Cox reduction and confidence sets of models: a theoretical elucidation

Rebecca M. Lewis[1] **and Heather S. Battey**[2]

Abstract. For sparse high-dimensional regression problems, Cox and Battey [2, 14] emphasised the need for confidence sets of models: an enumeration of those small sets of variables that fit the data equivalently well in a suitable statistical sense. This is to be contrasted with the single model returned by penalised regression procedures, effective for prediction but potentially misleading for subject-matter understanding. The proposed construction of such sets relied on preliminary reduction of the full set of variables. While various possibilities could be considered for this, [14] proposed a succession of regression fits based on incomplete block designs. The purpose of the present paper is to provide insight on both aspects of [14]. For an unspecified reduction strategy, we begin by characterising models that are likely to be retained in the model confidence set, emphasising geometric aspects. We then evaluate possible reduction schemes based on penalised regression or marginal screening, before theoretically elucidating the reduction of [14]. We identify features of the covariate matrix that may reduce its efficacy, and indicate improvements to the original proposal. An advantage of the approach is its ability to reveal its own stability or fragility for the data at hand.

Key words and phrases: genomics; high-dimensional regression; model uncertainty; partially balanced incomplete block arrangements; sparsity.

1. INTRODUCTION

In the context of regression with a large number p of potential explanatory features on $n \ll p$ independent individuals, usual practice is to identify a single low-dimensional model. Motivated by scientific studies in which subject-matter understanding is sought rather than immediate predictive success, [14] argued that when several reasonable explanations are statistically indistinguishable, one should aim to specify a confidence set of models. The arguments for this are in our view compelling, yet there may be some flexibility in how the broad goal is operationalised. The following approach was suggested in [14].

A *reduction phase* starts with the full set of p variables and aims to identify a much smaller set $\hat{\mathcal{S}} \subset \{1, \dots, p\}$ indexing variables with apparent explanatory power. The set $\hat{\mathcal{S}}$ is called the *comprehensive model*. Following this reduction, the *model assessment phase* assesses all submodels of $\hat{\mathcal{S}}$ for their compatibility with the data, having taken necessary measures to prevent miscalibration due to

Department of Mathematics, Imperial College London, 180 Queen's Gate, London SW7.

[1]rebecca.lewis15@imperial.ac.uk.

[2]h.battey@imperial.ac.uk.

double use of the data. The *confidence set of models* $\mathcal{M}$ is an enumeration of those small sets of variables that fit the data equivalently well in a suitable statistical sense.

The reduction phase as outlined in [14] is termed *Cox reduction* throughout the present paper and proceeds as follows.

- Arrange the p variable indices at random either in a $k \times k$ square or a $k \times k \times k$ cube, where preferably $k \leq 15$. Extensions to four or more dimensions are possible. There is no loss of generality in assuming that p is a perfect cube, as otherwise some positions are left unoccupied. A visualisation is in Supplementary Material Figure S1.
- Any intrinsic biological or physiological characteristics, known to be important, are either used to partition the sample or are included in all regressions and should not be arranged in the cube.
- Traverse the cube from its three directions, the rows, columns and "corridors", specifying $3k^2$ sets of k variables. Fit a least squares regression to each set.
- From each regression, provisionally select a small number of variables. This might be the two variables with most significant effect, or all those variables, if any, whose Student t statistics exceed a

threshold. Cox and Battey [14] recommended that variables never selected or selected only once be discarded, except in the absence of strong prior counter evidence.

- The next step depends on the number of variables remaining. If this remains large (in excess of 30, say), a second phase is carried out, similar to the first, typically based on a square rather than a cube. The resulting set of variables from this two-stage reduction is $\hat{\mathcal{S}}$.

A form of this procedure is presented in pseudo-code in Section S1 of the Supplementary Material.

The arrangement of variables into $3k^2$ distinct blocks of size k is the partially balanced incomplete block design of Yates [46] used in a design setting where the number of treatments exceeds the number of experimental units per block. The context of Cox reduction is fundamentally different in that there is no blocking of units and no randomisation of treatments to units. The analogy is that the design aims to minimise loss or redundancy of information. As contrasted with, say, fitting $3k^2$ regressions on totally random sets of k variables, Cox reduction ensures that every variable is assessed an equal number times, and never alongside the same set of variables. A practical advantage of Cox reduction is the reassurance it offers, through rerandomisation of the variable indices in the cube, over the security or otherwise of the conclusions. In particular, unstable sets $\hat{\mathcal{S}}$ point to fragility of the method on the data at hand and thereby guide the choice of tuning parameters.

While inspiration came from [46], motivation for Cox reduction is from Bradford Hill's [4] discussion of the circumstances under which an effect obtained in an observational study is relatively likely to have a causal interpretation. Such conditions include that the effect is reproduced in independent studies and behaves appropriately when the potential cause is applied, removed and then reinstated [13]. See [15, p.165–6] for further discussion. These aspects point to Cox reduction as at least a broadly appropriate approach for an initial phase of analysis, to be considered alongside some more established procedures such as marginal screening or penalised regression. We will show in Section 4 that the set of variables retained by Cox reduction is better suited than some other reduction strategies for assessment of models in the second phase. A further motivation for theoretical elucidation is that the procedure appears to be in use, as indicated by the download logs for the R package HCmodelSets [31]. Some published applications of Cox reduction include [3], [5], [33], [45].

Among the most relevant procedures proposed elsewhere in the high-dimensional literature is the multi-environment knockoff filter discussed by [36], although this generally comes with a different type of theoretical guarantee. The confidence set of [14] identifies a number of models that are by definition statistically indistinguishable, whereas the knockoff procedure returns a single model with a control on the false discovery rate [1, 9].

In a low-dimensional context, confidence sets of models have been emphasised repeatedly by Cox (see e.g., [12], [16], [17, Appendix A.2.5]). They have been considered from a different perspective in [30], which begins with a collection of candidate models and aims to identify those with the smallest expected loss at a given significance level. In [30], the relative explanatory power of pairs of models in the candidate set are compared, whereas [14] uses the comprehensive model as a reference set against which to gauge the adequacy of each submodel. If no submodel achieves comparable fit, the model confidence set of [14] is empty, whereas the set of [30] contains several models with equally poor fit. Bayesian approaches to model selection and the resulting credible sets of models are in the same vein, with obvious operational differences. For example, the posterior distributions in [37] and [29] assign non-zero weight to multiple models.

The purpose of the present paper is to provide theoretical insight into the construction of the model confidence set proposed by [14], revealing situations that lead the procedure to partially fail, either by discarding a genuinely relevant variable in the reduction phase, or by discarding the true model at the model assessment phase. Several modifications to the original proposal are possible, of which some were mentioned in [2] but others are new to the present paper, having emerged from theoretical analyses contained herein.

A presentational quirk of the paper is that a theoretical analysis of the model assessment phase of the procedure, relating to the model confidence set $\mathcal{M}$, is discussed first (Section 4) conditional on a comprehensive model $\hat{\mathcal{S}}$ having already been isolated. This is to emphasise the possibility of using other reduction strategies besides Cox reduction for the construction of $\hat{\mathcal{S}}$. Such alternatives are discussed in Sections 5. An analysis of Cox reduction is presented in Section 6.

Two sources of randomness will be used in the forthcoming analysis. When considering the model assessment phase (Section 4), the randomness is induced through the generative model for the outcome. For the theoretical analysis of Cox reduction (Section 6), the outcome is treated as fixed and the only source of randomness comes from the arrangement of variables in the cube. In Section 6, it is convenient to think of the outcome $Y \in \mathbb{R}^n$ as generated according to the linear model $Y = X\theta^0 + \epsilon$ where $X \in \mathbb{R}^{n \times p}$ is a design matrix, $\theta^0 \in \mathbb{R}^p$ is a sparse vector satisfying $\|\theta^0\|_0 = s \ll p$ and ϵ consists of independent and identically distributed random entries with mean zero and variance σ^2. Such a modelling assumption,

while convenient for sharpening terminology and notation, plays a negligible role in results of Section 6. A referee indicated p. 106 of Scheffé [39], which emphasises the role of randomisation to ensure statistical inferences from a notional normal-theory linear model are fair approximations under more realistic generative models. See also the comments on binary outcomes in Section 9.4.

2. A SIMPLE EXPLICIT EXAMPLE

The following explicit example helps motivate the confidence set of models perspective. Data were generated according to the the model

$$(1) \qquad Y = X\beta^* + \epsilon, \qquad \epsilon \sim N_n(\mathbf{0}_n, \mathbb{I}_{n \times n})$$

where $\beta^* = (1, 1, 0, \ldots, 0)^T \in \mathbb{R}^p$ with $p = 25$ and $n = 100$. Since p is relatively small, there is no need for the reduction phase; confidence sets can be constructed directly. Rows of X were generated independently from a normal distribution of zero mean and covariance

$$\Sigma = \begin{pmatrix} 1 & 0.99 & 0.99 & 0.5 & 0.5 & \ldots & 0.5 \\ 0.99 & 1 & 0.99 & 0.5 & 0.5 & \ldots & 0.5 \\ 0.99 & 0.99 & 1 & 0.5 & 0.5 & \ldots & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 & 0.5 & \ldots & 0.5 \\ \vdots & & & & \ddots & & \vdots \\ 0.5 & 0.5 & 0.5 & \ldots & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & \ldots & 0.5 & 0.5 & 1 \end{pmatrix}.$$

This setting was constructed so that the covariates indexed by $\{1, 2, 3\}$ are statistically indistinguishable. A linear regression model fitted with a LASSO penalty by cross-validation selected the single model indexed by covariate $\{2\}$. A likelihood ratio test of each model $\mathcal{S}_m \subseteq \{1, 2, 3\}$ against $[p] = \{1, \ldots, p\}$ declared $\{2\}$, $\{1, 2\}$, $\{2, 3\}$ and $\{1, 2, 3\}$ as statistically indistinguishable from model $[p]$ at the 1% level. These models constitute the confidence set $\mathcal{M}$, together with any other low-dimensional sub-models of $[p]$ not rejected at the same significance level. All four models produced similar predicted values (the correlations between their fitted values and those obtained from the full model were at least 0.97). While for prediction the choice between well-fitting models is rather arbitrary, if understanding is the goal, an arbitrary choice among statistically indistinguishable models is a misleading portrayal of the information in the data, particularly if the different models have fundamentally different scientific interpretations. For more elaborate examples based on real data see Section 8 and the empirical work in [14].

3. NOTATION

Let X be an $(n \times p)$-dimensional matrix of covariate observations with rows $(x_i^T)_{i=1}^n$ corresponding to n observational units. The letter i is reserved for indexing the transposed rows in this way, with other letters specifying columns. Let $Y = (Y_1, \ldots, Y_n)^T$ be the vector of outcomes. Define $\theta^0 \in \mathbb{R}^p$ to be the maximiser of an expected log likelihood function, whose sample version is

$$\ell(\theta) = \sum_{i=1}^n \ell_i(\theta, Y_i), \qquad \theta \in \mathbb{R}^p$$

where $\ell_i(\theta, Y_i)$ is a function of $x_i^T \theta$. Let $\mathcal{S} = \{j : \theta_j^0 \neq 0\}$ be the set of signal variables and any indices not included in this set specify noise variables. We refer to $X\theta^0$ as the signal.

Upper-case and lower-case letters denote matrices and vectors respectively. With the context ensuring no ambiguity, capital letters are also used for sets. For a vector $u \in \mathbb{R}^p$, a matrix $X \in \mathbb{R}^{n \times p}$ and a set $A \subseteq \{1, \ldots, p\}$, u_A denotes the vector of entries of u indexed by the set A. The columns of X indexed by A are written X_A if $|A| > 1$ and x_A if $|A| = 1$. For any $a \in A$, A_{-a} is the set $A \backslash \{a\}$ and A^c its complement $\{1, \ldots, p\} \backslash A$. Let $P_X = X(X^T X)^{-1} X^T$, i.e. the projection matrix onto the column span of X. To avoid double subscripting we write $P_A = X_A (X_A^T X_A)^{-1} X_A^T$. The vector e_j for $j \in \mathbb{N} \backslash \{0\}$ denotes the j-th standard basis vector whose dimension will be clear from the context. Let $\|\cdot\|_0$, $\|\cdot\|_2$ and $\|\cdot\|_\infty$ be the ℓ_0, ℓ_2 and ℓ_∞ vector norms. When the argument is a matrix, $\|\cdot\|_2$ refers to the spectral norm and $\|\cdot\|_\infty$ to the infinity matrix norm given by the maximum absolute row sum of the matrix. The maximum eigenvalue of a square matrix M is written $\lambda_{\max}(M)$. The sub-Gaussian norm of a univariate random variable Z is

$$\|Z\|_{\psi_2} = \sup_{q \geq 1} q^{-1/2} (\mathbb{E}|Z|^q)^{1/q}.$$

The sample correlation coefficient between two centred vectors u and v is written

$$R(u, v) = \frac{u^T v}{\|u\|_2 \|v\|_2}.$$

The sample multiple correlation coefficient between a centred vector u and a matrix X with centred columns is

$$R(u, X) = \max_{\alpha \in \mathbb{R}^p} \frac{u^T X \alpha}{\|u\|_2 \|X\alpha\|_2} = R(u, P_X u).$$

For two matrices X_A and X_B with centred columns,

$$R(X_A, X_B) = \max_{\alpha, \beta} \frac{\alpha^T X_A^T X_B \beta}{\|X_A \alpha\|_2 \|X_B \beta\|_2} = \|P_A P_B\|_2,$$

where the final equality follows from the proof of Lemma S5 in Section S5.

4. MODEL ASSESSMENT PHASE

4.1 Framework

Given a comprehensive model $\hat{\mathcal{S}}$, the model assessment phase enumerates all sub-models of $\hat{\mathcal{S}}$ that fit the data

equivalently well. As suggested by [14], this may be obtained by identifying all submodels of $\hat{S}$ that are not rejected in a likelihood ratio test at a given significance level ϑ.

For any $\mathcal{S}_m \subseteq \hat{S}$, define $\mathcal{U} = \mathcal{S}_m^c \cap \hat{S}$ and

$$\Theta_0^{(m)} = \{\theta \in \mathbb{R}^p : \theta_{\mathcal{S}_m^c} = 0\},$$

$$\Theta_1^{(m)} = \{\theta \in \mathbb{R}^p : \theta_{\mathcal{U}} \neq 0, \theta_{\hat{S}^c} = 0\}.$$

Let $\hat{r}$ and r_m be the column ranks of $\widetilde{X}$ and $\widetilde{X}_m$, where $\widetilde{X}_m$ consists of the columns of X indexed by $\mathcal{S}_m$. The likelihood ratio test of $H_0 : \theta^0 \in \Theta_0^{(m)}$ against $H_1 : \theta^0 \in \Theta_1^{(m)}$ is

$$(2) \qquad \psi_m(Y, X) = \mathbb{I}\{w(\mathcal{S}_m) \leq \chi^2_{\hat{r}-r_m}(1-\vartheta)\},$$

where $\chi^2_{\hat{r}-r_m}(1-\vartheta)$ is the $1-\vartheta$ quantile of the χ^2 distribution with $\hat{r} - r_m$ degrees of freedom and

$$w(\mathcal{S}_m) = 2\left\{ \sup_{\theta \in \Theta_1^{(m)} \cup \Theta_0^{(m)}} \ell(\theta) - \sup_{\theta \in \Theta_0^{(m)}} \ell(\theta) \right\}.$$

A model $\mathcal{S}_m$ is included in the model confidence set $\mathcal{M}$ if $\psi_m(Y, X) = 1$. To reduce the computational burden, it is reasonable to construct the confidence set of models by only testing submodels of $\hat{S}$ of size less than $s^\#$ say, chosen independently of n. The uncheckable assumption $s^\# \geq s$ is needed, otherwise the true model is necessarily excluded. Algorithm 1 in Section S1 of the supplementary material provides pseudo-code for the model assessment phase.

The likelihood ratio test has greater power to distinguish between some models than others. This section uncovers geometric features associated with models in the confidence set. This is considered under the assumptions that $\mathcal{S} \subseteq \hat{S}$ where $\mathcal{S}$ is the set of signal variables, and $\hat{S}$ and $\mathcal{M}$ are independent. It follows that $\mathcal{S}$ is contained in the model confidence set with probability converging to the chosen level $1 - \vartheta$ as $n \to \infty$. We begin by describing the local behaviour of the test using Le Cam's formulation of asymptotic normality. We then extend the results by examining more general departures from the null hypothesis.

For the rest of this section, let γ^0 be the entries of θ^0 indexed by the comprehensive model $\hat{S}$ and let $\widetilde{X}$ be the corresponding columns of X. Rearranging covariates if necessary, assume that γ^0 corresponds to the first few entries of θ^0. The notation $\ell(\gamma^0, 0)$ is used to represent $\ell(\theta^0)$ when $\theta^0 = (\gamma^0, 0)$.

4.2 Behaviour under contiguous alternatives

Interpretation of the log-likelihood ratio test $\psi_m(Y, X)$ in general models is aided through appeal to Le Cam's formulation of local asymptotic normality [34]. Under standard parametric regularity conditions, notably that the statistical model is differentiable in quadratic mean at γ^0, [44, Theorem 7.2] gives the following local asymptotic expansion of the log likelihood, which holds under model $\hat{S}$ or any submodel thereof:

$$(3) \qquad \ell(\gamma^0 + h_n, 0) - \ell(\gamma^0, 0) = h^T Z_n - \frac{1}{2} h^T H_0 h + o_p(1).$$

In equation (3), H_0 is the Fisher information per observation, $Z_n = n^{-1/2} \nabla_\gamma \ell(\gamma^0, 0)$ converges in distribution to a normal random vector of mean zero and covariance matrix H_0, and $n^{1/2} h_n \to h$. Sufficient conditions for differentiability in quadratic mean are given by [44, Proposition 7.6]. These are satisfied, for example, by most exponential family regression models. It follows from (3) that

$$\ell(\gamma^0 + h_n, 0) - \ell(\gamma^0, 0) \xrightarrow{\gamma^0} N\left(-\frac{1}{2} h^T H_0 h, h^T H_0 h \right),$$

where $\xrightarrow{\gamma^0}$ means converges weakly under the true distribution defined by $\theta^0 = (\gamma^0, 0)$.

Evans [23] generalised this result to allow two models, whose components are both in a $O(n^{-1/2})$ neighbourhood of the true distribution. This is a restriction to so-called contiguous alternatives, which in the context of equation (2) would imply that the model indexed by $\mathcal{S}_m$, if false, only omits variables whose associated signal strengths are dominated by the estimation error of the maximum likelihood estimator. This characterisation belongs to formal theory but indicates the directions from γ^0 in which the likelihood ratio test is expected to have relatively high or low power. As in [23], introduce h_n and $\widetilde{h}_n$ such that $h_n \to 0$, $\widetilde{h}_n \to 0$ and $n^{1/2}(h_n - \widetilde{h}_n) \to k$. Provided that the true distribution belongs to a model whose density with respect to an appropriate dominating measure is doubly differentiable in quadratic mean, in the sense of [23, Definition 2.8], then

$$\ell(\gamma^0 + h_n, 0) - \ell(\gamma^0 + \widetilde{h}_n, 0)$$

$$= \frac{k^T}{\sqrt{n}} \nabla_\gamma \ell(\gamma^0 + \widetilde{h}_n, 0) - \frac{1}{2} k^T H_0 k + o_p(1)$$

$$(4) \qquad \xrightarrow{\gamma^0} N\left(-\frac{1}{2} k^T H_0 k, k^T H_0 k \right).$$

To interpret this result in general models, suppose that k is a scalar multiple of a unit eigenvector v of H_0 with associated eigenvalue λ, so that $k = av$, and $k^T H_0 k = (a\lambda)^2 v^T v = a^2 \lambda^2$. Since eigenvectors of H_0 indicate orthogonal directions in which the Fisher information varies from highest to lowest, equation (4) shows that the false hypotheses in a local neighbourhood of γ^0 that are most likely to be excluded from the confidence set $\mathcal{M}$ are those for which k coincides with directions of high curvature of the log likelihood function.

4.3 Extensions to other alternatives

The results in the previous section focus on models $\mathcal{S}_m$ where the unknown parameter γ^0 can be consistently estimated. Depending on the form of the log-likelihood, consistent estimation of γ^0 is not necessary for accurate characterisation of the log-likelihood $\ell(\gamma^0, 0)$. Consider the normal theory linear model with covariate matrix $\widetilde{X}$ and known variance. Let $\widetilde{X}_m$ consist of the columns of $\widetilde{X}$ indexed by $\mathcal{S}_m$. Then,

$$w(\mathcal{S}_m) = \sigma^{-2} \|(P_{\widetilde{X}} - P_{\widetilde{X}_m})Y\|_2^2 \sim \chi^2_{\hat{r}-r_m}(\lambda),$$

$$\lambda = \sigma^{-2}\|(I - P_m)\widetilde{X}\gamma^0\|_2^2$$

when $\widetilde{X}$ has full rank and P_m is the projection matrix onto $\widetilde{X}_m$. A model $\mathcal{S}_m$ is included in the model confidence set with probability converging to $1 - \vartheta$ whenever $\|(I - P_m)\widetilde{X}\gamma^0\|_2$ converges to zero, that is, whenever the portion of the signal $\widetilde{X}\gamma^0$ that is orthogonal to the column span of $\widetilde{X}_m$ converges to zero. This result applies irrespective of how well γ^0 is estimated by $\mathcal{S}_m$.

To extend the results in the previous section, we consider likelihoods that depend on γ only through $\widetilde{X}\gamma$ for $\gamma \in \mathbb{R}^{\hat{s}}$. The following result shows that models satisfying assumptions (a)-(d) below are included in the model confidence set with probability converging to $1 - \vartheta$ as the sample size grows.

PROPOSITION 1.   *Suppose the log-likelihood function $\ell(\gamma, 0)$ depends on γ only through $\eta = \widetilde{X}\gamma$ and write $\bar{\ell}(\eta) = \ell(\gamma, 0)$. Let $\eta^0 = \widetilde{X}\gamma^0$ and suppose there exist unique maximisers*

$$\hat{\eta}_m = argmax_{\eta \in Col\text{-}Sp(\widetilde{X}_m)}\bar{\ell}(\eta)$$

$$\hat{\eta} = argmax_{\eta \in Col\text{-}Sp(\widetilde{X})}\bar{\ell}(\eta).$$

Assume the following:

(a) *Weak omitted signal: $\|(I - P_m)\eta^0\|_2 = o(n^{-1/2})$,*
(b) *Predictive consistency under both models: $\|\hat{\eta} - \eta^0\|_3^3 = O_P(n^{-1/2})$ and $\|\hat{\eta}_m - \eta^0\|_3^3 = O_P(n^{-1/2})$.*
(c) *A local asymptotic expansion: for $M_n = o(1)$,*

$$\sup_{h \in B_n} |\bar{\ell}(\eta^0 + h) - f(\eta^0, h)| = o_P(1)$$

where $B_n = \{h \in \mathbb{R}^n : \|h\|_3 \leq M_n\}$ and

$$f(\eta^0, h) = \bar{\ell}(\eta^0) + h^T \nabla_\eta \bar{\ell}(\eta^0) + \frac{1}{2}h^T\nabla^2_{\eta\eta}\bar{\ell}(\eta^0)h.$$

Further, $\|n^{-1/2}\nabla_\eta\bar{\ell}(\eta^0)\|_2$ and $\|\nabla^2_{\eta\eta}\bar{\ell}(\eta^0)\|_2$ are $O_P(1)$.

