are the BLUE (Case 2) and continuous OLSE (ordinary least squares estimator of θ), which is $\widehat{\theta}_{OLSE} = \int y(x)\zeta_{OLSE}(dx)$, where $\zeta_{OLSE}(dx) = M^{-1}f(x)dx$ with $M = \int f(x)f^T(x)dx$. For continuous OLSE, we obtain

$$p_{\zeta_{OLSE},\theta} = \Phi(t_{\zeta_{OLSE},\theta}/\sigma),\;\; t_{\zeta_{OLSE},\theta} = \frac{\theta^T\theta}{\sqrt{\theta^T M^{-1}\left[\int\int K(x,x')f(x)f^T(x')dx\,dx'\right]M^{-1}\theta}}\,.$$

Case 4: Finite number of observations. Assume $\mathcal{X} = \{x_1,\dots,x_N\}$, where all points $x_j \in \mathbb{R}^d$ are different. The model (1) can be written as $y_j = \theta^T f(x_j) + \varepsilon(x_j),\;\; j = 1,\dots,N$. Set $Y = (y_1,\dots,y_N)^T$, $X = (f_i(x_j))_{j,i=1}^{N,m}$ and $W = (K(x_i,x_j))_{i,j=1}^{N,N}$. Since K is strictly positive definite, the matrix W is non-degenerate. Assume also that the matrix $C = X^T W^{-1}X$ is non-degenerate. Then the BLUE of θ is $\widehat{\theta}_{\zeta_{BLUE}} = C^{-1}X^T W^{-1}Y$ and the optimal directional statistic $\widehat{\theta}_\xi$ becomes $\widehat{\theta}_\xi = X^T W^{-1}Y$. The formulae of Case 1 and Case 2 stay exactly as they are with $C = X^T W^{-1}X$. Specification of formulae of Case 3 to the N-point observation scheme is straightforward. In particular, for the OLSE we have $t_{\zeta_{OLSE},\theta} = \theta^T\theta/\sqrt{\theta^T M^{-1}X^T WXM^{-1}\theta}\,$, where $M = X^T X$.

2.3 Stochastic domination over the BLUE

The BLUE $\widehat{\theta}_{BLUE}$ defined in (5) is the best linear unbiased estimator of θ. If $\widehat{\theta}$ is any other unbiased estimator of θ, then $\mathbb{E}\theta^T\widehat{\theta} = \mathbb{E}\theta^T\widehat{\theta}_{BLUE} = \theta^T\theta$ but $\mathrm{var}(\theta^T\widehat{\theta}) \geq \mathrm{var}(\theta^T\widehat{\theta}_{BLUE})$ which shows that $\widehat{\theta}_{BLUE}$ dominates all other unbiased estimators from the point of view of direction estimation.

If the estimator $\widehat{\theta}$ is not an unbiased estimator of θ then above arguments are not valid. As we are only interested in linear estimators of θ, $\theta^T\widehat{\theta}$ is always Gaussian with some mean and variance. To compare $\theta^T\widehat{\theta}$ with $\theta^T\widehat{\theta}_{BLUE}$, let us normalize $\theta^T\widehat{\theta}$ so that the variance of the normalized version of $\theta^T\widehat{\theta}$ would coincide with $\mathrm{var}(\theta^T\widehat{\theta}_{BLUE})$, which is $\mathrm{var}(\theta^T\widehat{\theta}_{BLUE}) = \theta^T C^{-1}\theta$, see Case 2 in Section 2.2. Then the quality of the direction estimator will only be expressed through the mean of the corresponding normal distribution.

