PROCEEDINGS A

rspa.royalsocietypublishing.org

Research

Check for updates

Cite this article: Battey HS, Cox DR. 2018 Large numbers of explanatory variables: a probabilistic assessment. *Proc. R. Soc. A* **474**: 20170631.
http://dx.doi.org/10.1098/rspa.2017.0631

Received: 7 September 2017
Accepted: 4 June 2018

Subject Areas:
statistics, bioinformatics, mathematical modelling

Keywords:
genomics, regression analysis, sparsity

Author for correspondence:
H. S. Battey
e-mail: h.battey@imperial.ac.uk

Large numbers of explanatory variables: a probabilistic assessment

H. S. Battey[1] and D. R. Cox[2]

[1]Department of Mathematics, Imperial College London, London SW7 2AZ, UK
[2]Nuffield College, University of Oxford, Oxford OX1 1NF, UK

HSB, 0000-0001-9387-4628

Recently, Cox and Battey (2017 *Proc. Natl Acad. Sci. USA* **114**, 8592–8595 (doi:10.1073/pnas.1703764114)) outlined a procedure for regression analysis when there are a small number of study individuals and a large number of potential explanatory variables, but relatively few of the latter have a real effect. The present paper reports more formal statistical properties. The results are intended primarily to guide the choice of key tuning parameters.

1. Introduction

We consider studies of dependence in which there are relatively few individuals on each of which a large number of potential explanatory variables are measured. Genetic microarray experiments are a prime example. For progress, an implicit or explicit assumption of sparsity is required, namely that the great majority of the explanatory variables have no effect. We call explanatory variables with an effect on response signal variables, and those with no effect noise variables.

Powerful procedures for analysis emanate from the lasso [1] (for a discussion of the mathematical aspects, see [2]). They aim to uncover a single small set of signal variables effective for prediction. It is possible, however, that many different choices of explanatory variables are essentially equally effective and have quite different subject–matter implications. Cox & Battey [3] outline a different approach aiming to specify

THE ROYAL SOCIETY
PUBLISHING

those small sets of potential signal variables that give essentially the same fit; choice between different sets requires either more data or specific subject–matter information. Their procedure also allowed informal checks standard in much statistical work, such as those for nonlinearity, interaction or anomalous individuals.

In the present paper, we outline a probabilistic base for such an analysis. The emphasis is on how the procedure would perform under various idealized scenarios, as such providing guidance on the choice of key tuning parameters. The focus is on two key aspects at each stage and in the overall process of analysis. What is the probability that a signal variable is falsely discarded? How many of the variables ultimately suggested as potentially important are in fact noise variables? Our objective is essentially to calibrate the analytical procedure by specifying its behaviour under idealized conditions, not to set up a procedure to achieve preassigned error rates.

The ideas involved apply rather generally, for example, to likelihood-based fitting of logistic models for binary data.

2. Broad outline of procedure

Suppose that v variables are to be assessed for their explanatory power for a response variable. It is assumed that v is roughly k^d for $k \leq 15$ and d a small positive integer. In the method as outlined by Cox & Battey [3], the indices of these variables are arranged in a $k \times k \times k$ cube for $d = 3$ or $k \times k \times k \times k$ hypercube for $d = 4$, etc. Sets of variables are then selected for test k at a time by traversing the cube by rows, columns, etc. Every variable thus has associated with it d sets of $(k-1)$ companion variables from each time the traversal of the cube passes through that variable. The explanatory power of each variable is assessed alongside its d sets of companions and, on the basis of those analyses, variables are either discarded or retained according to a suitable decision rule. For instance, a variable might be retained if it is among the two most significant in at least $d/2$ of the d analyses in which it appears.

This is repeated with successively lower dimensional hypercubes until ideally fewer than 20 variables remain on which informal diagnostic checks are then performed.

The final phase of the analysis is to find small subsets of variables among the augmented set of retained variables and any square or interaction terms suggested at the informal phase. See Cox & Battey [3] for a more detailed description of the exploratory and subset selection phases.

