Statistics seminar, University of York

Confidence sets of models and some inferential separations

Heather Battey
Department of Mathematics, Imperial College London

November 26, 2025

Some old ideas from some new perspectives

Bardorff-Nielsen and Cox, 1994, p. 32

Consider a population of individuals and an event A of interest, for instance that an individual dies of heart disease before age 70. . . . Now suppose that a series of new individuals is drawn randomly from the population under study and for each it is required to calculate the probability of event A . . . . If each probability is to be relevant to the individual in question, it must be conditional on observed relevant features, such as age, sex, smoking habits and blood pressure. . . .

. . . Note, however, that, especially if we condition directly, we must limit the conditioning: otherwise we would reach the position where each individual is not only unique, but also uninformative about other individuals . . . .

Where should we limit the conditioning?

Two types of conditioning

- Conditioning by model formulation: conditioning synonymous with specification of the model.
- Technical conditioning: abstract (model+data)-based partitioning of the sample space.

Two types of conditioning

- Conditioning by model formulation: conditioning synonymous with specification of the model.
- Technical conditioning: abstract (model+data)-based partitioning of the sample space.

Fisherian inferential separations specify where to limit the conditioning to ensure relevance while avoiding degeneracy.

Some definitions

Notation

Model for random variable Y parametrised by θ and provisionally assumed true:

$$f_Y(y;\theta) = \prod_{i=1}^{n} f_{Y_i}(y_i;\theta)$$

Arbitrary evaluation point $y = (y_1, \ldots, y_n)$.

Sufficiency reduction, e.g. $s(y) = \sum_i y_i$.

Observed outcome y^o.

Sufficient statistic $S = s(Y)$.

Observed value $s^o = s(y^o)$.

Sufficiency reduction

All information in Y relevant for inference on θ is encapsulated in $S = s(Y)$.

$$f_Y(y;\theta) = \prod_{i=1}^{n} f_{Y_i}(y_i;\theta) = g(s(y);\theta)h(y)$$

Take S to be minimal sufficient, i.e. of lowest dimension.

Minimal sufficiency

Let d be the dimension of S. Let d_θ be the dimension of θ.

If $d > d_\theta$, then any estimator of θ must sacrifice information on θ by the definition of minimal sufficiency.

A useful starting point: determine a one-to-one transformation of the minimal sufficient statistic $S \cong (\hat{\theta}, A)$ where A is an ancillary statistic.

Without loss of generality, take $S = (\hat{\theta}, A)$.

The ancillary/co-ancillary separation

Separations within the minimal sufficient statistic

Likelihood function depends on the data only through S.

Realisable separation $S = (\hat{\theta}, A)$.

Notional idealised separation $S = (C(A), A)$.

Separates the information in S into components of dimensions d_θ and d_A without loss or redundancy.

Notional idealised separation

Notional idealised separation $S = (C(A), A)$.

Ancillary A; "maximal co-ancillary" $C(A)$

$$C(a^o) \stackrel{d}{=} S \mid \{A = a^o\}.$$

The observed value $a^o = a(y^o) = a(s^o)$ leaves $d_\theta = d - d_A$ degrees of freedom of variation of S consistent with the constraint $a(s) = a^o$.

Think of $C(a^o)$ as having a distribution on the d_θ-dimensional co-ancillary manifold:

$$\mathcal{C}(a^o) = \{s \in \mathbb{R}^d : a(s) = a^o\} \subset \mathbb{R}^d.$$

Ancillary statistic A

Ancillary A is defined through its properties w.r.t. θ.

Several property-based definitions have been put forward of varying stringency (e.g. B-N & Cox, 1994, p. 38).

Idealised situation: distribution of A does not depend on θ.

That does not mean that A is irrelevant for inference on θ (A is part of the minimal sufficient statistic).

It means that A, by itself, carries no info on the value of θ.

A vague but practically useful definition

Ancillary statistic: *A is ancillary for θ if, from observation of A alone, no information about the value of θ can in general be extracted.*

Implicit definition used by Fisher.

Relevance through conditioning

The conditioning event $\{A = a^o\}$ isolates hypothetical samples for which $s^o = (\hat{\theta}^o, a^o)$ is one realisation, retaining only the variability in S that is relevant for determining the horizontal position of the normed log-likelihood function, rather than its shape, the latter being fixed by $\{A = a^o\}$.

