Biometrika (2026), **113**, 4, asag045 https://doi.org/10.1093/biomet/asag045
Advance Access publication 7 July 2026

Post-reduction inference for confidence sets of models

By H. S. BATTEY

Department of Mathematics, Imperial College London,
180 Queen's Gate, London SW7 2AZ, U.K.
h.battey@imperial.ac.uk

D. G. RASINES

CUNEF Universidad, Calle Almansa 101, Madrid 28040, Spain
daniel.garciarasines@cunef.edu

and Y. TANG

Department of Mathematics, Imperial College London,
180 Queen's Gate, London SW7 2AZ, U.K.
yanbo.tang@imperial.ac.uk

Summary

Sparsity in a regression context makes the model itself an object of interest, pointing to a confidence set of models as the appropriate presentation of the evidence. A difficulty in areas such as genomics, where the number of candidate variables is vast, arises from the need for preliminary reduction before the assessment of models. The present paper considers a resolution using inferential separations fundamental to the Fisherian approach to conditional inference, namely, the sufficiency/co-sufficiency separation and the ancillary/co-ancillary separation. Tests of model adequacy based on such separations do not involve specifying a direction of departure from any postulated model, avoiding calibration issues that would arise in directed tests from using the same data for reduction and for model assessment. In idealized cases with no nuisance parameters, the separations extract the relevant information without loss or redundancy. The extent to which estimation of nuisance parameters affects this idealization is illustrated in detail for the normal-theory linear regression model, extending immediately to a lognormal accelerated-life model for time-to-event outcomes. As part of the analysis, we introduce a modified version of the refitted cross-validation estimator of Fan et al. (2012), whose distribution theory is tractable in the appropriate conditional sense. The paper concludes, among other things, that in settings where reduction on a reduced sample is unproblematic, sample splitting has high efficiency relative to our conditional analysis, echoing Cox (1975); otherwise, it gives miscalibrated confidence sets.

Some key words: Ancillarity; Co-sufficiency; Model uncertainty; Regression; Selective inference; Sparsity.

© 2026 Biometrika Trust
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from academic.oup.com/biomet/article/113/4/asag045/8726260 by Imperial College of Science Technology and Medicine user on 21 October 2026

1. Introduction

In the context of regression with many potential explanatory variables, Cox (1968) noted that multiple low-dimensional combinations of variables are often compatible with the data. An arbitrary choice between well-fitting models, while reasonable for prediction, will typically be misleading for scientific understanding, as different models correspond to different scientific explanations. This perspective, which was emphasized repeatedly over many years (e.g., Cox, 1977; Cox & Snell, 1974, 1989), has acquired renewed importance in the light of several modern areas, notably genomics, in which the number of variables for consideration is vast. Routine practice is to report a single model identified by a variable-selection algorithm, the most widely used being the lasso (Tibshirani, 1996). Cox & Battey (2017) highlighted concerns with the single-conclusion decision-theoretic nature of the lasso, recommending instead a presentation of evidence in the form of a confidence set of models, consisting of all relatively simple representations that are compatible with the data at a chosen significance. The notion of a confidence set of models is yet to gain traction, with statisticians tending to report a single conclusion. In two exceptions, Kartsonaki et al. (2022) and Prosser et al. (2026) presented analyses of epidemiological examples via a confidence set of models. Some isolated recent papers making similar arguments to those above have also emerged in the more theoretical literature (see § 2.2).

A set of all sparse models compatible with the data, while superficially less appealing than a single model, conveys the information in the data more honestly. In principle, from the sets of compatible models identified, the conclusions might be narrowed down through a carefully designed experiment, or by way of the contradiction argument laid out in a particular context by Battey & Reid (2023, § 5). Cox & Battey (2017) discussed how one might extract compact messages from large confidence sets of models. Specifically, among the large sets of models reported in their supplementary material, two variables, call them v_1 and v_2, are present in 96% and 94% of cases. In 78% of the models in which v_2 is not present, another variable, v_3, is present in its place, and only 1% of models include neither v_2 nor v_3.