3. Specification

At the ith stage of the process, let there be respectively v_{Si} signal variables and v_{Ni} noise variables. In the specific example below, $i = 0, 1, 2$. Thus of the v_{S0} signal variables initially present, v_{S2} are chosen at the end for detailed study. Given v_{S0}, the numbers of signal variables chosen at the first and second stage have binomial distributions with parameters θ_{S1} and $\theta_{S2} = \theta_{S1}\theta_{S2.1}$, where, in particular, $\theta_{S2.1}$ is the conditional probability, given that a specific signal variable is chosen at the first stage, that it is chosen again at the second stage. A different specification is needed for the noise variable because, initially at least, there are a large number of them. In particular, the v_{Ni} have Poisson distributions with means μ_{Ni}. We write $\phi_{N2.1} = \mu_{N2}/\mu_{N1}$ for the probability that a noise variable retained after step 1 is still retained at step 2.

A summary of the properties of a two-stage procedure is thus provided by θ_{S2} and μ_{N2}, the probability that a signal variable is retained and the expected number of noise variables not rejected.

We now study these in order to compare different strategies of analysis.

4. Some reduction strategies

We compare three possible approaches to each analysis of the first stage reduction. Thus, we may

— take the single variable with highest score (most significant);
— take the two variables with highest scores; and
— take all those variables, if any, whose scores exceed a threshold.

Significance tests based on the Wald, score or likelihood ratio statistics are natural choices. The former does, however, have the disadvantage of not being parameterization invariant.

In a set of k variables chosen at random for test, the number of signal variables has a Poisson distribution of mean kv_{S0}/v, where $v = v_{S0} + v_{N0}$ and v_{S0} is assumed modest, and hence has two or more signal variables with probability approximately $k^2 v_{S0}^2/(2v^2)$. This is small for the situations to be considered and does not affect the qualitative comparison of procedures. It is, therefore, ignored in later calculations.

It is convenient to phrase the initial discussion in terms of significance tests and their associated p-values. For noise variables, these are uniformly distributed on $(0, 1)$, and for signal variables their density can be modelled as $(1-\gamma)x^{-\gamma}$, where $0 < \gamma < 1$.

In the following calculations of the probability $\vartheta^{(j)}$ that the jth procedure listed above chooses the signal variable, it is convenient to use Stirling's formula in the form that for large k and fixed a the ratio $\Gamma(k+a)/\Gamma(k)$ is close to k^a.

If only the most significant out of k is chosen, the probability that the signal variable is taken is

$$\vartheta^{(1)} = \int_0^1 \mathrm{d}x(1-\gamma)x^{-\gamma}(1-x)^{k-1} = \frac{\Gamma(2-\gamma)}{k^{1-\gamma}}, \tag{4.1}$$

after simplifying by Stirling's formula. Note that the Gamma function is close to 1 over the range of interest. If $\gamma = 0$, there is no distinction between signal and noise variables and the notional signal variable is selected with probability $1/k$, that is, essentially at random.

If now we take the two most significant values out of k, then $\vartheta^{(2)} = \vartheta^{(1)} + \vartheta^{(2.1)}$, where

$$\vartheta^{(2.1)} = \int_0^1 \mathrm{d}x(k-1)x(1-\gamma)x^{-\gamma}(1-x)^{(k-2)}.$$

This is approximately

$$\frac{(1-\gamma)\Gamma(2-\gamma)}{k^{1-\gamma}}. \tag{4.2}$$

That is, approximately, $\vartheta^{(2.1)}/\vartheta^{(1)} = 1-\gamma$. The difference between $\vartheta^{(2)}$ and $\vartheta^{(1)}$ is maximized at approximately $\gamma_0 = (\log k - 1)/\log k$, at which $\vartheta^{(2.1)}$ is $\Gamma\{(\log k)^{-1}\}/(e \log k)$ (figure 1).