Let ξ satisfy (3). According to Case 1 in Section 2.2, $\mathbb{E}\theta^T\widehat{\theta}_\xi = \mathrm{var}(\theta^T\widehat{\theta}_\xi) = \theta^T C\theta$. Hence for the normalised estimator

$$\tilde{\theta}_\xi = \sqrt{\frac{\theta^T C^{-1}\theta}{\theta^T C\theta}}\widehat{\theta}_\xi\,.$$

we have $\mathrm{var}(\theta^T\tilde{\theta}_\xi) = \mathrm{var}(\theta^T\widehat{\theta}_{BLUE}) = \sigma^2\theta^T C^{-1}\theta$ and $\mathbb{E}\theta^T\tilde{\theta}_\xi = \sqrt{\theta^T C^{-1}\theta \cdot \theta^T C\theta} \geq \theta^T\theta = \mathbb{E}\theta^T\widehat{\theta}_{BLUE}$ for any θ. This yields that the direction of $\tilde{\theta}_\xi$ stochastically dominates the direction created by the BLUE, in the sense of domination of the corresponding normal distributions.

2.4 General scalar products

Assume that the scalar product in $\mathbb{R}^m$ is defined by

$$(a,b)_S = a^T Sb\,, \quad a,b \in \mathbb{R}^m\,,$$

where S is an arbitrary positive definite $m \times m$ matrix so that we are seeking for a linear statistic maximizing the probability $\Pr\{\theta^T S\widehat{\theta}_\zeta > 0\}$ in the class of linear statistics (2). Then similar arguments to the above yield that the optimal directional statistics is $\widehat{\theta}_{\xi,S} = S^{-1}\widehat{\theta}_\xi$, where $\widehat{\theta}_\xi$ is the directional statistic defined in Introduction.

3 Proof of Theorem 1.1

For any two signed measures $\nu, \nu' \in \Xi_1$, we define $\phi(\nu,\nu') := \int\int K(x,x')\nu(dx)\nu'(dx')$ so that $\varphi(\nu) = \phi(\nu,\nu)$. Since the kernel $K(x,x')$ is ISPD, $\varphi(\nu) > 0$ for any signed measure ν which is not a zero measure. The function $\phi(\cdot,\cdot)$ defines a scalar product on the space Ξ_1 of all finite signed measures on $\mathcal{X}$ so that (Ξ_1,ϕ) is a Hilbert space; in general, if the kernel K is not necessarily ISPD, then (Ξ_1,ϕ) is a pre-Hilbert space, see [1, 8].

The Cauchy-Schwarz inequality for $\nu_1, \nu_2 \in \Xi_1$ gives

$$|B_{12}| \leq \sqrt{B_1 B_2}\,, \tag{11}$$

where $B_1 = \varphi(\nu_1)$, $B_2 = \varphi(\nu_2)$ and $B_{12} = \phi(\nu_1, \nu_2)$. If the signed measure ν_2 is proportional to ν_1 so that $\nu_2(dx) = c\,\nu_1(dx)$ for some $c \in \mathbb{R}$, then the inequality (11) becomes an equality. Otherwise, if the measures ν_1 and ν_2 are not proportional, the inequality (11) is strict.

(a) From Lemma 2.1, for any $\theta \neq 0$ and any two signed vector-measures ζ and ζ' in Ξ_m, $p_{\zeta',\theta} \geq p_{\zeta,\theta}$ if and only if $t_{\zeta',\theta} \geq t_{\zeta,\theta}$. Assume $\zeta' = \xi$, where ξ satisfies (3) and $\zeta \in \Xi_m$ arbitrary. Set $\nu_1(dx) = \theta^T\xi(dx)$, $\nu_2(dx) = \theta^T\zeta(dx)$. Using (4) and (10) we have

$$B_1 = \varphi(\theta^T\xi) = \theta^T \int\int K(x,x')\xi(dx)\xi^T(dx')\theta = \theta^T C\theta = t^2_{\xi,\theta}\,.$$

Since matrix C is non-degenerate, $t_{\xi,\theta} = \sqrt{B_1} > 0$.

If $\nu_2 = 0$ then $t_{\zeta,\theta} = 0$ and the statement of the theorem follows. Assume that $\nu_2 \neq 0$. Using (3) and (8) we get

$$B_{12} = \phi(\theta^T\xi, \theta^T\zeta) = \theta^T \int\int K(x,x')\xi(dx)\zeta^T(dx')\theta = \theta^T \int f(x)\zeta^T(dx)\theta\,,$$

and

$$B_2 = \varphi(\theta^T\zeta) = \theta^T \int\int K(x,x')\zeta(dx)\zeta^T(dx')\theta > 0\,.$$

From (9), $t_{\zeta,\theta} = B_{12}/\sqrt{B_2}$.