Hypothetical replication

Inferential statements about θ inevitably involve hypothetical replication.

Two samples of the same size can produce log-likelihood functions that differ appreciably in shape, and yet are maximized at the same point.

Example: linear regression. Relevant precision characterised by $X^{\mathrm{T}}X$, not $\mathbb{E}(X^{\mathrm{T}}X)$: $X^{\mathrm{T}}X$ is ancillary when X is considered random.

The ancillary A separates samples of the same size according to their information content.

An exact conditional analysis with nuisance parameters

2×2 table in original and standardised form

	0 failure	1 success	
0 control	$N_{0\|0}$	$N_{1\|0}$	$N_{\cdot\|0}$
1 treated	$N_{0\|1}$	$N_{1\|1}$	$N_{\cdot\|1}$
	$N_{0\|\cdot}$	$N_{1\|\cdot}$	N

	0 failure	1 success	
0 control	$\hat{p}_{0\|0}$	$\hat{p}_{1\|0}$	$\hat{p}_{\cdot\|0}$
1 treated	$\hat{p}_{0\|1}$	$\hat{p}_{1\|1}$	$\hat{p}_{\cdot\|1}$
	$\hat{p}_{0\|\cdot}$	$\hat{p}_{1\|\cdot}$	1

Degrees of freedom for 2×2 table

	0 failure	1 success	
0 control			
1 treated			
			1

If the row and column totals are ignored, there are three degrees of freedom for variation of the entries of the table: $(\hat{p}_{0|0}, \hat{p}_{1|0}, \hat{p}_{0|1}, \hat{p}_{1|1})$ belong to the unit simplex in $\mathbb{R}^4$.

Degrees of freedom for 2×2 table

	0 failure	1 success	
0 control			$\hat{p}_{\cdot\|0}$
1 treated			$\hat{p}_{\cdot\|1}$
			1

Knowledge of (one of the) row totals leaves 2 degrees of freedom for how the table can be filled in.

Degrees of freedom for 2×2 table

	0 failure	1 success	
0 control			$\hat{p}_{\cdot\|0}$
1 treated			$\hat{p}_{\cdot\|1}$
	$\hat{p}_{0\|\cdot}$	$\hat{p}_{1\|\cdot}$	1

Knowledge of row and column totals leaves 1 degree of freedom for how the table can be filled in.

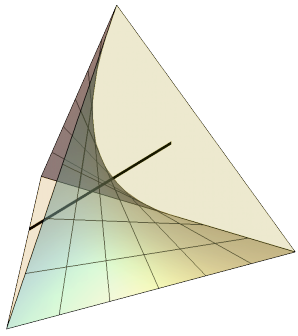
Conditioning in the 2×2 table

	0 failure	1 success	
0 control	$\hat{p}_{0\|0}$	$\hat{p}_{1\|0}$	$\hat{p}_{\cdot\|0}$
1 treated	$\hat{p}_{0\|1}$	$\hat{p}_{1\|1}$	$\hat{p}_{\cdot\|1}$
	$\hat{p}_{0\|\cdot}$	$\hat{p}_{1\|\cdot}$	1

Fisher argued that is it appropriate to condition on row and column totals in the analysis, these being ancillary.

After conditioning, the values of $(\hat{p}_{0|0}, \hat{p}_{1|0}, \hat{p}_{0|1}, \hat{p}_{1|1})$ have a distribution constrained to a one-dimensional subspace of the unit simplex.

Geometric exposition of Fisher's (1922) conditional analysis



Curved manifold (Feinberg & Gilbert, 1970): the set of true multinomial probabilities consistent with independence of the two binary variables.

Black line (co-ancillary manifold): constraint within the simplex (sample space for the standardised table) imposed by the marginal totals $\hat{p}_{1|\bullet} = 0.6$, $\hat{p}_{\bullet|1} = 0.4$.

Fisher's analysis: based on the distribution of $(\hat{p}_{0|0}, \hat{p}_{1|0}, \hat{p}_{0|1}, \hat{p}_{1|1})$ constrained to the line.

An example with many nuisance parameters (Cox, 1958)

One individual from each of n pairs is randomised to treatment, the other is the untreated control. Pairwise table:

	0 failure	1 success	
0 control			1
1 treated			1
			2

The design fixes the row totals.