(d) *Asymptotic normality of the score function: there exists a matrix $\widetilde{Q}$ of full rank whose columns span the column space of $\widetilde{X}$ and the first r_m columns, denoted $\widetilde{Q}_m$, span the column space of $\widetilde{X}_m$ where*

- $\max_{i=1}^n \|\widetilde{q}_i\|_2 = O(1)$ *where $\widetilde{q}_i^T$ denotes the i-th row of $\widetilde{Q}$,*
- *the eigenvalues of $\widetilde{Q}^T\widetilde{Q}/n$ are asymptotically bounded above and away from zero,*
- *the following asymptotic limits hold*

$$(\widetilde{Q}^T J^0 \widetilde{Q})^{-1/2}\widetilde{Q}^T U^0 \xrightarrow{d} N(0, \mathbb{I}),$$

$$(\widetilde{Q}_m^T J^0 \widetilde{Q}_m)^{-1/2}\widetilde{Q}_m^T U^0 \xrightarrow{d} N(0, \mathbb{I}),$$

with $U^0 = \nabla_\eta \bar{\ell}(\eta^0)$ and $J^0 = -\nabla^2_{\eta\eta}\bar{\ell}(\eta^0)$.

Then,

$$\mathbb{P}\{\psi_m(Y, X) = 1\} \to 1 - \vartheta$$

as $n \to \infty$.

The required $\|\cdot\|_3$-consistency appearing in assumptions (b) and (c) of Proposition 1 is unusual but arises naturally when considering the asymptotic expansion of the log-likelihood as a function of $\eta = \widetilde{X}\gamma$ instead of γ. Let $h = \widetilde{X}v$ with $n^{1/2}\|v\|_2 = O(1)$. Standard arguments show that under regularity conditions,

$$\begin{aligned}
\bar{\ell}(\eta^0 + h) &= \ell(\gamma^0 + v, 0) \\
&= \ell(\gamma^0, 0) + v^T\widetilde{X}^T\nabla_\eta\bar{\ell}(\eta^0) \\
&\quad + \frac{1}{2}v^T\widetilde{X}^T\nabla^2_{\eta\eta}\bar{\ell}(\eta^0)\widetilde{X}v + o_P(1) \\
&= \bar{\ell}(\eta^0) + h^T\nabla_\eta\bar{\ell}(\eta^0) \\
&\quad + \frac{1}{2}h^T\nabla^2_{\eta\eta}\bar{\ell}(\eta^0)h + o_P(1).
\end{aligned}$$

This is an asymptotic expansion of the form given in assumption (c) of Proposition 1. However, the condition $n^{1/2}\|v\|_2 = O(1)$ is strong as there exist cases where $n^{1/2}\|v\|_2$ is unbounded and the asymptotic expansion remains valid. To avoid this, we obtain a similar expansion about η^0 by identifying conditions on h for which

$$h^T\{\nabla^2_{\eta\eta}\bar{\ell}(\eta^0) - \nabla^2_{\eta\eta}\bar{\ell}(\eta)\}h = o_P(1),$$

where with slight notational inaccuracy, η lies on the line joining $\eta^0 + h$ and η^0, and may differ in each entry of the matrix. In certain generalised linear models, this term is of order $\|h\|_3^3$ which motivates assumptions (b) and (c). For further details, see the proof of Lemma S10 in the supplementary material.

Section S6 of the supplementary material gives conditions under which the assumptions in Proposition 1 are satisfied by canonical generalised linear models. These conditions are met by the linear and logistic regression models when $\mathcal{S}_m$ satisfies $\|(I - P_m)\widetilde{X}\gamma^0\|_2 = o(n^{-1/2})$ and so those models that only exclude a small portion of the signal $\widetilde{X}\gamma^0$ will be included in the model confidence set. It is not necessary for $\widetilde{X}$ to be full rank;

see the proof of Proposition 1 for details. The expected size of the model confidence set is at least $M(1 - \vartheta)$ where M denotes the number of models $\mathcal{S}_m$ satisfying $\|(I - P_m)\widetilde{X}\gamma^0\| = o(n^{-1/2})$. If there are many sets of covariates that are highly correlated in sample with signal variables, then the model confidence set will contain a large number of models on average.

5. REDUCTION PHASE: EVALUATION OF POSSIBLE STRATEGIES

The construction of a model confidence set hinges on a preliminary reduction to a set $\hat{\mathcal{S}}$ of manageable size that contains the true model with probability converging to one as $n, p \to \infty$. Possible reduction strategies include penalised regression procedures such as the LASSO [43] or marginal screening [26]. We briefly discuss some of the considerations involved, with an emphasis on the linear model.

5.1 Penalised regression

Penalised regression performs variable selection by minimising the least-squares or negative log-likelihood function subject to a constraint on the magnitudes of the entries of the parameter vector. An estimate of θ^0 is first obtained satisfying $\hat{\theta}^{(\lambda)} = \text{argmin}_{\theta \in \mathbb{R}^p}\{-\ell(\theta) + \lambda p(\theta)\}$, where $p(\theta)$ is typically of the form

$$p(\theta) = \sum_{j=1}^{p} p_j(|\theta_j|)$$

and the set

$$(5) \qquad \hat{\mathcal{S}} = \{j : \hat{\theta}_j^{(\lambda)} \neq 0\},$$

may be used to specify the comprehensive model. The tuning parameter λ determines the size of the set $\hat{\mathcal{S}}$. The LASSO [43], SCAD [25] and MCP [47] procedures arise from particular choices of $p(\theta)$.

As $p(\theta)$ is typically non-decreasing in the magnitudes $|\theta_j|$ for $j \in \{1, \ldots, p\}$, the comprehensive model obtained through penalised regression will rarely detect signal variables that are weakly correlated with the response variable. Decreasing the size of the tuning parameter introduces further covariates, however those that are correlated with the error term will be prioritised over those variables with a weak signal, resulting in an overfitted model. Cox reduction aims to avoid overfitting by performing many low-dimensional regressions and retaining only those variables that are consistently statistically significant at a chosen level.

Penalised regression may also fail to select all signal variables even when they are highly correlated with the response variable. This occurrence is explained in [22] for the linear model with LASSO solution $\hat{\theta}^{(\lambda)}$ obtained using the LARS algorithm. Briefly, given a current predicted response, the algorithm sequentially adds the covariate that is most highly correlated with the residuals to update the predicted response. The sequence of solutions obtained at each step of the algorithm corresponds to the solutions of the LASSO problem for decreasing λ [22, Theorem 1]. When two covariates are highly but not perfectly correlated in sample, including one of them in a LARS step reduces the correlation between the other and the residual, thus making it difficult for the second variable to enter the model. This can lead the LASSO to select unimportant noise variables over signal variables. The simulation results summarised at the start of Section 8 provide examples of this, showing that when signal strengths are weak and correlations among covariates are high, the comprehensive model determined by an undertuned LASSO often omits signal variables.

The situation is in principle less problematic when there are perfectly correlated signal variables. If a LASSO solution includes at least one of the perfectly correlated signal variables, then there must exist another solution that includes all of these variables. This applies to general loss functions and other penalty functions too. Unfortunately, commonly used optimisation algorithms such as coordinate descent [28, 6] only return one arbitrarily chosen solution, leaving open the possibility that some signal variables are discarded.

For these reasons, an undertuned penalised regression procedure is not recommended for construction of the comprehensive model.

5.2 Marginal screening

For sets E and F, let

$$\hat{\theta}_{Y:E} = (X_E^T X_E)^{-1} X_E^T Y$$
$$\hat{\theta}_{F:E} = (X_E^T X_E)^{-1} X_E^T X_F$$

and $\hat{\theta}_{Y:E.F}$ be the entries of $\hat{\theta}_{Y:K}$ corresponding to E when $K = E \cup F$. The notation closely follows [19]. Marginal screening [26] retains variables with the largest absolute marginal correlation with the outcome. To explore the limitations of marginal screening for construction of the comprehensive model, consider the decomposition

$$(6) \qquad \hat{\theta}_{Y:E} = \hat{\theta}_{Y:E.F} + (\hat{\theta}_{F:E})^T \hat{\theta}_{Y:F.E}.$$

where $E, F \subset \{1, \ldots, p\}$ with $|E| = 1$. The result (6) was first noted in [10] for sets F of size one and has been used subsequently in [11, 20]. A derivation for $|F| > 1$ in the linear model is given in Section S4 of the supplementary material (proof of Lemma S1). An asymptotic analogue of equation (6) for general regression models was given by [18]. It follows from (6) that $\hat{\theta}_{Y:E} = \hat{\theta}_{Y:E.F}$ if and only if either $\hat{\theta}_{F:E} = 0$ or $\hat{\theta}_{Y:F.E} = 0$ or the two

vectors are orthogonal. In particular, if this condition is violated when E indexes a signal variable and $F = \mathcal{S}\setminus E$, it is possible for the marginal effect to be inflated or diminished. This improves or curtails the ability of marginal screening to detect E. If there is total cancellation, i.e. $\hat{\theta}_{Y:E.F} \approx -(\hat{\theta}_{F:E})^T \hat{\theta}_{Y:F.E}$, marginal screening would be unable to detect the signal variable indexed by E.

The situation described above is also sometimes challenging for Cox reduction, as will become clear in Section 6. Differences between the two procedures are most apparent when there is partial cancellation, so that the signal variable indexed by E is deemed ineffective in the relatively strong reduction effectuated by marginal screening, but survives the weaker first round of Cox reduction, giving it the opportunity to be assessed in the presence of other strong variables in the second round of reduction.

A more subtle point is that the noise variables retained by Cox reduction facilitate model discrimination at the model assessment phase to a greater extent than those retained by marginal screening. The explanation is that marginal screening retains covariates that are highly correlated in sample with Y, and so the variables included in the comprehensive model generally span a relatively low-dimensional subspace of $\mathbb{R}^n$. Since Cox reduction requires that any apparent effect is not explained away by the companion variables, the angles between Y and the resulting n-dimensional vectors of observations on retained noise covariates need not be small. The implication is that $X_{\hat{\mathcal{S}}}$ typically spans a larger subspace of $\mathbb{R}^n$. In this way, Cox reduction is expected to identify a comprehensive model that fits the data better than that identified by marginal screening, making it harder for submodels to pass the likelihood ratio test. Furthermore, submodels of the comprehensive model obtained from marginal screening span a similar space to $X_{\hat{\mathcal{S}}}$ by construction, and so are unlikely to be rejected by a likelihood ratio test (see Section 4).

While the advantages of Cox reduction over marginal screening in particular circumstances are intuitively clear based on (6), there are also situations in which marginal screening outperforms Cox reduction. For this reason, Section 6 studies a setting in which marginal screening would be an obvious candidate for the construction of $\hat{\mathcal{S}}$. This is to highlight limitations of Cox reduction and point to potential improvements.

6. SOME THEORETICAL ANALYSIS OF COX REDUCTION

6.1 Framework

Cox reduction forms a comprehensive model by analysing variables alongside randomly selected sets of companion variables and retaining those whose apparent effects are not explained away by other variables. Some analysis of Cox reduction was provided in [2], where the focus was on the comparison of decision rules, legitimising certain simplifying assumptions. Based on their results, we consider the following version of Cox reduction.

- In the first round of Cox reduction, variable indices are randomly arranged in a $k \times k \times k$ cube. Variables are retained for further analysis if they are among the two most significant in at least two out of the three regressions in which they appear. See Algorithm 2 in Section S1 of the supplementary material.
- In the second round of Cox reduction, the indices corresponding to variables retained after the first round are randomly arranged in a $k \times k$ square. The comprehensive model indexes those variables that are significant at level α in at least one regression. See Algorithm 3 in Section S1 of the supplementary material.

The present section identifies potentially problematic situations that lead Cox reduction to retain noise variables over signal variables. For this analysis, we treat Y and X as fixed, with the only source of randomness coming from the arrangement of variable indices in the cube. Each of the regressions that are fitted are linear with variable significance determined by the magnitude of a Wald statistic. This is most natural when the generative model is of the form $Y = X\theta^0 + \epsilon$, where ϵ consists of centred, independent entries with known variance σ^2. However, with the exception of Proposition 2, the conclusions are free of modelling assumptions.

6.2 Assumptions

We make the following assumptions on the design matrix and response vector. This will allow the relative significance of variables to be approximated by more interpretable quantities.

CONDITION 1. *Suppose Y and all columns of X are centred, and X_K is of full rank for all $K \subset \{1, \ldots, p\}$ of cardinality k.*

CONDITION 2. *Suppose the indices $\{1, \ldots, p\}$ can be partitioned into disjoint sets $\mathcal{A}$ and $\mathcal{B} = \{1, \ldots, p\}\setminus \mathcal{A}$ such that*

$$\min_{A \subseteq \mathcal{A}, |A|=1} \{R(Y, x_A)\}^{-1} = O(1)$$

and the maximum sample correlation

$$\Delta(a) := \max_{\substack{A \subseteq \mathcal{A}, B \subset \mathcal{B}, \\ |A \cup B|=k, |A| \leq a}} \max\{R^2(X_A, X_B), R^2(Y, X_B)\}$$

satisfies $\Delta(a) = o(1)$ when a is of moderate size (to be specified where necessary). When $|A| = 0$ or $|B| = 0$, we let $R^2(X_A, X_B) = 0$ or $R(Y, X_B) = 0$ so that $\Delta(a)$ is well-defined.

Given our decision to treat Y and X as fixed, the terms $\Delta(a)$ or $R(Y, x_A)$ are, for present purposes, non-random quantities that vary as $n, p \to \infty$. Condition 1 is minor and ensures the Wald statistics are well-defined and can be related to sample correlation coefficients as defined in Section 3. To interpret and justify Condition 2, it will be useful to temporarily view the rows of the design matrix as independent and identically distributed random quantities with $\mathcal{A}$ indexing signal variables and noise variables correlated in population with signal variables, and $\mathcal{B}$ indexing the remaining noise variables. Under suitable conditions on the random process, $R(Y, x_A)$ is a sample correlation between two quantities that are related at the population level, and so will likely be bounded away from zero asymptotically. In contrast, $R(X_A, X_B)$ and $R(Y, X_B)$ are sample correlations between unrelated random variables and so these are expected to decay to zero as $n \to \infty$ provided the cardinality of the sets A and B are of moderate size. This behaviour coincides with that required in Condition 2. We focus on sets such that $|A \cup B| = k$, this being the block size in each traversal of the cube or square. The quantity a in the definition of $\Delta(a)$ reflects a restriction on the number of variables in $\mathcal{A}$ appearing in each regression, and we will soon see that during the first round of Cox reduction, it is sufficient to consider $a = 1$. Based on this interpretation, we refer to $\mathcal{A}$ as the set of *pseudo-signal variables* and $\mathcal{B}$ as the set of *pseudo-noise variables*.

Proposition 2 provides a more precise justification for Condition 2 in the case $a = 1$ when the entries of the design matrix and response vector are random observations from sub-Gaussian distributions. The result shows that $\Delta(1) = o_P(1)$ by making use of Theorem 3.1 in [27]. The behaviour of $\Delta(a)$ for $a > 1$ was not considered by [27].

PROPOSITION 2. *Assume the following:*

- *Each entry of Y is an independent observation of a centred sub-Gaussian random variable,*
- *For $a \in \mathcal{A}$, each entry of x_a is an independent observation of a centred sub-Gaussian random variable.*
- *Each row of $X_{\mathcal{B}}$ is an independent sample from the distribution of U where $U \in \mathbb{R}^{|\mathcal{B}|}$ is a random vector with independent and centred entries satisfying*

$$\sup_{\alpha \in \mathbb{R}^{|\mathcal{B}|} : \|\alpha\|_2 = 1} \|\alpha^T U\|_{\psi_2} < \infty.$$

- *$X_{\mathcal{A}}$ and Y are both independent of $X_{\mathcal{B}}$ (but possibly dependent on each other).*

Then,

$$\Delta(1) = O_P\{\log(|\mathcal{B}|)/n\} = o_P(1)$$

as $p, n \to \infty$ with $k, k^{-1} = O(1)$, $\log(|\mathcal{B}|) = o(n)$, $|\mathcal{A}| = o\{\log(|\mathcal{B}|)\}$ and $|\mathcal{A}|^{8/7} \log\{|\mathcal{B}|n\} = o(n^{1/7})$.

When $|\mathcal{A}| = O(1)$ and $|\mathcal{B}| = O(p)$, Proposition 2 shows that $\Delta(1) = o_P(n^{-6/7}) = o_P(1)$ as long as $\log p = o(n^{1/7})$. When this assumption is violated, it is possible for noise variables to be retained instead of signal variables. This is discussed further in Section 6.6 and some of the modifications outlined in Section 7 are designed to mitigate the issue.

Our formulation is deliberately constructed so that when $\mathcal{A}$ includes all signal variables, Cox reduction offers no obvious advantage over marginal screening as far as retention of signal variables is concerned. Limitations of Cox reduction may then be identified and potential improvements proposed.

6.3 Preliminary insights

For a set $K \subset \{1, \ldots, p\}$ of size k indexing variables in a given k-dimensional regression, suppose Cox reduction retains variables according to the Wald statistic

$$T_{Y:K} = \sigma^{-1} D_K^{-1/2} \hat{\theta}_{Y:K}$$

where D_K is a diagonal matrix with entries given by the diagonal entries of $(X_K^T X_K)^{-1}$ and $\hat{\theta}_{Y:K}$ was defined in Section 5.2. Lemma S2 shows that the entry of $T_{Y:K}$ corresponding to variable index $e \in K$ is

(7) $$\sigma^{-1} \|Y\|_2 R(Y, (I - P_{K_{-e}}) x_e)$$

and so the relative sizes of the Wald statistics depend on the sample correlation between the response variable Y and the projected covariates $(I - P_{K_{-e}}) x_e$. If a variable has a real effect on the response, this correlation will be large for many companion sets K_{-e}. Cox reduction is designed to exploit this. For conciseness, correlation in the present section means sample correlation unless indicated otherwise.

For a more precise analysis, let $A \subseteq \mathcal{A}$ and $B \subseteq \mathcal{B}$ index the variables appearing in a given traversal of the hypercube with $|A \cup B| = k$. Let $T_{Y:A.B}$ be the entries of $T_{Y:K}$ corresponding to A when $K = A \cup B$. The following result allows us to characterise the limiting behaviour of the statistics $T_{Y:A.B}$ and $T_{Y:B.A}$ under Conditions 1 and 2.

PROPOSITION 3. *Suppose $R(X_A, X_B) \leq 1 - c$ for some $c > 0$. Then there exists $C > 0$, depending only on c, such that*

$$T_{Y:A.B} = T_{Y:A} + \delta_{A.B}$$
$$T_{Y:B.A} = T_{(I-P_A)Y:B} + \delta_{B.A}^{(2)}$$
$$= \delta_{B.A}^{(1)} + \delta_{B.A}^{(2)}$$

where

$$\|\delta_{A.B}\|_\infty \leq C\sigma^{-1} \|Y\| \{R(X_A, X_B) + R(Y, X_B)\},$$
$$\|\delta_{B.A}^{(1)}\|_\infty \leq \sigma^{-1} \|Y\| \{R(X_A, X_B) + R(Y, X_B)\}$$
$$\|\delta_{B.A}^{(2)}\|_\infty \leq C\sigma^{-1} \|(I - P_A)Y\|_2 R(X_A, X_B)\}.$$

Proposition 3 shows that as $R(X_A, X_B) \to 0$ and $R(Y, X_B) \to 0$, the entries of the scaled statistic

$$\|Y\|^{-1} T_{Y:A.B}$$

converge to the corresponding entries of the scaled t-statistic obtained by regressing Y on X_A only. In contrast, the entries of $\|Y\|^{-1} T_{Y:B.A}$ converge to the entries of the scaled t-statistic obtained by regressing $(I - P_A)Y$ on X_B. The latter is negligible compared to the magnitudes of the entries of $\|Y\|^{-1} T_{Y:A.B}$. These observations will be of use when analysing each round of reduction in the following sections.

6.4 First reduction

In the first round of Cox reduction, variable $e \in \{1, \ldots, p\}$ is assessed alongside variables indexed by K_1, K_2 and K_3, where $K_1 \cap K_2 \cap K_3 = \{e\}$. Let $\mathcal{E}_{e:K}$ be the event that the absolute entry of $T_{Y:K}$ corresponding to $e \in K$ is among the two largest values in the set

$$\{|e_j^T T_{Y:K}| : j \in \{1, \ldots, k\}\}.$$

Then the variable indexed by e is retained through the first reduction if at least two of the three events $\mathcal{E}_{e:K_1}$, $\mathcal{E}_{e:K_2}$ and $\mathcal{E}_{e:K_3}$ occur. See Algorithm 2 in Section S1 of the Supplementary material.

When the number of pseudo-signal variables is small relative to the dimension $p = k^3$, it is likely that pseudo-signal variables are unaccompanied by other variables from $\mathcal{A}$ in at least 2/3 of the regressions in which they appear. By Proposition 3, pseudo-signal variables are retained through the first reduction based on their marginal relationship with the response. The following result establishes that as long as the marginal correlation between the response variable and a given pseudo-signal variable is sufficiently large, Cox reduction retains the pseudo-signal variable with probability close to one, where the randomness comes only from the arrangement of variable indices in the hypercube, the responses and covariates being treated as fixed at their realised values.

PROPOSITION 4. *Suppose Conditions 1 and 2 hold with $a = 1$ when $p, n \to \infty$ with $k = O(1)$. Then, for n large enough, the covariates indexed by $\mathcal{A}$ are retained after the first round of reduction with probability at least*

$$(8) \qquad 1 - \frac{|\mathcal{A}|(|\mathcal{A}| - 1)(|\mathcal{A}| - 2)(k-1)^2}{(k^3 - 1)(k^3 - 2)}.$$

The lower bound on the probability given in Proposition 4 is a lower bound on the probability that pseudo-signal variables appear unaccompanied by other pseudo-signal variables in at least 2/3 regressions in which they appear. This bound is plotted in Figure 1. When $|\mathcal{A}|$ is of moderate size relative to $p = k^3$, the probability is close to one. In light of this result, highly correlated variables may
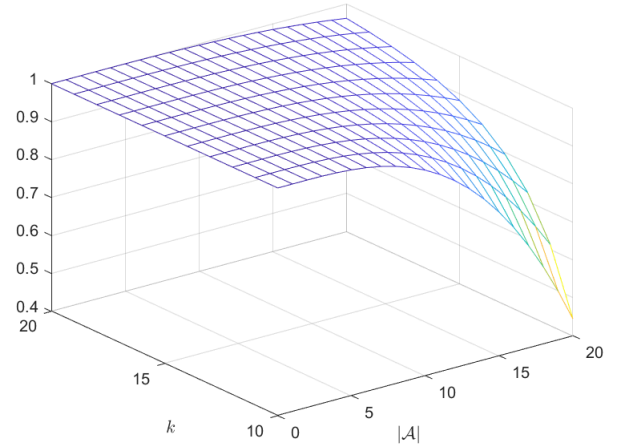

Fig 1: Plot of equation (8) for various values of k and $|\mathcal{A}|$.

be paired, with one representative of both, during the first round of reduction. This serves to reduce the cardinality of the the the set $\mathcal{A}$, thus increasing probability (8).

The result in Proposition 4 is conservative in that it focuses on the case where each pseudo-signal variable is the most significant in at least 2/3 regressions in which it appears. Cox reduction retains a variable if it is among the two most significant.

The first reduction retains all pseudo-signal variables whose marginal correlation with the response is sufficiently large and so mimics a conservative version of marginal screening that retains covariates with the largest $\hat{s}$ marginal correlations with the response, for $\hat{s}$ large. The two procedures differ in the set of retained pseudo-noise variables. The second round of Cox reduction, at which point the two procedures diverge more substantially, considers the joint explanatory power of sets of pseudo-signal variables.

6.5 Second reduction

Suppose that the index e, retained through the first reduction, is randomly arranged in a $k \times k$ square, where the dimension k may be different from that of the first reduction. Let K_1 and K_2 be the variables that share a row or column with e in the square and redefine $\mathcal{E}_{e:K}$ to be the event that the entry of $T_{Y:K}$ corresponding to $e \in K$ is significant at level α. Then, the second round of reduction retains the variable indexed by e on the event $\mathcal{E}_{e:K_1} \cup \mathcal{E}_{e:K_2}$. See Algorithm 3 in Section S1 of the Supplementary material.