From inequality (11) it follows that

$$t_{\xi,\theta} = \sqrt{B_1} \geq \frac{B_{12}}{\sqrt{B_2}} = t_{\zeta,\theta}\,. \tag{12}$$

This inequality holds for any $\theta \neq 0$ and any signed measure $\zeta \in \Xi_m$.

(b) If the signed measures $\theta^T\xi(\cdot)$ and $\theta^T\zeta(\cdot)$ are not proportional (so that there is no $c \in \mathbb{R}$ such that $\theta^T\zeta(\cdot) = c\,\theta^T\xi(\cdot)$) then there is strict inequality in (11) and therefore the inequality in (12) is strict implying $p_{\xi,\theta} > p_{\zeta,\theta}$. If $c = 0$ and hence $\nu_2 = \theta^T\zeta = 0$ then $p_{\xi,\theta} > \frac{1}{2} = p_{\zeta,\theta}$. If $\theta^T\zeta(\cdot) = c\,\theta^T\xi(\cdot)$ and $c < 0$ then $p_{\xi,\theta} > \frac{1}{2}$ but $p_{\zeta,\theta} = -p_{\xi,\theta} < \frac{1}{2}$.

□

4 Examples

4.1 Example 1

Assume that the regression model is discrete and has the form

$$y_j = \theta_0 + \theta_1 x_{1,j} + \theta_2 x_{2,j} + \varepsilon_j, \quad j = 1, \ldots, N \tag{13}$$

with $m = 3$, $N = 4$, $\mathbb{E}\varepsilon_i\varepsilon_j = 0$ for $i \neq j$, $\mathbb{E}\varepsilon_1^2 = \mathbb{E}\varepsilon_2^2 = 1$ and $\mathbb{E}\varepsilon_3^2 = \mathbb{E}\varepsilon_4^2 = \sigma^2$. Choose the observation points from the standard 2×2 full factorial design as follows:

$$(x_{1,1}, x_{2,1}) = (-1,-1), \quad (x_{1,2}, x_{2,2}) = (1,-1), \quad (x_{1,3}, x_{2,3}) = (1,1), \quad (x_{1,4}, x_{2,4}) = (-1,1)\,.$$

In this setup, $X^T X = 4\, I_3$, where I_3 is the identity 3×3-matrix,

$$C = X^T W^{-1} X = \begin{pmatrix} 2+\frac{2}{\sigma^2} & 0 & -2+\frac{2}{\sigma^2} \\ 0 & 2+\frac{2}{\sigma^2} & 0 \\ -2+\frac{2}{\sigma^2} & 0 & 2+\frac{2}{\sigma^2} \end{pmatrix}, \quad C^{-1} = \frac{1}{8} \begin{pmatrix} \sigma^2+1 & 0 & \sigma^2-1 \\ 0 & \frac{4\sigma^2}{\sigma^2+1} & 0 \\ \sigma^2-1 & 0 & \sigma^2+1 \end{pmatrix}.$$

For $\theta = (1,1,1)^T$, we obtain

$$t_{\xi,\theta} = \sqrt{2+\frac{10}{\sigma^2}}, \quad t_{\zeta_{BLUE},\theta} = 3\sqrt{\frac{2(\sigma^2+1)}{\sigma^2(\sigma^2+2)}}, \quad t_{\zeta_{OLSE},\theta} = \frac{12}{\sqrt{2+10/\sigma^2}}. \tag{14}$$