Logistic model for the probabilities

Binary outcomes on n matched pairs. For the ith pair the model is

$$p_{1|0}^{(i)} = \text{pr}(\text{success} \mid \text{control}) = \frac{e^{\alpha_i}}{1+e^{\alpha_i}}, \qquad p_{0|0}^{(i)} = 1 - p_{1|0}^{(i)}$$

$$p_{1|1}^{(i)} = \text{pr}(\text{success} \mid \text{treated}) = \frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}, \quad p_{0|1}^{(i)} = 1 - p_{1|1}^{(i)}$$

The logistic model is intermediate between a general multinomial representation and one in two independent binomials.

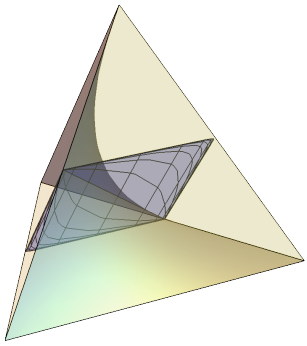
Remarks on the formulation

Allowing one nuisance parameter per pair encapsulates arbitrary dependence on (perhaps unmeasured) covariates.

Looks more restrictive, but it is much more general than a typical semiparametric model.

The problem of assessing a null treatment effect (or unit odds ratio) is broadly analogous (geometrically) to assessing independence in a pure contingency table, but the interpretation is different.

Logistic parametrisation of matched pair problem



Flat plane: subspace compatible with row totals $(\frac{1}{2}, \frac{1}{2})$ from matched pair design.

Curved contours of plane contours of equal β in the logistic parametrisation $(\alpha, \beta) \mapsto e^{\alpha+\beta}/(1 + e^{\alpha+\beta}) = \text{pr}(\text{success}|\text{treated})$.

Four possible pairwise tables

Because there are pair-specific nuisance parameters, we start by considering n separate pairwise tables. Four possibilities:

	F	S
C	1	0
T	1	0

	F	S
C	1	0
T	0	1

	F	S
C	0	1
T	1	0

	F	S
C	0	1
T	0	1

Number of tables of each type: R^{00}, R^{01}, R^{10}, R^{11}.

Four possible pairwise tables

Because there are pair-specific nuisance parameters, we start by considering n separate pairwise tables. Four possibilities:

	F	S
C	1	0
T	1	0

	F	S
C	1	0
T	0	1

	F	S
C	0	1
T	1	0

	F	S
C	0	1
T	0	1

Number of tables of each type: R^{00}, R^{01}, R^{10}, R^{11}.

Four possible pairwise tables

Because there are pair-specific nuisance parameters, we start by considering n separate pairwise tables. Four possibilities:

	F	S
C	1	0
T	1	0

	F	S
C	1	0
T	0	1

	F	S
C	0	1
T	1	0

	F	S
C	0	1
T	0	1

Number of tables of each type: R^{00}, R^{01}, R^{10}, R^{11}.

Four possible pairwise tables

Because there are pair-specific nuisance parameters, we start by considering n separate pairwise tables. Four possibilities:

	F	S
C	1	0
T	1	0

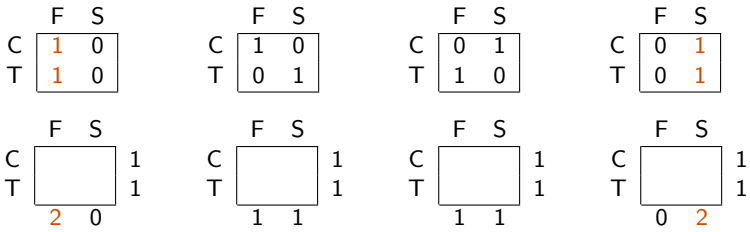
	F	S
C	1	0
T	0	1

	F	S
C	0	1
T	1	0

	F	S
C	0	1
T	0	1

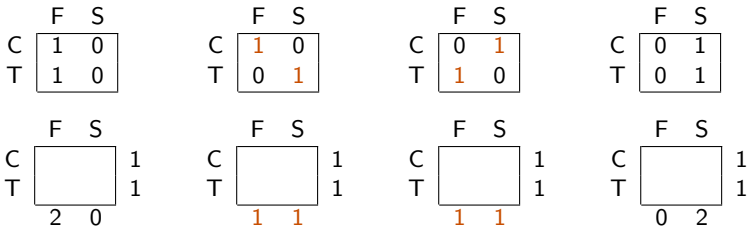
Number of tables of each type: R^{00}, R^{01}, R^{10}, R^{11}.