Suppose that a regression model in p variables is correctly specified, but overconditioned, in the sense that the vector of regression coefficients contains many zero entries, with the positions of the nonzero entries unknown. In the remainder of the paper, we therefore equate the specification of a model with a set of variable indices and write $\mathcal{E}^* \subset [p]$ for the sparse model in $s^* < p$ variables that generated the data. When p is not so large as to make the procedure practically infeasible or theoretically unsound, a simple way of constructing confidence sets of models is via a likelihood-ratio test of every low-dimensional subset of variables $\mathcal{E}_m \subset [p] := \{1, \ldots, p\}$ against the comprehensive model $[p]$. All models $\mathcal{E}_m$, $m = 1, \ldots, 2^p$, not rejected at level α comprise the confidence set of models $\mathcal{M}$, and this, by standard arguments, contains $\mathcal{E}^*$ with asymptotic probability $1 - \alpha$ in hypothetical repeated use. It is sensible in practice to restrict attention only to those models of reasonable size, substantially reducing the number of assessments to be made.

For the genomics applications we have in mind, p is typically of several orders of magnitude larger than the number of study individuals n. The above construction of $\mathcal{M}$ is therefore infeasible without preliminary reduction from $[p]$ to a reduced set, $\hat{\mathcal{E}}$ say, of more manageable size. The circumflex emphasizes that $\hat{\mathcal{E}}$ is constructed from the data; it is not intended as an estimator of $\mathcal{E}^*$, a much smaller subset. Any reasonable variable-screening procedure can be used for the construction of $\hat{\mathcal{E}}$; see Lewis & Battey (2025) for a theoretical discussion of the implications of that choice when the ultimate objective is the construction of $\mathcal{M}$.

Downloaded from academic.oup.com/biomet/article/113/4/asag045/8726260 by Imperial College of Science Technology and Medicine user on 21 October 2026

The key point as far as the present work is concerned is that, since the comprehensive model $\hat{\mathcal{E}}$ is selected in the light of the data, it fits the observed data better than an arbitrary model of the same size encompassing the one to be tested. For this reason, a likelihood-ratio test of $\mathcal{E}_m \subset \hat{\mathcal{E}}$ against $\hat{\mathcal{E}}$ typically rejects too often in hypothetical repeated use, so that the resulting model confidence set $\mathcal{M}$ has a lower coverage probability than suggested by the nominal level of the test. The problem is most pronounced in the context of logistic regression, although we do not address logistic regression in the present paper and refer to it only as an aid to understanding the difficulties with directed tests of model adequacy after reduction. In high dimensions there are typically many low-dimensional separating hyperplanes among the candidate variables that perfectly separate cases from controls. If just one such set was retained through reduction, a logistic regression fitted to $\hat{\mathcal{E}}$ would have perfect fit, and a likelihood-ratio test would be unable to detect any $\mathcal{E}^*$ that did not also form a separating hyperplane in the same set of data.

Battey & Cox (2018) used sample splitting as a simple device for restoring the nominal coverage, although it was noted that a potentially better approach would use a minimal sufficiency separation (Bartlett, 1937) to assess model adequacy, bypassing the likelihood-ratio test and obviating sample splitting. The same inferential separation was suggested for simulation-based model assessment by Engen & Lillegård (1997), Lindqvist & Taraldsen (2005), Lockhart et al. (2007) and Barber & Janson (2022). The approaches taken in the above references are based on Monte Carlo algorithms to approximately sample from the appropriate subset of the sample space after suitable conditioning. Markov chain Monte Carlo methods are too computationally expensive for the present purposes, as construction of $\mathcal{M}$ entails assessing all low-dimensional subsets of $\hat{\mathcal{E}}$ for their compatibility with the data, a combinatorially large problem.

