Information Geometry (2024) 7:355–369
https://doi.org/10.1007/s41884-024-00144-1

PERSPECTIVE

# Maximal co-ancillarity and maximal co-sufficiency

**H. S. Battey**[1]

Received: 10 June 2024 / Revised: 10 August 2024 / Accepted: 12 August 2024 /
Published online: 17 September 2024
© The Author(s) 2024

**Abstract**
The purpose of this exposition is to provide some new perspectives on conditional inference through a notional idealised separation within the minimal sufficient statistic, allowing a geometric account of key ideas from the Fisherian position. The notional idealised separation, in terms of an ancillary statistic and what I call a maximal co-ancillary statistic, provides conceptual insight and clarifies what is sought from an approximate conditional analysis, where exact calculations may not be available. A parallel framework applies in the Fisherian assessment of model adequacy. Both aspects are discussed and illustrated geometrically through examples.

**Keywords** Ancillary · Conditional inference · Inferential separations · Information · Minimal sufficiency · Model adequacy

## 1 Introduction

The question of where to limit the conditioning emerges in many guises. Barndorff-Nielsen and Cox [7, p. 32] gave a simple motivating example concerning the probability that an individual dies of heart disease before the age of 70, a large data base of past cases allowing estimation of the probability of interest. Relevance of such a calculation for any new individual requires that it is conditional on intrinsic features such as sex, smoking habits and so on. In principle, such conditioning could be extended to tens of thousands of genetic traits but this has to be limited, otherwise a situation is reached in which each individual is unique and uninformative about others.

The previous situation exemplifies the more tangible form of conditioning synonymous with specification of a suitable regression model. Standard terminology of conditional inference instead refers to a more abstract conditioning in which the observed data, in combination with a model (provisionally assumed true) dictate an abstract partitioning of the sample space to be used in inference. Cox [20] refers to the

Communicated by Tomonari Sei.

✉ H. S. Battey
h.battey@imperial.ac.uk

[1] Department of Mathematics, Imperial College London, 180 Queen's Gate, London SW7 2AZ, UK

two types as *conditioning by model formulation* and *technical conditioning* respectively. In both cases, Fisherian inferential separations, when available, specify where to limit the conditioning to ensure relevance while avoiding degeneracy. It is these separations and associated ideas that I aim to elucidate here, with an emphasis on geometric insight through two notional random variables which I call the maximal co-ancillary statistic and the maximal co-sufficient statistic. The latter terminology expands on one used by Barber and Jansen [2], who introduced an approach for sampling from the conditional distribution of an outcome variable given the realised value of a sufficient statistic, referring to this as *co-sufficient sampling*. While conditioning also plays an important role in the elimination of nuisance parameters, this is conceptually distinct from its core role in Fisherian inference, although the two motivations sometimes give rise to the same operational procedures.

The only novelty in the present work comes from seeing an old problem from a new perspective, which has some advantages of conceptualisation relative to more algebraic approaches.

Besides some brief remarks in Sects. 2.2 and 4, no attempt is made to survey a rather large literature on approximate conditional inference, to which important early contributions were made, inter alia, by Efron and Hinkley [26], Barndorff-Nielsen and Cox [6], Barndorff-Nielsen [4, 5, 18], McCullagh [38], Fraser [32, 33], Fraser and Reid [34], and Skovgaard [43]. The account serves primarily to provide insight into the foundations of Fisherian inference, and partly to illustrate what is sought from an approximate analysis when exact calculations are unavailable.

## 2 Ancillarity and maximal co-ancillarity

### 2.1 Separations within the minimal sufficient statistic

Consider a model for a vector of independent or conditionally independent random variables $Y = (Y_1, \ldots, Y_n)$, parametrised by $\theta$ and provisionally assumed true; this is specified by the joint density function $f_Y(y;\theta) = \prod_{i=1}^n f_{Y_i}(y_i;\theta)$ where $y = (y_1, \ldots, y_n)$ is an arbitrary evaluation point. Particularly relevant to this discussion are models for which there is a sufficiency reduction from $n$ to $d < n$. Specifically, if $s(y)$ is a sufficient statistic, then the joint density function at $y$ factorises as

$$f_Y(y;\theta) = \prod_{i=1}^{n} f_{Y_i}(y_i;\theta) = g(s(y);\theta)h(y), \tag{2.1}$$

showing that all information in the observed data $y^o$ relevant for inference on $\theta$ is contained in the observed value of the sufficient statistic $s^o = s(y^o)$. In the remainder of this exposition, the random variable $S = s(Y)$ denotes the minimal sufficient statistic, i.e. a statistic of smallest dimension $d$ among all those satisfying equation (2.1).

Suppose that the dimension of $S$ is $d > d_\theta$, where $d_\theta$ is the dimension of $\theta$, the difference in dimensions being $d_A = d - d_\theta$. Then from any estimator of $\theta$,

necessarily constructed from part of the minimal sufficient statistic, there must be a complementary component of dimension $d_A$ to complete the information contained in $S$, and any estimator of $\theta$ must sacrifice information on $\theta$ by the definition of minimal sufficiency. From a given minimal sufficient statistic $S$, particular interest lies in one-to-one transformations of the form $S \cong (T, A)$, where $\cong$ denotes one-to-one equivalence, $A$ is an ancillary statistic, in a sense to be clarified in Sect. 2.2, and $T$ is a one-to-one transformation of the maximum likelihood estimator. If $S$ is minimal sufficient, then so is $S' = (T, A) \cong S$, so without loss of generality we take $S = (T, A)$. In fact, in some discussions of exact and approximate conditional inference, including that of Fisher [31], $T$ is taken to be the maximum likelihood estimator $\hat{\theta}$, and it is conceptually helpful to keep that convention in mind, although it raises questions over the existence and construction of the ancillary complement $A$. That matter is covered to some extent in chapter 7 of Barndorff-Nielsen and Cox [7].