If $\sigma^2 \to \infty$ then

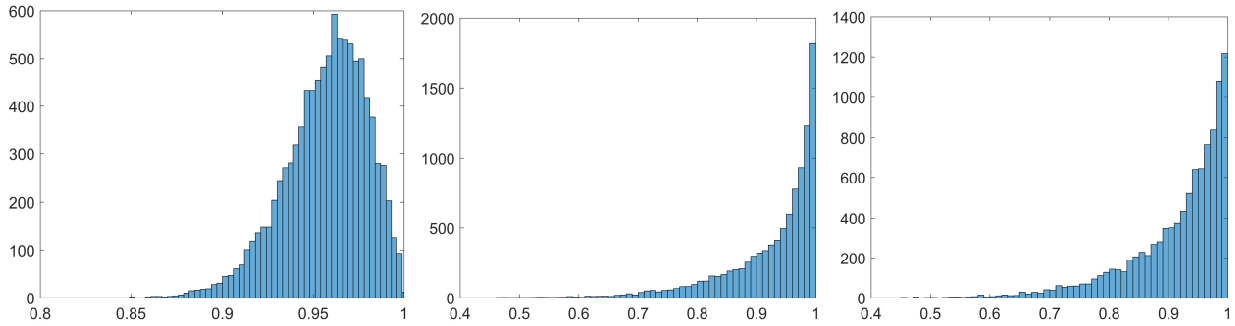
$$t_{\xi,\theta} = \sqrt{2} + \frac{5\sqrt{2}}{2\sigma^2} + O\left(\sigma^{-4}\right), \quad t_{\zeta_{BLUE},\theta} = \frac{3\sqrt{2}}{\sigma} + O\left(\sigma^{-3}\right), \quad t_{\zeta_{OLSE},\theta} = \frac{6\sqrt{2}}{\sqrt{5}\sigma} + O\left(\sigma^{-3}\right).$$

For large σ, the probability $p_{\xi,\theta}$ defined by (7) with $\zeta = \xi$ is approximately $\Phi(\sqrt{2}) \simeq 0.92135$ whereas the probability $p_{\zeta_{BLUE},\theta}$ is approximately 0.5. This example can be easily changed so that the probability $p_{\xi,\theta}$ for the directional statistic $\widehat{\theta}_\xi$ gets arbitrarily close to 1 by keeping the probability $p_{\zeta_{BLUE},\theta}$ close to 0.5.

4.2 Example 1, simulation results

We take 10000 simulations of the model (13) and compute the cosines (6) for three different statistics: (i) optimal directional statistics (ODS) $\widehat{\theta}_\xi$, (ii) the BLUE $\widehat{\theta}_{\zeta_{BLUE}}$, and (iii) OLSE, see Case 4 in Section 2.2. The value of σ^2 is taken to be 0.1 and 5.

Figure 1 contains histograms of (6) in the case of small noise, $\sigma^2 = 0.1$. Corresponding summary statistics are given in Table 1.



(a) ODS (b) BLUE (c) OLSE

Figure 1: Histograms of cosines $c_{\zeta,\theta}$ defined in (6) for different statistics, with $\sigma^2 = 0.1$.

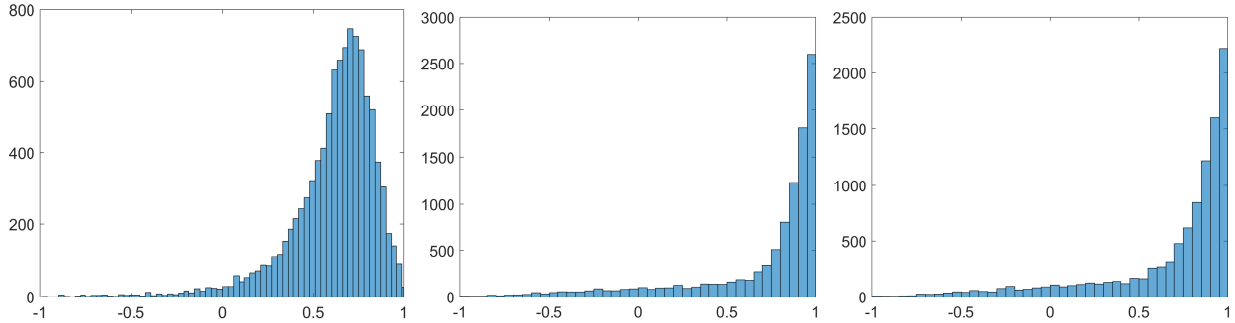
We make the following comments. The average value of the cosines $c_{\zeta,\theta}$ is largest for the ODS and its variance is much smaller than for OLSE and BLUE. Note also the different range of values of the cosines $c_{\zeta,\theta}$ obtained by each of the estimators (Fig. 1). Comparison of the