If we were to set a critical level at α, then the corresponding probability for a signal is

$$\vartheta^{(3)} = \int_0^\alpha (1-\gamma)x^{-\gamma}\,\mathrm{d}x = \alpha^{1-\gamma},$$

so that the equalities $\vartheta^{(3)} = \vartheta^{(1)}$ and $\vartheta^{(3)} = \vartheta^{(2)}$ are achieved for any γ by setting, respectively, $\alpha = k^{-1}\Gamma(2-\gamma)^{1/(1-\gamma)} \approx k^{-1}$ and

$$\alpha = k^{-1}\{\Gamma(2-\gamma)(2-\gamma)\}^{1/(1-\gamma)} \approx k^{-1}(2-\gamma)^{1/(1-\gamma)}.$$

This shows that strategy 3 with $\alpha = 1/k$ and strategy 1 are equivalent as this choice also makes $\mu^{(3)} = \mu^{(1)}$. By contrast, $(2-\gamma)^{1/(1-\gamma)} > 2$ for all $0 < \gamma < 1$, meaning that, for the same probability of selecting a signal variable, strategy 3 necessarily selects more noise variables than strategy 2 and, therefore, is strictly inferior.

In replacing the second strategy by the first, μ_{N1} and consequently μ_{N2} are reduced by roughly a factor of four independently of γ. However, the penalties of including noise variables are mainly incidental in view of the exhaustive search over models that must later take place. The associated computational burdens are reasonable unless μ_{N2} is, for example, 20 or more, and so strategy 2 would normally be preferred for its higher θ_{S1}.

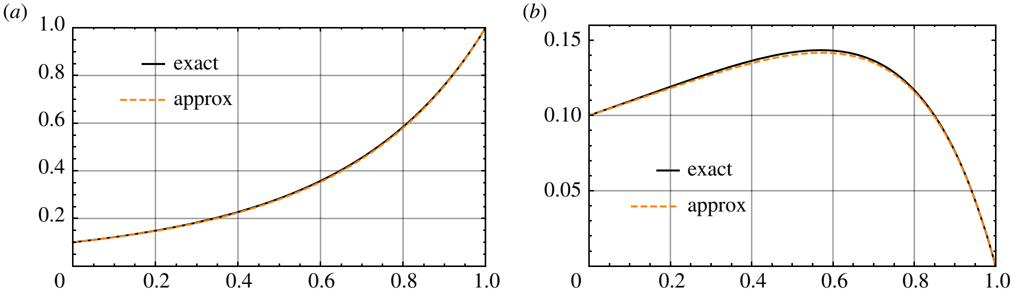

Figure 1. Exact and approximate values of $\vartheta^{(1)}$ (*a*) and $\vartheta^{(2.1)}$ (*b*) over $\gamma \in (0, 1)$ for $k = 10$. (Online version in colour.)

The probability $\theta_{S1}^{(j)}$ that the *j*th procedure retains a signal variable through the first stage is $\theta_{S1}^{(j)} = \vartheta^{(j)3} + 3\vartheta^{(j)2}(1 - \vartheta^{(j)})$ and so, approximately,

$$\theta_{S1}^{(1)} - \theta_{S1}^{(2)} = (1-\gamma)k^{2\gamma-3}[3(\gamma-3)k + \{7 + (\gamma-5)\gamma\}k^{\gamma}] \leq 0, \tag{4.3}$$

with equality at $\gamma = 1$, is the reduction in survival probability of signal variables from replacing the second procedure by the first. Here we have approximated by treating analyses in which the same variable appears as independent. The difference (4.3) is approximately

$$\frac{1 + 3\log k\{1 - e(1 + 2\log k) + \log k\}}{(e\log k)^3}, \tag{4.4}$$

at the point of maximum distinction $\gamma = \gamma_0$, and as the slowest decaying term in (4.4) is of the order of $1/(e^2 \log k)$ this shows the potential advantages of strategy 2 over strategy 1 to decrease only very slowly with *k*.