In contrast to the first round, pseudo-signal variables are likely to appear together in second-round regressions. Proposition 5 derives the expected number of pseudo-signal variables that share a row or column with a given $a \in \mathcal{A}$ in terms of the dimension of the square k and the
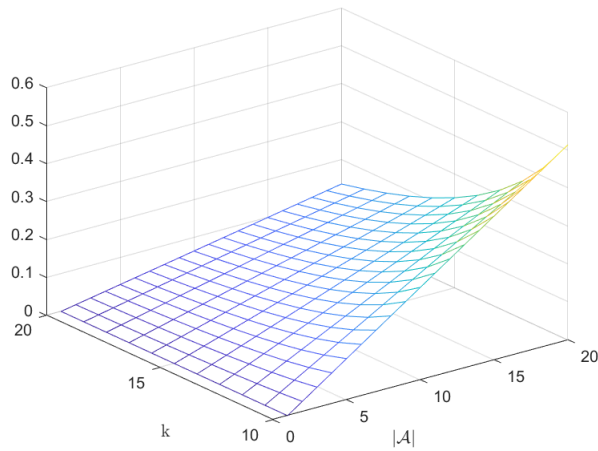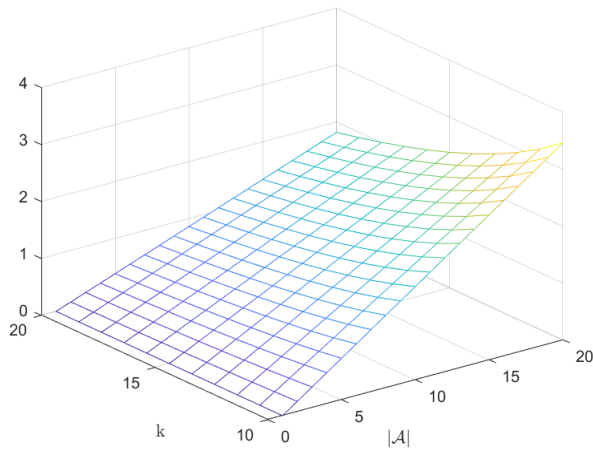
(a) Round 1: $k \times k \times k$ cube

(b) Round 2: $k \times k$ square

Fig 2: Expected number of indices from $\mathcal{A}_{-a}$ sharing a fibre with a for various values of k and $|\mathcal{A}|$.

size of the set $\mathcal{A}$. For comparison, the analogous calculation for the cube is also provided. Figure 2 shows that this expectation is substantially larger in the second round.

PROPOSITION 5. *Suppose the indices in $\mathcal{A}$ are randomised to positions in a $k \times k$ square. For a given index $a \in \mathcal{A}$, the expected number of indices from $\mathcal{A}_{-a}$ sharing a row or column with a is*

$$\frac{2(|\mathcal{A}| - 1)}{k + 1}.$$

In contrast, if the indices are randomised to positions of a $k \times k \times k$ cube, the expected number of indices from $\mathcal{A}_{-a}$ sharing a row, column or corridor with a is

$$\frac{3(|\mathcal{A}| - 1)}{k^2 + k + 1}.$$

Provided Conditions 1 and 2 hold for $a = k$, Propositions 3 and 5 imply that pseudo-signal variables are re-

tained through the second reduction based on the significance of statistics of approximate form

(9) $$\sigma^{-1} \|Y\|_2 R(Y, (I - P_{A_{-a}}) x_a)$$

where $A \subseteq \mathcal{A}$ denotes the pseudo-signal variables appearing in a given row or column and $a \in A$. Since the second reduction compares the magnitude of these statistics to a pre-defined threshold, the retention of pseudo-signal variables is approximately independent of all pseudo-noise variables. Provided the dependence between pseudo-signal variables does not make $R(Y, (I - P_{A_{-a}}) x_a)$ small for many possible sets A and indices a, Cox reduction will retain pseudo-signal variables.

In contrast, any pseudo-noise variables that survived the first reduction are retained through the second based on the magnitudes of statistics of approximate form

$$\sigma^{-1} \|(I - P_A)Y\|_2 R\{(I - P_A)Y, (I - P_{B_{-b}}) x_b\}$$

where $B \subseteq \mathcal{B}$ and $A \subseteq \mathcal{A}$ denote the pseudo-noise and pseudo-signal variables appearing in the same regression as $b \in B$. These statistics are bounded in magnitude by

$$\sigma^{-1} \|Y\|_2 \Delta(|A|)^{1/2}$$

up to a constant and so their relative sizes compared to (9) are asymptotically small. Cox reduction is not expected to prioritise these indices for retention. Proposition 6 formalises these observations, showing that when the sample size is large enough, pseudo-signal variables included in a set $\mathcal{A}_1 \subseteq \mathcal{A}$ are ranked more highly by Cox reduction than pseudo-noise variables. Provided the threshold for retention is chosen appropriately, Cox reduction will succesfully distinguish between these two types of variables.

PROPOSITION 6. *Suppose Conditions 1 and 2 hold with $a = k$. Define $\mathcal{A}_1 \subseteq \mathcal{A}$ to satisfy*

$$\max_{a \in \mathcal{A}_1} \max_{A \subseteq \mathcal{A} \setminus \{a\}, |A| \leq k-1} \{|R(Y, (I - P_A) x_a)|\}^{-1} = O(1).$$

Then there exist $N, \alpha > 0$ such that when $n \geq N$ the indices in $\mathcal{A}_1$ are retained by the second reduction but those in $\mathcal{B}$ that survived the first reduction, are not. This holds irrespective of the random arrangement of indices in the hypercube.

It is possible that $\mathcal{A}_1$ is empty or does not include the full set of pseudo-signal variables. In this case, the variables in $\mathcal{A} \setminus \mathcal{A}_1$ may be retained by Cox reduction, although this will depend on the random arrangement of the indices in the hypercube.

The second reduction requires any apparent explanatory power to persist when the variable in question is accompanied by relatively strong companion variables. Its intended purpose was to differentiate variables in $\mathcal{S}$ from those in $\mathcal{S}^c \cap \mathcal{A}$, when $\mathcal{S} \subseteq \mathcal{A}$. Consider a regression of Y on X_A where $A \subseteq \mathcal{A}$ contains signal and noise indices.

Lemma S3 shows that the entry of $\sigma T_{Y:A}$ corresponding to $a \in A$ is given by

$$(10) \qquad \Delta_1 + \Delta_2$$

where

$$\Delta_1 = \sqrt{1 - R^2(x_a, X_{A_{-a}})}\theta_a^0 \|x_a\|_2$$

$$\Delta_2 = \|Y - X_A\theta_A^0\|_2 R(Y - X_A\theta_A^0, (I - P_{A_{-a}})x_a).$$

The term Δ_1 describes the amount of the true signal corresponding to each variable that is recovered, and is equal to zero when a indexes a noise variable. The term Δ_2 characterises how well the variables in A are able to recover the omitted signal $Y - X_A\theta_A^0$ and is bounded in magnitude by $\|Y - X_A\theta_A^0\|_2$. A regression of Y on X_A is able to distinguish between the noise and signal variables in A if the term $\min_{a \in \mathcal{S} \cap A} |\Delta_1|$ is large in comparison to $\max_{a \in A} |\Delta_2|$. If a indexes a noise variable, then $\Delta_1 = 0$ and $|\Delta_2|$ is minimised (up to terms involving ϵ) when A includes the signal variables that are not orthogonal to x_a. In contrast, Δ_1 is largest for $a \in \mathcal{S} \cap A$ when colinearity between signal variables and other variables in A is absent. Whether or not Cox reduction accurately differentiates variables in $\mathcal{S}$ from those in $\mathcal{S}^c \cap \mathcal{A}$ when $\mathcal{S} \subseteq \mathcal{A}$ therefore depends on the balance of colinearity among the variables present, and the proportion of signal that they contain.

In practice, the threshold α is chosen so that the set of retained variables is relatively stable across repeated rerandomisation of the indices in the two hypercubes. In view of the observations above, the set of pseudo-noise variables retained due to high spurious correlation $R((I - P_A)Y, (I - P_{B_{-b}})x_b)$ is expected to differ upon repeated rerandomisation of the variable indices in the hypercubes. In contrast, provided the dependence between pseudo-signal variables does not cause $R(Y, (I - P_{A_{-a}})x_a)$ to be uniformly small over many sets A and indices $a \in A$, pseudo-signal variables will be retained by Cox reduction irrespective of the arrangement of variable indices. Thus, a relatively stable set of retained variables over repeated rerandomisation suggests presence of variables with genuine explanatory power. If no threshold α results in a stable set of manageable size, this is a warning against the procedure for the data at hand.

6.6 Problematic situations

The analysis above points to situations that may lead Cox reduction to exclude relevant variables. For example, a signal variable with a weak sample correlation with the outcome is relatively unlikely to be retained by Cox reduction. Section 5.2 discusses how such a variable might arise. Most variable selection procedures exclude the possibility that signal variables have weak marginal or partial correlations with the response. See Condition 3 in [26],

and the definition of partial faithfulness or Assumption 4 in [7].

In the second round of Cox reduction, variables from $\mathcal{A}$ appear together with frequency depending on the dimension of the square. When a signal variable has a weak partial correlation with the response given many different variables from $\mathcal{A}$, it may be declared unimportant in both of its second-round regressions and hence discarded by Cox reduction. Rerandomisation or a more systematic arrangement of the variable indices in the second reduction improve the situation. This is discussed in Sections 7 and 9.3.

Another unfavourable situation, more problematic for marginal screening than for Cox reduction, is when a noise variable that is uncorrelated with all signal variables has a large sample correlation with the response purely by chance. Simulations in [26] pointed to large spurious correlations for large p. This was formalised in [27] which established the limiting distribution of the maximum spurious correlation. The distributions of other order statistics were not discussed. Any spurious noise variable would be contained in $\mathcal{A}$ and likely survive the first reduction. Depending on the random arrangement of variables in the square, it may also survive the second reduction. Similarly, noise variables may be retained when multiple correlation coefficients between uncorrelated variables are spuriously large.

In view of the model assessment phase, inclusion of noise variables is less problematic than omission of signal variables provided there are not so many as to make assessment of models practically infeasible. Sample splitting and rerandomisation, discussed in Section 7, reduce the survival probability of noise variables.

7. RECOMMENDED IMPROVEMENTS TO COX REDUCTION

7.1 Alternating subsamples

We propose using a different portion of the sample for each reduction round, breaking the dependence between analyses in the first and second reductions and thereby reducing the probability that irrelevant noise variables appear correlated with the response in both rounds. Let $\mathcal{I}$ be a subset of $\{1, \ldots, n\}$. The split-sample version of Cox reduction uses observations indexed by $\mathcal{I}$ for the first reduction and the model assessment phase, and those indexed by $\mathcal{I}^c$ for the second reduction. For more than two rounds, alternation should continue in this way. We recommend including $30 - 40\%$ of observations in $\mathcal{I}$ for the two-stage procedure, based on the relative difficulty of the stages.

7.2 Rerandomisation

A benefit of Cox reduction is the external source of randomness from the arrangement of variable indices in

successive hypercubes. This provides a way of internally calibrating the procedure by rerandomising the variable indices. Battey and Cox [2] suggested this as a check on the stability of $\hat{\mathcal{S}}$. An elaboration is to rerandomise repeatedly, resulting in sets $\hat{\mathcal{S}}_1, \hat{\mathcal{S}}_2, \ldots$, and only retain variables in the final set $\hat{\mathcal{S}}$ if they are present most of the time, e.g. in at least half the rerandomisation outcomes. With this adaptation, Cox reduction is less dependent on the particular set of variables appearing in regressions together, reducing the chances that signal variables are omitted due to small partial correlations or that noise variables are retained due to high spurious correlation.

A natural and partially refutable criticism is that the procedure entails at least one tuning parameter (from the second-round reduction), increased to two when rerandomisation is introduced. In practice, selection of tuning parameters is guided by stability of the set $\hat{\mathcal{S}}$ and is rarely problematic. If there appears to be no choice of tuning pair that delivers a stable outcome, that is a warning against the procedure. The ability of Cox reduction to reveal its own fragility on the data at hand is an advantage over the alternatives discussed in Section 5.

8. NUMERICAL PERFORMANCE

Section S2 of the supplementary material shows the numerical performance of Cox reduction and the proposed construction of the model confidence set. Simulated data were generated to compare marginal screening and LASSO penalised regression to Cox reduction in the setting laid out in Section 6. Cox reduction and marginal screening both performed well, retaining all signal variables in the comprehensive model and yielding a model confidence set with high simulated coverage probability. In contrast, the undertuned LASSO failed to include all signal variables in the comprehensive model when correlations among variables were high and signal strengths were low. The size of the model confidence set constructed from marginal screening contained more models on average than that obtained from Cox reduction, as expected based on the discussion in Section 5. The recommended improvements to Cox reduction from Section 7 performed favourably.

To test the performance of the procedure in the presence of more complicated dependence structures, an analysis based on real data was performed. These were from the online supplement of [8] and consisted of the logarithm of the expression levels of $p = 4088$ genes, measured for $n = 71$ observational units. An artificial response was generated as $Y = \mathbb{1}_n + X\theta^0 + \epsilon$, with X from the genomics data, $\mathbb{1}_n$ an n-dimensional vector of ones, and $\epsilon \sim N(0, 1)$. Four cases were considered to illustrate our theoretical insights by adjusting the set $\mathcal{S}$ of signal variables. In all cases $\theta^0_{\mathcal{S}} = 2\mathbb{1}_3$.

To aid construction and understanding of the examples, we standardized the columns of X to have unit sample standard deviation. This would not be appropriate in a genuine statistical analysis as it destroys the interpretation of the regression coefficients.

In each of 500 Monte Carlo replications an artificial response was generated anew and a confidence set of models was constructed using the most competitive of the reduction procedures from the analysis in Section S2 of the supplementary material, namely marginal screening (MS), Cox reduction with rerandomisation (CR-R) and Cox reduction with rerandomisation and sample splitting (CR-RSS). A set of indices $\mathcal{I} \subset \{1, \ldots, n\}$ of size $|\mathcal{I}| = 29$ was drawn at random and fixed across Monte Carlo replications. This was used for model assessment for all procedures. The observations indexed by $\mathcal{I}^c$ were used for reduction for MS and CR-R, while CR-RSS used $\mathcal{I}$ for its first reduction and $\mathcal{I}^c$ for its second. For each approach, we recorded the proportion of times each signal variable was included in $\hat{\mathcal{S}}$, the proportion of times $\hat{\mathcal{S}}$ included the full set of signal variables, the proportion of times the set of signal variables was included in the model confidence set $\mathcal{M}$ and the size of $\mathcal{M}$.

For a fair comparison, the comprehensive model was taken to consist of 15 variables for all procedures. Thus in the rerandomised versions of Cox reduction we retained in $\hat{\mathcal{S}}$ the 15 variables that were suggested most often, where the tuning parameter in the second reduction was chosen such that stability of the model confidence set was achieved. It is possible to construct examples in which a large significance level is needed in the second round in order that the same variables are retained frequently over successive rerandomisation. This is usually indicative of a weak signal to noise ratio and signals caution. Alternatively, there may be a sharp transition in the significance level at which the number of retained variables in all randomisations jumps sharply. Such instability is again a warning of partial fragility.

A favourable situation for both marginal screening and Cox reduction is when the estimated absolute effect appears larger than it is due to strong dependence between signal variables of an appropriate sign. This situation was explored and the results were similar to those reported in the analysis in Section S2 and are therefore omitted. The aim of the following three examples is to highlight limitations of the methodology in the presence of certain covariate dependencies.

Example: effect cancellation A signal variable that is marginally uncorrelated in sample with the response variable is unlikely to be retained by marginal screening or Cox reduction. Section 5.2 shows how this can happen for signal variables. In particular, it entails fairly strong sample correlation between signal variables of an appropriate sign. In this example, the set of signal variables was

chosen to be the variables $\{1852, 3862, 4088\}$ with strong positive and negative correlations given by

$$\text{Corr}(X_\mathcal{S}, X_\mathcal{S}) = \begin{pmatrix} 1.00 & -0.73 & -0.72 \\ -0.73 & 1.00 & 0.73 \\ -0.72 & 0.73 & 1.00 \end{pmatrix}.$$

Table 1 displays the results. The situation is not improved by increasing the signal to noise ratio, with both marginal screening and Cox reduction rarely including the first signal variable in the comprehensive model, as expected. If the effect cancellation was only partial and in a suitable range, Cox reduction will sometimes dominate over marginal screening, as discussed in Section 5.

	MS	CR-R	CR-RSS
$\mathbb{P}(1852 \in \hat{\mathcal{S}})$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
$\mathbb{P}(3862 \in \hat{\mathcal{S}})$	1.00 (0.00)	1.00 (0.06)	1.00 (0.00)
$\mathbb{P}(4088 \in \hat{\mathcal{S}})$	1.00 (0.00)	1.00 (0.00)	0.99 (0.08)
$\mathbb{P}(\mathcal{S} \subseteq \hat{\mathcal{S}})$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
$|\hat{\mathcal{S}}|$	15	15	15
$\mathbb{P}(\mathcal{S} \in \mathcal{M})$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
$\mathbb{E}(|\mathcal{M}\backslash\mathcal{S}|)$	4280 (635)	2907 (1437)	3273 (1429)

Table 1: Simulation results using real data as the design matrix. The set of signal variables is $\mathcal{S} = \{1852, 3862, 4088\}$, $\hat{\mathcal{S}}$ is the comprehensive model, $\mathcal{M}$ is the set of models of size at most $s^\# = 5$ that are significant at level $\alpha = 0.01$. Empirical standard errors are given in brackets.

Example: noise variables with stronger marginal correlations than signal A second example illustrates a situation in which Cox reduction outperforms marginal screening. The set of signal variables was taken as $\{5, 1812, 1861\}$ with correlation structure

$$\text{Corr}(X_\mathcal{S}, X_\mathcal{S}) = \begin{pmatrix} 1.00 & 0.18 & 0.32 \\ 0.18 & 1.00 & 0.28 \\ 0.32 & 0.28 & 1.00 \end{pmatrix}.$$

Several noise variables were correlated with all three signal variables so that these appeared marginally stronger than the signal variables themselves. The signal variables nevertheless had sufficient marginal strength to be retained through Cox reduction. Results are in Table 2.

	MS	CR-R	CR-RSS
$\mathbb{P}(5 \in \hat{\mathcal{S}})$	0.00 (0.00)	0.53 (0.50)	0.95 (0.21)
$\mathbb{P}(1812 \in \hat{\mathcal{S}})$	0.00 (0.06)	0.91 (0.28)	1.00 (0.00)
$\mathbb{P}(1861 \in \hat{\mathcal{S}})$	0.19 (0.39)	0.65 (0.48)	1.00 (0.05)
$\mathbb{P}(\mathcal{S} \subseteq \hat{\mathcal{S}})$	0.00 (0.00)	0.28 (0.45)	0.95 (0.22)
$|\hat{\mathcal{S}}|$	15	15	15
$\mathbb{P}(\mathcal{S} \in \mathcal{M})$	0.00 (0.00)	0.27 (0.45)	0.93 (0.23)
$\mathbb{E}(|\mathcal{M}\backslash\mathcal{S}|)$	4854 (484)	2065 (1886)	555 (739)

Table 2: As Table 1 with $\mathcal{S}$ replaced by $\{5, 1812, 1861\}$

Example: weakly correlated signal variables This final example illustrates a setting where marginal screening outperforms Cox reduction. The set of signal variables was chosen to be a group of three weakly correlated variables indexed by $\{10, 2027, 2923\}$. The sample correlations among these signal variables were

$$\text{Corr}(X_\mathcal{S}, X_\mathcal{S}) = \begin{pmatrix} 1.00 & 0.10 & 0.08 \\ 0.10 & 1.00 & -0.04 \\ 0.08 & -0.04 & 1.00 \end{pmatrix}.$$

The results are reported in Table 3.

	MS	CR-R	CR-RSS
$\mathbb{P}(10 \in \hat{\mathcal{S}})$	0.95 (0.21)	0.34 (0.47)	1.00 (0.00)
$\mathbb{P}(2027 \in \hat{\mathcal{S}})$	1.00 (0.00)	0.03 (0.16)	1.00 (0.21)
$\mathbb{P}(2923 \in \hat{\mathcal{S}})$	1.00 (0.00)	0.72 (0.45)	0.04 (0.10)
$\mathbb{P}(\mathcal{S} \subseteq \hat{\mathcal{S}})$	0.95 (0.21)	0.00 (0.05)	0.04 (0.20)
$|\hat{\mathcal{S}}|$	15	15	15
$\mathbb{P}(\mathcal{S} \in \mathcal{M})$	0.94 (0.23)	0.00 (0.05)	0.04 (0.20)
$\mathbb{E}(|\mathcal{M}\backslash\mathcal{S}|)$	525 (935)	3653 (1615)	732 (647)

Table 3: As Table 1 with $\mathcal{S}$ replaced by $\{10, 2027, 2923\}$

Marginal screening successfully included all signal variables in the comprehensive model and produced a model confidence set of moderate size on average over Monte Carlo replicates. Cox reduction performed less well, with at least one signal variable being frequently omitted from $\hat{\mathcal{S}}$. In these rerandomised versions of Cox reduction, sample splitting increased the probability that the first two signal variables were retained but decreased the survival probability of the third.

In all three examples, the omission of signal variables at the reduction stage does not affect the usefulness of the model confidence set for identifying models that fit the data equally well.

9. FURTHER DISCUSSION

9.1 Bootstrap confidence sets of models

A natural attempt to construct confidence sets of models uses the LASSO or similar on B bootstrap samples, obtained by sampling n observations with replacement. This suggestion can be refuted in the light of [42, 41].

Let the true positive rate be the proportion of signal variables selected by the LASSO and the false discovery rate be the proportion of selected variables that are noise variables. [42] showed that even under idealised conditions with large signal strengths and uncorrelated variables, the false discovery rate is lower bounded by a function of the true positive rate with probability one. Further, [41] shows that the first noise variable is selected earlier on the LASSO path than the final signal variable. Hence, each model either includes the full set of signal variables

contaminated by noise variables, or does not include the full set of signal variables. These conclusions hold with probability tending to one in the notional double asymptotic regime in which p grows with n. As a result, with probability close to one, the true model is never selected and a model confidence set constructed in this way has coverage probability close to zero.

9.2 Cox reduction with unknown variance

An unknown error variance σ can either be estimated prior to Cox reduction based on all p covariates or estimated anew in each of the regressions that are run. The former approach results in an unbiased estimate, whilst the latter typically produces biased estimates from regressions omitting signal variables. Nevertheless, the latter is preferable as a severely biased estimate indicates that the joint explanatory power of covariates appearing in a regression together is weak, making it less likely that these covariates are retained through the second round of Cox reduction. The associated Wald statistic is

$$T_{Y:K} = \hat{\sigma}_K^{-1} D_K^{-1/2} \hat{\theta}_{Y:K},$$

$$\hat{\sigma}_K^2 = \frac{\|Y - X_K \hat{\theta}_{Y:K}\|_2^2}{n - |K|}$$

where D_K is a diagonal matrix with entries given by the diagonal entries of $(X_K^T X_K)^{-1}$. This choice does not affect the first-round reduction, which is based on the relative significance of variables in a given regression. Its effect in the second round is to de-emphasise covariates appearing in regressions with a weak signal.

If it is instead decided to estimate σ prior to Cox reduction, suitable estimators are in [24], [40] and [21]. See [38] for a numerical comparison of their performance. As the relative significance of covariates are unaffected by estimation of σ, an equivalent approach sets σ to be an arbitrary value, adjusting the second-round significance level to give a stable comprehensive model of manageable size.

9.3 Systematic arrangement of variables

Expression (10) shows that differentiation between Cox reduction and marginal screening can be achieved by forcing second-round regressions to contain correlated signal and noise variables from $\mathcal{A}$. This suggests a more systematic arrangement of variables in the square in which indices E and F for which $|R(x_E, x_F)| \gg 0$ appear together in rows or columns. Further analysis is needed to ascertain any optimal arrangement of variables in the second round to maximise the probability of retaining signal variables.