While $S = (\hat{\theta}, A)$ is an observable separation, insight is obtained through consideration of a notional idealised separation $S = (C(A), A)$, where $A$ is the same ancillary statistic. In the idealised separation, $C(A)$ is not a function of $A$ in the conventional sense; it can be thought of as the part of the information in $S$ not contained in $A$, so that $S = (C(A), A)$ separates the information into components of dimension $d_\theta$ and $d_A$ without loss or redundancy. The maximal co-ancillary statistic $C(A)$ is notional in the sense that it is typically not directly expressible in terms of the original data even if $A$ has an explicit representation; it is maximal in the sense that it completes the information in $S$ without loss. Once $A = a^o$ is observed, $C(a^o)$ has the conditional distribution of $S$ given $A = a^o$, and this 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 $a(s) = a^o$. The maximal co-ancillary $C(a^o)$ can therefore be thought of as having a distribution on the $d_\theta$-dimensional co-ancillary manifold embedded in $\mathbb{R}^d$:

$$\mathcal{C}(a^o) = \{s \in \mathbb{R}^d : a(s) = a^o\} \subset \mathbb{R}^d. \tag{2.2}$$

The geometry of conditional inference is illustrated through examples in Sect. 2.2 after some discussion of the definition of ancillarity and the motivation for conditioning. The considerations involved in constructing an approximate conditional analysis when the relevant separations are not available are discussed in Sect. 4.

### 2.2 On the definition of ancillarity

Usage of the term *ancillary statistic*, even in its strictest form, has differed in the literature in an important respect: some authors refer to any statistic as ancillary whose distribution is free of (or depends very weakly on) the parameter $\theta$. Other authors, including R. A. Fisher, have included in the definition of ancillarity that the statistic be part of the minimal sufficient statistic. Throughout this paper, the latter convention is adopted.

Several property-based definitions of ancillarity of varying stringency have been put forward by Barndorff-Nielsen and Cox [7, p. 38]. The idealised situation is when the distribution of $A$ does not depend on $\theta$, in which case the joint density function

factorises as

$$f_{T,A}(t,a;\theta)dtda = f_{T|A}(t \mid a;\theta) f_A(a)dtda \tag{2.3}$$

showing that all the information about $\theta$ lost by using $T$ in place of $S$ is recovered by conditioning on $A = a^o$. This formulation makes it clear that $A$ is not irrelevant for inference on $\theta$, rather that $A$, by itself, carries no information on the value of $\theta$.

The idealised definition is too strong for most settings and is complicated by the presence of nuisance parameters. A vague but practically useful definition, implicit in some of the constructions used by Fisher, is to specify that $A$ is ancillary for $\theta$ if, from observation of $A$ alone, no information about the value of $\theta$ can in general be extracted. Formalised constructions along these lines have been proposed, e.g. by Barndorff-Nielsen [3]. Other constructions have been used in connection with higher-order approximate conditional inference via Barndorff-Nielsen's $p^*$ formula [4, 5]. Skovgaard [43] discusses the accuracy of the approximation induced by second-order relative to first-order ancillary statistics such as that of Efron and Hinkley [26].

### 2.3 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 $\ell(\cdot) - \ell(\hat{\theta})$, rather than its shape, the latter being fixed by $A = a^o$. The separation thereby achieves relevance without degeneracy, the broad definition of ancillarity guaranteeing an exact or approximate decoupling of location and shape and $d \ll n$ ensuring non-degeneracy.

That such a conditional analysis should be sought is not uncontroversial. Cox [14] provided a compelling example and brought clarity to several sources of contention. In two examples of common relevance, ancillarity is routinely invoked and rarely questioned: the sample size $n$ is almost always treated as non-random even when it is not fixed in advance; in a regression context with covariate matrix $X$ treated as random, it is widely acknowledged that the appropriate measure of precision depends on the observed value of $X^{\mathrm{T}}X$, rather than its expectation. Yates [45] gave a historical account of the debate in the context of $2 \times 2$ tables.

In response to the usual argument against conditioning (e.g. Brown, [12]) that conditional tests lose power, the Fisherian position is that unconditional power comparisons are irrelevant once the data have been observed, a point emphasised particularly by Cox (e.g. [14, 19, 22]). As noted by a referee, power can be used for planning, i.e. in specifying a sample size, although it is arguably more appropriate to set the sample size to attain a particular unconditional standard error, which requires specification of one parameter rather than three (a power-based calculation requires specification of the level of the test, the distance from the null hypothesis value, and the required unconditional power at that distance).

Birnbaum [11] highlighted an apparent paradox of conditional inference, purporting to show that adoption of the Fisherian position invalidates the use of repeated sampling considerations as part of the inference procedure, essentially leading to the Bayesian

paradigm, or to other approaches that are in conformity with the likelihood principle. Durbin [25] noted that the requirement that an ancillary statistic be part of the minimal sufficient statistic voids the result of Birnbaum [11]. See Evans [27] and Evans and Frangakis [28] for an extensive discussion.