	ODS	BLUE	OLSE
Mean	0.95660	0.92779	0.91719
Variance	0.00052	0.00595	0.00646

Table 1: Mean and variance of cosines $c_{\zeta,\theta}$ taken over 10000 simulations with $\sigma^2 = 0.1$

cosines $c_{\zeta,\theta}$ between ODS and BLUE shows that there are a number of occasions where BLUE performs badly, with the angle between the estimate and true value of θ being large. Over the 10,000 simulations, BLUE gave the wrong sign of at least one component for approximately 5% of the simulations, whilst ODS maintained the correct sign for all components across all simulations.

Figure 2 contain histograms of the cosines $c_{\zeta,\theta}$ with $\sigma^2 = 5$. Corresponding summary statistics are given in Table 2. Scatterplots comparing values of the cosines $c_{\zeta,\theta}$ obtained by ODS, BLUE and OLSE are given in Figure 3. The number of points in each quadrant of the scatterplot in Figure 3(a), starting in the top left corner and going clockwise is 64, 9535, 241, 160. Analogously, 172, 9427, 77, 324 for Figure 3(b). These numbers are in agreement with (14) for $\sigma^2 = 5$.



(a) ODS (b) BLUE (c) OLSE

Figure 2: Histograms of cosines $c_{\zeta,\theta}$ for different statistics, $\sigma^2 = 5$.

	ODS	BLUE	OLSE
Mean	0.61504	0.70955	0.69019
Variance	0.05477	0.15048	0.14751

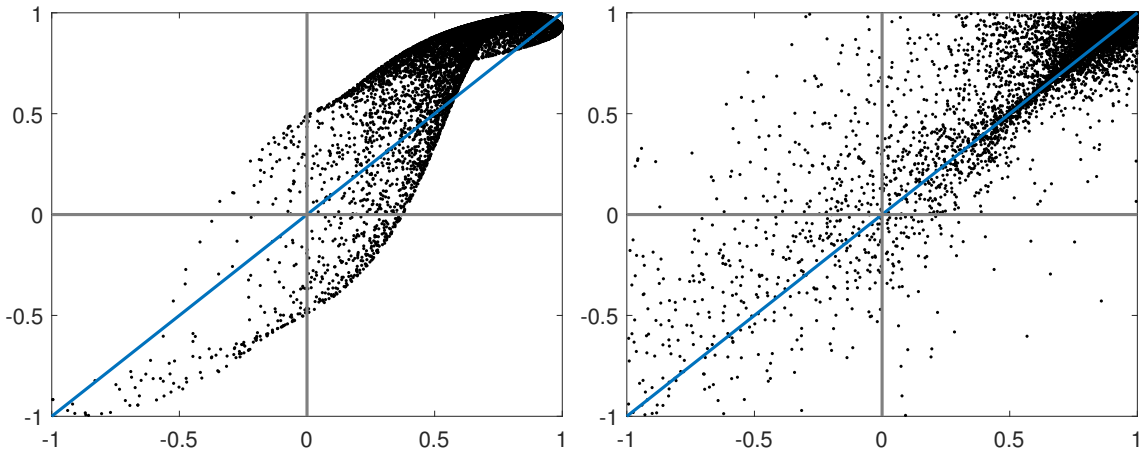
Table 2: Mean and variance of cosines $c_{\zeta,\theta}$ taken over 10000 simulations with $\sigma^2 = 5$

Notice in this example, that the mean (Table 2) is smaller for ODS than OLSE and BLUE. However, the variance of ODS is significantly smaller and indeed ODS has the smallest mean square error.

