UNIVERSITY OF EDINBURGH STATISTICS SEMINAR

Large numbers of explanatory variables

H. S. Battey
Department of Mathematics, Imperial College London

3 May 2019

- Regression, broadly defined.
- Response variable Y_i, e.g., blood pressure, disease status, survival time, ...
- Vector X_i of v potential explanatory variables.
- Observed on units, e.g. patients, $i = 1, \ldots, n$.
- $v \gg n$, e.g. X_i arising from gene expression analysis.

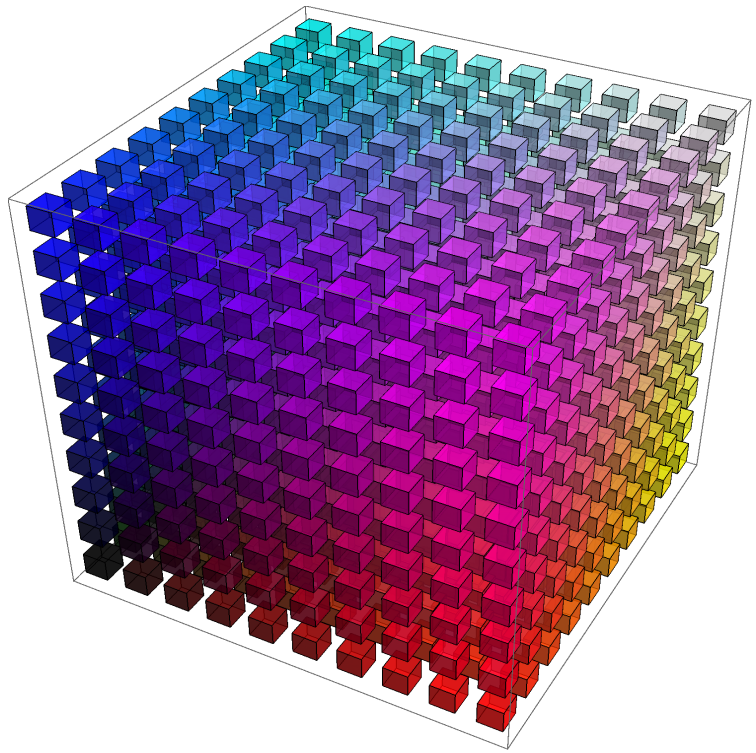
- Goal: scientific understanding.
- Sparsity is critical for interpretability and statistical stability.
- Lasso (Tibshirani, 1996): penalized LS:
  Minimize $\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$.
- More generally penalized MLE.
- There results a single model.
- Cox, D. R. and Battey, H. S. (2017), *Proc. Nat. Acad. Sci.*, 114, 8592–8595: aim for a confidence set of models.

- Goal: scientific understanding.
- Sparsity is critical for interpretability and statistical stability.
- Lasso (Tibshirani, 1996): penalized LS:
  Minimize $\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$.
- More generally penalized MLE.
- There results a single model.
- Cox, D. R. and Battey, H. S. (2017), *Proc. Nat. Acad. Sci.*, 114, 8592–8595: aim for a confidence set of models.

- Arrange variable indices in a $k \times k \times k \times \cdots$ hypercube where $k \leq 15$.
- Traverse the cube: rows, columns, etc. Assess each variable several times alongside $k - 1$ different companions.

- Arrange variable indices in a $k \times k \times k \times \cdots$ hypercube where $k \leq 15$.
- Traverse the cube: rows, columns, etc. Assess each variable several times alongside $k - 1$ different companions.

- Arrange variable indices in a $k \times k \times k \times \cdots$ hypercube where $k \leq 15$.
- Traverse the cube: rows, columns, etc. Assess each variable several times alongside $k - 1$ different companions.



- Arrange variable indices in a $k \times k \times k \times \cdots$ hypercube where $k \leq 15$.
- Traverse the cube: rows, columns, etc. Assess each variable several times alongside $k - 1$ different companions.
- Retain or discard variables according to a decision rule.
- Repeat.
- Informal checks.
- Test low dimensional subsets for compatibility with the data. A "confidence set" of models.

- Arrange variable indices in a $k \times k \times k \times \cdots$ hypercube where $k \leq 15$.
- Traverse the cube: rows, columns, etc. Assess each variable several times alongside $k - 1$ different companions.
- Retain or discard variables according to a decision rule.
- Repeat.
- Informal checks.
- Test low dimensional subsets for compatibility with the data. A "confidence set" of models.

- Arrange variable indices in a $k \times k \times k \times \cdots$ hypercube where $k \leq 15$.
- Traverse the cube: rows, columns, etc. Assess each variable several times alongside $k - 1$ different companions.
- Retain or discard variables according to a decision rule.
- Repeat.
- Informal checks.
- Test low dimensional subsets for compatibility with the data. A "confidence set" of models.

THREE PHASES

- A reduction phase uses significance tests as an informal guide to discarding variables.
- On the reduced set, an exploratory phase allows assessment of anomalies.
- A model selection phase assesses candidate models for their compatibility with the data.

TYPICAL OUTPUT

Model	Recoded variable indices																				
1	1	16	3	8	-	5	-	19	-	-	-	-	-	-	-	-	-	-	-	-	-
2	1	16	3	8	-	5	-	19	-	-	-	-	-	-	-	-	-	-	-	-	2
3	1	16	3	8	-	-	-	-	-	10	-	-	-	-	-	-	-	-	4	-	-
4	1	16	3	8	-	5	-	19	-	-	-	-	-	-	-	-	-	-	4	-	-
5	1	16	3	8	-	5	-	-	-	10	-	-	-	-	-	-	-	-	-	-	-
6	1	16	3	8	-	5	-	19	-	-	-	-	-	-	-	-	-	-	-	-	-
7	1	16	3	-	15	5	-	-	-	10	-	-	-	-	-	-	-	-	-	-	-
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

- This arose from an example with n =129 individuals: 105 with osteoarthritis, 24 controls. $v \sim 50,000$ genetic variables.
- Any choice between such models would require additional data or subject matter expertise.