### 2.4 The co-ancillary manifold in two exact conditional analyses

The primary role of the notional separation $S = (C(A), A)$ is to aid geometric interpretation via the maximal co-ancillary statistic $C(a^o)$, distributed on the co-ancillary manifold $\mathcal{C}(a^o)$ of equation (2.2). The present section provides an illustration based on two related examples: the conditional analysis of the $2 \times 2$ contingency table, as given by Fisher [30], and the binary matched pair analysis in the logistic parametrisation given by Cox [15]. In neither of these examples does the exact ancillarity property (2.3) hold: in both cases the distribution of $A$ depends very slightly on the parameter of interest in such a way that from observation of $A$ alone, no information about the value of the interest parameter $\theta$ can in general be extracted. In the second of the two examples, $n$ pair-specific nuisance parameters are eliminated from the analysis, which also illustrates the second role for technical conditioning. A different exposition of parts of this section appeared in Battey [9]; the development here emphasises the role of the co-ancillary manifold, which is new to the present paper, and omits other details of less fundamental importance.

Consider a $2 \times 2$ contingency 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 |

(2.4)

In the leftmost table, $N_{c|r}$ counts the number of individuals with column outcome $c$ and row outcome $r$ using the convenient coding $c \in \{0, 1\}$ and $r \in \{0, 1\}$. Column and row totals are indicated in an obvious notation, e.g. $N_{\cdot|r} = N_{0|r} + N_{1|r}$. The standardised table is obtained from the original by division by $N$.

For concreteness, the chosen example with the two binary variables {failure, success} and {control, treated} illustrates a situation most naturally thought of as a binary regression in a single binary covariate. The original formulation dating at least to Pearson [41] was a so-called pure contingency table in which the two binary variables are on an equal footing. Conceptually, the two situations are very different, but there are close parallels in the algebraic definitions of the primary objects of interest. Specifically, the cross-product ratio $\theta = (p_{0|0}p_{1|1})/(p_{0|1}p_{1|0})$ used as a measure of dependence in the pure contingency table is algebraically equal to the odds ratio $\theta = (p_{1|1}/p_{0|1})/(p_{1|0}/p_{0|0})$ when one variable is treated as potentially explanatory for the other. Here, $p_{c|r}$ are the probabilities associated with each pair of binary outcomes.

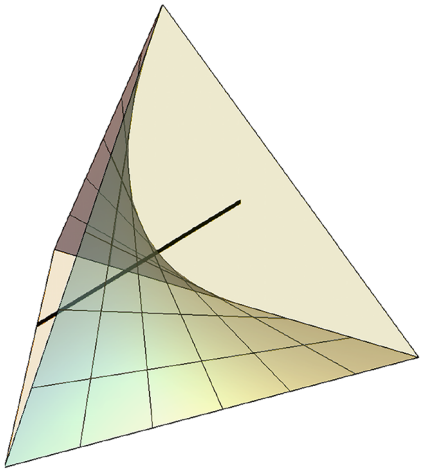

**Fig. 1** The triangular pyramid is the simplex $\mathcal{S}_3 \subset \mathbb{R}^4$; the curved manifold is the independence surface in the pure contingency table; the thick black line is the co-ancillary manifold $\mathcal{C}(a^o)$

The convenience of this example for geometric visualisation is that the sample space for the standardised table is on the scale of probabilities, a scale on which statements concerning parameter values can also be represented. The simplex in $\mathbb{R}^4$ is a three-dimensional object, allowing visualisation, and is represented in Fig. 1 by a triangular pyramid. If the row and column totals in the standardised table are ignored, there are three degrees of freedom for variation of the entries of the table and $(\hat{p}_{0|0}, \hat{p}_{1|0}, \hat{p}_{0|1}, \hat{p}_{1|1})$ belongs to the unit simplex in $\mathbb{R}^4$. Knowledge of one of the row totals (and therefore both) leaves two degrees of freedom for how the entries of the table can be filled in, while further knowledge of the column totals leaves only one. Fisher [30] argued that it is appropriate to condition on the row and column totals in the analysis of the $2 \times 2$ table, these being ancillary in the broad sense of Sect. 2.2. After conditioning, the values $(\hat{p}_{0|0}, \hat{p}_{1|0}, \hat{p}_{0|1}, \hat{p}_{1|1})$, viewed as random variables, have a distribution constrained to a one-dimensional subspace of the unit simplex.

In Fig. 1 the curved manifold is the set of true multinomial probabilities consistent with independence of the two binary variables, viewed on an equal footing in a pure contingency table, and was determined algebraically by Fienberg and Gilbert [29]. Specifically, it can be checked that with $r, s \in [0, 1]^2$ and $(p_{0|0}, p_{1|0}, p_{0|1}, p_{1|1}) = (rs, r(1-s), (1-r)s, (1-r)(1-s))$, the cross product ratio $\theta$ is equal to 1, which quantifies the independence condition that the probability of every joint event is equal to the product of the corresponding marginal events.

