Maximal co-ancillarity and maximal co-sufficiency

Heather Battey
Department of Mathematics, Imperial College London

EPFL, June 14, 2024

The problem of conditioning

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 . . . .*

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

(If S is minimal sufficient, then so is $S' = (\hat{\theta}, A) \cong S$, so 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" $C(A)$

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

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.*

This appears to be the implicit definition used by Fisher.

Formalised constructions along these lines have been proposed
e.g. Barndorff-Nielsen (1973). On *M*-ancillarity. *Biometrika*, 60, 447–455.

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	1	
	failure	success	
0 control	$N_{0\mid 0}$	$N_{1\mid 0}$	$N_{\bullet\mid 0}$
1 treated	$N_{0\mid 1}$	$N_{1\mid 1}$	$N_{\bullet\mid 1}$
	$N_{0\mid\bullet}$	$N_{1\mid\bullet}$	N

	0	1	
	failure	success	
0 control	$\hat{p}_{0\mid 0}$	$\hat{p}_{1\mid 0}$	$\hat{p}_{\bullet\mid 0}$
1 treated	$\hat{p}_{0\mid 1}$	$\hat{p}_{1\mid 1}$	$\hat{p}_{\bullet\mid 1}$
	$\hat{p}_{0\mid\bullet}$	$\hat{p}_{1\mid\bullet}$	1

Degrees of freedom for 2×2 table

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

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

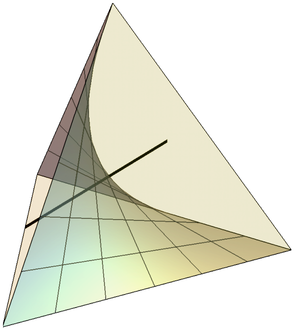
Conditioning in the 2×2 table

	0	1	
	failure	success	
0 control	$\hat{p}_{0\mid 0}$	$\hat{p}_{1\mid 0}$	$\hat{p}_{\bullet\mid 0}$
1 treated	$\hat{p}_{0\mid 1}$	$\hat{p}_{1\mid 1}$	$\hat{p}_{\bullet\mid 1}$
	$\hat{p}_{0\mid \bullet}$	$\hat{p}_{1\mid \bullet}$	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 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	1	
	failure	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 *i*th pair the model is

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

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

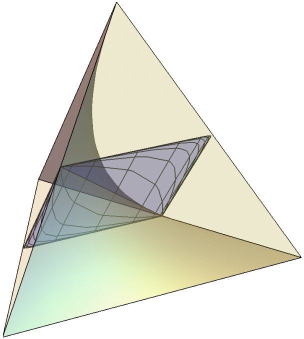
Remarks on the formulation

Allowing one nuisance parameter per pair encapsulates arbitrary covariate dependence.

In that context, the implicit assumption is that the outcomes on treated and untreated individuals are conditionally independent given (unmeasured) covariates and treatment indicator.

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 differs considerably.

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

	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

	F	S	
C			1
T			1
	2	0	

	F	S	
C			1
T			1
	1	1	

	F	S	
C			1
T			1
	1	1	

	F	S	
C			1
T			1
	0	2	

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

Four possible pairwise tables

	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

	F	S	
C			1
T			1
	2	0	

	F	S	
C			1
T			1
	1	1	

	F	S	
C			1
T			1
	1	1	

	F	S	
C			1
T			1
	0	2	

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

Discordant pair table:

$$\begin{array}{c|cc|c} & F & S & \\ \hline C & R^{01} & R^{10} & m \\ T & R^{10} & R^{01} & m \\ \hline & m & m & \end{array}$$

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

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

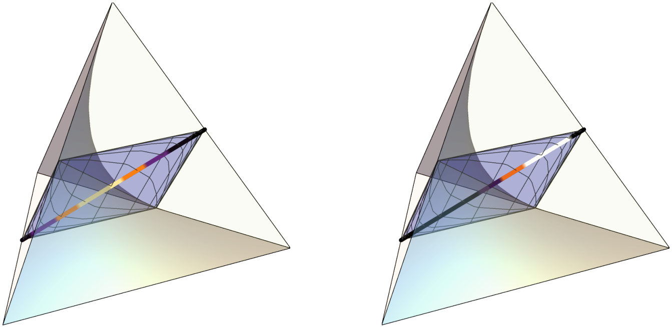
Have eliminated all the nuisance parameters $\alpha_1, \ldots, \alpha_n$.