Four possible pairwise tables



In the leftmost and rightmost tables (concordant pairs), conditioning on column totals leaves no degrees of freedom.

Four possible pairwise tables



In the leftmost and rightmost tables (concordant pairs), conditioning on column totals leaves no degrees of freedom.

In the two inner tables (discordant pairs) there remains one degree of freedom after conditioning.

Conditional analysis based on discordant pairs

Conditioning in the pairwise tables leads us to discard concordant pairs.

- R^{01} tables of type $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ contribute $\begin{bmatrix} R^{01} & 0 \\ 0 & R^{01} \end{bmatrix}$
- R^{10} tables of type $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ contribute $\begin{bmatrix} 0 & R^{10} \\ R^{10} & 0 \end{bmatrix}$

Discordant pair table:

	F	S	
C	R^{01}	R^{10}	m
T	R^{10}	R^{01}	m
	m	m	

Conditional on row and column totals $m = R^{01} + R^{10}$

$$R^{01} \sim \text{Bin}(m, e^{\beta}/(1 + e^{\beta})).$$

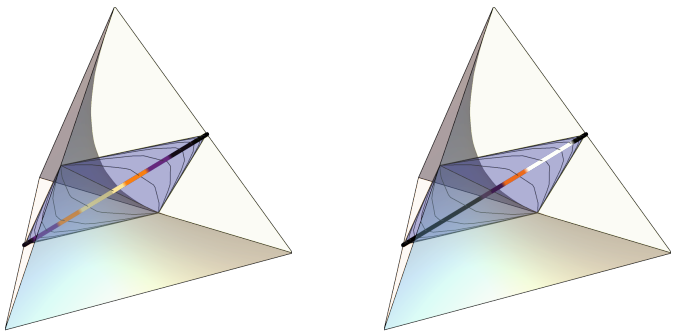
Have eliminated all the nuisance parameters $\alpha_1, \ldots, \alpha_n$.

A little more detail

Let T_i and C_i be the binary outcomes on the treated and untreated individuals respectively.

$$\text{pr}(T_i = 1, C_i = 0 \mid \underbrace{T_i + C_i = 1}_{\text{discordant pair}}) = \frac{\left(\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}\right)\left(\frac{1}{1+e^{\alpha_i}}\right)}{\left(\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}\right)\left(\frac{1}{1+e^{\alpha_i}}\right) + \left(\frac{e^{\alpha_i}}{1+e^{\alpha_i}}\right)\left(\frac{1}{1+e^{\alpha_i+\beta}}\right)} = \underbrace{\frac{e^{\beta}}{1+e^{\beta}}}_{\text{no nuisance param}}$$

Binomial distribution on the co-ancillary manifold



Induced discrete distributions on the co-ancillary manifold $C(a^o)$ (straight line) corresponding to $\beta = 0$ (left) and $\beta = 2$ (right) from $m = 7$ discordant pairs.

Two roles of conditioning

The example illustrates two roles of conditioning:

- Relevance (because the conditioning statistics are ancillary for the interest parameter in the broad sense).
- Elimination of nuisance parameters.

The sufficiency/co-sufficiency separation

Bardorff-Nielsen and Cox, 1994, p. 29

The motivation for regarding sufficiency as important is that it represents a separation of the information in the data into two types, that concerned with inference about θ given the model and that concerned with adequacy of the model. To make this separation vivid, consider the following.

1. *Suppose that the investigator observes that $S = s$. Then some inference can be drawn about θ, assuming the adequacy of the family and using in some way the distribution of S as a function of θ.*
2. *Suppose in a second stage that the investigator learns that $Y = y$. The additional information in the second stage is derived in effect by observing one realization of the conditional distribution of Y given $S = s$. Since this distribution does not involve θ it can throw no additional light on the value of θ. If, however, the observation is extreme in some relevant sense it can throw doubt on the adequacy of the family.*

Notional idealised separation

Let $d < n$ be the dimension of the minimal sufficient statistic.