The black line is the co-ancillary manifold $\mathcal{C}(a^o)$ of equation (2.2) determined by the observed values of the ancillary statistics $a^o = (\hat{p}^o_{1|\cdot}, \hat{p}^o_{\cdot|1})$, which was chosen as $a^o = (0.6, 0.4)$ for the purpose of Fig. 1. All standardised tables on the line have the same marginal totals, where the two extremes are the tables that solve $p_{c|r} = \min\{\hat{p}^o_{c|\cdot}, \hat{p}^o_{\cdot|r}\}$ for all combinations $\{c, r\} \in \{0, 1\}$, the other three entries of each table being determined by the marginal totals, and two pairs among the four tables being equal by construction. Fisher's [30] exact conditional analysis can be thought of as being based on the distribution of $(\hat{p}_{0|0}, \hat{p}_{1|0}, \hat{p}_{0|1}, \hat{p}_{1|1})$ constrained to this line.

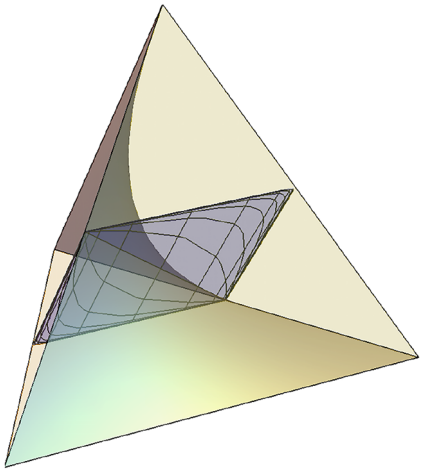

**Fig. 2** The flat plane is the subspace of entries of the standardised table compatible with the constraint $(\frac{1}{2}, \frac{1}{2})$ on the row totals implied by the matched pair design. The curved contours of the plane are contours of equal $\beta$ in the logistic parametrisation $(\alpha, \beta) \mapsto e^{\alpha+\beta}/(1+e^{\alpha+\beta}) =$ pr(success|treated), while the vertical straight lines are contours of equal $\alpha$

The analysis can be extended to a setting of common relevance in biomedical research through consideration of matched comparisons in binary outcomes, formalised by Cox [15, 16]. Given $n$ pairs of matched individuals (e.g. monozygotic twins, or left and right sides of the same individual, etc.) one unit from each pair is randomised to receive treatment, the other being the untreated control. The pairwise table of counts thus has row totals both equal to 1 by design. The logistic parametrisation of the row probabilities for the $i$th pair are

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

which are converted to the scale of the standardised pairwise table with true probabilities $(p^{(i)}_{0|0}, p^{(i)}_{1|0}, p^{(i)}_{0|1}, p^{(i)}_{1|1})$ by division by 2. By the pairing, all tables resulting from the design have row totals $(\frac{1}{2}, \frac{1}{2})$, and if $\alpha_i = \alpha$ for all $i$ the possibilities are constrained to the flat plane in Fig. 2. The curved contours of the plane are contours of equal $\beta$ in the logistic parametrisation $(\alpha, \beta) \mapsto e^{\alpha+\beta}/(1+e^{\alpha+\beta}) =$ pr(success|treated), while the vertical straight lines are contours of equal $\alpha$.

The logistic model with pair-specific odds at baseline is intermediate between a general multinomial representation and a representation in two independent binomials. This can be seen on noting that the pair specific-nuisance parameter $\alpha_i$ allows arbitrary dependence on (unmeasured) covariates, the implicit regression model specifying that outcomes are conditionally independent given the treatment indicator and covariates. Here, $\beta$ has the interpretation of an additive treatment effect on the log-odds scale, or equivalently $\theta = e^{\beta}$ has the interpretation of a multiplicative treatment effect on the odds scale, where, by definition

$$(q^{(i)}_{1|1}/q^{(i)}_{0|1}) = e^{\alpha_i} e^{\beta}, \quad (q^{(i)}_{1|0}/q^{(i)}_{0|0}) = e^{\alpha_i}.$$

Because of the pair-specific nuisance parameters, direct analysis of a table in $n$ pairs of the form (2.4) does not allow inference on the parameter $\beta$ of interest and Fig. 2 does not apply without a preliminary step to be described next.

A standard analysis can in principle be applied to the pairwise tables, although with some degeneracy. Consider the four possible pairwise tables and their marginal totals:

| 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 | |

and let $R^{00}$, $R^{01}$, $R^{10}$, $R^{11}$ denote the number of tables (i.e. number of pairs) of each type. In the leftmost and rightmost tables (concordant pairs), conditioning on column totals leaves no degrees of freedom for how the table can be filled in. In the two inner tables (discordant pairs) there remains one degree of freedom after conditioning. Conditioning in the pairwise tables thus leads us to discard concordant pairs, retaining $R^{01}$ tables of type $(1, 0, 0, 1)$ and $R^{10}$ tables of type $(0, 1, 1, 0)$. The discordant pair table is therefore

| F | S | |
|---|---|---|
| C: $R^{01}$ | $R^{10}$ | $m$ |
| T: $R^{10}$ | $R^{01}$ | $m$ |
| $m$ | $m$ | |

(2.5)

Importantly for the analysis, the joint probability that any pairwise table is of type $(1, 0, 0, 1)$ given that it is either of type $(1, 0, 0, 1)$ or of type $(0, 1, 1, 0)$ is

$$\frac{q^{(i)}_{0|0}q^{(i)}_{1|1}}{q^{(i)}_{0|0}q^{(i)}_{1|1} + q^{(i)}_{1|0}q^{(i)}_{0|1}} = \frac{e^{\beta}}{1 + e^{\beta}}$$

which is free of the pair-specific nuisance parameters. It follows that a conditional analysis using the discordant pair table (2.5) is based on the conditional distribution of, say, $R^{01}$ given the column total $m$, which is binomial of index $m$ and parameter $e^{\beta}/(1 + e^{\beta})$. The induced discrete distributions on the co-ancillary manifold $\mathcal{C}(a^o)$ is depicted in Fig. 3 for $m = 7$ and two different values of $\beta$.