A less general but more conventional formulation is to assume a normal theory linear model. Define Δ to be signal strength multiplied by the square root of the sample size. The values of $\vartheta^{(j)}$ are then

$$\begin{aligned}
\vartheta^{(1)} &= \int \phi(x-\Delta)\Phi^{k-1}(x)\,\mathrm{d}x, \\
\vartheta^{(2)} &= \vartheta^{(1)} + \int (k-1)\phi(x-\Delta)\Phi^{k-2}(x)\Phi(-x)\,\mathrm{d}x, \\
\vartheta^{(3)} &= \Phi(\Delta - \kappa^*),
\end{aligned}$$

where $\phi(x)$ and $\Phi(x)$ are the standard normal probability density and cumulative distribution function at *x*, respectively. The threshold in the third strategy is, by convention, calibrated as the upper quantile κ^* of the standard normal distribution. We thereby approximate $\vartheta^{(1)}$ and $\vartheta^{(2.1)} \triangleq (k-1)\int \phi(x-\Delta)\Phi^{k-2}(x)\Phi(-x)\,\mathrm{d}x$ and show that the two formulations are qualitatively equivalent.

The integral in $\vartheta^{(2.1)}$ is negligible over $\mathbb{R}^-$ for $\Delta > 0$ and over $\mathbb{R}^+$ is of the form of a generalized Laplace integral for Δ fixed, that is,

$$\int_0^{\infty} g(x,\nu)\exp\{\nu h(x,\nu)\}\,\mathrm{d}x, \tag{4.5}$$

where $g(x, \nu)$ is uniformly bounded in *x* as $\nu \to \infty$ and *h* has a single maximum, $x_0(\nu)$, which varies with ν. Specifically we take $g(x) = \phi(x - \Delta)$, $\nu = k - 2$ and $h(x, \nu) = \log\Phi(x) + \log\Phi(-x)/\nu$. The integral $\vartheta^{(2.1)}$ is thus approximable by the method of Laplace (e.g. [4, pp. 60–65]). The idea in outline is that, for large ν, by far the greatest contribution to the integral comes from

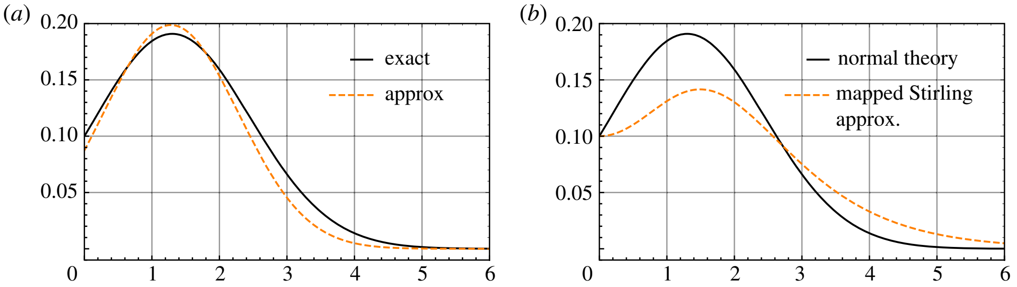

Figure 2. (*a*) Exact and approximate values of $\vartheta^{(2.1)}$ as functions of Δ for $k = 10$; (*b*) $\vartheta^{(2.1)}$ as a function of Δ for $k = 10$ in the normal theory formulation and in the *p*-value formulation with $\gamma = \{\cosh(\Delta) - 1\}/\cosh(\Delta)$. (Online version in colour.)

a neighbourhood of x_0. The integral is evaluated by expanding g and h in a neighbourhood of $x = x_0$, leading to

$$\vartheta^{(2.1)} = (k-1)g(x_0, \Delta)\exp\{h(x_0, \nu)\nu\}\left\{-\frac{2\pi}{h''(x_0,\nu)\nu}\right\}^{1/2} + O\left(\frac{1}{\nu^2}\right), \tag{4.6}$$

where the symbol $'$ denotes the partial derivative with respect to the first argument. It is customary when constructing Laplace integrals to define h such that its maximum is independent of ν. This is not possible here as, without the contribution of the $\Phi(-x)$ term, h has no sharp maximum. That the form in (4.5) is permissible is discussed by Copson [5, pp. 42–47].