The principles developed here, specifically for the purpose of confidence sets of sparse regression models, are based on the sufficiency/co-sufficiency separation and the ancillary/co-ancillary separation fundamental to the Fisherian approach to conditional inference and reviewed by Barndorff-Nielsen & Cox (1994, Chs. 2.3 and 2.5). In the absence of a proposed direction of departure from the postulated model, a combination of the ancillary and co-sufficient components in principle achieves the maximal extraction of information relevant for the assessment of model adequacy, having extracted the information relevant for estimation and reduction. In practice, there is considerable flexibility in how such information is used, with the most naïve approaches being ineffectual for detecting relevant departures.

Methods powerful against specific alternatives are often valuable in small samples. This raises the possibility that there is no advantage to approaches based on inferential separations over sample splitting. In particular, an approach using a large portion of the data for reduction and a small portion for model assessment with a directed test may be expected to perform well. For inference on a collection of means selected on the basis of data from normal distributions, Cox (1975) showed through a theoretical analysis that the efficiency of sample splitting relative to an exact conditional approach is surprisingly high. In the present context, there is a clearer separation of information in the data relevant to the different phases of analysis, so a greater efficiency gain might be expected, although this has to be counterbalanced by the lack of directionality for the full-sample test. These considerations are probed by simulation in § 6 for the normal-theory linear regression model. The normal-theory setting is one in which sample splitting has a relatively high efficiency due to the availability of the F-test for assessing

Downloaded from academic.oup.com/biomet/article/113/4/asag045/8726260 by Imperial College of Science Technology and Medicine user on 21 October 2026

model adequacy, whose distribution theory is exact, and therefore applies at small sample sizes.

2. Related literature and summary of contributions

2.1. *Post-selection versus post-reduction inference*

The only commonality between the post-reduction inference problem considered here and the more familiar post-selection inference problem of, e.g., Lee et al. (2016) and Fithian et al. (2017) is that an uncritical re-use of the data leads to erroneous calibration. The most important conceptual difference from the post-selection inference literature is that our inferential goal, the true model, is not data dependent, as the reduction phase serves only to focus the analysis, and does not dictate the object of inference. Although the analysis is performed under the assumption that $\{\mathcal{E}^* \subset \hat{\mathcal{E}}\}$, conditioning on this event barely affects the distribution of test statistics for model assessment, as the latter satisfies, by construction, $\text{pr}(\mathcal{E}^* \subset \hat{\mathcal{E}}) \approx 1$.

2.2. *Model uncertainty*

Hansen et al. (2011) proposed something called a model confidence set. Their construction does not, however, have the usual properties associated with confidence sets. In particular, Hansen et al. (2011) considered the relative explanatory power of pairs of models among an initial candidate set, leading to the specification of a reduced set of models with equally good or equally poor fit. By contrast, if no sparse model is compatible with the data, a confidence set of models in the sense of Cox & Battey (2017) would be empty. The definition used by Ferrari & Yang (2015) is the same as ours; since they considered a low-dimensional setting, there are no issues of post-reduction inference.

Another strand of work in this vein has recently developed. Fisher et al. (2019) formalized the notion of a *Rashomon set* as the collection of predictive models whose empirical loss lies within a specified tolerance of the best-performing model. Guan & Tibshirani (2020) acknowledged the difficulties associated with the reporting of a single model and proposed a procedure that delivers a set of models with strong predictive performance. Lei (2020) proposed an approach based on cross-validation that returns a set of well-fitting candidate models and has a coverage guarantee for the model that is optimal in a predictive sense. A highly cited paper by Wang et al. (2020), intended for application to genetic fine-mapping, seeks Bayesian credible sets of variables, defined as subsets of variables with a given minimal probability of containing at least one effect variable. The papers of Kim & Ramdas (2025), Wang et al. (2026) and Zhang et al. (2026) appear to be in a conceptually similar vein to the Cox strand, although very different in formulation and detail.