The example illustrates two roles of conditioning: relevance, because the conditioning statistics are ancillary for the interest parameter in the broad sense of Sect. 2.2; and elimination of nuisance parameters. The latter situation only happens in particular classes of model for which the jointly sufficient statistic for all parameters factorises appropriately under the chosen conditioning.

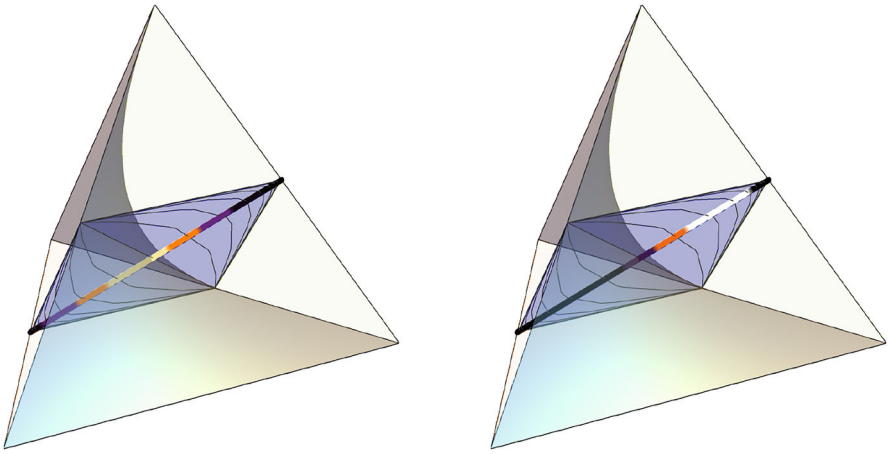

**Fig. 3** The diagonal straight line on the plane is the co-ancillary manifold $\mathcal{C}(a^o)$ specified by the column totals in table (2.5); the shading on this line is the induced discrete distribution over entries of the standardised table compatible with the constraints. This distribution is binomial of index $m = 7$ and parameter $e^\beta/(1+e^\beta)$ for $\beta = 0$ (left) and $\beta = 2$ (right)

## 3 Sufficiency and maximal co-sufficiency

### 3.1 The sufficiency/co-sufficiency separation

Broadly paralleling the discussion of Sect. 2 is an exposition of conditioning for the assessment of model adequacy. From a model (2.1), parametrised by $\theta$ and provisionally assumed true, the relevant information for inference on $\theta$ is contained in $S = s(Y)$. Provided that $d < n$, a portion the information in $Y$ is not used for inference on $\theta$ and is available for the assessment of model adequacy. The idea is implicit in some of Fisher's work but is set out with clarity by Barndorff-Nielsen and Cox [7, p. 29]. Specifically, if $y^o$ is extreme when calibrated against the conditional distribution of $Y$ given $S = s^o$, then this casts doubt on the adequacy of the model. This falsification approach to inference is appropriate when there is no clear notion of what types of departure from the null hypothesis are likely to arise.

In the notional idealised separation $Y \cong (S, Q(S))$, $Q(S)$ is not a function of $S$ in the conventional sense but $Q(s^o)$ has the distribution of $Y$ (or some one-to-one transformation of $Y$) given $S = s^o$. The maximal co-sufficient statistic $Q(S)$ is notional in the same sense that $C(A)$ is, both typically being incapable of direct expression in terms of the original data; it is maximal in the sense that it completes the information in $Y$ without loss once $S$ is given. The term *co-sufficient sampling* was used by Barber and Jansen [2] to describe sampling from the conditional distribution of $Y$ given $S = s^o$. Their work is the first serious attempt that I have seen to make explicit use of the sufficiency/co-sufficiency separation for the assessment of model adequacy.

The realisation of the sufficient statistic $S = s(Y)$ fixes the co-sufficient manifold

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

leaving $n - d$ degrees of freedom for variation of $y$ consistent with the constraint $s(y) = s^o$. Thus $\mathcal{Q}(s^o)$ is a manifold of dimension $n - d$ embedded in $\mathbb{R}^n$ and, once $s^o$ is observed, $Q(s^o)$ is a random variable constrained to $\mathcal{Q}(s^o)$, whose distribution is induced by that of $Y$.

As noted by a referee, when $d_\theta < d$ the marginal distribution of an ancillary statistic $A$ of the form (2.3) is also available for model checking.