The second derivative at x is

$$h''(x,\nu) = x\phi(x)\left\{\frac{\Phi(x) - \Phi(-x)\nu}{\Phi(-x)\Phi(x)\nu}\right\} - \phi^2(x)\left\{\frac{\Phi^2(-x)\nu + \Phi^2(x)}{\{\Phi(-x)\Phi(x)\}^2\nu}\right\}$$

and the unique maximizer x_0 of h satisfies

$$0 = \phi(x_0)\left\{\frac{1}{\Phi(x_0)} - \frac{1}{\Phi(-x_0)\nu}\right\}. \tag{4.7}$$

Let $\tilde{x}_0 \triangleq -\Phi^{-1}(1/k)$. As $\Phi(\tilde{x}_0) = (k-1)/k = 1 + O(k^{-1})$, we have $x_0 \simeq \tilde{x}_0$, leading to

$$h''(x_0,\nu) \simeq -\phi^2\left\{\Phi\left(\frac{1}{k}\right)\right\}k, \quad (k \to \infty), \tag{4.8}$$

and an asymptotic approximation to $\vartheta^{(2.1)}$ is

$$\frac{\sqrt{(2\pi)}(k-1)^k}{k^{k+1}}\frac{\phi(x_0 - \Delta)}{\phi\{\Phi^{-1}(1/k)\}} \tag{4.9}$$

$$= 2\pi\frac{(k-1)^k}{k^{k+1}}\phi(\Delta)\exp\left\{-\Delta\Phi^{-1}\left(\frac{1}{k}\right)\right\}, \quad (k \to \infty). \tag{4.10}$$

Equation (4.9) has a unique maximum at $\Delta_0(k) = -\Phi^{-1}(1/k)$ of

$$\exp\left[\frac{1}{2}\left\{\Phi^{-1}\left(\frac{1}{k}\right)\right\}^2\right]\frac{(k-1)^k}{k^{k+1}}\sqrt{(2\pi)}. \tag{4.11}$$

Figure 2*a* shows the accuracy of the approximation (4.9) for $k = 10$.

The qualitative implications of the normal theory formulation are equivalent to those of the formulation in terms of *p*-values. For instance, at $\Delta = 0$, which corresponds to $\gamma = 0$, equation (4.9) is $\sqrt{(2\pi)}(k-1)^k k^{-(k+1)}$. This is $1/k$ to two decimal places for k sufficiently large. If γ is redefined as $\gamma = \{\cosh(\Delta) - 1\}/\cosh(\Delta)$, then equation (4.2) has a unique maximum at $\text{sech}^{-1}(1/\log k)$, which, like $\Delta_0(k)$, is a very slowly growing function of k. Figure 2*b* shows the approximation in equation (4.2) as a function of Δ with γ redefined as described. Any mapping $\Delta \mapsto g(\Delta)$ such that $g : \mathbb{R}^+ \to$

$(0,1)$ is monotonically increasing with $g(0)=0$ and $g(x)\approx 1$ for $x>4$ gives essentially the same qualitative conclusions.

5. Relative severity of each stage

We now consider the relative advantages of

- a mild first stage and severe second stage
- a severe first stage and a mild second stage,

where, for clarity of exposition, ‘mild’ and ‘severe’ are associated with the same quantitative values in both versions. In particular, let $1>\vartheta_{\mathrm{H}}>\vartheta_{\mathrm{L}}>0$, where ϑ_{H} and ϑ_{L} are the probabilities that a signal variable is retained in a single analysis of a mild and severe reduction, respectively. Let $\theta_{S2}^{\mathrm{HL}}$ and $\theta_{S2}^{\mathrm{LH}}$ denote the overall survival probabilities of the first and the second of these procedures, respectively. To demonstrate that

$$\begin{aligned}\theta_{S2}^{\mathrm{HL}} &= \{\vartheta_{\mathrm{H}}^3+3\vartheta_{\mathrm{H}}^2(1-\vartheta_{\mathrm{H}})\}\{\vartheta_{\mathrm{L}}^2+2\vartheta_{\mathrm{L}}(1-\vartheta_{\mathrm{L}})\}\\ &> \{\vartheta_{\mathrm{L}}^3+3\vartheta_{\mathrm{L}}^2(1-\vartheta_{\mathrm{L}})\}\{\vartheta_{\mathrm{H}}^2+2\vartheta_{\mathrm{H}}(1-\vartheta_{\mathrm{H}})\}=\theta_{S2}^{\mathrm{LH}},\end{aligned}$$