A little more detail

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

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

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.

Approximate conditional inference (1/2)

Practical implementation of conditional inference is typically more difficult and context-dependent.

Exact conditional inference is not always available.

Can the idealised analysis be achieved approximately?

Two objects potentially requiring approximation: the co-ancillary manifold $\mathcal{C}(a^o)$; the distribution of $C(a^o)$ on the (approximate) co-ancillary manifold.

Approximate conditional inference (2/2)

Two objects potentially requiring approximation: the co-ancillary manifold $\mathcal{C}(a^o)$; the distribution of $C(a^o)$ on the (approximate) co-ancillary manifold.

Approximate conditional inference via the tangent exponential model (Fraser, 1988, 1990; Fraser and Reid, 1988; Davison and Reid, 2024) seems in this vein, but. . .

Connection not immediately transparent. No preliminary reduction by sufficiency.

The geometric underpinnings of p^* in curved exponential families (Barndorff-Nielsen, 1980) appear compatible.

The sufficient/co-sufficient 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 $y = (y_1, \ldots, y_n)$:

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

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 $\mathcal{X} = \text{col-span}(X)$.

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

$\mathcal{Q}(s^o) \subset \mathbb{R}^n$ is the $(n-d)$-dimensional subspace spanned by the columns of U, a matrix of eigenvectors of the projection matrix $I - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$.

$$UU^{\mathrm{T}} = I - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}, \quad U^{\mathrm{T}}U = I_{n-d}.$$

The distribution of $Q(s^o)$ on $\mathcal{Q}(s^o)$ is that induced by the distribution of Y after projection on $\mathcal{Q}(s^o)$.

Conditioning is equivalent to projection in this example.

Linear regression with unknown dispersion

Suppose now that, in the context of a linear regression, the dispersion parameter σ^2 is unknown.

Minimal sufficient statistic $S \cong (X^{\mathrm{T}}Y, \hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon})$ where $\hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon} = Y^{\mathrm{T}}UU^{\mathrm{T}}Y$ is the residual sum of squares.

Previous notation needs adjusting on account of S having been enlarged: Let $\mathcal{W}(X^{\mathrm{T}}y^o)$ be the $(n - d_\theta)$-dimensional flat manifold from the conditioning event $X^{\mathrm{T}}Y = X^{\mathrm{T}}y^o$, the space to which $W := X^{\mathrm{T}}Y$ belongs.

Further conditioning $\hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon} = (\hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon})^o = (y^o)^{\mathrm{T}}UU^{\mathrm{T}}y^o$ fixes the length of W and defines a spherical hypersurface within $\mathcal{W}(X^{\mathrm{T}}y^o) \subset \mathbb{R}^n$. This is $\mathcal{Q}(s^o)$.

Connection to maximal invariant tests (1/2)

We may equivalently consider the statistic of unit length $V := U^{\mathrm{T}} Y / \| U^{\mathrm{T}} Y \|$.

Then $\mathcal{Q}(s^o)$ is the $(n - d_\theta)$-dimensional unit sphere.

The statistic V is the maximal invariant used in the invariant testing literature, and is uniformly distributed on the surface of the unit sphere under correct specification of the model.

A practical difficulty: all possible realisations of V are compatible with uniformity on the sphere even when the model assumption is violated.

Connection to maximal invariant tests (2/2)

The Fisherian falsification framework fails here without some form of pseudo-replication (power to detect departures from uniformity on the sphere requires at least two observations).

Tests based on departures from the hypothesised mean model in particular directions are essentially equivalent to an F test of $\gamma = 0$ in the extended model $Y = X\theta + Z\gamma + \varepsilon$ (all invariant tests are functions of the data only through the maximal invariant).

Connection to ongoing research

At least one context where specification of an alternative direction is ideally to be avoided: post-selection inference for confidence sets of models, crudely addressed by sample splitting in Battey and Cox (2018). Ongoing work. . .

The end