9.4 Binary responses

Scheffé [39] emphasised the approximate validity of inference based on a notional linear model, owing to the randomisation. Lewis and Battey [35] formalise this intuition beyond the context of Cox reduction when the outcomes are generated from a linear logistic model. Generative binary models, when fitted by maximum likelihood, are problematic in the second stage of Cox reduction, where many combinations of strong variables form a separating hyperplane, that is, produce no classification errors within sample. Ordinary least squares fitting overcomes the difficulties.

9.5 Prediction

The argument for confidence sets of models is somewhat weakened when prediction is the primary goal. However, stability of the predictor over time and in different contexts is important and more likely with a model that has an underlying interpretation as well as immediate predictive success. By consideration of prediction intervals for each model in $\mathcal{M}$, we account both for model uncertainty and statistical uncertainty. Overlapping prediction intervals are reassuring, while those differing considerably point to instability of the predictor under alternative circumstances. See Section S3 of the supplementary material for an example.

Availability of source code. Source code for implementing Cox reduction and constructing confidence sets of models is available from www.ma.imperial.ac.uk/∼hbattey/softwareCube.html. The more user-interfaced R package HCmodelSets [31] is accompanied by a detailed guide to usage [32]. Source code implementing some of the refinements discussed in the present work, such as those used in the simulations in Section S2.1 of the supplementary material, is available from https://github.com/rm-lewis.

REFERENCES

[1] BARBER, R. F. and CANDÈS, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.* **43** 2055-2085.

[2] BATTEY, H. S. and COX, D. R. (2018). Large numbers of explanatory variables: a probabilistic assessment. *Proc. R. Soc. Lond. A.* **474**.

[3] BIND, M. and RUBIN, D. B. (2023). High-dimensional randomization-based inference capitalizing on classical design and modern computing. *Behaviormetrika* **50** 9-29.

[4] BRADFORD HILL, A. (1965). The environment and disease: association or causation. *Proc. R. Soc. Med.* **58** 295-300.

[5] BRAGG, F., KARTSONAKI, C. and GUO, Y. E. A. (2022). The role of NMR-based circulating metabolic biomarkers in development and risk prediction of new onset type 2 diabetes. *Nature Scientific Reports* **12** 15071.

[6] BREHENY, P. and HUANG, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Stat.* **5** 232-253.

[7] BÜHLMANN, P., KALISCH, M. and MAATHUIS, M. H. (2010). Variable selection in high-dimensional linear models: partially faithful distributions and the PC-simple algorithm. *Biometrika* **97** 261-278.

[8] BÜHLMANN, P., KALISCH, M. and MEIER, L. (2014). High-dimensional statistics with a view toward applications in biology. *Annu. Rev. Stat. Appl.* **1** 255-278.

[9] CANDÈS, E., FAN, Y., JANSON, L. and LV, J. (2018). Panning for Gold: 'Model-X' Knockoffs for High Dimensional Controlled Variable Selection. *J. Roy. Statist. Soc., B* **80** 551-577.

[10] COCHRAN, W. G. (1938). The omission or addition of an independent variable in multiple linear regression. *J. R. Statist. Soc. Suppl.* **5** 171–176.

[11] COX, D. R. (1960). Regression analysis when there is prior information about supplementary variables. *J. Roy. Statist. Soc., B* **22** 172–176.

[12] COX, D. R. (1968). Notes on some aspects of regression analysis (with discussion). *J. Roy. Statist. Soc., A* **131** 265–279.

[13] COX, D. R. (1992). Causality: some statistical aspects. *J. Roy. Statist. Soc., A* **155** 291-301.

[14] COX, D. R. and BATTEY, H. S. (2017). Large numbers of explanatory variables, a semi-descriptive analysis. *Proc. Natl. Acad. Sci. USA* **114** 8592-8595.

[15] COX, D. R. and DONNELLY, C. A. (2011). *Principles of Applied Statistics*. Cambridge University Press.

[16] COX, D. R. and SNELL, E. J. (1974). The choice of variables in observational studies. *J. Roy. Statist. Soc., C* **23** 51–59.

[17] COX, D. R. and SNELL, E. J. (1989). *Analysis of binary data* **32**, Second ed. Chapman & Hall, London.

[18] COX, D. R. and WERMUTH, N. (1990). An approximation to maximum likelihood estimates in reduced models. *Biometrika* **77** 747–761.

[19] COX, D. R. and WERMUTH, N. (1996). *Multivariate Dependencies*. Chapman & Hall, London.

[20] COX, D. R. and WERMUTH, N. (2003). A condition for avoiding effect reversal after marginalization. *J. R. Statist. Soc. B* **65** 937-941.

[21] DICKER, L. (2014). Variance estimation in high-dimensional linear models. *Biometrika* **101** 269-284.

[22] EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407-499.

[23] EVANS, R. (2020). Model selection and local geometry. *Ann. Statist.* **48** 3513-3544.

[24] FAN, J., GUO, S. and HAO, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. R. Statist. Soc. B* **74** 37-65.

[25] FAN, J. and LI, R. (2011). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348-1360.

[26] FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Statist. Soc. B* **70** 849-911.

[27] FAN, J., SHAO, Q. and ZHOU, W. (2018). Are discoveries spurious? Distributions of maximum spurious correlations and their applications. *Ann. Statist.* **46** 989-1017.

[28] FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1-22.

[29] GEORGE, E. I. and MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881-889.

[30] HANSEN, P., LUNDE, A. and NASON, J. (2011). The model confidence set. *Econometrica* **79** 453-497.

[31] HOELTGEBAUM, H. H. (2023). HCmodelSets: Regression with a Large Number of Potential Explanatory Variables. R package version 1.1.3.

[32] HOELTGEBAUM, H. H. and BATTEY, H. S. (2019). HCmodelSets: An R package for specifying sets of well-fitting models in high dimensions. *The R Journal* **11** 370–379.

[33] KARTSONAKI, C., PANG, Y., MILLWOOD, I., YANG, L., GUO, Y., WALTERS, R., LV, J., HILL, M., YU, C., CHEN, Y., CHEN, X., O'NEILL, E., CHEN, J., TRAVIS, R. C., CLARKE, R., LI, L., CHEN, Z. and HOLMES, M. V. (2022). Circulating proteins and risk of pancreatic cancer: a case-subcohort study among Chinese adults. *International Journal of Epidemiology* **51** 817-829.

[34] LE CAM, L. (1960). Locally asymptotically normal families of distributions. Certain approximations to families of distributions and their use in the theory of estimation and testing hypotheses. *Univ. California Publ. Statist.* **3** 37–98.

[35] LEWIS, R. and BATTEY, H. S. (2023). On inference in high-dimensional logistic regression models with separated data. *Biometrika*.

[36] LI, S., SESIA, M., ROMANO, Y., CANDÈS, E. and SABATTI, C. (2022). Searching for consistent associations with a multi-environment knockoff filter. *Biometrika* **109** 611-629.

[37] MITCHELL, T. J. and BEAUCHAMP, J. J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* **83** 1023-1032.

[38] REID, S., TIBSHIRANI, R. and J., F. (2016). A study of error variance estimation in lasso regression. *Statist. Sinica* **26** 35-67.

[39] SHEFFÉ, H. (1959). *Analysis of Variance*. Wiley.

[40] STÄDLER, N., BÜHLMANN, P. and VAN DE GEER, S. (2010). ℓ_1-penalization for mixture regression models. *Test* **19** 209-256.

[41] SU, W. (2018). When is the first spurious variable selected by sequential regression procedures? *Biometrika* **105** 517-527.

[42] SU, W., BOGDAN, M. and CANDES, E. (2017). False discoveries occur early on the lasso path. *Ann. Statist.* **45** 2133-2150.

[43] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288.

[44] VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

[45] Y., P., KARTSONAKI, C., LV, J., MILLWOOD, I. Y., FAIRHURST-HUNTER, Z., TURNBULL, I., BRAGG, F., HILL, M. R., YU, C., GUO, Y., CHEN, Y., YANG, L., CLARKE, R., WALTERS, R. G., WU, M., CHEN, J., LI, L., CHEN, Z. and HOLMES, M. V. (2022). Adiposity, metabolomic biomarkers, and risk of nonalcoholic fatty liver disease: a case-cohort study. *Am. J. Clin. Nutr.* **115** 799-810.

[46] YATES, F. (1936). A new method of arranging variety trials involving a large number of varieties. *The Journal of Agricultural Science* **26** 424-455.

[47] ZHANG, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894-942.

SUPPLEMENTARY MATERIAL

This document provides the supplementary material for the paper "Cox reduction and confidence sets of models: a theoretical elucidation".

S1. PSEUDO-CODE

This section provides pseudo-code for the reduction and model assessment phases of [1]. The model confidence set may be constructed based on an arbitrary comprehensive model using Algorithm 1. This approach is discussed in Section 4 of the main text. Algorithms 2 and 3 outline the construction of the comprehensive model based on Cox reduction. Algorithm 2 focuses on the first round of reduction in a cube and a depiction of the partially balanced incomplete block arrangement is given in Figure S1. Algorithm 3 considers the second round in a square.

For a design matrix X and response vector Y, the following commands may be used to apply both rounds of Cox reduction to construct the comprehensive model and form the model confidence set.

Load the data X and Y;
Initialise α, ϑ and $s^{\#}$;
$\hat{\mathcal{S}}_1 \leftarrow$ CR_CUBE(X, Y)
$\hat{\mathcal{S}} \leftarrow$ CR_SQUARE$(X, Y, \hat{\mathcal{S}}_1, \alpha)$
$\mathcal{M} \leftarrow$ MCS$(X, Y, \hat{\mathcal{S}}, \vartheta, s^{\#})$
Return $\mathcal{M}$, the model confidence set.

Algorithm 1 *An algorithm for constructing the model confidence set based on a comprehensive model $\hat{\mathcal{S}}$.*

---

procedure MCS$(X, Y, \hat{\mathcal{S}}, \vartheta, s^{\#})$

    Input: design matrix $X \in \mathbb{R}^{n \times p}$, response vector $Y \in \mathbb{R}^{n \times 1}$, comprehensive model $\hat{\mathcal{S}}$, significance level $\vartheta \in (0, 1)$, maximum model size $s^{\#} \in \mathbb{N}$;

    Initialise $\mathcal{M} \leftarrow \emptyset$;

    for each $\mathcal{S}_m \subset \hat{\mathcal{S}}$ with $|\mathcal{S}_m| \leq s^{\#}$ **do**

        Compute the likelihood ratio test statistic $w(\mathcal{S}_m)$ for testing $\mathcal{S}_m$ against $\hat{\mathcal{S}}$,

        if $w(\mathcal{S}_m)$ is not significant at level ϑ **then**

            $\mathcal{M} \leftarrow \mathcal{M} \cup \mathcal{S}_m$;

        end if

    end for

    Return $\mathcal{M}$.

end procedure

Algorithm 2 *An algorithm for performing the first round of Cox reduction in a cube. Variables are selected if they are among the two most significant in a regression. Variables are retained if they are selected in at least two out of three regressions. The algorithm returns the indices of selected variables.*

---

procedure CR_CUBE(X, Y)

    Input: design matrix $X \in \mathbb{R}^{n \times p}$, response vector $Y \in \mathbb{R}^{n \times 1}$;

    Do: randomly arrange the indices $1, \ldots, p$ in a cube of size $k \times k \times k$ where $k \geq p^{1/3}$;

    for each row, column or "corridor" in the cube **do**

        Identify the indices $\mathcal{K} \subset \{1, \ldots, p\}$ in that row, column or corridor,

        Regress Y against $X_{\mathcal{K}}$,

        Select the indices corresponding to the two most significant variables;

    end for

    $\hat{\mathcal{S}}_1 \leftarrow$ the set of indices selected in at least two out of three regressions;

    Return $\hat{\mathcal{S}}_1$.

end procedure

Algorithm 3 *An algorithm for performing the second round of Cox reduction in a square. Variables are selected if they are significant at level α. Variables are retained if they are selected in at least one out of two regressions. The algorithm returns the comprehensive model.*

---

procedure CR_SQUARE$(X, Y, \hat{\mathcal{S}}_1, \alpha)$

    Input: design matrix $X \in \mathbb{R}^{n \times p}$, response vector $Y \in \mathbb{R}^{n \times 1}$, set of selected indices $\hat{\mathcal{S}}_1 \subseteq \{1, \ldots, p\}$, significance level $\alpha \in (0, 1)$;

    Do: randomly arrange the indices in $\hat{\mathcal{S}}$ in a square of size $k \times k$ where $k \geq |\hat{\mathcal{S}}|^{1/2}$;

    for each row or column in the square **do**

        Identify the indices $\mathcal{K} \subset \hat{\mathcal{S}}_1$ in that row/column,

        Regress Y against $X_{\mathcal{K}}$,

        Select the indices corresponding to variables that are significant at level α;

    end for

    $\hat{S} \leftarrow$ the set of indices selected in at least one out of two regressions;

    Return $\hat{S}$.

end procedure

S2. NUMERICAL PERFORMANCE

This section analyses the numerical performance of Cox reduction and the construction of the model confidence set. Simulated data were generated to compare marginal screening and LASSO penalised regression to Cox reduction in the setting laid out in Section 6 of the original paper.
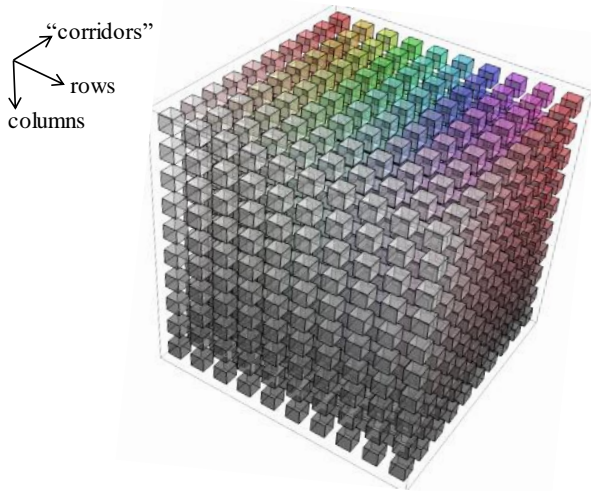
Fig S1: Partially balanced incomplete block arrangement of $p = 1000$ variables in blocks of size $k = 10$. Each small cube represents a variable index and appears with $k - 1$ companion variables. The colour is for visualisation only and plays no role in the analysis.

S2.1 Monte Carlo experiments

In each of 500 Monte Carlo replications, $n = 100$ copies of independent, multivariate normal variables were generated of dimension $p = 1000$ and covariance matrix $P\Sigma P^{-1}$, where P is a permutation matrix and Σ is an identity matrix with a block of size $|\mathcal{A}| = s + c$ replaced by an equi-correlation matrix with correlations ρ. The covariates corresponding to s correlated variables were added to randomly generated Gaussian noise with unit variance to produce a response variable. The sample was split into two portions containing 40% and 60% of the observations respectively. Comprehensive models $\hat{\mathcal{S}}$ were obtained using the following reduction strategies:

- Marginal screening (MS): $\hat{\mathcal{S}}$ was taken to be the 15 covariates with the highest sample correlation with the response based on the largest portion of data.
- Undertuned LASSO: $\hat{\mathcal{S}}$ was taken as the set (5) defined in Section 5 of the original paper based on the largest portion of data with λ chosen as the smallest such that a model of size at most 15 was selected.
- Cox Reduction (CR): $\hat{\mathcal{S}}$ was the set of retained variables after two rounds of Cox reduction on the largest portion of data. In the first round, variables that were among the two most significant in at least 2/3 regressions were retained. In the second round, all variables significant at level α were retained. The second round was performed initially with $\alpha = 0.05$ and repeated, reducing α by 0.001 each time, until at most 15 variables remained.
- Cox Reduction with sample splitting (CR-SS): as CR but instead of using the largest portion of data

for both reduction rounds, the smallest portion was used for the first round and the largest for the second.

- Cox Reduction with rerandomisation (CR-R): as CR but with rerandomisation of variable indices in the cube. For each randomisation, α was initialised at 0.05 and reduced by 0.001 until model of size at most 25 was obtained. The process was repeated 10 times and the comprehensive model $\hat{\mathcal{S}}$ consisted of all variables retained in at least 40% of repetitions. The restriction to 25 variables was made to ensure that $|\hat{\mathcal{S}}|$ was roughly 15, although this differs according to the randomisation. The largest portion of data was used for all repetitions.
- Cox Reduction with rerandomisation and sample splitting (CR-RSS): as CR-R but in all repetitions the smallest portion of data was used for the first reduction and the largest for the second, as in CR-SS.

A model confidence set was obtained for each comprehensive model using the smaller portion of data by including all submodels of size at most $s^{\#} = 5$ not rejected at level $\vartheta = 0.01$. Let $\mathcal{C}$ be the set of noise variables having non-zero populaton correlation with signal variables. For each Monte Carlo replicate, the proportion of variables in $\mathcal{S}$ and $\mathcal{C}$ retained was recorded, together with an indicator of whether the comprehensive model $\hat{\mathcal{S}}$ contained the full set of signal variables $\mathcal{S}$, whether $\mathcal{S}$ was included in the confidence set of models $\mathcal{M}$ and the number of false models included in $\mathcal{M}$. The process was repeated with various choices of s, c, ρ and signal to noise ratio. The results are summarised in Tables S1-S4, where expectations and probabilities refer to averages over Monte Carlo replication.

The model confidence set should ideally include a) the true model and b) a minimal number of other well-fitting models. All reduction strategies included the true set of signal variables in the comprehensive model in a large proportion of cases. The undertuned LASSO performed the worst, particularly when correlations among variables were high and the signal to noise ratio was low where it failed to include all signal variables in the comprehensive model in a large number of cases (see the LASSO component of the second chunk of each table). Whilst marginal screening and the variations of Cox Reduction performed similarly in terms of retention of signal variables at the reduction stage and selection of the true model at the model assessment stage, marginal screening generally included more false models in the confidence set. This is partly due to our retaining a fixed number of variables for marginal screening, which is higher on average than the number of variables retained by the other reduction procedures. It is also to be expected based on the discussion in Section 5.2

Table S1: Simulation results when variables are correlated with $n = 100$, $s = 3$ and $c = 3$. $\mathcal{S}$ is the set of signal variables, $\mathcal{C}$ is the set of noise variables correlated (in population) with signal variables, $\hat{\mathcal{S}}$ is the comprehensive model obtained after all reduction rounds, $\mathcal{M}$ is the set of models that are significant at level ϑ. Empirical standard errors are given in brackets.

s	c	ρ	$\frac{\text{signal}}{\text{noise}}$	Reduction Strategy	$\mathbb{E}\left(\frac{|\mathcal{S}\cap\hat{\mathcal{S}}|}{|\mathcal{S}|}\right)$	$\mathbb{E}\left(\frac{|\mathcal{C}\cap\hat{\mathcal{S}}|}{|\mathcal{C}|}\right)$	$\mathbb{E}|\hat{\mathcal{S}}|$	$\mathbb{P}(\mathcal{S}\subseteq\hat{\mathcal{S}})$	$\mathbb{P}(\mathcal{S}\in\mathcal{M})$ $(\vartheta=0.01)$	$\mathbb{E}(|\mathcal{M}\backslash\mathcal{S}|)$ $(\vartheta=0.01)$
3	3	0.9	1	LASSO	0.98 (0.09)	0.35 (0.27)	13.6 (1.4)	0.93 (0.26)	0.91 (0.28)	1715.1 (946.7)
				MS	1.00 (0.00)	1.00 (0.00)	15.0 (0.0)	1.00 (0.00)	0.98 (0.13)	2947.3 (1080.7)
				CR	1.00 (0.04)	0.96 (0.11)	14.4 (0.8)	0.99 (0.11)	0.98 (0.14)	2493.2 (1078.8)
				CR-SS	1.00 (0.02)	0.99 (0.06)	14.7 (0.7)	1.00 (0.06)	0.98 (0.14)	2570 (1101.1)
				CR-R	1.00 (0.00)	1.00 (0.00)	13.1 (2.0)	1.00 (0.00)	0.99 (0.11)	1893.9 (1810.7)
				CR-RSS	1.00 (0.00)	1.00 (0.00)	10.9 (2.0)	1.00 (0.00)	0.97 (0.18)	755.6 (859.7)
3	3	0.9	0.6	LASSO	0.84 (0.19)	0.29 (0.26)	13.6 (1.5)	0.56 (0.50)	0.55 (0.50)	2237.4 (1085.1)
				MS	1.00 (0.00)	1.00 (0.00)	15.0 (0.0)	1.00 (0.00)	0.98 (0.13)	4080.4 (822.9)
				CR	0.98 (0.08)	0.94 (0.13)	14.4 (0.9)	0.94 (0.24)	0.93 (0.26)	3396.8 (1042.6)
				CR-SS	0.99 (0.07)	0.95 (0.12)	14.7 (0.8)	0.97 (0.18)	0.92 (0.27)	3438.7 (1101.9)
				CR-R	1.00 (0.00)	1.00 (0.00)	13.7 (2.0)	1.00 (0.00)	1.00 (0.04)	3211.6 (2401.1)
				CR-RSS	1.00 (0.00)	1.00 (0.00)	11.5 (2.3)	1.00 (0.00)	0.94 (0.23)	1443.4 (1583.7)
3	3	0.5	1	LASSO	1.00 (0.00)	0.18 (0.24)	13.7 (1.3)	1.00 (0.00)	0.99 (0.10)	386.5 (316.4)
				MS	1.00 (0.00)	0.99 (0.07)	15.0 (0.0)	1.00 (0.00)	0.99 (0.11)	642.8 (477.1)
				CR	1.00 (0.00)	0.84 (0.22)	14.3 (1.0)	1.00 (0.00)	0.98 (0.13)	509.5 (411.6)
				CR-SS	1.00 (0.00)	0.96 (0.12)	14.7 (0.8)	1.00 (0.00)	0.96 (0.19)	494.9 (432.9)
				CR-R	1.00 (0.00)	0.96 (0.12)	13.6 (2.1)	1.00 (0.00)	0.99 (0.09)	505.5 (723.2)
				CR-RSS	1.00 (0.00)	0.99 (0.06)	11.6 (2.3)	1.00 (0.00)	0.96 (0.20)	212.6 (304.7)
3	3	0.5	0.6	LASSO	0.98 (0.08)	0.19 (0.23)	13.6 (1.4)	0.94 (0.23)	0.94 (0.24)	1385.6 (953.4)
				MS	1.00 (0.00)	0.96 (0.13)	15.0 (0.0)	1.00 (0.00)	0.99 (0.08)	2301.8 (1131.9)
				CR	0.99 (0.07)	0.74 (0.27)	14.3 (1.0)	0.96 (0.20)	0.95 (0.23)	1815.7 (1089.2)
				CR-SS	1.00 (0.03)	0.89 (0.19)	14.7 (0.8)	0.99 (0.10)	0.93 (0.26)	1698.9 (1141.1)
				CR-R	1.00 (0.01)	0.89 (0.21)	13.8 (2.0)	1.00 (0.04)	0.99 (0.10)	1771.4 (1798.7)
				CR-RSS	1.00 (0.00)	0.95 (0.12)	11.8 (2.4)	1.00 (0.00)	0.92 (0.27)	767.7 (1099.4)

Table S2: Simulation results when variables are correlated with $n = 100$, $s = 3$ and $c = 5$. $\mathcal{S}$ is the set of signal variables, $\mathcal{C}$ is the set of noise variables correlated (in population) with signal variables, $\hat{\mathcal{S}}$ is the comprehensive model obtained after all reduction rounds, $\mathcal{M}$ is the set of models that are significant at level ϑ. Empirical standard errors are given in brackets.