### 3.2 The co-sufficient manifold in canonical exponential family regression

Consider a canonical exponential family regression model for outcomes $Y = (Y_1, \ldots, Y_n)$, for which the conditional density or mass function at $y = (y_1, \ldots, y_n)$ given covariate vectors $x_1, \ldots, x_n$ is given by

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

The minimal sufficient statistic is $S = \sum_{i=1}^n x_i Y_i = X^{\mathrm{T}}Y$, where each $x_i$ has the same dimension $d_\theta$ as $\theta$ and $X$ is the $n \times d_\theta$ matrix (treated as non-random) with $x_i^{\mathrm{T}}$ as its $i$th row. In this setting $d = d_\theta$ and there no ancillary statistic as defined in Sect. 2.2. The normal directions to $\mathcal{Q}(s^o) \subset \mathbb{R}^n$ at $y^o$ are specified by the matrix of gradient vectors

$$\frac{\partial s^{\mathrm{T}}(y)}{\partial y}\bigg|_{y=y^o} = X,$$

and since there is no dependence on $y^o$, the co-sufficient manifold $\mathcal{Q}(s^o) \subset \mathbb{R}^n$ is flat. It is spanned by an orthogonal basis for $\mathcal{X}^\perp = \{v \in \mathbb{R}^n : v^{\mathrm{T}}x = 0, x \in \mathcal{X}\}$, where $\mathcal{X}$ is the column space of the $n \times d$ matrix of covariate data. This basis may be taken, for instance, as the $n - d$ columns of the matrix $U$ of eigenvectors of the projection $I_n - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$. By construction, these satisfy $UU^{\mathrm{T}} = I_n - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ and $U^{\mathrm{T}}U = I_{n-d}$.

### 3.3 Linear regression with unknown dispersion

Suppose now that, in the context of a normal-theory linear regression model, the dispersion parameter $\sigma^2$ is unknown. The minimal sufficient statistic is then $S \cong (X^{\mathrm{T}}Y, \hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon})$ where $\hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon} = Y^{\mathrm{T}}UU^{\mathrm{T}}Y$ is the residual sum of squares. The notation of Sect. 3.2 needs adjusting on account of the sufficient statistic having been enlarged. Let $\mathcal{W}(X^{\mathrm{T}}y^o)$ denote the $(n - d_\theta)$-dimensional flat manifold derived above from the conditioning event $X^{\mathrm{T}}Y = X^{\mathrm{T}}y^o$. This is the space to which $W := U^{\mathrm{T}}Y$ belongs. The additional conditioning on $\hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon} = (\hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon})^o = y^{o\mathrm{T}}UU^{\mathrm{T}}y^o$ fixes the length of $W$, thereby

defining a spherical hypersurface within $\mathcal{W}(X^{\mathrm{T}}y^o)$. The co-sufficient manifold $\mathcal{Q}(s^o)$ is a hypersphere when viewed as embedded within $\mathcal{W}(X^{\mathrm{T}}y^o)$, i.e. it has one dimension missing from the ambient space (recall that $d = d_\theta + 1$). Since $\mathcal{W}(X^{\mathrm{T}}y^o)$ is a $(n - d_\theta)$-dimensional subspace of $\mathbb{R}^n$, $\mathcal{Q}(s^o)$ is a $(n - d)$-dimensional subspace of $\mathbb{R}^n$, as stated in Sect. 3.1.

For the purpose of assessing the adequacy of a model that includes as covariates the columns of $X$ it is equivalent to consider the statistic $Q := U^{\mathrm{T}}Y/\|U^{\mathrm{T}}Y\|$ of unit length, in which case $\mathcal{Q}(s^o)$ is the $(n - d)$-dimensional unit sphere. The statistic $Q$ is the maximal invariant used in the invariant testing literature, which can be shown to be uniformly distributed on the surface of the unit sphere under correct specification of the model. A direct derivation of this result can be found in the supplementary material of Battey and McCullagh [10] but the result is well known and also follows from the stochastic representation of elliptically symmetric random variables [13, Theorem 1]. A difficulty in using this result in practice, is that all possible realisations of $Q$ are consistent with uniformity on the sphere, even when the model assumption is violated. If the distribution of $Q$ is to be used directly, this points to some form of pseudo-replication as the means for exploiting the sufficiency/co-sufficiency separation within the Fisherian falsification framework, for instance, by constructing $Q_1(s_1^o), \ldots, Q_m(s_m^o)$ from $m$ equally sized partitions of the data such that $n/m > d$. These $m$ statistics are then independently uniformly distributed on the unit sphere in $(n/m) - d$ dimensions under the null hypothesis.

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 being functions of the data only through the maximal invariant (e.g. Lehmann, [36]). There is, however, at least one context where specification of an alternative direction is ideally to be avoided: that of the construction of confidence sets of models in high-dimensional regression ([8, 21]). In that setting, a preliminary reduction from $p \gg n$ potential explanatory variables to $v < n$ means that an assessment of the adequacy of a model based on a low-dimensional subset of these variables cannot legitimately be based on a standard comparison to the encompassing model in $v$ variables. For instance, a likelihood ratio or $F$ test rejects too often in hypothetical repeated use due to the comparison model having been selected in the light of the data. It seems desirable to develop a falsification approach based on the null distribution of $Q(s^o)$ or related quantities such as $Q_1(s_1^o), \ldots, Q_m(s_m^o)$.

A referee has pointed out that $Q$ being uniform on the sphere does not in itself invalidate its use for model checking. Informal checks for the exclusion of a covariate are often based on systematic departures of residuals (or standardised residuals, which are closely related to $Q$) from total randomness. However, this seems to require a relatively clear idea of which variables might have been excluded from the model and, in the context of the previous paragraph, leads back to the difficulties highlighted there. If the regression mean model is correctly specified and the only departures to be assessed are from independent and identically distributed normal errors, then $k$-statistics can be constructed based on the least squares residuals and compared to the relevant cumulants under correct specification of the model. McCullagh ([39], chapter 4.7) discusses unbiased estimation of residual cumulants from $k$-statistics. See also Anscombe [1], Cox and Snell [23], McCullagh and Pregibon [40].