Notional idealised separation: $Y \cong (S, Q(S))$.

The "co-sufficient statistic" $Q(s^o)$ has the distribution of Y (or some one-to-one transformation thereof) given $S = s^o$.

The observed value $s^o = s(y^o)$ leaves $n - d$ degrees of freedom for variation of y consistent with the constraint $s(y) = s^o$.

Think of $Q(s^o)$ as having a distribution on the co-sufficient manifold

$$\mathcal{Q}(s^o) = \{y \in \mathbb{R}^n : s(y) = s^o\} \subset \mathbb{R}^n.$$

The manifold $\mathcal{Q}(s^o)$ in canonical exponential family regression

Regression model for outcomes $Y = (Y_1, \ldots, Y_n)$. Conditional density or mass function at $y = (y_1, \ldots, y_n)$:

$$f(y; x_1^{\mathrm{T}}\theta, \ldots, x_n^{\mathrm{T}}\theta) = \exp\Big[\phi^{-1}\Big\{\theta^T \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} K(x_i^T\theta)\Big\}\Big] \prod_{i=1}^{n} h(y_i, \phi^{-1}),$$

Sufficient statistic for θ assuming ϕ known: $S = \sum_{i=1}^{n} x_i Y_i = X^{\mathrm{T}} Y$.

The manifold $\mathcal{Q}(s^o)$ in canonical exponential family regression

Sufficient statistic for θ assuming ϕ known: $S = \sum_{i=1}^n x_i Y_i = X^{\mathrm{T}} Y$.

The normal directions to $\mathcal{Q}(s^o) \subset \mathbb{R}^n$ at y^o are specified by

$$\left.\frac{\partial s^{\mathrm{T}}(y)}{\partial y}\right|_{y=y^o} = X.$$

No dependence on y_o, therefore $\mathcal{Q}(s^o)$ is flat and spanned by an orthogonal basis for $\mathcal{X}^\perp = \{v \in \mathbb{R}^n : v^{\mathrm{T}} x = 0, x \in \mathcal{X}\}$, where $\mathcal{X} = \text{col-span}(X)$.

Modern relevance?

Context

- Regression, broadly defined. Dimension $p \gg$ study individuals n.
- Aim: scientific understanding.
- In many genomics contexts an assumption of sparsity is natural.
- Popular approaches based on penalised regression produce a single model.
- There are often several or many models that fit the data equivalently well.

Simple low-dimensional example

$$
\begin{aligned}
Y = X\beta + \varepsilon, \quad \varepsilon &= (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}, \quad \varepsilon_i \sim N(0,1), \quad n = 100, \\
\beta &= (1, 1, 0, \ldots, 0)^{\mathrm{T}} \in \mathbb{R}^p, \qquad\qquad p = 25.
\end{aligned}
$$

Rows of X drawn from $N_p(0, \Sigma)$: high correlation between first three variables.

Comprehensive model: $[p] := \{1, \ldots, p\}$; true model: $\mathcal{S} = \{1, 2\}$.

Lasso (with cross-validated tuning) selects a single model: $\{2\}$.

A likelihood-ratio test of each low-dimensional submodel $\mathcal{S}_m \subset \{1, \ldots, p\}$ against $[p]$ declared $\{2\}, \{1, 2\}$, $\{2, 3\}$ and $\{1, 2, 3\}$ as statistically indistinguishable from $[p]$.

Confidence set of models: $\mathcal{M} = \{\{2\}, \{1, 2\}, \{2, 3\}, \{1, 2, 3\}\}$

Low-dimensional formulation

- Full set of variables: $[p] = \{1, \ldots, p\}$. True model: $\mathcal{E}^* \subset [p]$.
  \# of possible models $= 2^p$.
- Sparsity $\Longrightarrow$ consider models with fewer variables, i.e. models of size s or less, of which there are $\bar{m} = \sum_{v=1}^{s} \binom{p}{v}$.
- For every possible model $\mathcal{E}_m$, $m = 1, \ldots, \bar{m}$, test $\mathcal{E}_m \subset [p]$ against $[p]$ by a likelihood ratio test. Confidence set of models:

$$\mathcal{M}_\alpha = \{\text{all models not rejected by a LR test of size } \alpha\}$$

- By standard arguments $\text{pr}(\mathcal{E}^* \in \mathcal{M}_\alpha) \to 1 - \alpha$.