4.3 Example 2

Assume $\mathcal{X} = [a, b]$, $\varepsilon(x)$ is a Brownian motion with covariance function $K(t, s) = \min\{t, s\}$. The covariance matrix of the BLUE is C^{-1} with

$$C = \int_a^b \dot{f}(s)\dot{f}^T(s)ds + \frac{1}{a}f(a)f^T(a)$$



(a) BLUE against ODS (b) BLUE against OLSE

Figure 3: Comparison of the cosines $c_{\zeta,\theta}$ from BLUE with ODS or OLSE, $\sigma^2 = 5$.

where $\dot{f}(s) = (f_1'(s), \ldots, f_m'(s))^T$, see for example formula (2.5) in [3].

Consider a particular case $m = 3$, $f(x) = (1, x, x^2)^T$, $a = 1$ and $\theta = (1, 1, 1)^T$. In this case, we obtain $\dot{f}(s) = (0, 1, 2s)^T$,

$$C = \begin{pmatrix} 1 & 1 & 1 \\ 1 & b & b^2 \\ 1 & b^2 & \frac{1}{3}(4b^3-1) \end{pmatrix}, \; C^{-1} = \frac{1}{(b-1)^3} \begin{pmatrix} (b^2+b+1)\,b & -(4\,b^2+b+1) & 3\,b \\ -(4\,b^2+b+1) & 4\,(b^2+b+1) & -3\,(1+b) \\ 3\,b & -3\,(1+b) & 3 \end{pmatrix}.$$

For large b, the probability $p_{\xi,\theta}$ defined by (7) with $\zeta = \xi$ quickly tends to 1 whereas the probability $p_{\zeta_{BLUE},\theta}$ is approximately $\Phi(3) < 1$.

Looking at the matrix C^{-1}, which is the covariance matrix of the BLUE, we can observe that the variances of the individual estimators $\widehat{\theta}_{i,BLUE}$ $(i = 1, 2)$ of the parameters θ_2 and θ_3 tend to zero as $b \to \infty$. However, the variance of the linear form $\theta^T \widehat{\theta}_{BLUE}$ is

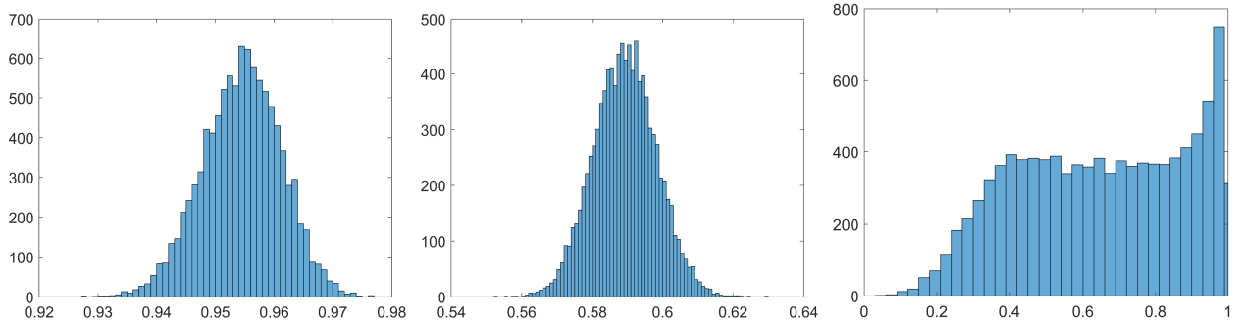
$$\mathrm{var}(\theta^T \widehat{\theta}_{BLUE}) = \theta^T C^{-1} \theta = \frac{3b^3 + 7b^2 + 13b + 13}{(b-1)^3} = 3 + \frac{16}{b} + O(b^{-2}), \quad b \to \infty,$$

and hence it does not tend to zero as $b \to \infty$.

4.4 Example 2, simulation results

We take 10000 simulations of a discrete version of the model defined in the previous subsection (we have considered a random walk on a grid with 100 points rather than the Brownian motion) and, like in Example 1, compute the cosines $c_{\zeta,\theta}$ for three different statistics: (i) ODS, (ii) BLUE and (iii) continuous OLSE. We take $b = 4$ and $N = 100$. Figure 4 contains histograms of the cosines $c_{\zeta,\theta}$.