A natural proposal would involve applying the lasso to bootstrapped samples. This, however, does not have the required coverage properties, as noise variables appear earlier on the lasso path than the last signal variable with high probability (Su et al., 2017; Su, 2018).

Breiman (2001) referred to the existence of many alternative predictive models as the *Rashomon effect*, and used this to challenge the relevance of statistical models, a view that we do not share.

2.3. *Summary of contributions*

The idea of using the sufficiency/co-sufficiency separation for the avoidance of post-reduction miscalibration of confidence sets of regression models was mentioned briefly elsewhere (Battey, 2022; Battey & Cox, 2022). The main contribution of the present paper is to make the relevant inferential separations explicit for regression models, to provide

Downloaded from academic.oup.com/biomet/article/113/4/asag045/8726260 by Imperial College of Science Technology and Medicine user on 21 October 2026

geometric insight into the nature of the inferential problem and to embed into this framework a generalization of the randomized inference strategy of Rasines & Young (2023). The resulting approach can be viewed as an inducement of internal replication for the assessment of model adequacy (Battey & Reid 2026) and, alongside a number of semiparametric examples, has informed that work.

3. Assessment of model adequacy

3.1. *The co-sufficient manifold*

The discussion to be presented in this section is, at this point, disconnected from the problem outlined in the introduction. The connection will be established in § 5 after the presentation of some explicit calculations for the Gaussian linear model in § 4.

The role of sufficiency in the assessment of model adequacy, noted by Bartlett (1937), is more easily approached via Barndorff-Nielsen & Cox (1994, p. 29). When the ideas can be operationalized, there are no difficulties associated with a double use of the data for model assessment and parametric inference. For notational simplicity throughout the present § 3, we drop the subscript m indexing the particular model under consideration. The ideas are at this point detached from the specifics of confidence sets of models and follow the geometric exposition of Battey (2024).

Consider an arbitrary model $\mathcal{E}$, provisionally assumed to contain the true unknown outcome distribution, and having an associated parameter vector θ. The sufficient statistic $S = s(Y)$ for θ is a function of $Y = (Y_1, \ldots, Y_n)$ with observed value $s^o = s(y^o) \in \mathbb{R}^d$, where $y^o \in \mathbb{R}^n$ is the observed value of Y. For n independent draws from the same distribution in model $\mathcal{E}$, the joint density or mass function at $y = (y_1, \ldots, y_n)$ separates as

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

by the factorization theorem (e.g., Barndorff-Nielsen & Cox 1994, Ch. 2.3). An important implication of (1), provided that the dimension d of the minimal sufficient statistic S is smaller than n, is that there is information in Y not used for inference on θ, and therefore available for assessment of the model.

It is convenient to write the information in Y as $\{S, Q(S)\}$, where, in this notional idealized separation, $Q(S)$ is not a function of S in the conventional sense, and typically does not have an analytic form in terms of the observations; rather, $Q(s^o)$ has the conditional distribution of Y, or some one-to-one transformation of Y, given $S = s^o$. The realization s^o 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 $s(y) = s^o$, i.e., $\mathcal{Q}(s^o)$ is a manifold of dimension $n - d$ embedded in $\mathbb{R}^n$. Once s^o is observed, the notional random variable $Q(s^o)$ captures the randomness in Y after conditioning on $S = s^o$ and is constrained to $\mathcal{Q}(s^o)$. The implicit general definition of $Q(s^o)$ is

$$\mathrm{pr}(Q(s^o) \in \mathcal{A}) = \mathrm{pr}(Y \in \mathcal{A} \mid S = s^o),$$

where $\mathcal{A} \subseteq \mathcal{Q}(s^o)$ by definition.