INFORMAL MOTIVATION

Write $Z = (Y_i, X_i)_{i=1}^n$.

Suppose we were to ignore computational constraints.

- For every low-dimensional model m, with non-zero parameter vector θ_m, identify the sufficient statistic S_m for θ_m.
- All models compatible with the data in the sense that z is not extreme when calibrated against the distribution of $Z|S_m = s_m$, should be reported as a confidence set of models, alongside the associated confidence statements for θ_m.
- Barndorff-Nielsen and Cox (1994). Cox and Snell (1974).

MORE ON THE REDUCTION PHASE

- No restriction to perfect (hyper)cubes.
- Partially balanced incomplete block designs (Yates, 1936).
- A more severe reduction than marginal screening.
- Prior assessment of importance.
- Arrangement rerandomization.
- A version of backward selection.
- Computation.

Battey, H.S. and Cox, D.R. (2018),
Proc. R. Soc. Lond. A., 474:

- Specify behaviour under idealized conditions.
- Provide guidance on decision rules.
- The goal is not to set up a procedure to achieve pre-assigned error rates.

SOME CANDIDATE FIRST-STAGE REDUCTION STRATEGIES

- Retain the single variable with highest score (lowest p-value);
- Retain the two variables with highest scores;
- Retain all variables, if any, whose scores exceed a threshold.

In the second stage of the procedure, the third strategy is always used.

KEY ASPECTS OF THE REDUCTION PHASE

- What is the probability that a signal variable is falsely discarded?
- How many of the variables ultimately suggested as potentially important are noise variables?

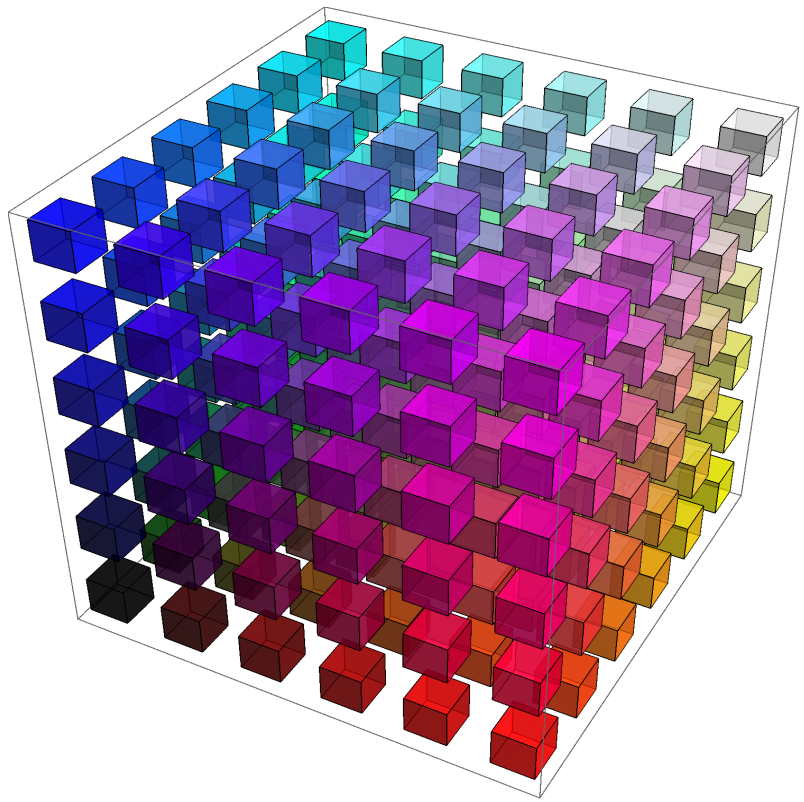
A SIMPLIFYING APPROXIMATION

Analyses involving the same variable are treated as independent.

- Slightly pessimistic assessment of the proportion of correctly retained signal variables.
- Slightly optimistic assessment of the number of falsely retained noise variables.

Signal variables are on an equal footing (same signal strength)

Behaviour is governed largely by θ, the probability that a particular signal variable survives reduction in any single analysis in which it appears. This depends on the reduction strategy used.



- In a set of k variables chosen at random for test, the number of signal variables has a Poisson distribution of mean $a \triangleq kv_{S0}/v$.
- $$\lim_{a \to 0} \frac{1 - e^{-a}(1+a)}{1 - e^{-a}} = 0.$$

  So $\theta \sim \vartheta$ for small a, where ϑ is the survival probability conditioned on the event that there is exactly one signal variable in the set.
- A similar argument applies for a modest number of noise variables correlated with signal variables.

SOME ANALYSIS OF FIRST-STAGE REDUCTION

To avoid distributional assumptions on the response variable we frame the discussion in terms of p-values.

- For noise variables these are uniformly distributed on $(0, 1)$.
- For signal variables, we model their density as $(1 - \gamma)x^{-\gamma}$, $0 < \gamma < 1$.

STRATEGY 1

$$\begin{aligned}
\vartheta^{(1)} &= \text{pr}\{\text{signal beats best of } (k-1) \text{ noise}\} \\
&= (1-\gamma)\int_0^1 dx(1-x)^{k-1}x^{-\gamma} \\
&= (1-\gamma)\frac{\Gamma(1-\gamma)\Gamma(k)}{\Gamma(1-\gamma+k)} \approx \Gamma(2-\gamma)/k^{1-\gamma}.
\end{aligned}$$

$\Gamma(\cdot)$ is close to one over the range of interest. If $\gamma = 0$ the notional signal variable is selected w.p. $1/k$, i.e. at random.