s	c	ρ	$\frac{\text{signal}}{\text{noise}}$	Reduction Strategy	$\mathbb{E}\left(\frac{\|\mathcal{S}\cap\hat{\mathcal{S}}\|}{\|\mathcal{S}\|}\right)$	$\mathbb{E}\left(\frac{\|\mathcal{C}\cap\hat{\mathcal{S}}\|}{\|\mathcal{C}\|}\right)$	$\mathbb{E}\|\hat{\mathcal{S}}\|$	$\mathbb{P}(\mathcal{S}\subseteq\hat{\mathcal{S}})$	$\mathbb{P}(\mathcal{S}\in\mathcal{M})$ $(\vartheta=0.01)$	$\mathbb{E}(\|\mathcal{M}\backslash\mathcal{S}\|)$ $(\vartheta=0.01)$
3	5	0.9	1	LASSO	0.97 (0.10)	0.31 (0.21)	13.7 (1.3)	0.91 (0.29)	0.90 (0.29)	1849.2 (1003.7)
				MS	1.00 (0.00)	1.00 (0.00)	15.0 (0.0)	1.00 (0.00)	1.00 (0.06)	3349.2 (1093.0)
				CR	1.00 (0.04)	0.94 (0.10)	14.5 (0.8)	0.99 (0.11)	0.98 (0.15)	2840.7 (1146.7)
				CR-SS	1.00 (0.03)	0.97 (0.08)	14.8 (0.4)	0.99 (0.09)	0.97 (0.18)	3034.9 (1174.2)
				CR-R	1.00 (0.00)	1.00 (0.01)	13.8 (1.8)	1.00 (0.00)	0.99 (0.10)	2630.3 (2179.5)
				CR-RSS	1.00 (0.00)	1.00 (0.00)	12.3 (1.8)	1.00 (0.00)	0.99 (0.12)	1400.7 (1259.5)
3	5	0.9	0.6	LASSO	0.81 (0.21)	0.30 (0.21)	13.5 (1.4)	0.50 (0.50)	0.50 (0.50)	2309.4 (1069.5)
				MS	1.00 (0.00)	1.00 (0.00)	15.0 (0.0)	1.00 (0.00)	0.99 (0.08)	4410.4 (775.4)
				CR	0.96 (0.12)	0.89 (0.13)	14.4 (0.8)	0.87 (0.34)	0.85 (0.35)	3547.5 (1146.3)
				CR-SS	0.98 (0.07)	0.92 (0.11)	14.8 (0.5)	0.95 (0.22)	0.91 (0.28)	3906.1 (1149.9)
				CR-R	1.00 (0.00)	1.00 (0.01)	14.6 (1.9)	1.00 (0.00)	0.98 (0.13)	4550.9 (3329.7)
				CR-RSS	1.00 (0.00)	1.00 (0.01)	13.0 (2.1)	1.00 (0.00)	0.94 (0.23)	2477.8 (2243.7)
3	5	0.5	1	LASSO	1.00 (0.00)	0.16 (0.17)	13.7 (1.4)	1.00 (0.00)	0.99 (0.12)	403.2 (333.5)
				MS	1.00 (0.00)	0.99 (0.05)	15.0 (0.0)	1.00 (0.00)	0.98 (0.13)	689.2 (516.4)
				CR	1.00 (0.01)	0.80 (0.20)	14.3 (0.9)	1.00 (0.04)	0.98 (0.13)	551.4 (455.2)
				CR-SS	1.00 (0.00)	0.95 (0.10)	14.8 (0.6)	1.00 (0.00)	0.97 (0.17)	558.3 (486.2)
				CR-R	1.00 (0.00)	0.96 (0.10)	14.2 (1.9)	1.00 (0.00)	0.98 (0.15)	593.1 (588.6)
				CR-RSS	1.00 (0.00)	0.99 (0.05)	13.2 (2.0)	1.00 (0.00)	0.97 (0.18)	408.8 (686.3)
3	5	0.5	0.6	LASSO	0.99 (0.06)	0.18 (0.18)	13.7 (1.3)	0.97 (0.18)	0.95 (0.21)	1450.1 (994.8)
				MS	1.00 (0.00)	0.94 (0.13)	15.0 (0.0)	1.00 (0.00)	0.99 (0.09)	2501.9 (1255.3)
				CR	0.98 (0.08)	0.65 (0.25)	14.3 (1.0)	0.95 (0.22)	0.93 (0.26)	1901.1 (1110.0)
				CR-SS	1.00 (0.01)	0.87 (0.16)	14.8 (0.6)	1.00 (0.04)	0.95 (0.21)	1982.2 (1209.3)
				CR-R	1.00 (0.01)	0.85 (0.20)	14.5 (2.1)	1.00 (0.04)	0.99 (0.11)	2680.5 (2676.4)
				CR-RSS	1.00 (0.00)	0.95 (0.11)	13.4 (2.2)	1.00 (0.00)	0.95 (0.22)	1472.8 (1908.0)

Table S3: Simulation results when variables are correlated with $n = 100$, $s = 5$ and $c = 3$. $\mathcal{S}$ is the set of signal variables, $\mathcal{C}$ is the set of noise variables correlated (in population) with signal variables, $\hat{\mathcal{S}}$ is the comprehensive model obtained after all reduction rounds, $\mathcal{M}$ is the set of models that are significant at level ϑ. Empirical standard errors are given in brackets.

s	c	ρ	$\frac{\text{signal}}{\text{noise}}$	Reduction Strategy	$\mathbb{E}\left(\frac{\lvert\mathcal{S}\cap\hat{\mathcal{S}}\rvert}{\lvert\mathcal{S}\rvert}\right)$	$\mathbb{E}\left(\frac{\lvert\mathcal{C}\cap\hat{\mathcal{S}}\rvert}{\lvert\mathcal{C}\rvert}\right)$	$\mathbb{E}\lvert\hat{\mathcal{S}}\rvert$	$\mathbb{P}(\mathcal{S}\subseteq\hat{\mathcal{S}})$	$\mathbb{P}(\mathcal{S}\in\mathcal{M})$ $(\vartheta=0.01)$	$\mathbb{E}(\lvert\mathcal{M}\backslash\mathcal{S}\rvert)$ $(\vartheta=0.01)$
5	3	0.9	1	LASSO	0.98 (0.06)	0.40 (0.28)	13.8 (1.3)	0.92 (0.28)	0.91 (0.29)	1033.1 (720.3)
				MS	1.00 (0.00)	1.00 (0.00)	15.0 (0.0)	1.00 (0.00)	0.99 (0.11)	1707.3 (964.7)
				CR	1.00 (0.00)	0.99 (0.05)	14.4 (0.9)	1.00 (0.00)	0.99 (0.12)	1455.6 (897.3)
				CR-SS	1.00 (0.00)	1.00 (0.03)	14.8 (0.5)	1.00 (0.00)	0.99 (0.11)	1594.3 (928.1)
				CR-R	1.00 (0.00)	1.00 (0.00)	13.2 (1.8)	1.00 (0.00)	0.99 (0.11)	1058.9 (991.4)
				CR-RSS	1.00 (0.00)	1.00 (0.00)	11.7 (1.8)	1.00 (0.00)	0.98 (0.13)	594.4 (608.1)
5	3	0.9	0.6	LASSO	0.87 (0.14)	0.37 (0.28)	13.7 (1.3)	0.47 (0.5)	0.47 (0.50)	2200 (1163.3)
				MS	1.00 (0.00)	1.00 (0.00)	15.0 (0.0)	1.00 (0.00)	0.99 (0.10)	3665.1 (1050.6)
				CR	0.99 (0.05)	0.94 (0.13)	14.4 (0.9)	0.93 (0.26)	0.92 (0.27)	3034.9 (1227.0)
				CR-SS	1.00 (0.03)	0.97 (0.10)	14.8 (0.5)	0.98 (0.13)	0.97 (0.18)	3300.3 (1144.9)
				CR-R	1.00 (0.00)	1.00 (0.00)	13.8 (1.8)	1.00 (0.00)	0.99 (0.08)	2831.6 (2213.0)
				CR-RSS	1.00 (0.00)	1.00 (0.00)	12.2 (1.9)	1.00 (0.00)	0.98 (0.13)	1483.4 (1261.2)
5	3	0.5	1	LASSO	1.00 (0.00)	0.26 (0.25)	13.9 (1.2)	1.00 (0.00)	0.99 (0.09)	60.7 (87.2)
				MS	1.00 (0.00)	1.00 (0.03)	15.0 (0.0)	1.00 (0.00)	0.99 (0.10)	88.9 (131.0)
				CR	1.00 (0.01)	0.91 (0.17)	14.3 (0.9)	1.00 (0.04)	0.99 (0.09)	80.2 (139.4)
				CR-SS	1.00 (0.00)	0.98 (0.08)	14.8 (0.5)	1.00 (0.00)	0.98 (0.13)	80.1 (120.4)
				CR-R	1.00 (0.00)	0.99 (0.05)	13.8 (1.9)	1.00 (0.00)	0.99 (0.09)	69.7 (115.1)
				CR-RSS	1.00 (0.00)	1.00 (0.03)	12.8 (2.0)	1.00 (0.00)	0.98 (0.13)	46.9 (78.0)
5	3	0.5	0.6	LASSO	0.99 (0.04)	0.27 (0.25)	13.8 (1.2)	0.97 (0.17)	0.96 (0.19)	587.8 (539.1)
				MS	1.00 (0.00)	0.99 (0.05)	15.0 (0.0)	1.00 (0.00)	0.99 (0.09)	953.4 (768.8)
				CR	0.99 (0.04)	0.87 (0.20)	14.4 (1.0)	0.96 (0.19)	0.95 (0.22)	769.8 (663.9)
				CR-SS	1.00 (0.01)	0.96 (0.12)	14.8 (0.5)	1.00 (0.04)	0.96 (0.19)	800.6 (714.3)
				CR-R	1.00 (0.00)	0.97 (0.10)	14.0 (1.8)	1.00 (0.00)	0.99 (0.10)	789.1 (927.3)
				CR-RSS	1.00 (0.00)	1.00 (0.04)	13.0 (2.1)	1.00 (0.00)	0.98 (0.15)	527.7 (1017.5)

Table S4: Simulation results when variables are correlated with $n = 100$, $s = 5$ and $c = 5$. $\mathcal{S}$ is the set of signal variables, $\mathcal{C}$ is the set of noise variables correlated (in population) with signal variables, $\hat{\mathcal{S}}$ is the comprehensive model obtained after all reduction rounds, $\mathcal{M}$ is the set of models that are significant at level ϑ. Empirical standard errors are given in brackets.

s	c	ρ	$\frac{\text{signal}}{\text{noise}}$	Reduction Strategy	$\mathbb{E}\left(\frac{|\mathcal{S}\cap\hat{\mathcal{S}}|}{|\mathcal{S}|}\right)$	$\mathbb{E}\left(\frac{|\mathcal{C}\cap\hat{\mathcal{S}}|}{|\mathcal{C}|}\right)$	$\mathbb{E}|\hat{\mathcal{S}}|$	$\mathbb{P}(\mathcal{S}\subseteq\hat{\mathcal{S}})$	$\mathbb{P}(\mathcal{S}\in\mathcal{M})$ $(\vartheta=0.01)$	$\mathbb{E}(|\mathcal{M}\backslash\mathcal{S}|)$ $(\vartheta=0.01)$
5	5	0.9	1	LASSO	0.98 (0.06)	0.36 (0.22)	13.7 (1.3)	0.89 (0.32)	0.88 (0.33)	1116.0 (805.7)
				MS	1.00 (0.00)	1.00 (0.00)	15.0 (0.0)	1.00 (0.00)	0.99 (0.09)	2050.1 (1132.2)
				CR	1.00 (0.03)	0.97 (0.08)	14.4 (0.8)	0.98 (0.13)	0.98 (0.15)	1736.2 (1039.4)
				CR-SS	1.00 (0.01)	0.99 (0.05)	14.8 (0.4)	1.00 (0.04)	0.99 (0.12)	1930.4 (1116.6)
				CR-R	1.00 (0.00)	1.00 (0.00)	13.9 (1.6)	1.00 (0.00)	0.99 (0.10)	1675.6 (1551.9)
				CR-RSS	1.00 (0.00)	1.00 (0.00)	13.0 (1.5)	1.00 (0.00)	0.99 (0.12)	1134.9 (1047.2)
5	5	0.9	0.6	LASSO	0.84 (0.14)	0.35 (0.22)	13.8 (1.3)	0.37 (0.48)	0.35 (0.48)	2292.1 (1198.4)
				MS	1.00 (0.00)	1.00 (0.00)	15.0 (0.0)	1.00 (0.00)	0.99 (0.09)	3991.9 (1016.1)
				CR	0.98 (0.07)	0.92 (0.12)	14.5 (0.8)	0.89 (0.31)	0.89 (0.32)	3427.4 (1222.9)
				CR-SS	0.99 (0.04)	0.95 (0.09)	14.9 (0.4)	0.95 (0.21)	0.94 (0.24)	3670.9 (1138.6)
				CR-R	1.00 (0.00)	1.00 (0.00)	14.5 (1.7)	1.00 (0.00)	0.99 (0.10)	3820.8 (2746.1)
				CR-RSS	1.00 (0.00)	1.00 (0.00)	13.5 (1.7)	1.00 (0.00)	0.99 (0.12)	2717.6 (2109.1)
5	5	0.5	1	LASSO	1.00 (0.00)	0.23 (0.19)	13.8 (1.3)	1.00 (0.00)	0.99 (0.10)	58.8 (92.3)
				MS	1.00 (0.00)	1.00 (0.03)	15.0 (0.0)	1.00 (0.00)	0.99 (0.12)	90.5 (115.5)
				CR	1.00 (0.02)	0.88 (0.16)	14.4 (0.9)	0.99 (0.09)	0.98 (0.15)	76.0 (108.6)
				CR-SS	1.00 (0.01)	0.97 (0.08)	14.9 (0.4)	1.00 (0.06)	0.98 (0.15)	79.8 (97.2)
				CR-R	1.00 (0.00)	0.99 (0.06)	14.7 (1.7)	1.00 (0.00)	0.99 (0.12)	89.2 (119.5)
				CR-RSS	1.00 (0.00)	0.99 (0.03)	14.3 (1.9)	1.00 (0.00)	0.98 (0.15)	75.7 (163.1)
5	5	0.5	0.6	LASSO	0.99 (0.05)	0.24 (0.20)	13.8 (1.3)	0.95 (0.21)	0.95 (0.23)	582.8 (526.2)
				MS	1.00 (0.00)	0.99 (0.05)	15.0 (0.0)	1.00 (0.00)	0.99 (0.08)	1051.2 (822.0)
				CR	0.98 (0.06)	0.81 (0.19)	14.4 (0.9)	0.92 (0.27)	0.91 (0.29)	856.8 (743.8)
				CR-SS	1.00 (0.01)	0.94 (0.10)	14.8 (0.4)	1.00 (0.06)	0.97 (0.18)	889.3 (747.0)
				CR-R	1.00 (0.01)	0.97 (0.08)	14.9 (1.7)	1.00 (0.04)	0.99 (0.10)	1159.4 (1189.3)
				CR-RSS	1.00 (0.00)	0.99 (0.05)	14.5 (1.9)	1.00 (0.00)	0.97 (0.16)	935.1 (1167.4)

of the main paper. We probe this further in the simulations in Section 8 of the main paper. With this caveat, the randomised versions of Cox Reduction appear to perform best, retaining the true set of signal variables with high probability and including fewer models in the confidence set. Sample splitting performed favourably. It was noted from a subsidiary empirical analysis (omitted) that in the second-round reduction with a fixed choice of significance level, CR-SS retained fewer noise variables than CR while retaining all signal variables with equally high probability. Marginal screening was also slightly more likely to include noise variables correlated with signal variables in the comprehensive model.

To facilitate the evaluation of each reduction strategy's performance across a range of scenarios, 10-15 variables were always included in $\hat{\mathcal{S}}$ and models of size at most five were included in $\mathcal{M}$ based on a fixed significance level ϑ. When the true model consisted of only three variables, more models were included in $\mathcal{M}$ than when the true model consisted of five variables. This is to be expected based on the analysis in Section 4 of the original paper. When correlations among variables were high and signal strengths low, numerous models produced similar fit so that $\mathcal{M}$ contained a larger number of models.

In most cases the number of models in $\mathcal{M}$ is very large. [2] gave a logical argument that can be used to narrow down these conclusions. This can be performed completely independently of Cox reduction and the construction of confidence sets of models, so that there are no difficulties of post-selection inference.

S3. COX REDUCTION AND STABILITY OF PREDICTIONS

We illustrate the ideas outlined in Section 9.5 of the original paper with data generated as in Table S3 using $s = 5$, $c = 3$, a signal to noise ratio of 1 and ρ equal to 0.5 and 0.9. A confidence set of models was obtained using the rerandomised version of Cox reduction with sample splitting and the significance level for the model selection phase was set to $\vartheta = 0.01$. The model confidence set consisted of 10 and 84 models respectively when $\rho = 0.5$ and $\rho = 0.9$, with $\mathcal{S} \in \mathcal{M}$ under both scenarios. Each model was used to generate confidence sets of predictions based on 100 newly generated observations. Figure S2 plots the union of the confidence sets of predictions against the true responses. For each observation, the interval is coloured according to the proportion of confidence sets containing each point.

Observations were generally included or were close to at least one confidence interval. The large region of overlap across models gives reassurance over predictive conclusions. For design matrices with more dependence between covariates, the procedure may suggest numerous models and the prediction intervals will sometimes be appreciably larger than those obtained from a single well-fitting model, reflecting uncertainty over the model.
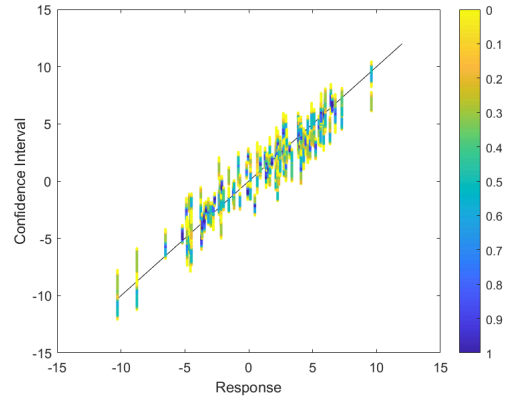
(a) $\rho = 0.5$: $|\mathcal{M}| = 10$
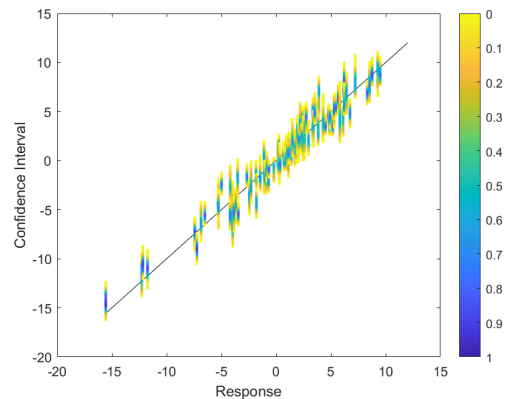
(b) $\rho = 0.9$: $|\mathcal{M}| = 84$

Fig S2: The union of all confidence sets of predictions obtained from the models in the confidence set are plotted against the true responses. Intervals are coloured based on the proportion of confidence sets that contain each point; sections contained in all confidence sets are dark blue. The black line $y = x$ represents perfect prediction.

S4. PROOFS OF MAIN RESULTS

Throughout this section, let $E, F \subset \{1, \ldots, p\}$ be such that $|E \cup F| < n$ with $E \cap F = \emptyset$. These sets will always be used to index columns of X. Define $K = E \cup F$ to be the indices appearing in a given traversal of the hypercube. In a regression of Y on X_K, let

$$\hat{\theta}_{Y:K} = (X_K^T X_K)^{-1} X_K^T Y,$$

and use $\hat{\theta}_{Y:E.F}$ and $\hat{\theta}_{Y:F.E}$ to denote the entries of $\hat{\theta}_{Y:K}$ corresponding to the sets E and F respectively. Let $\hat{\theta}_{F:E} = (X_E^T X_E)^{-1} X_E^T X_F$. Similarly, let $T_{Y:K}$ be the Wald-type statistic obtained from a linear regression of Y on X_K, and denote its components corresponding to E and F by $T_{Y:E.F}$ and $T_{Y:F.E}$ respectively. More pre-

cisely,

$$T_{Y:K} = \sigma^{-1} D_K^{-1/2} \hat{\theta}_{Y:K}$$

where D_K is a diagonal matrix with entries given by the diagonal entries of $(X_K^T X_K)^{-1}$. The notation

$$(X_K^T X_K)^{-1}_{(E,E)}$$

will be used to denote the rows and columns of the matrix $(X_K^T X_K)^{-1}$ corresponding to the set E.