or equivalently

$$\frac{\vartheta_{\mathrm{H}}(2-\vartheta_{\mathrm{H}})}{\vartheta_{\mathrm{L}}(2-\vartheta_{\mathrm{L}})}<\frac{\vartheta_{\mathrm{H}}\{\vartheta_{\mathrm{H}}(3-2\vartheta_{\mathrm{H}}\}}{\vartheta_{\mathrm{L}}\{\vartheta_{\mathrm{L}}(3-2\vartheta_{\mathrm{L}}\}},$$

suppose for a contradiction that

$$\frac{\vartheta_{\mathrm{H}}}{\vartheta_{\mathrm{L}}}\leq\frac{(2-\vartheta_{\mathrm{H}})}{(2-\vartheta_{\mathrm{L}})}\frac{(3-2\vartheta_{\mathrm{L}})}{(3-2\vartheta_{\mathrm{H}})}=1+\frac{\vartheta_{\mathrm{H}}-\vartheta_{\mathrm{L}}}{6-\{3\vartheta_{\mathrm{L}}+2\vartheta_{\mathrm{H}}(2-\vartheta_{\mathrm{L}})\}}. \tag{5.1}$$

The supremum of $3\vartheta_{\mathrm{L}}+2\vartheta_{\mathrm{H}}(2-\vartheta_{\mathrm{L}})$ is 5, attained only in the limit as $\vartheta_{\mathrm{L}}\to\vartheta_{\mathrm{H}}\to 1$. But $5<6-\vartheta_{\mathrm{L}}$ because $\vartheta_{\mathrm{L}}<1$ and so equation (5.1) implies

$$\frac{\vartheta_{\mathrm{H}}}{\vartheta_{\mathrm{L}}}<1+\frac{\vartheta_{\mathrm{H}}-\vartheta_{\mathrm{L}}}{\vartheta_{\mathrm{L}}}=\frac{\vartheta_{\mathrm{H}}}{\vartheta_{\mathrm{L}}},$$

a contradiction, and we conclude that $\theta_{S2}^{\mathrm{HL}}>\theta_{S2}^{\mathrm{LH}}$. That is, the first procedure gives a higher probability of a signal variable being retained than the second. The same argument shows that the first procedure results also in a higher probability of noise variables being retained.

6. A simple example

The performance of various stages of the procedure may be investigated as follows. Generate $n=10^2$ replicates of $v=10^3$ variables from a normal distribution of zero mean and covariance matrix $P\Sigma P^{-1}$, where P is a permutation matrix and Σ is an identity matrix with one diagonal block replaced by a correlation matrix of dimension $v_{S0}+v_{C0}$ and equal correlation ρ. For each replicate, v_{S0} of the $v_{S0}+v_{C0}$ correlated variables is multiplied by a constant signal and added, together with standard normal noise. Conditional on a realization x_S of the v_{S0} signal variables, the resulting response variable is then normally distributed of mean $\gamma^T x_S$ and unit variance, where γ is a v_{S0} vector of constants. Arrange the indices of the v variables in a 10 × 10 × 10 cube. The rows, the columns, etc. of the cube form $3k^2=300$ sets, each of $k=10$ variables. Reduction is performed by taking the top two scoring variables in each analysis in the first stage and by taking all those exceeding the threshold of a 0.1% level test in the second stage. Finally, find small subsets of variables that give adequate fit using a likelihood ratio test against the comprehensive model.

Summaries estimated from 500 Monte Carlo replications are reported in table 1. The general conclusion is that such correlation slightly degrades the probability of a signal variable being retained and increases the number of false models not rejected. The comprehensive model in the likelihood ratio test is taken as the model with all variables retained through the reduction phase. Having been selected in the light of the data, it achieves a better fit to the data than an

Table 1. $\mathcal{S}$ is the true set of signal variables, $\hat{\mathcal{S}}$ is the set of variables surviving the reduction phase, $\mathcal{M}$ is the set of low-dimensional models whose likelihood ratio test against the comprehensive model is not rejected at the 1% level. Empirical standard errors in parenthesis.