STRATEGY 2

$\vartheta^{(2)} = \text{pr(signal among 2 most significant)} = \vartheta^{(1)} + \vartheta^{(2.1)}$, where

$$\begin{aligned} \vartheta^{(2.1)} &= \text{pr(signal comes second)} \\ &= \int_0^1 dx(k-1)(1-x)^{k-2}x(1-\gamma)x^{-\gamma} \\ &\approx (1-\gamma)\Gamma(2-\gamma)/k^{1-\gamma}. \end{aligned}$$

$\vartheta^{(2)} - \vartheta^{(1)}$ is negligible for γ close to 1. If $\gamma = 0$, $\vartheta^{(2)} = 2/k$.

APPROXIMATION ERROR

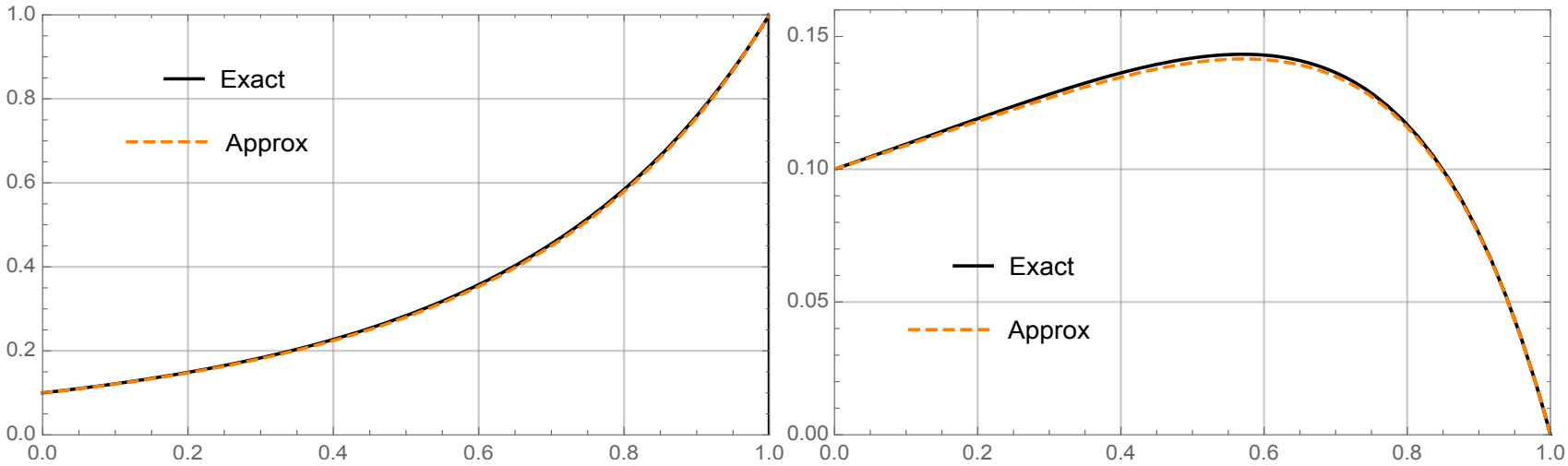


Figure: Exact and approximate values of $\vartheta^{(1)}$ (left) and $\vartheta^{(2.1)}$ (right) over $\gamma \in (0, 1)$ for $k = 10$.

STRATEGY 3

If we set a critical level α, then

$$\vartheta^{(3)} = \int_0^{\alpha} dx(1-\gamma)x^{-\gamma} = \alpha^{1-\gamma}.$$

COMPARISON OF STRATEGIES

- Choose α to equalize survival probabilities of signal variables and compare expected number of retained noise variables.
- Stategies 1 and 3 are equivalent at this α.
- 2 dominates 3, i.e. strategy 3 selects more noise variables on average for the same probability of selecting a signal variable.
- Strategy 2 increases number of retained noise variables by roughly a factor of four over strategy 1.

A DIFFERENT FORMULATION

- Instead of p-values, formulate as normal-theory linear model with signal strength Δ.
- Approximation of the integrals that define $\vartheta^{(j)}$, $j = 1, 2, 3$ leads to qualitatively the same conclusions, e.g., for large k,

$$\vartheta^{(2.1)} \simeq 2\pi \frac{(k-1)^k}{k^{k+1}} \phi(\Delta) \exp\{-\Delta \Phi^{-1}(1/k)\}$$

- $\Delta = 0$ (equivalent to $\gamma = 0$), $\vartheta^{(2.1)} \approx k^{-1}$ for large k.

A DIFFERENT FORMULATION

- Instead of p-values, formulate as normal-theory linear model with signal strength Δ.
- Approximation of the integrals that define $\vartheta^{(j)}$, $j = 1, 2, 3$ leads to qualitatively the same conclusions, e.g., for large k,

$$\vartheta^{(2.1)} \simeq 2\pi \frac{(k-1)^k}{k^{k+1}} \phi(\Delta) \exp\{-\Delta \Phi^{-1}(1/k)\}$$

- $\Delta = 0$ (equivalent to $\gamma = 0$), $\vartheta^{(2.1)} \approx k^{-1}$ for large k.