We make the following remarks. The average value of the cosines $c_{\zeta,\theta}$ is largest for the ODS estimator and is noticeably larger than that for the BLUE and continuous OLSE. The

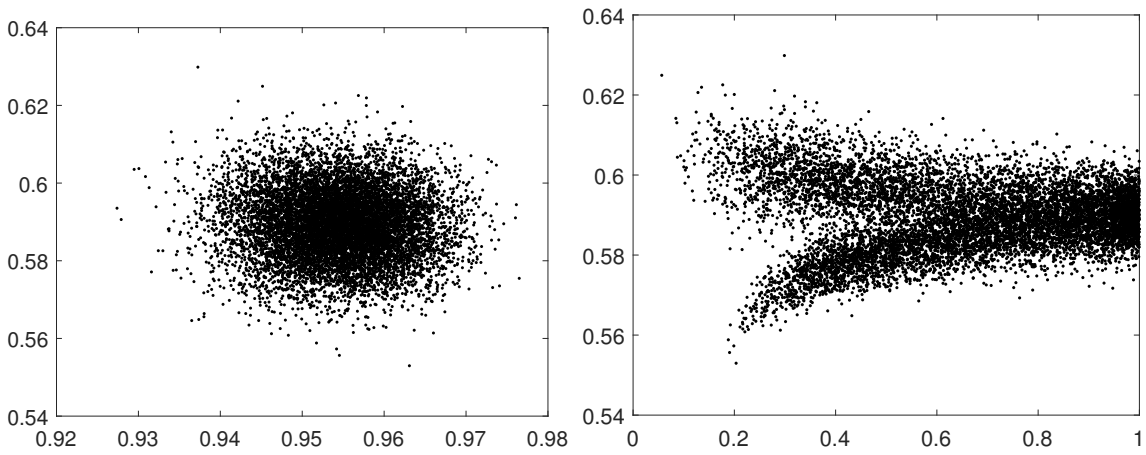


(a) ODS (b) BLUE (c) OLSE

Figure 4: Histograms of (6) for different statistics, $b = 4$ and $N = 100$.

shape of the histograms is similar for ODS and BLUE, whilst the shape of the histogram for continuous OLSE is very different. In this example, continuous OLSE performs very poorly.

Scatterplots comparing values of cosines (6) obtained by ODS, BLUE and continuous OLSE are given in Figure 5. There appears to be no relationship between the estimates obtained from ODS and BLUE.



(a) BLUE against ODS (b) BLUE against OLSE

Figure 5: Comparison of (6) from BLUE with ODS or OLSE.

5 Conclusions

We have considered a general linear regression model where we have derived an optimal directional statistic which maximizes the probability that the scalar product between the vector of unknown parameters and any linear estimator is larger than zero. We have provided arguments explaining why this statistics is better than the BLUE and have illustrate our results on two analytical and numerical examples. The results obtained are very general and could be applied to models where the number of parameters in the model exceeds the number of observations. It turns out that the form of the optimal directional statistic is rather simple can be easily computed even if the number of parameters is very large.

References

[1] A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

[2] G. E. P. Box and K. B. Wilson. On the experimental attainment of optimum conditions. In *Breakthroughs in Statistics*, pages 270–310. Springer, 1992.

[3] H. Dette, M. Konstantinou, and A. Zhigljavsky. A new approach to optimal designs for correlated observations. *The Annals of Statistics*, 45(4):1579–1608, 2017.

[4] H. Dette, A. Pepelyshev, and A. Zhigljavsky. The blue in continuous-time regression models with correlated errors. *submitted to the Annals of Statistics, arXiv preprint arXiv:1611.09804*, 2018.

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

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

[7] W. Näther. *Effective observation of random fields*. Teubner, 1985.

[8] B.K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G.R.G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(Apr):1517–1561, 2010.