## 4 Brief remarks on approximate conditional inference

This note sought to elucidate two types of notional idealised analyses from a geometric point of view and thereby provide insight into the foundations of Fisherian inference. The practical implementation of such ideas is typically more difficult and context-dependent, as illustrated in Sect. 2.4. The idea that one might achieve the idealised conditional analyses approximately led to a sizeable literature on approximate conditional inference. Section 2 of the present note emphasises that there are two objects potentially requiring approximation: the co-ancillary manifold $\mathcal{C}(a^o)$, and the distribution of $C(a^o)$ on the (approximate) co-ancillary manifold. Approximate conditional inference via the tangent exponential model (Fraser, [32, 33]; Fraser and Reid, [34]) seems positioned towards this, although the connection is not immediately transparent because there is no preliminary reduction to a minimal sufficient statistic. By conditioning on a statistic of larger dimension than $A$ when a minimal sufficiency reduction $S = (T, A)$ of dimension $d < n$ is available, conditioning on $A = a^o$ is implicit, so relevance is certainly achieved. The implicit definition of the co-ancillary manifold is analogous except insofar that it is viewed as a $d_\theta$-dimensional manifold embedded in $\mathbb{R}^n$ rather than in $\mathbb{R}^d$. It is not immediately clear, however, whether the additional constraints may produce some degeneracy in the induced distribution, and whether the resulting inference may depend on the data in ways other than through the minimal sufficient statistic. The latter question was raised by Skovgaard [44] in his discussion of Fraser [35] and rebutted in the rejoinder. See Davison and Reid [24] for an exposition closest in spirit to that of the present work; note however that the definition of ancillarity used there differs from that used here and in general captures the information in $A$ and in $Q(S)$. The geometric underpinnings of the $p^*$ formula in curved exponential families [4] appear compatible with the discussion of Sect. 2.

## 5 Closing discussion

The broad definition of ancillarity in Sect. 2.2 and implicit in Sect. 2.4 was deliberately vague: $A$, part of the minimal sufficient statistic, is ancillary for $\theta$ if, from observation of $A$ alone, no information about the value of $\theta$ can in general be extracted. When $d_\theta < d$ there might be many such statistics under consideration. Lloyd [37] raised the possibility that an exactly distribution-constant statistic may be virtually ineffectual at discriminating informative samples from uninformative ones, while a statistic whose distribution depends slightly on the parameter can be highly effective in this role. On the other hand, when some dependence on the parameter of interest is permitted in the definition of ancillarity, there is the possibility that, in some regions on the parameter space, appreciable information is discarded through conditioning. Senn ([42], p. 298) discusses a situation of this kind in the context of a pure $2 \times 2$ table. Correspondingly, the discarding of concordant pairs in the matched comparison study has been questioned on the grounds that a large number of such pairs is superficially suggestive of a null effect.

Many aspects of the choice between ancillary statistics are covered in the recent work of Evans and Frangakis [28]. In the Fraser and Reid work from Sect. 4 there is no

explicit construction of an ancillary statistic and the conditioning is achieved approximately through projection, which is unique to the order of approximation considered. It is not immediately clear how the implicit definition fits into the discussion of Evans and Frangakis [28].

The present paper, while offering some geometric insight, does not obviously lead to any resolutions regarding the choice of ancillary statistics. It may be that a single formalised notion of ancillarity is too restrictive to apply seamlessly across all situations and that some flexibility and scientific judgement is needed when it comes to operationalising the broad ideas in specific contexts.

**Acknowledgements** The work originated from an expository lecture given at the O. E. Barndorff Nielsen memorial conference at the University of Aarhus, May 2024. I thank the organisers for their invitation. I am grateful to Nancy Reid for alerting me to most of the references in Sect. 4 and for discussions about approximate conditional inference. It is a pleasure to thank two anonymous referees and the associate editor for their valuable comments.

**Author Contributions** This is a single-authored manuscript.

**Funding** The work was supported by the UK Engineering and Physical Sciences Research Council under grant number EP/T01864X/1.

**Data availability** No datasets were generated or analysed during the current study.

## Declarations

**Conflict of interest** The author declares no competing interests.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

## References

1. Anscombe, F.J.: Examination of residuals. Proc. Fourth Berkeley Symp. **1**, 1–36 (1961)
2. Barber, R.F., Janson, L.: Testing goodness of fit and conditional independence with approximate co-sufficient sampling. Ann. Stat. **50**, 2514–2544 (2022)
3. Barndorff-Nielsen, O.E.: On $M$-ancillarity. Biometrika **60**, 447–455 (1973)
4. Barndorff-Nielsen, O.E.: Conditionality resolutions. Biometrika **67**, 293–310 (1980)
5. Barndorff-Nielsen, O.E.: On a formula for the distribution of the maximum likelihood estimator. Biometrika **70**, 343–365 (1983)
6. Barndorff-Nielsen, O.E., Cox, D.R.: Edgeworth and saddle-point approximations with statistical applications (with discussion). J. R. Stat. Soc. B **41**, 279–312 (1979)
7. Barndorff-Nielsen, O.E., Cox, D.R.: Inference and Asymptotics. Chapman and Hall, London (1994)
8. Battey, H. S., Cox, D. R.: Large numbers of explanatory variables: a probabilistic assessment. *Proc. Roy. Soc. Lond. A*, 474, article number 20170631 (2018)
9. Battey, H. S.: D. R. Cox: aspects of scientific inference. *J. R. Statist. Soc. A*, to appear (2024)
10. Battey, H.S., McCullagh, P.: An anomaly arising in the analysis of processes with more than one source of variability. Biometrika **111**, 677–689 (2024)