A DIFFERENT FORMULATION

- Instead of p-values, formulate as normal-theory linear model with signal strength Δ.
- Approximation of the integrals that define $\vartheta^{(j)}$, $j = 1, 2, 3$ leads to qualitatively the same conclusions, e.g., for large k,

$$\vartheta^{(2.1)} \simeq 2\pi \frac{(k-1)^k}{k^{k+1}} \phi(\Delta) \exp\{-\Delta \Phi^{-1}(1/k)\}$$

- $\Delta = 0$ (equivalent to $\gamma = 0$), $\vartheta^{(2.1)} \approx k^{-1}$ for large k.

APPROXIMATION ERROR

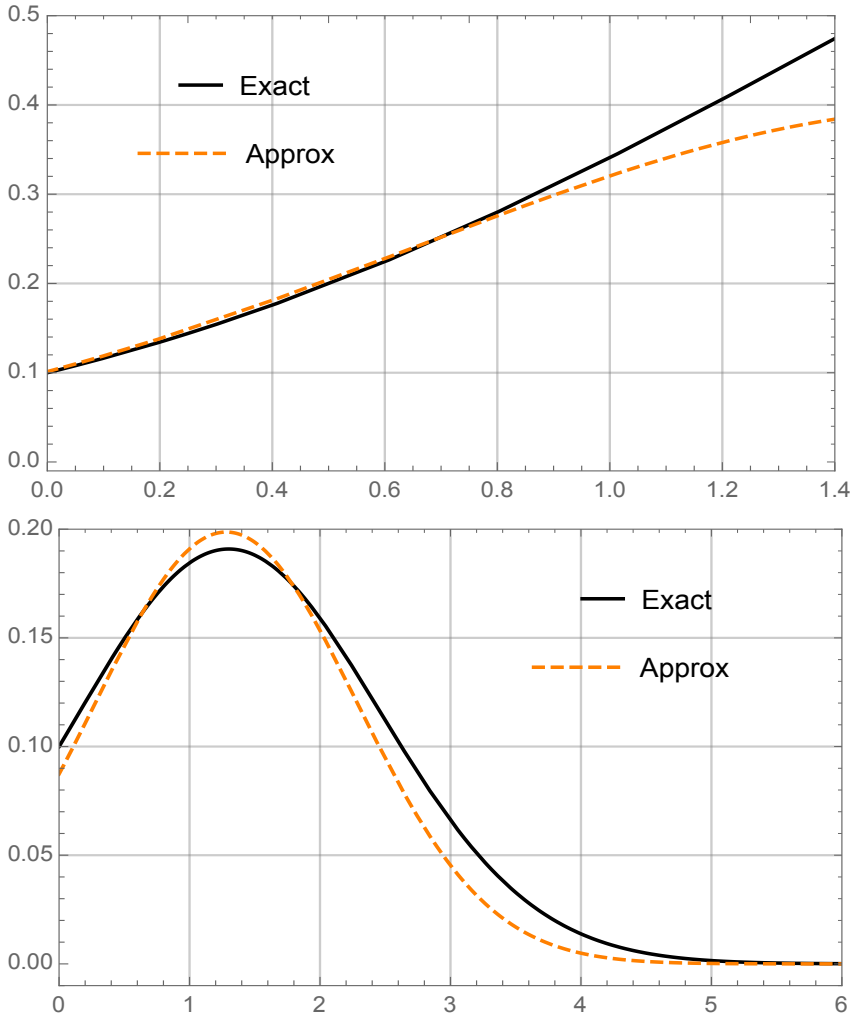


Figure: Exact and approximate values of $\vartheta^{(1)}$ (left) and $\vartheta^{(2.1)}$ (right) as functions of Δ for $k = 10$.

QUALITATIVE COMPARISON OF FORMULATIONS

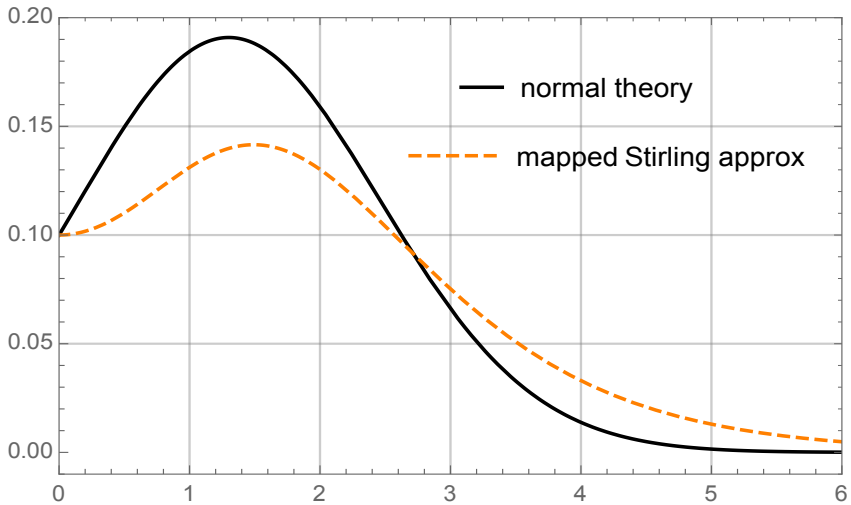


Figure: $\vartheta^{(2.1)}$ as a function of Δ in the Gaussian formulation and in the p-value formulation with $\gamma = \{\cosh(\Delta) - 1\}/\cosh(\Delta)$.

THE FINAL PHASE: "CONFIDENCE SETS" OF MODELS

- Sets comprise all low-dimensional subsets that pass a LR test against the comprehensive model.
- Post-selection inference. Sample splitting.
- Conditional coverage guaranteed in normal theory and asymptotically valid for ML.

ESTABLISHING UNCONDITIONAL COVERAGE

(a) Subject to a constraint on $\mathbb{E}|\widehat{\mathcal{S}}|$, what is pr$(\mathcal{S} \subseteq \widehat{\mathcal{S}})$?