				$\mathrm{pr}(\mathcal{S} \subseteq \hat{\mathcal{S}})$			$\mathrm{pr}(\mathcal{S} \in \mathcal{M})$		$\mathbb{E}\lvert\mathcal{M}\backslash\mathcal{S}\rvert$	
v_{S0}	v_{C0}	ρ	$\frac{\text{signal}}{\text{noise}}$	lasso	full sample	split sample	full sample	split sample	full sample	split sample
1	1	0.9	1	1.00 (0.04)	1.00 (0.00)	1.00 (0.00)	0.57 (0.50)	0.99 (0.08)	6.8 (9.0)	15.7 (30.0)
1	1	0.9	0.6	0.95 (0.21)	0.96 (0.21)	0.74 (0.44)	0.45 (0.50)	0.74 (0.44)	5.4 (6.5)	13.1 (25.6)
1	1	0.5	1	1.00 (0.00)	1.00 (0.00)	1.00 (0.04)	0.55 (0.50)	0.98 (0.13)	4.4 (5.9)	10.6 (49.8)
1	1	0.5	0.6	0.99 (0.12)	0.96 (0.19)	0.76 (0.43)	0.41 (0.49)	0.76 (0.43)	2.6 (3.7)	11.0 (27.8)
1	3	0.9	1	1.00 (0.04)	1.00 (0.06)	0.98 (0.13)	0.67 (0.47)	0.97 (0.16)	27.3 (27.3)	68.4 (108)
1	3	0.9	0.6	0.90 (0.30)	0.93 (0.26)	0.73 (0.45)	0.48 (0.50)	0.72 (0.45)	23.5 (21.1)	39.8 (90.2)
1	3	0.5	1	1.00 (0.00)	1.00 (0.00)	0.99 (0.08)	0.62 (0.49)	0.99 (0.12)	11.3 (17.4)	12.2 (30.0)
1	3	0.5	0.6	0.99 (0.10)	0.95 (0.22)	0.76 (0.43)	0.38 (0.49)	0.75 (0.43)	4.1 (6.2)	9.47 (20.6)
5	1	0.9	1	0.99 (0.08)	1.00 (0.00)	1.00 (0.04)	0.95 (0.21)	0.98 (0.13)	7.5 (8.1)	105 (143)
5	1	0.9	0.6	0.79 (0.41)	0.99 (0.09)	0.95 (0.22)	0.88 (0.33)	0.95 (0.23)	46.2 (39.8)	182 (255)
5	1	0.5	1	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.96 (0.19)	0.99 (0.09)	0.0 (0.0)	15.4 (26.1)
5	1	0.5	0.6	1.00 (0.04)	1.00 (0.00)	0.98 (0.14)	0.90 (0.31)	0.97 (0.17)	1.4 (2.6)	80.7 (120)
5	3	0.9	1	0.99 (0.10)	1.00 (0.00)	1.00 (0.04)	0.96 (0.19)	0.99 (0.11)	17.5 (14.5)	303 (317)
5	3	0.9	0.6	0.75 (0.43)	0.98 (0.13)	0.91 (0.28)	0.91 (0.29)	0.90 (0.30)	116 (93)	430 (405)
5	3	0.5	1	1.00 (0.00)	1.00 (0.00)	1.00 (0.04)	0.98 (0.15)	0.99 (0.11)	0.0 (0.4)	41.3 (79.9)
5	3	0.5	0.6	1.00 (0.04)	1.00 (0.00)	0.96 (0.19)	0.92 (0.28)	0.95 (0.21)	3.6 (5.5)	217 (325)

arbitrary model embedding the one to be tested and so the probability of the true model being accepted, conditional on it having been retained, is lower than the nominal coverage probability of the likelihood ratio test. A simple way to recover the conditional nominal coverage is to split the sample. Table 1 reports the improvement in coverage probability and the associated increase in the number of false models not rejected from using 70 observations for reduction and the remaining 30 to construct the final set of low-dimensional models.

7. Discussion

(a) On the choice of k

The combinatorial arrangements used are essentially the partially balanced incomplete block designs introduced 80 years ago in the context of plant breeding trials [6]. Our treatment has thus taken k as determined *a priori*, preferably below 15. An alternative would be to draw random sets of size k for each variable, leaving the choice of k undetermined. The argument against too large a value of k is that the precision of the resulting estimates is degraded by the correlations induced when there are many correlated variables among those selected for test.