High-dimensional complications

- Likelihood ratio test is infeasible when $p > n$.

  Natural resolution: eliminate variables that appear to have no effect. Conservative reduction procedures come with theoretical guarantees.

- But: the reduced set $\hat{\mathcal{E}}$ has been selected in the light of the data. Fits the data better than an arbitrary model of the same size.

  Implication: a likelihood ratio test of any subset of variables $\mathcal{A} \subset \hat{\mathcal{E}}$ rejects $\mathcal{A}$ too often in hypothetical repeated use. Thus

$$\lim_{n \to \infty} \text{pr}(\mathcal{E}^* \in \mathcal{M}_\alpha) \ll 1 - \alpha$$

A simple resolution: split the sample

- Sample splitting means that the reduced set of variables $\hat{\mathcal{E}}$ is independent of the data subsequently fitted to the variables indexed by $\hat{\mathcal{E}}$.

  The usual distribution theory for the likelihood ratio statistic is restored.

- But information is sacrificed at reduction and model assessment phases.

  The Fisherian inferential separations suggest a better way.

Example: high-dimensional linear regression

Formulation

- Outcome vector $Y \in \mathbb{R}^n$ normally distributed of mean $\mu = X\theta \in \mathcal{X} \subset \mathbb{R}^n$.

$$Y = X\theta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n). \tag{1}$$

- There is a larger matrix Z with column dimension $p \gg n$ such that $\mu = Z\beta \in \mathcal{X}$, where β has the same non-zero entries as θ in the positions $\mathcal{E}^*$ corresponding to X, and is zero elsewhere.
- Preliminary reduction to a set $\hat{\mathcal{E}}$: the resulting model

$$Y = X_{\hat{\mathcal{E}}}\theta_{\hat{\mathcal{E}}} + \varepsilon$$

fits better than an arbitrary model of the same size embedding (1) or any other lower-dimensional model to the tested.

Minimal sufficient statistic

$$\text{True model : } Y = X\theta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n).$$

Minimal sufficient statistic for θ when σ^2 is unknown is

$$S = (X^{\mathrm{T}} Y, \hat{\varepsilon}^{\mathrm{T}} \hat{\varepsilon}) \cong (\hat{\theta}, \hat{\varepsilon}^{\mathrm{T}} \hat{\varepsilon}) \cong (X\hat{\theta}, \hat{\varepsilon}^{\mathrm{T}} \hat{\varepsilon})$$

if X is treated as fixed, where $\hat{\theta}$ is the ordinary least squares estimator of θ and $\hat{\varepsilon}^{\mathrm{T}} \hat{\varepsilon}$ is the residual sum of squares.

The first conditioning: projection onto $\mathcal{X}^\perp$

- Variation in $Y \in \mathbb{R}^n$ conditional on $S = s^o$?
- Subspace $\mathcal{X}$: the d_θ-dimensional space spanned by the columns of X. Orthogonal complement:

$$\mathcal{X}^\perp = \{v \in \mathbb{R}^n : v^{\mathrm{T}} x = 0, x \in \mathcal{X}\} \subset \mathbb{R}^n.$$

  No loss or redundancy: $\mathcal{X} \oplus \mathcal{X}^\perp = \mathbb{R}^n$.
- Variation in $Y \in \mathbb{R}^n$ conditional on $X^{\mathrm{T}} Y = X^{\mathrm{T}} y^o$ lives in $\mathcal{X}^\perp$: a $(n - d_\theta)$-dimensional subspace of $\mathbb{R}^n$.
- $\mathcal{X}^\perp$ is spanned by e.g. the $n - d_\theta$ eigenvectors corresponding to the non-null eigenvalues of the projection matrix $M = I - X(X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}}$.
- Let U be this orthonormal matrix of eigenvectors. Then

$$UU^{\mathrm{T}} = M, \quad U^{\mathrm{T}} U = I_{n-d_\theta}.$$

The second conditioning

- Consider $W = U^{\mathrm{T}} Y$. This is normally distributed of zero mean and variance $\sigma^2 I_{n-d_\theta}$. It captures the variation in Y orthogonal to that in $\hat{\theta}$: conditioning on $\hat{\theta}$ is equivalent to projection onto $\mathcal{X}^\perp$ in this case. We also need to condition on $W^{\mathrm{T}} W = \hat{\varepsilon}^{\mathrm{T}} \hat{\varepsilon}$.
- Transform $W \mapsto (W/\|W\|, W^{\mathrm{T}} W) =: (Q, R^2)$. It can be shown that Q and R^2 are independent, so conditioning on $\hat{\varepsilon}^{\mathrm{T}} \hat{\varepsilon}$ produces the same conditional distribution as the marginal distribution of Q: this is our co-sufficient statistic.
- Q is uniformly distributed on the unit hypersphere embedded in $\mathcal{X}^{\mathrm{T}}$.