PROOF OF PROPOSITION 1. The log-likelihood ratio test statistic can be expressed as

$$w(\mathcal{S}_m)/2 = \sup_{\eta \in \text{Col-Sp}(\widetilde{X})} \bar{\ell}(\eta) - \sup_{\eta \in \text{Col-Sp}(\widetilde{X}_m)} \bar{\ell}(\eta)$$

$$= \sup_{h_1 \in \text{Col-Sp}(\widetilde{X})} \bar{\ell}(\eta^0 + h_1)$$

$$- \sup_{h_0 \in \{\text{Col-Sp}(\widetilde{X}_m) - \eta^0\}} \bar{\ell}(\eta^0 + h_0)$$

as η^0 is assumed to be in the column span of $\widetilde{X}$ but not necessarily in the column span of $\widetilde{X}_m$. Let $\hat{h}_0 = \hat{\eta}_m - \eta^0$ and $\hat{h}_1 = \hat{\eta} - \eta^0$ be the maximisers. By assumption, $\|\hat{h}_0\|_3^3, \|\hat{h}_1\|_3^3 = O_P(n^{-1/2})$. Define M_n to be a sequence converging to zero as $n \to \infty$ so that $\|\hat{h}_1\|_3/M_n = o_P(1)$ and $\|\hat{h}_0\|_3/M_n = o_P(1)$. For example, $M_n = n^{-1/7}$. Then, with probability converging to one, $\hat{h}_1, \hat{h}_0 \in B_n$, the ℓ_3-ball of radius M_n about zero, and so

$$w(\mathcal{S}_m)/2 = \Delta_1 - \Delta_0 + o_P(1)$$

where

$$\Delta_1 = \sup_{h_1 \in H_n^{(1)}} \bar{\ell}(\eta^0 + h_1) - \bar{\ell}(\eta^0)$$

$$\Delta_0 = \sup_{h_0 \in H_n^{(0)}} \bar{\ell}(\eta^0 + h_0) - \bar{\ell}(\eta^0),$$

$H_n^{(1)} = \text{Col-Sp}(\widetilde{X}) \cap B_n$ and

$$H_n^{(0)} = (\text{Col-Sp}(\widetilde{X}_m) - \eta^0) \cap B_n.$$

Let $U^0 = \nabla_\eta \bar{\ell}(\eta^0)$ and $J^0 = -\nabla_{\eta\eta}^2 \bar{\ell}(\eta^0)$. For $h_1 \in H_n^{(1)}$, $\|h_1\|_3 \leq M_n$ and $h_1 = \widetilde{Q}\beta$ for some $\beta \in \mathbb{R}^{\hat{r}}$ and $\widetilde{Q}$ defined in the statement of the proposition. So, applying the local asymptotic expansion,

$$\Delta_1 = \sup_{\beta : \widetilde{Q}\beta \in H_n^{(1)}} \{\beta^T \widetilde{Q}^T U^0 - \frac{1}{2}\beta^T \widetilde{Q}^T J^0 \widetilde{Q}\beta\}$$

$$+ o_P(1).$$

The supremum is achieved at

$$\beta = (\widetilde{Q}^T J^0 \widetilde{Q})^{-1} \widetilde{Q}^T U^0$$

satisfying

$$\|\widetilde{Q}\beta\|_3^3 \leq \|\widetilde{Q}\beta\|_\infty \|\widetilde{Q}\beta\|_2^2 = O_P(n^{-1/2}) = o_P(M_n^3)$$

because, letting $Z = (\widetilde{Q}^T J^0 \widetilde{Q})^{-1/2} \widetilde{Q}^T U^0$,

$$\|\beta\|_2 = n^{-1/2} \|(\widetilde{Q}^T J^0 \widetilde{Q}/n)^{-1/2} Z\|_2$$

$$\leq n^{-1/2} \|(\widetilde{Q}^T J^0 \widetilde{Q}/n)^{-1/2}\|_2 \|Z\|_2$$

$$= O_P(n^{-1/2})$$

by (c), (d) and $\|\widetilde{Q}\beta\|_\infty \leq \max_{i=1}^n \|\widetilde{q}_i\|_2 \|\beta\|_2$. This maximiser lies in B_n with probability converging to one and so

$$\Delta_1 = \|(\widetilde{Q}^T J^0 \widetilde{Q})^{-1/2} \widetilde{Q}^T U^0\|_2^2/2$$

$$(1) \qquad + o_P(1).$$

For $h_0 \in H_n^{(0)}$, $\|h_0\|_3 \leq M_n$ and $h_0 = \widetilde{Q}_m \alpha + v$ where $v = -(I - P_m)\eta^0$ and $\alpha \in \mathbb{R}^{r_m}$. So,

$$\Delta_0 = \sup_{\alpha : (\widetilde{Q}_m \alpha + v) \in H_n^{(0)}} \{(\widetilde{Q}_m \alpha + v)^T U^0$$

$$- \frac{1}{2}(\widetilde{Q}_m \alpha + v)^T J^0 (\widetilde{Q}_m \alpha + v)\}$$

$$+ o_P(1).$$

By assumption $\|v\|_2 = o(n^{-1/2})$ and

$$n^{-1/2} \|\widetilde{Q}_m \alpha\|_2 \leq n^{-1/2} \|h_0\|_2 \leq n^{-1/3} \|h_0\|_3 = o(1).$$

So,

$$|v^T U^0| \leq \sqrt{n} \|v\|_2 \|n^{-1/2} U_0\|_2 = o_P(1)$$

$$|v^T J^0 \widetilde{Q}_m \alpha| = \|\sqrt{n}v\|_2 \|n^{-1/2} \widetilde{Q}_m \alpha\|_2 \|J_0\|_2 = o_P(1)$$

$$|v^T J^0 v| \leq \|n^{-1} J^0\|_2 \|\sqrt{n}v\|_2^2 = o_P(1).$$

Thus,

$$\Delta_0 = \sup_{\alpha : (\widetilde{Q}_m \alpha + v) \in H_n^{(0)}} \{(\widetilde{Q}_m \alpha)^T U^0 - \frac{1}{2}(\widetilde{Q}_m \alpha)^T J^0 (\widetilde{Q}_m \alpha)\}$$

$$+ o_P(1)$$

and maximising this expression gives

$$\Delta_0 = \|(\widetilde{Q}_m^T J^0 \widetilde{Q}_m)^{-1/2} \widetilde{Q}_m^T U^0\|_2^2/2$$

$$(2) \qquad + o_P(1).$$

Combining equations (1) and (2)

$$w(\mathcal{S}_m) = \|(\widetilde{Q}^T J^0 \widetilde{Q})^{-1/2} \widetilde{Q}^T U^0\|_2^2$$

$$- \|(\widetilde{Q}_m^T J^0 \widetilde{Q}_m)^{-1/2} \widetilde{Q}_m^T U^0\|_2^2 + o_P(1)$$

$$= (\Sigma^{-1/2} Z)^T \Gamma \Sigma^{-1/2} Z + o_P(1)$$

where $Z = n^{-1/2} \widetilde{Q}^T U^0$, $n\Sigma = \widetilde{Q}^T J^0 \widetilde{Q}$, $n\Sigma_{mm} = \widetilde{Q}_m^T J^0 \widetilde{Q}_m$ and

$$\Gamma = \mathbb{I} - \Sigma^{1/2} \begin{pmatrix} \Sigma_{mm}^{-1} & 0 \\ 0 & 0 \end{pmatrix} \Sigma^{1/2}.$$

The matrix Γ is positive semi-definite of rank $\hat{r} - r_m$ with eigenvalues equal to one or zero. This is because

$$\lambda\left\{\Sigma^{1/2}\begin{pmatrix}\Sigma_{mm}^{-1} & 0\\ 0 & 0\end{pmatrix}\Sigma^{1/2}\right\} = \lambda\left\{\begin{pmatrix}\Sigma_{mm}^{-1} & 0\\ 0 & 0\end{pmatrix}\Sigma\right\}$$

where for any square matrix M, $\lambda(M)$ is the set of eigenvalues of M. As Σ_{mm} are the entries of Σ in the first r_m rows and r_m columns, this set consists of the eigenvalues of an upper triangular matrix with r_m ones and $\hat{r} - r_m$ zeroes on the diagonal. Thus, the set of eigenvalues of Γ consists of r_m zeroes and $\hat{r} - r_m$ ones. Write $\Gamma = P^T \Lambda P$ where P is an orthonormal matrix and Λ is a diagonal matrix with $\hat{r} - r_m$ ones and r_m zeroes on the diagonal. Then,

$$w(\mathcal{S}_m) = (P\Sigma^{-1/2}Z)^T\Lambda(P\Sigma^{-1/2}Z) + o_P(1)$$
$$= \sum_{i=1}^{\hat{r}-r_m}(P\Sigma^{-1/2}Z)_i^2 + o_P(1)$$

and by assumption, $P\Sigma^{-1/2}Z \xrightarrow{d} N(0, \mathbb{I})$. As asymptotically $w(\mathcal{S}_m)$ is the sum of the squares of $\hat{r} - r_m$ independent standard normal random variables

$$w(\mathcal{S}_m) \xrightarrow{d} \chi^2_{\hat{r}-r_m}$$

and so the result follows. $\qquad\square$

LEMMA S1.
$$\hat{\theta}_{Y:E} = \hat{\theta}_{Y:E.F} + (\hat{\theta}_{F:E})^T\hat{\theta}_{Y:F.E}.$$

PROOF. Let $K = E \cup F$. By definition,

$$\hat{\theta}_{Y:E} = (X_E^TX_E)^{-1}X_E^TP_KY$$
$$+(X_E^TX_E)^{-1}X_E^T(I - P_K)Y$$
$$= (X_E^TX_E)^{-1}X_E^TP_KY$$
$$= (X_E^TX_E)^{-1}X_E^TX_K\hat{\theta}_{Y:K}$$
$$= (I, (X_E^TX_E)^{-1}X_E^TX_F)\hat{\theta}_{Y:K}$$
$$= \hat{\theta}_{Y:E.F} + (\hat{\theta}_{F:E})^T\hat{\theta}_{Y:F.E}$$

where we have assumed without loss of generality that E indexes the first few columns of X_K. Further,

$$\hat{\theta}_{Y:E} = \hat{\theta}_{Y:E.F} \iff (\hat{\theta}_{F:E})^T\hat{\theta}_{Y:F.E} = 0.$$
$$\square$$

LEMMA S2. *Suppose $|E| = 1$. Then,*
$$T_{Y:E.F} = \sigma^{-1}\|Y\|_2R(Y, (I - P_F)x_E).$$

PROOF. By Lemma S4 and the definitions of marginal and multiple sample correlation coefficients,

$$\hat{\theta}_{Y:E.F} = \{x_E^T(I - P_F)x_E\}^{-1}x_E^T(I - P_F)Y$$
$$= \frac{R(Y, (I - P_F)x_E)\|Y\|_2}{\|x_E\|\sqrt{1 - R^2(x_E, X_F)}}.$$

By writing the inverse of a matrix in terms of a Schur complement,

$$(X_K^TX_K)_{(E,E)}^{-1} = (x_E^T(I - P_F)x_E)^{-1}$$
$$= \|x_E\|_2^{-2}(1 - R^2(x_E, X_F))^{-1}.$$

Combining these equations gives the desired result. $\qquad\square$

PROOF OF PROPOSITION 2. Let $\epsilon = M\log(|\mathcal{B}|)$ with $M > 0$. By a union bound,

$$\mathbb{P}\{n\Delta(1) > \epsilon\} \leq \mathbb{P}\left\{\max_{B\subset\mathcal{B},|B|=k}nR^2(Y, X_B) > \epsilon\right\}$$
$$+\mathbb{P}\left\{\max_{B\subset\mathcal{B},|B|=k-1}nR^2(Y, X_B) > \epsilon\right\}$$
$$+|\mathcal{A}|\max_{A\subseteq\mathcal{A},|A|=1}\mathbb{P}\left\{\max_{B\subset\mathcal{B},|B|=k-1}nR^2(X_A, X_B) > \epsilon\right\}$$

and so it is sufficient to consider the limiting distribution of the maximum spurious correlation between a vector and an independent random matrix consisting of k or $k - 1$ columns. We focus on the term $R^2(X_A, X_B)$ although the results follow identically for $R^2(Y, X_B)$. Under the assumptions in the proposition, Theorem 3.1 in [3] states that there exists a constant $C > 0$ independent of $|\mathcal{B}|$, k and n such that

$$\sup_{\epsilon\geq 0}|\mathbb{P}\{n\max_{B\subset\mathcal{B},|B|=k-1}R^2(X_A, X_B) > \epsilon\} - \mathbb{P}\{\widetilde{R}^2 > \epsilon\}|$$
$$\leq Cn^{-1/8}[\max\{\log(|\mathcal{B}|), \log n\}]^{7/8}$$

where

$$\widetilde{R}^2 = Z_{(|\mathcal{B}|)}^2 + \cdots + Z_{(|\mathcal{B}|-k+2)}^2$$

with $Z \sim N_{|\mathcal{B}|}(0, \mathbb{I})$ and

$$Z_{(1)}^2 \leq Z_{(2)}^2 \leq \cdots \leq Z_{(|\mathcal{B}|)}^2$$

are the order statistics of $\{Z_j^2\}_{j=1}^{|\mathcal{B}|}$. The aim is to obtain the rate at which $\mathbb{P}\{\widetilde{R}^2 > \epsilon\}$ converges to zero. Let $M_n = \max_{i=1,\ldots|\mathcal{B}|}Z_i$ and $L_n = \max_{i=1,\ldots|\mathcal{B}|}\{-Z_i\}$. As a maximum of i.i.d. standard Gaussian random variables, [4] shows that $a_n^{-1}(M_n - b_n)$ converges in distribution to a Gumbel distribution where $a_n = b_n^{-1}$ and

$$b_n = \sqrt{2\log(|\mathcal{B}|)} - \frac{\log\log(|\mathcal{B}|) + \log(4\pi)}{2\sqrt{2\log(|\mathcal{B}|)}}$$
$$+O\{(\log|\mathcal{B}|)^{-1}\}.$$

A similar result holds for L_n. The rate of convergence to a Gumbel distribution is $\{\log(|\mathcal{B}|)\}^{-1}$, and so

$$\mathbb{P}(\widetilde{R}^2 > \epsilon) \leq \mathbb{P}\{(k-1)Z_{(|\mathcal{B}|)}^2 > \epsilon\}$$
$$\leq \mathbb{P}\{(k-1)^{1/2}M_n > \sqrt{\epsilon}\}$$
$$+\mathbb{P}\{(k-1)^{1/2}L_n > \sqrt{\epsilon}\}$$
$$(3) \qquad \leq 2\left\{1 - \exp\left(-e^{-f_n}\right) + \eta_n\right\}$$

where, for $\epsilon = M \log(|\mathcal{B}|)$ with $M > 0$ large enough,

$$\max\{f_n^{-1}, \eta_n\} = O[\{\log(|\mathcal{B}|)\}^{-1}].$$

As $e^x \geq 1 + x$ for all x, we have

$$1 - \exp\left(-e^{-f_n}\right) \leq 1 - \exp\left\{\frac{-1}{f_n + 1}\right\}$$
$$\leq (1 + f_n)^{-1},$$

and so

$$\mathbb{P}\left\{n \max_{\substack{B \subset \mathcal{B}, \\ |B| = k-1}} R^2(X_A, X_B) > \epsilon\right\}$$
$$= O\left(\max\left\{\epsilon_n, \log(|\mathcal{B}|)^{-1}\right\}\right)$$

where $\epsilon_n = n^{-1/8} \max\{\log(|\mathcal{B}|), \log n\}^{7/8}$. Using similar arguments,

$$\mathbb{P}(n\Delta(1) > \epsilon) = O\left(|\mathcal{A}| \max\left\{\epsilon_n, \log(|\mathcal{B}|)^{-1}\right\}\right).$$

When $|\mathcal{A}| = o\{\log(|\mathcal{B}|)\}$ and $|\mathcal{A}|^{8/7} \log\{|\mathcal{B}|n\} = o(n^{1/7})$, all terms on the right hand side converge to zero and so

$$\Delta(1) = O_P(\log(|\mathcal{B}|)/n).$$

The result follows when $\log(|\mathcal{B}|)/n = o(1)$. $\qquad\square$

PROOF OF PROPOSITION 3. The result follows directly from Lemmas S7 and S8 in Section S5. $\qquad\square$

PROOF OF PROPOSITION 4. We can condition on the event $\mathcal{G}$ that each variable in $\mathcal{A}$ appears unaccompanied by other pseudo-signal variables in at least 2/3 regressions in which it appears. A lower bound on the probability that this event occurs is given in Lemma S9. On the event $\mathcal{G}$, every pseudo-signal variable will survive the first round of reduction if

$$(4) \quad \min_{A \subseteq \mathcal{A}, |A|=1} \min_{B \in \mathcal{B}_k} \{|T_{Y:A.B}| - \|T_{Y:B.A}\|_\infty\} > 0$$

where $\mathcal{B}_k = \{B \subseteq \mathcal{B} : |B| = k - 1\}$. By Proposition 3, Lemma S2 and Lemma S8,

$$T_{Y:A.B} = \sigma^{-1} \|Y\|_2 R(Y, x_A) + \delta_{A.B}$$
$$T_{Y:B.A} = \delta_{B.A}^{(1)} + \delta_{B.A}^{(2)}$$

where provided $R(X_A, X_B) < 1 - c$, there exists $C > 0$ depending only on c such that,

$$|\delta_{A.B}| \leq C\sigma^{-1} \|Y\|_2 \{R(X_A, X_B) + R(Y, X_B)\}$$
$$\|\delta_{B.A}^{(1)}\|_\infty \leq \sigma^{-1} \|Y\|_2 \{R(X_A, X_B) + R(Y, X_B)\}$$
$$\|\delta_{B.A}^{(2)}\|_\infty \leq C\sigma^{-1} \|\widetilde{Y}\|_2 R(X_A, X_B).$$

Thus, (4) occurs if for all $A \subseteq \mathcal{A}$ with $|A| = 1$, $B \in \mathcal{B}_k$,

$$\sigma^{-1} \|Y\|_2 |R(Y, x_A)| - |\delta_{A.B}|$$
$$(5) \qquad\qquad > \|\delta_{B.A}^{(1)}\|_\infty + \|\delta_{B.A}^{(2)}\|_\infty$$

By assumption, there exists $\epsilon > 0$ such that for n large enough,

$$\min_{A \subset \mathcal{A}, |A|=1} |R(Y, x_A)| > \sigma\epsilon$$

and

$$\max_{A \subset \mathcal{A}, |A|=1, B \in \mathcal{B}_k} \max\{|\delta_{A.B}|, \|\delta_{B.A}^{(1)}\|_\infty, \|\delta_{B.A}^{(2)}\|_\infty\} \leq \epsilon\|Y\|_2/4.$$

Then, condition (5) holds and the result follows. $\qquad\square$

PROOF OF PROPOSITION 5. First consider the $k \times k$ square. Let $\mathcal{G}_{a,a_0}$ be the event that indices a and a_0 share a row or column. The expected number of indices from $\mathcal{A}_{-a}$ sharing a row or column with a is given by

$$\mathbb{E}\left[\sum_{a_0 \in \mathcal{A}_{-a}} \mathbb{1}\{\mathcal{G}_{a,a_0}\}\right] = \frac{2(|\mathcal{A}| - 1)(k - 1)}{(k^2 - 1)}.$$

This follows because conditional on the location of a in the square, there are $2(k - 1)$ out of a total of $k^2 - 1$ locations for a_0 that ensure that a and a_0 share a row or column.

The arguments for the $k \times k \times k$ cube are similar. Now let $\mathcal{G}_{a,a_0}$ be the event that indices a and a_0 share a row, column or corridor. The expected number of indices from $\mathcal{A}_{-a}$ sharing a row, column or corridor with a is given by

$$\mathbb{E}\left[\sum_{a_0 \in \mathcal{A}_{-a}} \mathbb{1}\{\mathcal{G}_{a,a_0}\}\right] = \frac{3(|\mathcal{A}| - 1)(k - 1)}{(k^3 - 1)}$$

as conditional on the location of a in the cube, there are $3(k - 1)/(k^3 - 1)$ locations for a_0 to ensure that a and a_0 share a row, column or corridor. $\qquad\square$

PROOF OF PROPOSITION 6. By assumption, there exists $N_1, C_1 > 0$ such that for all $n \geq N_1$,

$$\Delta(k) \leq 1/2$$

$$\max_{a \in \mathcal{A}_1} \max_{A \subseteq \mathcal{A} \setminus \{a\}, |A| \leq k-1} \{|R(y, (I - P_A)x_a)|\}^{-1} \leq C_1^{-1}.$$

Consider a regression of Y on the covariates indexed by $A \cup B$ where $A \subseteq \mathcal{A}$, $B \subseteq \mathcal{B}$ and $A \cap \mathcal{A}_1 \neq \emptyset$. By Lemma S2 and Proposition 3, when $n \geq N_1$, the entry of $|T_{Y:A.B}|$ corresponding to $a \in A \cap \mathcal{A}_1$ is given by

$$\sigma^{-1} \|Y\|_2 R(Y, (I - P_{A_{-a}})x_a) + \delta_1$$

where there exists a universal constant $C_2 > 0$ such that

$$|\delta_1| \leq C_2 \sigma^{-1} \|Y\|_2 \Delta^{1/2}(k).$$

In particular, for $n \geq N_1$, this entry is at least as large in magnitude as

$$(6) \qquad \sigma^{-1} \|Y\|_2 \{C_1 - C_2 \Delta^{1/2}(k)\}.$$

In contrast, consider an arbitrary regression in Cox reduction that includes a pseudo-noise variable $b \in \mathcal{B}$. The entry of its corresponding Wald statistic is bounded above in magnitude by

$$(7) \qquad \sigma^{-1} \|y\|_2 (2 + C_2) \Delta^{1/2}(k),$$

when $n \geq N_1$, where we have used the second equation in Proposition 3. In light of (6) and (7), there exists α such that the second round of Cox reduction retains those indices $a \in \mathcal{A}_1$ but not those in $\mathcal{B}$ when

$$(8) \qquad (2 + C_2) \Delta^{1/2}(k) < \{C_1 - C_2 \Delta^{1/2}(k)\}.$$

By assumption, there exists N_2 such that whenever $n \geq N_2$,

$$\Delta^{1/2}(k) < \frac{C_1}{2(1 + C_2)}.$$

Choosing $N = \max\{N_1, N_2\}$, it follows that (8) holds for $n \geq N$ and so the result follows. $\qquad \square$

LEMMA S3. *Suppose $|E| = 1$. Then*

$$\sigma T_{Y:E.F} = \Delta_1 + \Delta_2$$

where

$$\Delta_1 = \sqrt{1 - R^2(x_E, X_F)} \, \theta_E^0 \|x_E\|_2$$
$$\Delta_2 = \|Y - X_K \theta_K^0\|_2 R(Y - X_K \theta_K^0, (I - P_F) x_E)$$

where $K = E \cup F$.

PROOF. By definition,

$$(X_K^T X_K)^{-1} X_K^T Y = \theta_K^0 + (X_K^T X_K)^{-1} X_K^T (Y - X_K \theta_K^0).$$

Further,

$$(X_K^T X_K)^{-1}_{(E,E)} = \|x_E\|_2^{-2} (1 - R^2(x_E, X_F))^{-1},$$

so

$$\sigma T_{Y:E.F} = \sqrt{1 - R^2(x_E, X_F)} \, \theta_E^0 \|x_E\|_2$$
$$+ \sigma T_{(Y - X_K \theta_K^0):E.F}.$$

The result follows by applying Lemma S2 to the second term. $\qquad \square$

S5. PROOFS OF ADDITIONAL RESULTS

The notation in this section follows the notation in Section S4.

LEMMA S4.

$$\hat{\theta}_{Y:E.F} = \{X_E^T (I - P_F) X_E\}^{-1} X_E^T (I - P_F) Y$$
$$\hat{\theta}_{Y:F.E} = \{X_F^T (I - P_E) X_F\}^{-1} X_F^T (I - P_E) Y.$$

PROOF. By definition

$$\begin{pmatrix} X_E^T X_E & X_E^T X_F \\ X_F^T X_E & X_F^T X_F \end{pmatrix} \begin{pmatrix} \hat{\theta}_{Y:E.F} \\ \hat{\theta}_{Y:F.E} \end{pmatrix} = \begin{pmatrix} X_E^T Y \\ X_F^T Y \end{pmatrix}.$$

Using the second row, $\hat{\theta}_{Y:F.E} = (X_F^T X_F)^{-1} X_F^T (Y - X_E \hat{\theta}_{Y:E.F})$. Substituting this expression into the first row

$$X_E^T \{I - X_F (X_F^T X_F)^{-1} X_F^T\} X_E \hat{\theta}_{Y:E.F}$$
$$= X_E^T \{I - X_F (X_F^T X_F)^{-1} X_F^T\} Y.$$

Similar arguments can be used to derive an analogous result for $\hat{\theta}_{Y:F.E}$. $\qquad \square$

LEMMA S5. *Suppose $X_K = (X_E, X_F) \in \mathbb{R}^{n \times k}$ is a matrix of full rank. Then,*

$$\|P_E P_F\|_2 = \|P_F P_E\|_2 < 1.$$

PROOF. Define

$$R(X_E, X_F) = \max_{\alpha \in \mathbb{R}^{|E|}} \max_{\beta \in \mathbb{R}^{|F|}} \frac{\alpha^T X_E^T X_F \beta}{\|X_E \alpha\|_2 \|X_F \beta\|_2}.$$

The term $R(X_E, X_F)$ is non-negative by definition. Further, by the Cauchy-Schwarz inequality, $R(X_E, X_F) \leq 1$ with equality if and only if $X_E \alpha$ and $X_F \beta$ are linearly dependent. Thus, when X is full-rank $0 \leq R(X_E, X_F) < 1$. As $R(X_E, X_F) = R(X_F, X_E)$, it is sufficient to show that

$$\|P_E P_F\|_2 = R(X_E, X_F).$$

First note that for a given α,

$$\max_{\beta \in \mathbb{R}^{|F|}} \frac{\alpha^T X_E^T X_F \beta}{\|X_E \alpha\|_2 \|X_F \beta\|_2} = R(X_E \alpha, X_F) = \frac{\|P_F X_E \alpha\|_2}{\|X_E \alpha\|_2}.$$

Indeed, following the arguments in [5, p. 164-165],

$$\max_{\beta \in \mathbb{R}^{|F|}} \frac{\alpha^T X_E^T X_F \beta}{\|X_E \alpha\|_2 \|X_F \beta\|_2}$$
$$= \max_{\beta \in \mathbb{R}^{|F|}} \frac{\alpha^T X_E^T X_F (X_F^T X_F)^{-1/2} (X_F^T X_F)^{1/2} \beta}{\|X_E \alpha\|_2 \|X_F \beta\|_2}$$
$$\leq \max_{\beta \in \mathbb{R}^{|F|}} \frac{\|(X_F^T X_F)^{-1/2} X_F^T X_E \alpha\|_2 \|(X_F^T X_F)^{1/2} \beta\|_2}{\|X_E \alpha\|_2 \|X_F \beta\|_2}$$
$$= \frac{\|P_F X_E \alpha\|_2}{\|X_E \alpha\|_2}$$

where the Cauchy-Schwarz inequality was used in the second line. Then,

$$R^2(X_E, X_F) = \max_{\alpha \in \mathbb{R}^{|E|}} \frac{\|P_F X_E \alpha\|_2^2}{\|X_E \alpha\|_2^2}$$
$$= \max_{u \in \mathbb{R}^{|E|}} \frac{\|P_F X_E (X_E^T X_E)^{-1/2} u\|_2^2}{\|u\|_2^2}$$
$$= \lambda_{\max} \{R_E P_F R_E^T\}.$$

where $R_E = (X_E^T X_E)^{-1/2} X_E^T$. For any $m, n \in \mathbb{N} \backslash \{0\}$ and two matrices $M \in \mathbb{R}^{m \times n}$ and $N \in \mathbb{R}^{n \times m}$, the non-zero eigenvalues of MN are equal to the non-zero eigenvalues of NM. Then, as the eigenvalues of $R_E P_F R_E^T$ are non-negative,

$$R^2(X_E, X_F) = \lambda_{\max}(P_F P_E)$$
$$= \lambda_{\max}(P_F P_E P_F)$$
$$= \|P_E P_F\|_2^2.$$

$\square$

LEMMA S6. *For $e \in E$, define Q_e to be the projection matrix onto $(I - P_F) X_{E_{-e}}$. Then,*

$$(I - P_F - Q_e) = (I - P_F - \Delta_e)(I - P_{E_{-e}})$$

where

$$\max_{e \in E} \|\Delta_e\|_2 \le (1 + c^{-1}) R(X_E, X_F)$$

when $R(X_E, X_F) \le 1 - c$ for some $c > 0$.