11. Birnbaum, A.: On the foundations of statistical inference (with discussion). J. Am. Stat. Assoc. **57**, 269–332 (1962)
12. Brown, L.D.: An ancillary paradox which appears in multiple regression. Ann. Stat. **18**, 471–493 (1990)
13. Cambanis, S., Huang, S., Simons, G.: On the theory of elliptically contoured distributions. J. Multivariate Anal. **11**, 368–385 (1981)
14. Cox, D.R.: Some problems connected with statistical inference. Ann. Math. Stat. **29**, 357–372 (1958)
15. Cox, D.R.: Two further applications of a model for binary regression. Biometrika **45**, 562–65 (1958)
16. Cox, D.R.: The Analysis of Binary Data. Methuen, London (1970)
17. Cox, D.R.: The choice between alternative ancillary statistics. J. R. Stat. Soc. B **33**, 251–255 (1971)
18. Cox, D.R.: Local ancillarity. Biometrika **67**, 279–286 (1980)
19. Cox, D.R.: Discussion of 'Tests of significance of $2 \times 2$ contingency tables' by Yates, p. 147. J. R, Statist. Soc. A (1984)
20. Cox, D.R.: Principles of Statistical Inference. Cambridge University Press, Cambridge (2006)
21. Cox, D.R., Battey, H.S.: Large numbers of explanatory variables, a semi-descriptive analysis. Proc. Nat. Acad. Sci. **114**, 8592–8595 (2017)
22. Cox, D.R.: Statistical significance. Ann. Rev. Stat. Appl. **7**, 1–10 (2020)
23. Cox, D.R., Snell, E.J.: A general definition of residuals (with discussion). J. R. Stat. Soc. B **30**, 248–275 (1968)
24. Davison, A. C., Reid, R.: The tangent exponential model. *Handbook of Bayesian, Fiducial, and Frequentist Inference*, 210–237 (2022)
25. Durbin, J.: On Birnbaum's theorem on the relation between sufficiency, conditionality and likelihood. J. Am. Stat. Assoc. **65**, 395–398 (1970)
26. Efron, B., Hinkley, D.V.: Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. Biometrika **65**, 457–487 (1978)
27. Evans, M.A.: What does the proof of Birnbaum's theorem prove? Elec. J. Stat. **7**, 2645–2655 (2013)
28. Evans, M., Frangakis, C.: On resolving problems with conditionality and its implications for characterizing statistical evidence. Sankya A **85**, 1103–1126 (2023)
29. Fienberg, S.E., Gilbert, J.P.: The geometry of a two by two contingency table. J. Am. Stat. Assoc. **65**, 694–701 (1970)
30. Fisher, R.A.: On the interpretation of $\chi^2$ from contingency tables, and the calculation of $P$. Biometrika **13**, 211–230 (1922)
31. Fisher, R.A.: Two new properties of mathematical likelihood. Proc. R. Soc. Lond A **144**, 285–307 (1934)
32. Fraser, D.A.S.: Normed likelihood as saddlepoint approximation. J. Multivar. Anal. **27**, 181–193 (1988)
33. Fraser, D.A.S.: Tail probabilities from observed likelihoods. Biometrika **77**, 65–76 (1990)
34. Fraser, D.A.S., Reid, N.: On conditional inference for a real parameter: a differential approach on the sample space. Biometrika **75**, 251–264 (1988)
35. Fraser, D.A.S.: Ancillaries and conditional inference (with discussion). Stat. Sci. **19**, 333–369 (2004)
36. Lehmann, E.L.: Testing Statistical Hypotheses. Chapman & Hall, London (1959)
37. Lloyd, C.J.: Effective conditioning. Aust. J. Stat. **34**, 241–260 (1992)
38. McCullagh, P.: Local sufficiency. Biometrika **71**, 233–244 (1984)
39. McCullagh, P.: Tensor Methods in Statistics. Chapman and Hall, London (1987)
40. McCullagh, P., Pregibon, D.: k-Statistics and dispersion effects in regression. Ann. Stat. **15**, 202–219 (1987)
41. Pearson, K.: On the criticism that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Phil. Mag. **50**, 157–175 (1900)
42. Senn, S.: Lessons from TGN1412 and TARGET: implications for observational studies and meta-analysis. Pharmaceut. Stat. **7**, 294–301 (2008)
43. Skovgaard, I.: On the density of minimum contrast estimators. Ann. Stat. **18**, 779–789 (1990)
44. Skovgaard, I.: Discussion of 'Ancillaries and conditional inference' by D. A. S. Fraser. Stat. Sci. **19**, 355–358 (2004)
45. Yates, F.: Tests of significance of $2 \times 2$ contingency tables (with discussion). J. R. Stat. Soc. A **147**, 426–463 (1984)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.