The following is a simplistic formulation outlining the effect of k on the estimated standard errors and ignoring other aspects.

Suppose $Z_1, \ldots, Z_k$ have zero mean, unit variance and are weakly correlated with average correlation $\bar{\rho}$. Then the ratio of the variance of $l^T Z$, an estimated effect, to that assuming independence is approximately $1 - \bar{\rho} + \bar{\rho}(\Sigma l_i)^2 / \Sigma l_i^2$.

The worst case is where, if the correlations are all of the same sign, the l_i are approximately equal and the ratio becomes $1 + (k-1)\bar{\rho}$. Thus, the calculation of standard errors assuming independence is not unduly distorted so long as this ratio does not exceed 2, that is, the average correlation does not exceed about $1/k$. If correlations of the order 0.05–0.1 are likely to be present, k of the order 10–20 is a reasonable choice, justifying the choice described here and in the previous paper.

(b) Arrangement randomization

Strong reassurance of the security of one's conclusions is given if, upon re-randomization of the arrangement of the variable indices in the cube, the outcome is relatively stable. An unstable outcome would most likely indicate that too severe a reduction has been used. The probabilistic properties set forth in §4 guarantee robustness to re-randomization under idealized conditions provided that the decision rules used are not too severe, and empirical experience with microarray data supports this for continuous responses. Some applications with binary outcome require caution. In particular, when v is very large, there may be an appreciable number of relatively low-dimensional sets of variables perfectly predicting the outcome. If a set of variables forms such a separating hyperplane, then so does any larger set. The likelihood would, in such a case, be theoretically unbounded and consequently all sets of k variables containing a primitive set of separating variables would be retained, leading to too mild a reduction.

The special features of the procedure with binary responses will not be explored in the present paper. Note, however, that agreement of the outcome with the lasso solution should not be expected with binary responses for the reasons outlined by Cox & Battey [3].

(c) Further remarks

The analysis discussed in the paper is intended to be largely exploratory. The object of the formal probabilistic discussion is to calibrate the procedure to show how it performs under idealized conditions. It is not aimed to justify an explicit probabilistically based assessment, such as a confidence coefficient, attached to a specific analysis. To develop such an assessment, a Bayesian analysis might be considered. For this, meaningful prior probabilities have to be attached to key unknown features, such as the true number of signal variables. A flat or so-called indifference prior would in this case be inappropriate in that it would put overwhelming probability on large values; a Poisson prior distribution of modest mean might be more suitable. Aspects describing the correlation structure of errors also need explicit formulation. At this stage of the work, we have not followed that route.

Data accessibility. This article has no additional data. The procedure as described in §2 and the previous paper is implemented in the R package 'HCmodelSets', written by H. H. Hoeltgebaum and available on the Comprehensive R Archive Network.

Authors' contributions. H.S.B and D.R.C designed the research, performed the research and wrote the paper.

Competing interests. We declare we have no competing interests.

Funding. The work was supported by a UK Engineering and Physical Sciences research fellowship (EP/P002757/1) to H.S.B.

References

1. Tibshirani R. 1996 Regression shrinkage and selection via the LASSO. *J. R. Stat. Soc. B* **58**, 267–288.
2. van de Geer S. 2016 *Estimation and testing under sparsity*. Cham, Switzerland: Springer.
3. Cox DR, Battey HS. 2017 Large numbers of explanatory variables, a semi-descriptive analysis. *Proc. Natl Acad. Sci. USA* **114**, 8592–8595. (doi:10.1073/pnas.1703764114)
4. de Bruijn NG. 1981 *Asymptotic methods in analysis*. New York, NY: Dover Publications. Corrected reprint of the third edition.
5. Copson ET. 1965 *Asymptotic expansions*. Cambridge, UK: Cambridge University Press.
6. Yates F. 1936 A new method of arranging variety trials involving a large number of varieties. *J. Agric. Sci.* **26**, 424–455. (doi:10.1017/S0021859600022760)