PROOF. Expressing Q_e in terms of $(I - P_F) X_{E_{-e}}$ and factorising shows that

$$Q_e = (I - P_F) X_{E_{-e}} H^{-1/2} M^{-1} H^{-1/2} X_{E_{-e}}^T (I - P_F)$$

where $H = X_{E_{-e}}^T X_{E_{-e}}$ and

$$M = I - H^{-1/2} X_{E_{-e}}^T P_F X_{E_{-e}} H^{-1/2}.$$

As

$$\|H^{-1/2} X_{E_{-e}}^T P_F X_{E_{-e}} H^{-1/2}\|_2 \le \|P_{E_{-e}} P_F\|_2 < 1$$

by Lemma S5, we can use the geometric series for matrices to invert M and obtain

$$Q_e = (I - P_F) P_{E_{-e}} (I - P_F)$$
$$+ (I - P_F) \sum_{m=1}^{\infty} (P_{E_{-e}} P_F)^m P_{E_{-e}} (I - P_F)$$
$$= P_{E_{-e}} - P_F$$
$$+ (I - P_{E_{-e}}) \sum_{m=0}^{\infty} (P_F P_{E_{-e}})^m P_F (I - P_{E_{-e}}).$$

Then $(I - P_F - Q_e) = (I - P_F - \Delta)(I - P_{E_{-e}})$ where

$$\Delta_e = (I - P_{E_{-e}}) \sum_{m=1}^{\infty} (P_F P_{E_{-e}})^m P_F - P_{E_{-e}} P_F$$

and when $R(X_E, X_F) \le 1 - c$,

$$\max_{e \in E} \|\Delta_e\|_2 \le \max_{e \in E} \left\{ \frac{\|P_F P_{E_{-e}}\|_2}{1 - \|P_F P_{E_{-e}}\|_2} + \|P_F P_{E_{-e}}\|_2 \right\}$$
$$= \max_{e \in E} \left\{ \frac{R(X_F, X_{E_{-e}})}{1 - R(X_F, X_{E_{-e}})} + R(X_F, X_{E_{-e}}) \right\}$$
$$\le \frac{R(X_F, X_E)}{1 - R(X_F, X_E)} + R(X_F, X_E)$$
$$\le (1 + c^{-1}) R(X_E, X_F).$$

LEMMA S7. *Suppose $R(X_E, X_F) \le 1 - c$ for some $c > 0$. Then, there exists $C > 0$ depending only on c such that*

$$\|T_{Y:E.F} - T_{Y:E}\|_\infty \le C\sigma^{-1}\|Y\|_2$$
$$\times \{R(X_E, X_F) + R(Y, X_F)\}$$
$$\|T_{Y:E.F} - T_{\widetilde{Y}:E}\|_\infty \le C\sigma^{-1}\|\widetilde{Y}\|_2 R(X_E, X_F).$$

where $\widetilde{Y} = (I - P_F) Y$.

PROOF. Without loss of generality, consider the first entry of $T_{Y:E.F}$. Let $j = 1$ and e be the first entry of the set E. By Lemma S4 and further applications of the Schur complement,

$$\hat{\theta}_{Y:E.F} = \{X_E^T (I - P_F) X_E\}^{-1} X_E^T (I - P_F) Y$$
$$(X_K^T X_K)^{-1}_{(EE)} = \{X_E^T (I - P_F) X_E\}^{-1},$$

where $K = E \cup F$. The first row of the matrix $\{X_E^T (I - P_F) X_E\}^{-1}$ is

$$(S^{-1}, -S^{-1} x_e^T (I - P_F) X_{E_{-e}} \{X_{E_{-e}}^T (I - P_F) X_{E_{-e}}\}^{-1}).$$

where Q_e is the projection matrix onto $(I - P_F) X_{E_{-e}}$ and

$$S^{-1} = \{x_e^T (I - P_F - Q_e) x_e\}^{-1}.$$

It can be shown that $I - P_F - Q_e$ is a projection matrix using the fact that $(I - P_F) Q_e = Q_e$. Let

$$u_e = (I - P_F - Q_e) x_e.$$

The first entry of $\hat{\theta}_{Y:E.F}$ is given by $\{u_e^T u_e\}^{-1} u_e^T Y$ and so, the first entry of $\sigma T_{Y:E.F}$ is

$$(9) \quad \{u_e^T u_e\}^{-1/2} u_e^T Y = \{u_e^T u_e\}^{-1/2} u_e^T (I - P_F) Y.$$

The two expressions given in the statement of the lemma arise from re-writing each of the two expressions given in the equality above. Consider the first expression in (9). By Lemma S6,

$$(I - P_F - Q_e) = (I - P_F - \Delta_e)(I - P_{E_{-e}})$$

where for $c' = 1 + c^{-1}$,

$$\max_{e \in E} \|\Delta_e\|_2 \le \frac{R(X_F, X_E)}{1 - R(X_F, X_E)} + R(X_F, X_E)$$
$$\le c' R(X_E, X_F).$$

Then, $u_e = (I - P_F - \Delta_e)(I - P_{E_{-e}}) x_e$ and so,

$$\frac{u_e^T Y}{\{x_e^T (I - P_{E_{-e}}) x_e\}^{1/2}} = \frac{x_e^T (I - P_{E_{-e}}) Y}{\{x_e^T (I - P_{E_{-e}}) x_e\}^{1/2}} + \delta_{e,1}$$
$$(10) \qquad = \sigma e_j^T T_{Y:E} + \delta_{e,1}$$

by the proof of Lemma S2, where

$$\max_{e \in E} |\delta_{e,1}| = \sup_{e \in E} \left| \frac{x_e^T (I - P_{E_{-e}})(P_F + \Delta_e^T)Y}{\{x_e^T(I - P_{E_{-e}})x_e\}^{1/2}} \right|$$

$$\leq \|Y\|_2 \{\max_{e \in E} \|\Delta_e\|_2 + R(Y, X_F)\}$$

$$\leq \|Y\|_2 \left\{ c'R(X_E, X_F) + R(Y, X_F) \right\}$$

As $(I - P_F - \Delta_e)(I - P_{E_{-e}}) = (I - P_F - Q_e)$ is a projection matrix and $(I - P_{E_{-e}})(P_F + \Delta_e) = (P_F + \Delta_e)$, it can be shown that $(P_F + \Delta_e)(I - P_{E_{-e}})$ is also a projection matrix and so

$$u_e^T u_e = x_e^T(I - P_F - Q_e)x_e$$

$$= x_e^T(I - P_F - \Delta_e)(I - P_{E_{-e}})x_e$$

$$(11) \qquad = \|(I - P_{E_{-e}})x_e\|_2^2(1 - \delta_{e,2}^2)$$

where

$$\delta_{e,2} = \frac{\|(P_F + \Delta_e)(I - P_{E_{-e}})x_e\|_2}{\|(I - P_{E_{-e}})x_e\|_2}$$

$$\leq \frac{\|P_F(I - P_{E_{-e}})x_e\|_2}{\|(I - P_{E_{-e}})x_e\|_2} + \frac{\|\Delta_e(I - P_{E_{-e}})x_e\|_2}{\|(I - P_{E_{-e}})x_e\|_2}$$

$$\leq R(X_E, X_F) + \|\Delta_e\|_2$$

as $(I - P_{E_{-e}})x_e$ is in the column span of X_E. So,

$$\max_{e \in E} |\delta_{e,2}| \leq (1 + c')R(X_E, X_F).$$

Combining (9), (10) and (11),

$$\sigma|e_j^T(T_{Y:E.F} - T_{Y:E})| \leq |\sigma e_j^T T_{Y:E}\{(1 - \delta_{e,2}^2)^{-1/2} - 1\}|$$

$$+ |\delta_{e,1}(1 - \delta_{e,2}^2)^{-1/2}|$$

$$\leq C\|Y\|_2\{R(X_E, X_F) + R(Y, X_F)\}$$

for some $C > 0$ depending only on c. This establishes the first expression in the statement of the lemma.

To obtain the second expression, replace (10) by the following,

$$\frac{u_e^T(I - P_F)Y}{\{x_e^T(I - P_{E_{-e}})x_e\}^{1/2}} = \frac{x_e^T(I - P_{E_{-e}})(I - P_F)Y}{\{x_e^T(I - P_{E_{-e}})x_e\}^{1/2}} + \delta_{e,1}$$

$$(12) \qquad = \sigma e_j^T T_{(I-P_F)Y:E} + \delta_{e,1}$$

by the proof of Lemma S2, where

$$\max_{e \in E} |\delta_{e,1}| = \sup_{e \in E} \left| \frac{x_e^T(I - P_{E_{-e}})\Delta_e^T(I - P_F)Y}{\{x_e^T(I - P_{E_{-e}})x_e\}^{1/2}} \right|$$

$$\leq \|(I - P_F)Y\|_2 \max_{e \in E} \|\Delta_e\|_2$$

$$\leq c'\|(I - P_F)Y\|_2 R(X_E, X_F).$$

Combining (9), (11) and (12),

$$\sigma|e_j^T(T_{Y:E.F} - T_{\widetilde{Y}:E})| \leq |\sigma e_j^T T_{\widetilde{Y}:E}\{(1 - \delta_{e,2}^2)^{-1/2} - 1\}|$$

$$+ |\delta_{e,1}(1 - \delta_{e,2}^2)^{-1/2}|$$

$$\leq C\|\widetilde{Y}\|_2 R(X_E, X_F)$$

where $\widetilde{Y} = (I - P_F Y)$. $\qquad \square$

LEMMA S8.

$$\|T_{(I-P_F)Y:E}\|_\infty \leq \sigma^{-1}\|Y\|_2\{R(Y, X_E) + R(X_E, X_F)\}.$$

PROOF. Let $\widetilde{Y} = (I - P_F)Y$, $j = 1$ and suppose e is the first index in E. By Lemma S2,

$$\sigma e_j^T T_{\widetilde{Y}:E} = \|\widetilde{Y}\|_2 R(\widetilde{Y}, (I - P_{E_{-e}})x_e)$$

$$= \frac{\widetilde{Y}^T(I - P_{E_{-e}})x_e}{\{x_e^T(I - P_{E_{-e}})x_e\}^{1/2}}.$$

Let $\alpha = (1, -x_e^T X_{E_{-e}}(X_{E_{-e}}^T X_{E_{-e}})^{-1})^T$. Then,

$$|\sigma e_j^T T_{\widetilde{Y}:E}| = \left| \frac{\widetilde{Y}^T X_E \alpha}{\|X_E \alpha\|_2} \right|$$

$$(13) \qquad \leq \|\widetilde{Y}\|_2 R(\widetilde{Y}, X_E)$$

by definition of the multiple correlation coefficient. Further,

$$\|\widetilde{Y}\|_2 R(\widetilde{Y}, X_E) = \|P_E(I - P_F)Y\|_2$$

$$\leq \|P_E Y\|_2 + \|P_E P_F Y\|_2$$

$$\leq \|Y\|_2\{R(Y, X_E) + R(X_E, X_F)\}$$

by the proof of Lemma S5. The result follows by combining this with (13) and noting that the bound does not depend on j. $\qquad \square$

LEMMA S9. *Suppose the indices $\{1, \ldots, k^3\}$ are randomly arranged in a $k \times k \times k$ cube and let $\mathcal{A} \subseteq \{1, \ldots, k^3\}$. Let $\mathcal{G}$ be the event that every index in $\mathcal{A}$ is unaccompanied by other indices in $\mathcal{A}$ in at least two out of three of the row, column or corridor in which it appears. Then,*

$$\mathbb{P}(\mathcal{G}) \geq 1 - \frac{|\mathcal{A}|(|\mathcal{A}| - 1)(|\mathcal{A}| - 2)(k - 1)^2}{(k^3 - 1)(k^3 - 2)}.$$

PROOF. Consider the probability of the complement event $\mathcal{G}^c$: that there exists an index in $\mathcal{A}$ that appears unaccompanied in at most one regression. For $\mathcal{G}^c$ to occur there must exist an element of $\mathcal{A}$, call it index a, where two of its row, column or corridor contain at least one other element from $\mathcal{A}$. Let a_1 and a_2 be the two elements. Given the position of index a, the probability that a_1 and a share a row, column or corridor is $3(k - 1)/(k^3 - 1)$. Then, the probability that a_2 also shares such a block with a but not with a_1 is $2(k - 1)/(k^3 - 2)$. Finally, there are $\binom{|\mathcal{A}|}{3}$ ways of choosing indices a, a_1 and a_2. Applying a union bound, the probability of $\mathcal{G}^c$ is bounded above by

$$\mathbb{P}(\mathcal{G}^c) \leq \frac{6\binom{|\mathcal{A}|}{3}(k - 1)^2}{(k^3 - 1)(k^3 - 2)}$$

$$= \frac{|\mathcal{A}|(|\mathcal{A}| - 1)(|\mathcal{A}| - 2)(k - 1)^2}{(k^3 - 1)(k^3 - 2)}.$$

□

S6. CONFIDENCE SET OF MODELS IN THE GENERALISED LINEAR MODEL

In the generalised linear model with canonical link function, the log-likelihood is of the form

$$\bar{\ell}(\eta) = \sum_{i=1}^{n} \frac{y_i \eta_i - b(\eta_i)}{\phi} - c(y_i)$$

where ϕ is a known constant, $b : \mathbb{R} \mapsto \mathbb{R}$ is a three-times differentiable function satisfying $b''(x) \geq 0$ for all $x \in \mathbb{R}$ and $c : \mathbb{R} \mapsto \mathbb{R}$. The following result establishes conditions under which the assumptions in Proposition 1 of the main paper hold for this log-likelihood. These conditions are satisfied for the linear and logistic regression models.

PROPOSITION S1. *Let P_m be the projection matrix onto the column span of $\widetilde{X}_m$. Suppose $\|(I - P_m)\eta^0\|_2 = o(n^{-1/2})$ and make the following assumptions:*

(A1) there exists a constant $c_1 > 0$ such that $b'''(x) \leq c_1$ for all $x \in \mathbb{R}$,

(A2) there exist constants $c_2, C_2 > 0$ independent of n such that for all $i = 1, \dots, n$

$$c_2 \leq Var(Y_i) \leq C_2, \qquad \mathbb{E}\{|Y_i - \mathbb{E}(Y_i)|^3\} \leq C_2,$$

(A3) for any $\delta > 0$, there exists a constant $c_3 > 0$ such that $b''(x_i) > c_3$ for all $i = 1, \dots, n$ and any x satisfying $\|x - \eta^0\|_2 \leq \delta$,

(A4) there exists a matrix $\widetilde{Q}$ of full rank whose columns span the column space of $\widetilde{X}$ and the first r_m columns, written $\widetilde{Q}_m$, span the column space of $\widetilde{X}_m$ such that there exists constants $c_4, c_5, c_6 > 0$ with

$$c_4 n \leq \lambda_{min}(\widetilde{Q}^T \widetilde{Q}) \leq \lambda_{max}(\widetilde{Q}^T \widetilde{Q}) \leq c_5 n$$

for n large enough and $\max_{i=1,\dots n} \|\tilde{q}_i\|_\infty \leq c_6$ where $\tilde{q}_i^T$ denotes the i-th row of $\widetilde{Q}$.

Then the conditions in Proposition 1 are met.

The first three conditions in Proposition S1 relate to the assumed form of the generalised linear model. Assumption (A4) places conditions only on the design matrix $\widetilde{X}$ as in Proposition 1 of the main text. The eigenvalue condition in (A4) would be satisfied if the non-zero eigenvalues of $n^{-1}\widetilde{X}^T\widetilde{X}$ were asymptotically bounded above and away from zero.

In this section, we provide a proof of Proposition S1. Before proceding, it will be useful to summarise key notation and simple consequences of assumptions (A1)-(A4) that will be used on various occasions throughout the proof. First, it is well-known that for $i = 1, \dots, n$,

$$\mathbb{E}(Y_i) = b'(\eta_i^0), \qquad Var(Y_i) = \phi b''(\eta_i^0)$$

where $\eta^0 = \widetilde{X}\gamma^0$ is the true signal. In particular,

$$\eta^0 = \operatorname{argmax}_{\eta \in \text{Col-Sp}(\widetilde{X})} \mathbb{E}\{\bar{\ell}(\eta)\}.$$

Define

$$\hat{\eta} = \operatorname{argmax}_{\eta \in \text{Col-Sp}(\widetilde{X})} \bar{\ell}(\eta)$$

$$\hat{\eta}_m = \operatorname{argmax}_{\eta \in \text{Col-Sp}(\widetilde{X}_m)} \bar{\ell}(\eta)$$

$$\eta_m^0 = \operatorname{argmax}_{\eta \in \text{Col-Sp}(\widetilde{X}_m)} \mathbb{E}\{\bar{\ell}(\eta)\}.$$

We assume these maximisers are unique and set the relevant score equations to zero. Let $U^0 = \nabla_\eta \bar{\ell}(\eta^0)$ and $J^0 = -\nabla_{\eta\eta}^2 \bar{\ell}(\eta^0)$. For the generalised linear model with canonical link, $U^0 = \phi^{-1}(Y - \mathbb{E}(Y))$ and J^0 is the diagonal matrix with (i,i)-th entry given by $\phi^{-2}Var(Y_i) = \phi^{-1}b''(\eta_i^0)$. Extend these definitions to

$$U_m^0 = \nabla_\eta \bar{\ell}(\eta_m^0), \qquad J_m^0 = \nabla_{\eta\eta}^2 \bar{\ell}(\eta_m^0).$$

It will be useful to define $D(\eta_1, \eta_2)$ to be a diagonal matrix whose (i,i)-th entry is given by $b''(\eta_i^{(i)})$ where $\eta^{(i)}$ is a vector on the line joining η_1 and η_2. The following are useful consequences of assumptions (A1)-(A4).

(C1) Let $n\Sigma = \widetilde{Q}^T J^0 \widetilde{Q}$. Then,

$$\lambda_{max}(\Sigma) \leq \lambda_{max}(\widetilde{Q}^T \widetilde{Q}/n)\lambda_{max}(J^0) = O(1)$$

by assumption (A2) and (A4). Similarly, $\lambda_{min}^{-1}(\Sigma) = O(1)$.

(C2) Suppose η is such that $\|\eta - \eta^0\|_\infty = o(1)$. By assumption (A1) and the mean value theorem,

$$\|D(\eta, \eta^0) - \phi J^0\|_2 = \|D(\eta, \eta^0) - D(\eta^0, \eta^0)\|_2$$
$$\leq c_1 \|\eta - \eta^0\|_\infty$$
$$= o(1).$$

Further, by assumption (A4),

$$n^{-1}\|\widetilde{Q}^T\{D(\eta, \eta^0) - \phi J^0\}\widetilde{Q}\|_2 = o(1).$$

(C3) For some β and β^0, write $\eta = \widetilde{Q}\beta$, $\eta^0 = \widetilde{Q}\beta^0$ and define $U_\eta = \nabla_\eta \bar{\ell}(\eta)$. Let $D = D(\eta, \eta^0)$. Then, assuming $n^{1/2}\|\beta - \beta^0\|_2 = o(1)$ and using a Taylor expansion,

$$\phi\|\widetilde{Q}_m^T(U_\eta - U^0)\|_2 = \|\widetilde{Q}_m^T\{b'(\eta) - b'(\eta^0)\}\|_2$$
$$= \|\widetilde{Q}_m^T D\widetilde{Q}(\beta^0 - \beta)\|_2$$
$$\leq \|\widetilde{Q}^T D\widetilde{Q}(\beta^0 - \beta)\|_2$$
$$\leq \|\widetilde{Q}^T D\widetilde{Q}\|_2\|\beta^0 - \beta\|_2$$
$$= o(n^{1/2})$$

where we have used the notation $b'(\eta)$ to denote the vector with i-th entry equal to $b'(\eta_i)$. This results follows from (C1), (C2) and the fact that $\|\eta - \eta^0\|_\infty \leq \max_{i=1,\dots n} \|\tilde{q}_i\|_2\|\beta - \beta^0\|_2 = o(1)$ by assumption (A4).

To prove Proposition S1, we break the proof into three parts. First we establish a local asymptotic expansion of the log-likelihood function about η^0 (Lemma S10), then we show that the score function is asymptotically normally distributed (Lemma S11). Finally, we prove that the estimated predicted values $\hat{\eta}_m$ are consistent estimators of η^0 (Lemma S15). A proof that $\hat{\eta}$ is also a consistent estimator of η^0 is omitted as the ideas resemble those in Lemma S15 and related results. Some of these results follow classical results closely, which we include for completeness.