(b) Subject to a lower bound on pr$(\mathcal{S} \subseteq \widehat{\mathcal{S}})$, what is $\mathbb{E}|\widehat{\mathcal{S}}|$?

(c) Same questions for undertuned lasso.

A SIMPLE EXPERIMENT

In each of 500 Monte Carlo replications:

- Generate $n = 10^2$ replicates of $v = 10^3$ variables from a $N(0, P\Sigma P^{-1})$, P a permutation matrix.
- Σ is an identity matrix with one diag block replaced by a correlation matrix of dim $v_{S0} + v_{C0}$ and equal correlation ρ.

A SIMPLE EXPERIMENT (CONTINUED)

- For each $i = 1, \ldots, n$, v_{S0} of the $v_{S0} + v_{C0}$ correlated variables are multiplied by a constant signal and added, together with standard normal noise. This is Y_i.
- Arrange variable indices in $10 \times 10 \times 10$ cube and reduce by strategy 2 followed by strategy 3 at the 0.1% level.
- LR test against the comprehensive model.

MC ESTIMATES: 2^4 FACTORIAL EXPERIMENT

v_{S0}	v_{C0}	ρ	$\frac{\text{signal}}{\text{noise}}$	$\text{pr}(\mathcal{S} \subseteq \widehat{\mathcal{S}})$ undertuned lasso (full)	undertuned lasso (split)	CB (full)	CB (split)	$\text{pr}(\mathcal{S} \in \mathcal{M})$ CB (full)	CB (split)	$\mathbb{E}\lvert\mathcal{M}\backslash\mathcal{S}\rvert$ CB (full)	CB (split)
1	1	0.9	1	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.57 (0.50)	0.99 (0.10)	6.71 (9.35)	18.53 (43.53)
1	1	0.9	0.6	0.95 (0.22)	0.86 (0.35)	0.98 (0.14)	0.84 (0.37)	0.43 (0.50)	0.83 (0.38)	4.98 (4.87)	17.44 (33.48)
1	1	0.5	1	0.96 (0.20)	0.90 (0.30)	1.00 (0.00)	1.00 (0.00)	0.54 (0.50)	1.00 (0.00)	4.62 (4.91)	9.15 (19.64)
1	1	0.5	0.6	0.88 (0.33)	0.69 (0.46)	0.99 (0.10)	0.82 (0.39)	0.32 (0.47)	0.82 (0.39)	2.28 (2.53)	8.14 (15.49)
1	3	0.9	1	1.00 (0.00)	0.99 (0.10)	1.00 (0.00)	0.99 (0.10)	0.62 (0.49)	0.98 (0.14)	23.67 (17.98)	78.93 (97.47)
1	3	0.9	0.6	0.92 (0.27)	0.75 (0.44)	0.98 (0.14)	0.81 (0.39)	0.47 (0.50)	0.81 (0.39)	26.56 (24.78)	57.39 (128.02)
1	3	0.5	1	0.98 (0.14)	0.88 (0.33)	1.00 (0.00)	1.00 (0.00)	0.59 (0.49)	0.98 (0.14)	12.51 (14.05)	16.61 (45.29)
1	3	0.5	0.6	0.87 (0.34)	0.79 (0.41)	0.98 (0.14)	0.83 (0.38)	0.43 (0.50)	0.82 (0.39)	3.63 (4.37)	13.76 (40.87)
5	1	0.9	1	0.97 (0.17)	0.96 (0.20)	1.00 (0.00)	1.00 (0.00)	0.94 (0.24)	0.99 (0.10)	7.82 (8.40)	102.05 (123.82)
5	1	0.9	0.6	0.80 (0.40)	0.59 (0.49)	0.99 (0.10)	0.99 (0.10)	0.92 (0.27)	0.99 (0.10)	41.32 (36.96)	176.80 (166.64)
5	1	0.5	1	1.00 (0.00)	0.95 (0.22)	1.00 (0.00)	1.00 (0.00)	0.96 (0.20)	1.00 (0.00)	0.04 (0.40)	16.85 (25.66)
5	1	0.5	0.6	0.99 (0.10)	0.94 (0.24)	1.00 (0.00)	0.99 (0.10)	0.87 (0.34)	0.98 (0.14)	1.50 (2.63)	78.01 (118.03)
5	3	0.9	1	0.98 (0.14)	0.94 (0.24)	1.00 (0.00)	1.00 (0.00)	0.97 (0.17)	0.99 (0.10)	18.23 (15.62)	381.76 (382.79)
5	3	0.9	0.6	0.83 (0.38)	0.52 (0.50)	1.00 (0.00)	0.96 (0.20)	0.89 (0.31)	0.95 (0.22)	119.59 (103.24)	578.29 (492.43)
5	3	0.5	1	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.99 (0.10)	0.99 (0.10)	0.04 (0.40)	36.95 (53.05)
5	3	0.5	0.6	0.99 (0.10)	0.97 (0.17)	1.00 (0.00)	0.96 (0.20)	0.90 (0.30)	0.94 (0.24)	3.83 (5.64)	241.18 (261.09)

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

The lasso is tuned to pick at least as many variables as are retained through the reduction phase.