Implications

- The previous conclusion is true only if we project onto $\mathcal{X}^{\perp}$. If the postulated model is wrong, we project on the wrong space.
- Force a contradiction: for every subset of variables $\mathcal{E}_m \subset \hat{\mathcal{E}}$, project on the corresponding $\mathcal{X}_m^{\perp}$, and calculate the corresponding $Q^{(m)}$. If it is not uniformly distributed on the unit hypersphere, that is evidence against the model.
- This avoids any comparison of fit relative to the larger model $\hat{\mathcal{E}}$.
- A problem: from one observation of Q there is no power to reject a false model $\mathcal{E}_m$.

Synthetic replication

- In the vein of Rasines and Young (2023), generate pseudo-replicates of Y with common distribution.
- Let L be an $n \times (k-1)$ matrix of independent $N(0,1)$ entries. There exists a $k \times k$ deterministic matrix Γ such that

$$[\tilde{Y}^{(1)} \dots \tilde{Y}^{(k)}] = [Y \; L]\Gamma \overset{\text{indep}}{\sim} N(\mu, k\sigma^2 I_n).$$

- These yield independent pseudo-replicates of the co-sufficient statistics $\tilde{Q}^{(1)}, \dots, \tilde{Q}^{(k)}$, uniformly distributed on the unit hypersphere in $\mathcal{X}^{\perp}$ if and only if the postulated model is the correct one.

Confidence sets of models: theoretical guarantees

- An α-level confidence set of models $\mathcal{M}$ is all low-dimensional subsets of variables in $\hat{\mathcal{E}}$ for which a test of uniformity of $\tilde{Q}^{(1)}, \ldots, \tilde{Q}^{(k)}$ does not reject at level α.
- Provided that $\mathcal{E}^* \subset \hat{\mathcal{E}}$, i.e. all variables in the true model survive reduction, $\text{pr}(\mathcal{E}^* \in \mathcal{M}) = 1 - \alpha$.

Usage of confidence sets of models

Confidence sets usually contain large numbers of models when $p \gg n$.
This is an honest reflection of the information in the data.

Any choice between statistically indistinguishable models requires either additional data or subject-matter expertise.

Compact messages can be extracted. In the example of Cox & Battey (2017):

- Two variables, v_1 and v_2, are present in 96% and 94% of models.
- In 78% of the models in which v_2 is absent, another variable, v_3, is present in its place.
- Only 1% of models include neither v_2 nor v_3.

Some references

- Original proposal of confidence set of models emphasising conceptual aspects:
  - Cox, D. R. and Battey. H. S. (2017). Large numbers of explanatory variables: a semi-descriptive analysis. *Proc. Nat. Acad. Sci.*, 114 (32), 8592–8595
- Some theory:
  - Battey, H. S. and Cox, D. R. (2018). Large numbers of explanatory variables: a probabilistic assessment. *Proc. Roy. Soc. Lond. A: Math. Phys. Sci.*, 474, 20170631.
  - Lewis, R. and Battey, H. S. (2025). Cox reduction and confidence sets of models: a theoretical elucidation. *Statist. Sci.*, 40, 313–328.
- Post-reduction inference:
  - Battey, H. S., Rasines, D. G. and Tang, Y. (2025). Post-reduction inference for confidence sets of models. *arXiv: 2507.2507.10373*.
- Based on the geometric perspectives in:
  - Battey, H. S. (2024). Maximal co-ancillarity and maximal co-sufficiency. *Information Geometry*, 7, 355–369.

  and the randomised inference perspective in
  - Rasines, D. G. and Young, G. A. (2023). Splitting strategies for post-selection inference. *Biometrika*, 110, 597–614.

The end