S6.0.1 Local asymptotic expansion

LEMMA S10. *Suppose conditions (A1)-(A4) hold. Then, for any $M_n = o(1)$,*

$$\sup_{h \in B_n} |\bar{\ell}(\eta^0 + h) - f(\eta^0, h)| = o(1)$$

where $B_n = \{h \in \mathbb{R}^n : \|h\|_3 \le M_n\}$ and

$$f(\eta^0, h) = \bar{\ell}(\eta^0) + h^T \nabla_\eta \bar{\ell}(\eta^0) + \frac{1}{2} h^T \nabla_{\eta\eta}^2 \bar{\ell}(\eta^0) h.$$

Further, $n^{-1/2} U^0 = n^{-1/2} \nabla_\eta \bar{\ell}(\eta^0)$ and $J^0 = -\nabla_{\eta\eta}^2 \bar{\ell}(\eta^0)$ satisfy

$$\|n^{-1/2} U^0\|_2 = O_P(1), \qquad \|J^0\|_2 = O_P(1).$$

PROOF. Suppose h satisfies $\|h\|_3 \le M_n$. Using a Taylor expansion,

$$\bar{\ell}(\eta^0 + h) - \bar{\ell}(\eta^0) = h^T U^0 - \frac{h^T J^0 h}{2}$$
$$- \frac{1}{2\phi} h^T \{D(\eta^0, \eta^0 + h) - \phi J^0\} h.$$

Let $D = D(\eta^0, \eta^0 + h)$. By assumption (A1) and the mean value theorem,

$$|(D - \phi J^0)_{ii}| = |b''(\eta_i^{(i)}) - b''(\eta_i^0)| \le c_1 |h_i|$$

where $\eta^{(i)}$ is a vector on the line joining $\eta^0 + h$ and η^0. Then, as $\|h\|_3 \le M_n$,

$$|h^T(D - \phi J^0)h| = \left| \sum_{i=1}^n h_i^2 (D - \phi J^0)_{ii} \right|$$
$$\le c_1 \sum_{i=1}^n |h_i|^3$$
$$\le c_1 M_n^3.$$

As this upper bound holds for any $h \in B_n$, it follows that

$$\sup_{h \in B_n} |\bar{\ell}(\eta^0 + h) - f(\eta^0, h)| \le c_1 M_n^3 / (2\phi) = o(1).$$

Further,

$$\|n^{-1/2} U^0\|_2^2 = \sum_{i=1}^n \frac{\{Y_i - \mathbb{E}(Y_i)\}^2}{\phi^2 n}$$

Let $Z_i = \{Y_i - \mathbb{E}(Y_i)\}^2$ and $\sigma_i^2 = \text{Var}(Y_i)$. Then,

$$\mathbb{E} \left| \sum_{i=1}^n \frac{(Z_i - \sigma_i^2)}{n} \right|^{3/2} \le \sum_{i=1}^n \mathbb{E}\{|Z_i - \sigma_i^2|^{3/2}\}/n$$
$$\le \frac{2^{3/2}}{n} \sum_{i=1}^n \mathbb{E} \left\{ \left(\frac{|Z_i| + |\sigma_i^2|}{2} \right)^{3/2} \right\}$$
$$\le \frac{2^{1/2}}{n} \sum_{i=1}^n \mathbb{E}(|Z_i|^{3/2} + |\sigma_i^2|^{3/2})$$

by convexity of the function $|x|^{3/2}$ for all $x \in \mathbb{R}$. By assumption (A2),

$$\mathbb{E}|Z_i|^{3/2} \le C_2, \qquad |\sigma_i^2|^{3/2} \le C_2^{3/2}$$

and so

$$\mathbb{E} \left| \sum_{i=1}^n (Z_i - \sigma_i^2)/n \right|^{3/2} \le 2^{1/2} C_2 (1 + \sqrt{C_2})$$

Thus, for $\epsilon > 0$ and $M^{3/2} = \epsilon^{-1} \phi^{-3} \sqrt{2} C_2 (1 + \sqrt{C_2})$,

$$\mathbb{P} \left(\left| \sum_{i=1}^n Z_i - \sigma_i^2 \right| \ge M\phi^2 n \right) \le \frac{\mathbb{E} \left| \sum_{i=1}^n (Z_i - \sigma_i^2) \right|^{3/2}}{M^{3/2} \phi^3 n^{3/2}}$$
$$\le \epsilon.$$

As σ_i^2 is bounded, this implies that $\|n^{-1/2} U^0\|_2 = O_P(1)$. Also,

$$\|\nabla_{\eta\eta}^2 \bar{\ell}(\eta^0)\|_2^2 = \phi^{-1} \max_i |b''(\eta_i^0)| = \phi^{-2} \max_i |\text{Var}(Y_i)|$$

which is $O(1)$ by assumption (A2). □

S6.0.2 Asymptotic normality of the score function We establish the asymptotic normality of $(\widetilde{Q}^T J^0 \widetilde{Q})^{-1/2} \widetilde{Q}^T U^0$. The case that replaces $\widetilde{Q}$ by $\widetilde{Q}_m$ follows closely.

LEMMA S11. *Suppose conditions (A1)-(A4) hold. Then*

$$(\widetilde{Q}^T J^0 \widetilde{Q})^{-1/2} \widetilde{Q}^T U^0 \xrightarrow{d} N(0, \mathbb{I}),$$

with $U^0 = \nabla_\eta \bar{\ell}(\eta^0)$ and $J^0 = -\nabla_{\eta\eta}^2 \bar{\ell}(\eta^0)$.

PROOF. Consider any unit vector $u \in \mathbb{R}^{\hat{r}}$ and let $n\Sigma = \widetilde{Q}^T J^0 \widetilde{Q}$. Then

$$u^T (\widetilde{Q}^T J^0 \widetilde{Q})^{-1/2} \widetilde{Q}^T U^0 = \sum_{i=1}^n z_i$$

where

$$z_i = n^{-1/2} \phi^{-1} u^T \Sigma^{-1/2} \widetilde{q}_i \{Y_i - \mathbb{E}(Y_i)\}.$$

This is a sum of independent, centred random variables whose combined variance is one. Further, by (A2),

$$\sum_{i=1}^n \mathbb{E}(|z_i|^3) \leq C_2 n^{-3/2} \sum_{i=1}^n |\phi^{-1} u^T \Sigma^{-1/2} \widetilde{q}_i|^3$$

$$\leq C_2 \phi^{-3} n^{-3/2} \sum_{i=1}^n \|\Sigma^{-1/2} \widetilde{q}_i\|_2^3.$$

By assumption (A4) and (C1),

$$\max_{i=1}^n \|\Sigma^{-1/2} \widetilde{q}_i\|_2^2 \leq \max_{i=1}^n \|\tilde{q}_i\|_2^2 \{\lambda_{min}(\Sigma)\}^{-1} = O(1).$$

Thus

$$\sum_{i=1}^n \mathbb{E}(|z_i|^3) \to 0.$$

The asymptotic normality of the score function follows by Lyapunov's central limit theorem. □

S6.0.3 Consistency of predictions The following results are used to show that $\hat{\eta}$ and $\hat{\eta}_m$ are consistent estimators of η^0. To avoid repeated statements, we only derive the results for $\hat{\eta}_m$. The results for $\hat{\eta}$ can be derived without the need for Lemmas S12 and S13.

LEMMA S12. *Suppose* $\|\eta - \eta^0\|_2 > \phi^{-1} c_2/(2c_1)$. *Then,*

$$|M(\eta^0) - M(\eta)| \geq \frac{c_2^3}{16 c_1^2 \phi^4 n}$$

where $M(\eta) = n^{-1} \mathbb{E}\{\bar{\ell}(\eta)\}$.

PROOF. Let $\delta = \phi^{-1} c_2/(2c_1)$ and $B_\delta(\eta^0) = \{x \in \mathbb{R}^n : \|x - \eta^0\|_2 \leq \delta\}$ be the ball of radius δ about η^0. If $x \in B_\delta(\eta^0)$ then $\|x - \eta^0\|_\infty \leq \delta$ and so, by assumption (A1),

$$|b''(x_i) - b''(\eta_i^0)| \leq c_1 \delta$$

for all $i = 1, \ldots, n$. Assumption (A2) then implies that

$$b''(x_i) \geq b''(\eta_i^0) - \phi^{-1} c_2/2 \geq \phi^{-1} c_2/2.$$

We use this to show that $M(x)$ is bounded above by a quadratic form for $x \in B_\delta(\eta^0)$. By a Taylor expansion and the fact that $\nabla_\eta M(\eta^0) = 0$,

$$M(x) - M(\eta^0) = \frac{-(x - \eta^0)^T D(x, \eta^0)(x - \eta^0)}{2\phi n}$$

In particular, as the entries of $D(x, \eta^0)$ are bounded below by $\phi^{-1} c_2/2$,

$$(14) \quad M(x) - M(\eta^0) \leq \frac{-c_2 \|x - \eta^0\|_2^2}{4\phi^2 n}, \quad \forall x \in B_\delta(\eta^0).$$

We now show that for $\eta \notin B_\delta(\eta^0)$ we have $M(\eta) \leq M(x^0)$ where $x^0 = t\eta + (1-t)\eta^0$ for some $t \in [0,1]$ such that $\|x^0 - \eta^0\|_2 = \delta$. By a Taylor expansion,

$$M(\eta) - M(x^0) = \sum_{i=1}^n \{\nabla_{\eta_i} M(v^{(i)})\}(\eta_i - x_i^0)$$

$$= (1-t) \sum_{i=1}^n \{\nabla_{\eta_i} M(v^{(i)})\}(\eta_i - \eta_i^0)$$

where each $v^{(i)}$ is on the line joining η and x^0. So, there exists $s_i \in [0,1]$ such that $v^{(i)} = \{s_i + t(1 - s_i)\}\eta + (1 - s_i)(1-t)\eta^0$. Then, as $v^{(i)} - \eta^0 = \{s_i + t(1 - s_i)\}(\eta - \eta^0)$ we can rewrite each term in the summation as

$$\{\nabla_{\eta_i} M(v^{(i)})\}(\eta_i - \eta_i^0) = \frac{\{\nabla_{\eta_i} M(v^{(i)})\}(v^{(i)} - \eta_i^0)}{s_i + t(1 - s_i)}$$

$$= \frac{\{b'(\eta_i^0) - b'(v_i^{(i)})\}(v_i^{(i)} - \eta_i^0)}{\phi n \{s_i + t(1 - s_i)\}}$$

$$\leq 0$$

where the last inequality follows by a further Taylor expansion of $b'(v_i^{(i)})$ about η_i^0 and noting that $b''(\cdot)$ is a non-negative function. Thus, for $\eta \notin B_\delta(\eta^0)$ there exists x^0 satisfying $\|x^0 - \eta^0\|_2 = \delta$ with

$$M(\eta) \leq M(x^0).$$

The result then follows as $x^0 \in B_\delta(\eta^0)$ and so, by (14),

$$M(\eta) \leq M(x^0) \leq M(\eta^0) - \frac{c_2 \delta^2}{4\phi^2 n}.$$

□

LEMMA S13. *Define*

$$\alpha_m^0 = argmax_{\alpha \in \mathbb{R}^{r_m}} \mathbb{E}\{\bar{\ell}(\widetilde{Q}_m \alpha)\}$$

$$\beta^0 = argmax_{\beta \in \mathbb{R}^{\hat{r}}} \mathbb{E}\{\bar{\ell}(\widetilde{Q}\beta)\}$$

and $\beta_m^0 = ((\alpha_m^0)^T, 0)^T \in \mathbb{R}^{\hat{r}}$. *Suppose conditions (A1)-(A4) hold and* $\|(I - P_m)\eta^0\|_2 = o(1)$. *Then,*

$$\sqrt{n}\|\beta_m^0 - \beta^0\|_2 = o(1).$$

PROOF. Define $M(\eta) = n^{-1} \mathbb{E}\{\bar{\ell}(\eta)\}$. A Taylor expansion reveals that $M(\widetilde{Q}\beta_m^0) - M(\widetilde{Q}\beta^0)$ is equal to

$$\frac{-1}{2\phi n}(\beta^0 - \beta_m^0)^T \widetilde{Q}^T D(\eta^0, \eta_m^0)\widetilde{Q}(\beta^0 - \beta_m^0)$$

and so

$$(15) \quad \|\beta_m^0 - \beta^0\|_2^2 \leq \frac{2\phi|M(\widetilde{Q}\beta_m^0) - M(\widetilde{Q}\beta^0)|}{\lambda_{min}\{n^{-1}\widetilde{Q}^T D(\eta^0, \eta_m^0)\widetilde{Q}\}}$$

where $\eta_m^0 = \widetilde{Q}\beta_m^0$ and $\eta^0 = \widetilde{Q}\beta^0$. Our aim is to show that $|M(\widetilde{Q}\beta_m^0) - M(\widetilde{Q}\beta^0)| = o(n^{-1})$. Note that

$$M(P_m \widetilde{Q}\beta^0) \leq M(\widetilde{Q}\beta_m^0) \leq M(\widetilde{Q}\beta^0)$$

by definition of β_m^0 and β^0. Then,

$$|M(\widetilde{Q}\beta_m^0) - M(\widetilde{Q}\beta^0)| \le |M(P_m\eta^0) - M(\eta^0)|$$
$$= \frac{\{(I - P_m)\eta^0\}^T D\{(I - P_m)\eta^0\}}{2n\phi}$$
$$\le \frac{\|(I - P_m)\eta^0\|_2^2 \lambda_{max}(D)}{2n\phi}.$$

where $D = D(\eta^0, P_m\eta^0)$. As $\|(I - P_m)\eta^0\|_\infty \le \|(I - P_m)\eta^0\|_2 = o(1)$, (C2) and assumption (A2) imply that $\lambda_{max}(D) = O(1)$. Then,

$$|M(\widetilde{Q}\beta_m^0) - M(\widetilde{Q}\beta^0)| = O(\|(I - P_m)\eta^0\|_2^2/n)$$
$$\text{(16)} \qquad\qquad = o(n^{-1}).$$

By Lemma S12, equation (16) implies that $\widetilde{Q}\beta_m^0$ must be in the ℓ_2-ball of radius $\delta = c_2/(2c_1\phi)$ about η^0 for large enough n. Hence, $\lambda_{min}\{n^{-1}\widetilde{Q}^T D(\eta^0, \eta_m^0)\widetilde{Q}\}$ is bounded away from zero by assumption (A3) and (A4). Combining this with (15) and (16),

$$\|\beta_m^0 - \beta^0\|_2^2 = o(n^{-1}).$$

$\square$

LEMMA S14. *Define*

$$\hat{\alpha}_m = argmax_{\alpha_m \in \mathbb{R}^{r_m}} \bar{\ell}(\widetilde{Q}_m\alpha_m)$$
$$\alpha_m^0 = argmax_{\alpha_m \in \mathbb{R}^{r_m}} \mathbb{E}\{\bar{\ell}(\widetilde{Q}_m\alpha_m)\}$$

Suppose conditions (A1)-(A4) hold and assume that

$$n^{1/2}\|(I - P_m)\eta^0\|_2 = o(1).$$

Then,

$$\|\hat{\alpha}_m - \alpha_m^0\|_2 = O_P(n^{-1/2}).$$

The proof of this result closely follows the arguments in Theorem 4.17 of [6] with minor modifications to account for the differences between α^0 and α_m^0.

PROOF. Fix $\epsilon > 0$. Our aim is to show that there exists $M > 0$ such that for n large enough,

$$\mathbb{P}(\sqrt{n}\|\hat{\alpha}_m - \alpha_m^0\|_2 > M) \le \epsilon.$$

For $\delta > 0$, define $A_n(\delta) = \{\alpha \in \mathbb{R}^{r_m} : \|(\widetilde{Q}_m^T J^0 \widetilde{Q}_m)^{1/2}(\alpha - \alpha_m^0)\|_2 \le \delta\}$ and $\partial A_n(\delta)$ to be the boundary of $A_n(\delta)$. Let E_δ be the event that

$$E_\delta = \{\forall \alpha \in \partial A_n(\delta), \bar{\ell}(\widetilde{Q}_m\alpha) < \bar{\ell}(\widetilde{Q}_m\alpha_m^0)\}.$$

We will show that for n large enough, $\mathbb{P}(E_\delta) \ge 1 - (1 + \epsilon')16\delta^{-2}r_m$ for any $\epsilon' > 0$, and when $M^2 = 2\delta^2\phi/(c_2c_4)$

it follows that $\mathbb{P}(\sqrt{n}\|\hat{\alpha}_m - \alpha_m^0\|_2 > M \mid E_\delta) = 0$. Then, for n large enough,

$$\mathbb{P}(\sqrt{n}\|\hat{\alpha}_m - \alpha_m^0\|_2 > M) \le \mathbb{P}(\sqrt{n}\|\hat{\alpha}_m - \alpha_m^0\|_2 > M \mid E_\delta)$$
$$+ \mathbb{P}(E_\delta^c)$$
$$= \mathbb{P}(E_\delta^c)$$
$$\le (1 + \epsilon')16\delta^{-2}r_m.$$

The result will follow on choosing $\delta^2 \ge 16r_m(1 + \epsilon')/\epsilon$ where we have used the fact that r_m is bounded above.

Let $n\Sigma_{mm} = \widetilde{Q}_m^T J^0 \widetilde{Q}_m$ and $\lambda = (n\Sigma_{mm})^{1/2}(\alpha - \alpha_m^0)$. Then, for n large enough,

$$\sup_{\alpha \in A_n(\delta) \cup \partial A_n(\delta)} \sqrt{n}\|\alpha - \alpha_m^0\|_2 \le \sup_{\|\lambda\|_2 \le \delta} \frac{\|\lambda\|_2}{\lambda_{min}^{1/2}(\Sigma_{mm})}$$
$$\text{(17)} \qquad\qquad \le \sqrt{\frac{\delta^2\phi}{c_2c_4}}.$$

by (A2) and (A4).

On the event E_δ, the maximum likelihood estimator is in the set $A_n(\delta)$. This is because if for all $\alpha \in \partial A_n(\delta)$,

$$\bar{\ell}(\widetilde{Q}_m\alpha) < \bar{\ell}(\widetilde{Q}_m\alpha_m^0)$$

then $\bar{\ell}(\widetilde{Q}_m\alpha)$ has a local maximum at a point $\hat{\alpha}_m$ inside $A_n(\delta)$ satisfying

$$\nabla_\alpha \bar{\ell}(\widetilde{Q}_m\hat{\alpha}_m) = 0.$$

Standard results imply the uniqueness of a local maximum, and so this point must be the MLE. Combining this with (17), when $M^2 = 2\delta^2\phi/(c_2c_4)$, it follows that for n large enough,

$$\mathbb{P}(\sqrt{n}\|\hat{\alpha}_m - \alpha_m^0\|_2 > M \mid E_\delta) = 0.$$

Finally we show that $\mathbb{P}(E_\delta) \ge 1 - (1 + \epsilon')16\delta^{-2}r_m$ for any $\epsilon' > 0$ and n large enough. For all $\alpha \in \partial A_n(\delta)$,

$$\bar{\ell}(\widetilde{Q}_m\alpha) - \bar{\ell}(\widetilde{Q}_m\alpha_m^0) = (\alpha - \alpha_m^0)^T \widetilde{Q}_m^T U_m^0$$
$$- \frac{1}{2\phi}(\alpha - \alpha_m^0)^T \widetilde{Q}_m^T D \widetilde{Q}_m(\alpha - \alpha_m^0)$$
$$= \lambda^T (\widetilde{Q}_m J^0 \widetilde{Q}_m)^{-1/2} \widetilde{Q}_m^T U^0 + \Delta_1$$
$$- \Delta_2 - \frac{\lambda^T \lambda}{2}$$

where $\lambda = (\widetilde{Q}_m J^0 \widetilde{Q}_m)^{1/2}(\alpha - \alpha_m^0)$, $D = D(\widetilde{Q}_m\alpha, \widetilde{Q}_m\alpha_m^0)$ and

$$\Delta_1 = \lambda^T (\widetilde{Q}_m J^0 \widetilde{Q}_m)^{-1/2} \widetilde{Q}_m^T (U_m^0 - U^0)$$
$$\Delta_2 = \frac{\|\{\widetilde{Q}_m^T(D - \phi J^0)\widetilde{Q}_m\}^{1/2}(\widetilde{Q}_m J^0 \widetilde{Q}_m)^{-1/2}\lambda\|_2^2}{2\phi}$$

By (C1), (C3) and Lemma S13,

$$\sup_{\alpha \in \partial A_n(\delta)} |\Delta_1| \le \delta\|(\widetilde{Q}_m^T J^0 \widetilde{Q}_m)^{-1/2}\|_2 \|\widetilde{Q}_m^T(U_m^0 - U^0)\|_2$$
$$= o(1).$$

By equation (17) and (A4),

$$\sup_{\alpha \in \partial A_n(\delta)} \|\widetilde{Q}_m^T(\alpha - \alpha_m^0)\|_\infty = o(1).$$

Then,

$$\sup_{\alpha \in \partial A_n(\delta)} |\Delta_2| \leq \sup_{\alpha \in \partial A_n(\delta)} \|\{n^{-1}\widetilde{Q}_m^T(D - \phi J^0)\widetilde{Q}_m\}^{1/2}\|_2^2 \times$$

$$\|\{n^{-1}\widetilde{Q}_m^T J^0 \widetilde{Q}_m\}^{-1/2}\|_2^2 \sup_{\alpha \in \partial A_n(\delta)} \|\lambda\|_2^2$$

$$= o(1)$$

by the arguments in (C1) and (C2). Thus,

$$\sup_{\alpha \in \partial A_n(\delta)} \bar{\ell}(\widetilde{Q}_m\alpha) - \bar{\ell}(\widetilde{Q}_m\alpha_m^0) = \sup_{\|\lambda\|=\delta} \lambda^T Z - \delta^2/2 + o(1)$$

$$\leq \delta\|Z\|_2 - \delta^2/2 + o(1).$$

where $Z = (\widetilde{Q}_m^T J^0 \widetilde{Q}_m)^{-1/2}\widetilde{Q}_m^T U^0$. By the arguments in Lemma S11 (replacing $\widetilde{Q}$ with $\widetilde{Q}_m$), $\|Z\|_2^2 \xrightarrow{d} \chi_{r_m}^2$ and so for any $\epsilon' > 0$,

$$\mathbb{P}(\|Z\|_2 \leq \delta/4) \geq 1 - (1 + \epsilon')\mathbb{P}(\chi_{r_m}^2 > \delta^2/16)$$

$$\geq 1 - (1 + \epsilon')16\delta^{-2}r_m$$

when n is large enough. When $\|Z\|_2 \leq \delta/4$ the event E_δ holds and so

$$\mathbb{P}(E_\delta) \geq 1 - (1 + \epsilon')16\delta^{-2}r_m$$

for n large enough. $\square$

LEMMA S15. *Suppose conditions (A1)-(A4) hold and* $n^{1/2}\|(I - P_m)\eta^0\|_2 = o(1)$. *Then,*

$$\|\hat{\eta}_m - \eta^0\|_3 = O_P(n^{-1/6}).$$

PROOF. By uniqueness of the maximum likelihood estimators,

$$\hat{\eta}_m = \widetilde{Q}_m\hat{\alpha}_m = \widetilde{Q}\hat{\beta}_m$$
$$\eta^0 = \widetilde{Q}\beta^0.$$

where $\hat{\alpha}_m$ is defined in Lemma S14, β^0 is defined in Lemma S13, and $(\hat{\beta}_m)^T = ((\hat{\alpha}_m)^T, 0)$. Similarly, write

$$\eta_m^0 = \widetilde{Q}_m\alpha_m^0 = \widetilde{Q}\beta_m^0$$

where α_m^0 is defined in Lemma S14 and $(\beta_m^0)^T = ((\alpha_m^0)^T, 0)$. By Lemma S14, $\|\sqrt{n}(\hat{\alpha}_m - \alpha_m^0)\|_2 = O_P(1)$. By Lemma S13, $\sqrt{n}\|\beta_m^0 - \beta^0\|_2 = o(1)$. Then,

$$\sqrt{n}\|\hat{\beta}_m - \beta^0\|_2 \leq \sqrt{n}(\|\hat{\beta}_m - \beta_m^0\|_2 + \|\beta_m^0 - \beta^0\|_2)$$

$$= \|\sqrt{n}(\hat{\alpha}_m - \alpha_m^0)\|_2 + o(1)$$

$$= O_P(1).$$

So, by assumption (A4),

$$\|\hat{\eta}_m - \eta^0\|_2^2 = (\hat{\beta}_m - \beta^0)^T \widetilde{Q}^T \widetilde{Q}(\hat{\beta}_m - \beta^0) = O_P(1).$$

Further, assuming $\|\widetilde{q}_i\|_\infty$ is bounded for all $i = 1, \ldots, n$,

$$\|\hat{\eta}_m - \eta^0\|_\infty \leq \max_{i=1,\ldots n} \|\widetilde{q}_i\|_2 \|\hat{\beta}_m - \beta^0\|_2 = O_P(n^{-1/2}).$$

Then,

$$\|\hat{\eta}_m - \eta^0\|_3^3 \leq \|\hat{\eta}_m - \eta^0\|_\infty \|\hat{\eta}_m - \eta^0\|_2^2 = O_P(n^{-1/2}).$$

$\square$

REFERENCES

[1] BATTEY, H. S. and COX, D. R. (2018). Large numbers of explanatory variables: a probabilistic assessment. *Proc. R. Soc. Lond. A.* **474**.

[2] BATTEY, H. S. and REID, N. (2021). Inference in high-dimensional linear regression. *arXiv:2106.12001*.

[3] FAN, J., SHAO, Q. and ZHOU, W. (2018). Are discoveries spurious? Distributions of maximum spurious correlations and their applications. *Ann. Statist.* **46** 989-1017.

[4] HALL, P. (1979). On the rate of convergence of normal extremes. *J. Appl. Probab.* **16** 433-439.

[5] MUIRHEAD, R. (2005). *Aspects of Multivariate Statistical Theory*. John Wiley & Sons, Inc., New Jersey.

[6] SHAO, J. (1999). *Mathematical statistics. Springer Texts in Statistics*. Springer New York.