ESTIMATED CONTRASTS ON LOGIT SCALE

Let $\xi_0(A)$ and $\xi_1(A)$ be the probabilities pr(outcome) for variable A at its low and high levels respectively. Let ψ_A be the treatment effect of A, multiplicative on the odds scale[1], where

$$\frac{\xi_1(A)}{1-\xi_1(A)} \triangleq \psi_A \lambda_A, \qquad \frac{\xi_0(A)}{1-\xi_0(A)} \triangleq \lambda_A$$

	A			
outcome	v_{S0}	v_{C0}	ρ	ss
$\mathcal{S} \subseteq \widehat{\mathcal{S}}$	6.573	0.823	0.758	51.83
$\mathcal{S} \in \mathcal{M}$	4.792	0.831	0.773	13.97

Table: Monte Carlo estimates of ψ_A. "ss" = signal strength

[1] Equivalently $\log \psi_A$ is an additive treatment effect on the logit scale

ANOTHER SIMPLE EXPERIMENT: SURVIVAL OUTCOMES

- Slightly larger sample size: $n = 150$ instead of $n = 100$.
- Covariates are generated as in the previous experiment.
- The outcomes are from a PH model with Weibull baseline hazard and exponentially distributed censoring times.
- Fitted by partial likelihood (Cox, 1972; 1975).

MC ESTIMATES: 2^4 FACTORIAL EXPERIMENT (SURVIVAL)

v_{S0}	v_{C0}	ρ	$\frac{\text{signal}}{\text{noise}}$	$\text{pr}(\mathcal{S} \subseteq \widehat{\mathcal{S}})$ undertuned lasso (full)	undertuned lasso (split)	CB (full)	CB (split)	$\text{pr}(\mathcal{S} \in \mathcal{M})$ CB (full)	CB (split)	$\mathbb{E}\lvert\mathcal{M}\setminus\mathcal{S}\rvert$ CB (full)	CB (split)
1	1	0.9	1	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.03 (0.17)	0.95 (0.23)	54.1 (92.2)	1273 (1490)
1	1	0.9	0.6	0.99 (0.12)	0.94 (0.24)	1.00 (0.04)	0.97 (0.17)	0.00 (0.04)	0.89 (0.31)	15.6 (57.3)	1863 (2264)
1	1	0.5	1	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.04 (0.21)	0.95 (0.21)	57.4 (97.4)	962 (1085)
1	1	0.5	0.6	1.00 (0.00)	0.98 (0.13)	0.99 (0.08)	0.96 (0.20)	0.00 (0.00)	0.90 (0.31)	13.0 (31.6)	1734 (2374)
1	3	0.9	1	1.00 (0.00)	0.99 (0.09)	1.00 (0.00)	1.00 (0.04)	0.07 (0.25)	0.95 (0.22)	102 (209)	2468 (2738)
1	3	0.9	0.6	0.97 (0.18)	0.90 (0.30)	0.98 (0.13)	0.95 (0.21)	0.01 (0.09)	0.91 (0.29)	45.0 (98.5)	3182 (3700)
1	3	0.5	1	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.07 (0.25)	0.95 (0.22)	105 (158)	1094 (1090)
1	3	0.5	0.6	1.00 (0.00)	1.00 (0.06)	1.00 (0.00)	0.97 (0.17)	0.00 (0.04)	0.91 (0.28)	18.6 (51.1)	1955 (2859)
5	1	0.9	1	0.98 (0.15)	0.90 (0.30)	1.00 (0.00)	1.00 (0.04)	0.78 (0.41)	0.91 (0.29)	30.9 (46.5)	916 (1165)
5	1	0.9	0.6	0.79 (0.41)	0.52 (0.50)	1.00 (0.00)	0.99 (0.08)	0.59 (0.49)	0.94 (0.24)	136 (180)	2216 (2390)
5	1	0.5	1	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.80 (0.40)	0.91 (0.28)	0.00 (0.09)	59.0 (118)
5	1	0.5	0.6	1.00 (0.00)	0.99 (0.12)	1.00 (0.00)	1.00 (0.04)	0.54 (0.50)	0.90 (0.31)	1.46 (4.22)	382 (572)
5	3	0.9	1	0.98 (0.13)	0.86 (0.35)	1.00 (0.00)	1.00 (0.04)	0.80 (0.40)	0.86 (0.35)	46.4 (66.2)	1383 (1682)
5	3	0.9	0.6	0.71 (0.45)	0.48 (0.50)	1.00 (0.00)	0.99 (0.11)	0.63 (0.48)	0.90 (0.30)	242 (310)	2846 (2603)
5	3	0.5	1	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.83 (0.38)	0.87 (0.34)	0.09 (1.03)	73.4 (175)
5	3	0.5	0.6	1.00 (0.00)	0.99 (0.11)	1.00 (0.00)	1.00 (0.04)	0.59 (0.49)	0.90 (0.30)	2.35 (5.25)	575 (925)

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

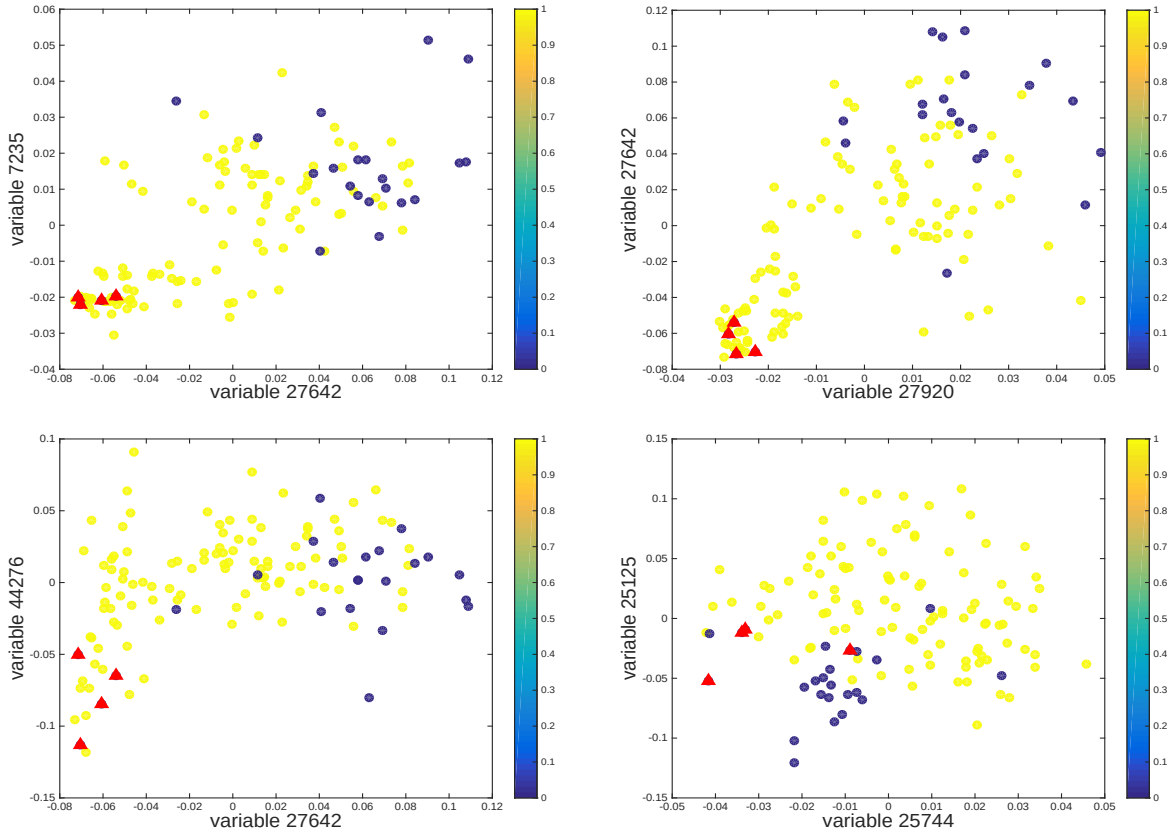
The lasso is tuned to pick at least as many variables as are retained through the reduction phase.

A REAL EXAMPLE

- n =129 individuals: 105 with osteoarthritis, 24 controls.
- $v \sim 50,000$ genetic variables.
- Traverse hypercube: fit a standard (i.e. low-dimensional) linear logistic regression to the corresponding sets of variables[a].
- A set of approx 20 variables results.

[a] For details of decision rules used, see Cox and Battey (2017).

EXPLORATORY PHASE: SOME INTERACTION PLOTS



The majority of suggested interactions can be discarded on the basis of such plots. The above are more strongly suggested.

A SET OF WELL-FITTING MODELS

Model	Recoded variable indices																				
1	1	16	3	8	-	5	-	19	-	-	-	-	-	-	-	-	-	-	-	-	-
2	1	16	3	8	-	5	-	19	-	-	-	-	-	-	-	-	-	-	-	-	2
3	1	16	3	8	-	-	-	-	-	10	-	-	-	-	-	-	-	-	-	4	-	-
4	1	16	3	8	-	5	-	19	-	-	-	-	-	-	-	-	-	-	-	4	-	-
5	1	16	3	8	-	5	-	-	-	10	-	-	-	-	-	-	-	-	-	-	-	-
6	1	16	3	8	-	5	-	19	-	-	-	-	-	-	-	-	-	-	-	-	-	-
7	1	16	3	-	15	5	-	-	-	10	-	-	-	-	-	-	-	-	-	-	-	-
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

- Statistics can take us no further.
- Any choice between such models would require additional data or subject matter expertise.

One *could* construct summaries in an attempt to extract compact messages.

A	B																			
	1	16	3	8	15	5	6	19	13	10	17	12	9	18	11	21	14	7	4	20	2
1		1	0.72	0.22	0.69	0.66	0.75	0.25	0.94	0.09	0.66	0.19	0.16	0	0.03	0.16	0.53	0.28	0.19	0.16	0.03
16	1		0.54	0.68	0.98	0.07	0.41	0.02	0.15	0.66	0.78	0.20	0.15	0.10	0.66	0.07	0.29	0.20	0.41	0.15	0.02
3	0.95	0.90		0.82	0.99	0.31	0.52	0	0.41	0.33	0.49	0.20	0.44	0.28	0.37	0.11	0.23	0.17	0.15	0.09	0.07
8	0.91	0.95	0.88		0.73	0.62	0.64	0.41	0.40	0.35	0.36	0.33	0.23	0.28	0.19	0.28	0.23	0.23	0.15	0.13	0.07
15	0.96	1	0.99	0.74		0.80	0.54	0.63	0.34	0.31	0.18	0.30	0.19	0.16	0.12	0.22	0.13	0.24	0.17	0.12	0.11
5	0.96	0.87	0.56	0.62	0.80		0.56	0	0.45	0.46	0.33	0.31	0.28	0.24	0.40	0.24	0.26	0.13	0.25	0.07	0.06
6	0.98	0.93	0.74	0.70	0.61	0.63		0.43	0.31	0.38	0.26	0.12	0.33	0	0.30	0	0.24	0.12	0.16	0	0.12
19	0.94	0.91	0.57	0.61	0.75	0.33	0.54		0.42	0.41	0.36	0.29	0.29	0.26	0.29	0.22	0.21	0.14	0.18	0.07	0.07
13	1	0.93	0.78	0.65	0.61	0.68	0.51	0.48		0.35	0.28	0.26	0.25	0.22	0.29	0.17	0.19	0.21	0.16	0.12	0.09
10	0.94	0.97	0.76	0.63	0.61	0.69	0.58	0.50	0.37		0.32	0.30	0.35	0.30	0.19	0.23	0.23	0.19	0.11	0.10	0.08
17	0.98	0.98	0.82	0.65	0.55	0.63	0.52	0.47	0.32	0.34		0.32	0.28	0.21	0.22	0.25	0.19	0.20	0.13	0.10	0.09
12	0.95	0.94	0.72	0.64	0.63	0.63	0.44	0.42	0.32	0.34	0.34		0.30	0.23	0.25	0	0.22	0.19	0.16	0.09	0.10
9	0.95	0.93	0.81	0.59	0.57	0.62	0.58	0.43	0.32	0.39	0.30	0.31		0.27	0.23	0.24	0.22	0.19	0.16	0.10	0.08
18	0.94	0.93	0.76	0.63	0.56	0.60	0.38	0.42	0.31	0.36	0.25	0.24	0.27		0.26	0.16	0.22	0.20	0.17	0.11	0.08
11	0.95	0.98	0.80	0.60	0.56	0.70	0.58	0.47	0.39	0.29	0.29	0.30	0.27	0.29		0.23	0.17	0.20	0.16	0.10	0.08
21	0.95	0.93	0.72	0.65	0.62	0.63	0.42	0.43	0.31	0.34	0.33	0.09	0.30	0.22	0.25		0.21	0.18	0.15	0.09	0.09
14	0.97	0.95	0.76	0.63	0.58	0.64	0.56	0.43	0.34	0.35	0.29	0.30	0.29	0.28	0.20	0.22		0.20	0.17	0.11	0.09
7	0.96	0.94	0.75	0.64	0.64	0.58	0.50	0.38	0.36	0.32	0.31	0.28	0.27	0.27	0.24	0.20	0.21		0.16	0	0.08
4	0.96	0.96	0.75	0.61	0.62	0.65	0.54	0.44	0.35	0.29	0.28	0.29	0.28	0.27	0.24	0.21	0.21	0.19		0.10	0.09
20	0.96	0.95	0.75	0.63	0.62	0.60	0.48	0.40	0.35	0.32	0.29	0.27	0.27	0.27	0.22	0.20	0.20	0.09	0.15		0.08
2	0.95	0.94	0.75	0.61	0.62	0.60	0.55	0.40	0.34	0.32	0.30	0.28	0.26	0.26	0.22	0.21	0.20	0.18	0.15	0.10	

Table: proportion of the models in the set of well-fitting models not containing variable B that contain variable A, i.e. $|\mathcal{M}(A \cap \neg B)|/|\mathcal{M}(\neg B)|$, where $\mathcal{M}(\neg B)$ is the set of models in the confidence set that do not contain variable B.

But the whole set of well-fitting models should also be reported.

Variable number (occurrence rate)	Gene name	Description and biological function
7235 (0.96)	ESYT2-007	Tethers the endoplasmic reticulum to the cell membrane. Plays a role in FGF signalling. May play a role in cellular lipid transport.
48433 (0.94)	LTBP1	Latent transforming growth factor beta binding protein. Diseases associated with LTBP1 include geleophysic dysplasia.
25125 (0.75)	PRR5L	Associates with the mTORC2 complex that regulates cellular processes including survival and organization of the cytoskeleton.
29679 (0.61)	-	mRNA.
48415 (0.61)	RP11-542K23.10	RNA Gene.
25744 (0.61)	NDEL1	Plays a role in multiple processes including cytoskeletal organization, cell signaling and neuron migration, outgrowth and maintenance.
27642 (0.53)	SRFBP1	Serum response factor binding protein. May play a role in biosynthesis and/or processing of SLC2A4 in adipose cells.
45991 (0.33)	MAZ	MYC associated zinc finger protein.
36409 (0.31)	SERTAD1	Stimulates E2F1/TFDP1 transcriptional activity.
48549 (0.29)	COL9A2	Collagen type IX alpha 2 chain. Mutations in this gene are associated with multiple epiphyseal dysplasia.
44276 (0.27)	GLS	Plays an essential role in generating energy for metabolism.
33385 (0.26)	LFNG	Encodes evolutionarily conserved glycosyltransferases. Mutations in this gene have been associated with autosomal recessive spondylocostal dysostosis 3.
37443 (0.22)	WDR20	Regulates the activity of the USP12-UAF1 deubiquitinating enzyme complex.
46771 (0.19)	PLAGL2	Zinc-finger protein that recognizes DNA and/or RNA.
27920 (0.18)	ANKRD24	Protein Coding gene.
25470 (0.14)	SPEN	Encodes a hormone inducible transcriptional repressor.
11643 (0.08)	NAT10	Protein coding gene with numerous biological functions.

Table: Gene function is obtained from GeneCards. Variables highlighted in orange have been associated with other bone abnormalities and bone diseases.

BINARY OUTCOMES: A WARNING

- Well-fitting models are found, but not all.
- Modifications are needed for theoretical guarantees.
- Lasso and separating hyperplanes.
- Overlap should not be expected.

SUMMARY

- It is misleading to report one model if statistics is unable to distinguish between many.
- This view is in contraposition to that implicit in the use of the lasso and related methods.
- We have outlined a different approach whose aim is essentially a confidence set of models.
- Any choice between well-fitting models must be based on subject-matter expertise or additional data.

REFERENCES

- Cox, D. R. and Battey, H. S. (2017) Large numbers of explanatory variables, a semi-descriptive analysis, *Proc. Nat. Acad. Sci.*, 114, 8592–8595.
- Battey, H. S. and Cox, D. R. (2018), Large numbers of explanatory variables: a probabilistic assessment, *Proc. R. Soc. Lond. A*, 474.

SOFTWARE

- Matlab code is available from my website.
- An R package `HCmodelSets` has been written by H. H. Hoeltgebaum.