Downloaded from academic.oup.com/biomet/article/113/4/asag045/8726260 by Imperial College of Science Technology and Medicine user on 21 October 2026

Barndorff-Nielsen & Cox (1994, p. 29) encouraged an understanding of the separation (1) for the assessment of model adequacy by hypothetically treating the data as arising in two stages. In the first stage we learn that $S = s^o$, which is used for inference on θ. In the second stage we learn that $Y = y^o$, which is effectively one draw from the conditional distribution of Y, given $S = s^o$. If y^o is extreme when calibrated against this conditional distribution, that casts doubt on the adequacy of the model. The logic is the usual analogue of proof by contradiction underpinning the core Fisherian hypothesis tests. By calibrating y^o against the conditional distribution under the model hypothesis of Y, given $S = s^o$, we avoid the need to compare a smaller model against a larger one, thereby avoiding the post-reduction miscalibration suffered by a likelihood-ratio test against a larger data-dependent model, as described in § 1.

For some models, notably the linear regression model and canonical generalized linear models, the co-sufficient manifold can be specified explicitly. When this is not the case, it is natural to seek local approximations to the manifold of interest, and since $s(y^o) = s^o$ by construction, it is reasonable to consider the tangent plane to $\mathcal{Q}(s^o)$ locally at y^o. Some discussion of this is provided in § F.4 in the Supplementary Material. We focus in the present paper on situations for which the co-sufficient manifold is readily deduced.

3.2. *The ancillary separation*

A second approach to the assessment of model adequacy is available when the dimension d of S is larger than d_θ, the dimension of θ. Write $S \cong (S_1, A)$, where S_1 estimates θ and A is an ancillary statistic. Ideally, A is ancillary in the strongest sense, having a distribution that does not depend on θ, which implies the factorization

$$f_S(s;\theta) = f_{S|A}(s \mid a;\theta) f_A(a). \tag{2}$$

More generally, definitions of ancillarity in the presence of nuisance parameters are collated on page 38 of Barndorff-Nielsen & Cox (1994). Since the information relevant for inference for θ is contained in the conditional distribution of S, given $A = a^o$, the marginal distribution of A is available for model checking.

To extract the maximal amount of information for model assessment, the random variables A and $Q(s^o)$ should, in principle, both be sought. Whether this leads to more powerful inference in practice than an approach based on sample splitting, or based on one or other of A and $Q(s^o)$, is unclear from this abstract discussion because of the possibility of using directed tests when the sample is partitioned, and because the detailed implementation may vary according to the structure of the regression model under consideration. Section 4 below gives analytic calculations for the Gaussian linear regression model.

4. Distributional calculations for Gaussian linear regression

4.1. *Distribution on the co-sufficient manifold*

Consider a linear regression model in which, under the postulated model, $Y \in \mathbb{R}^n$ is normal with mean $\mu = X\theta \in \mathcal{X}$ and covariance matrix $\Sigma = \sigma^2 I_n$. The space $\mathcal{X}$ is the d_θ-dimensional linear space spanned by the d_θ columns of X, with $(n - d_\theta)$-dimensional orthogonal complement

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

Downloaded from academic.oup.com/biomet/article/113/4/asag045/8726260 by Imperial College of Science Technology and Medicine user on 21 October 2026

A minimal sufficient statistic when σ^2 is unknown is either $S = (X^{\mathrm{T}}Y, \hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon})$ if the covariate matrix X is treated as fixed, or $S = (X^{\mathrm{T}}Y, \hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon}, X^{\mathrm{T}}X)$ if X is treated as random, where $\hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon}$ is the residual sum of squares. Equivalently, $X^{\mathrm{T}}Y$ can be replaced by $\hat{\theta}$, the ordinary least squares and maximum likelihood estimator of θ. Here we treat X as fixed so that $d = d_\theta + 1$. The information in the data that is not contained in S can be extracted through orthogonal projection of Y onto $\mathcal{X}^\perp$, followed by an additional conditioning on the realized value of $\hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon}$. The first conditioning implicit in the projection restricts to the $(n - d_\theta)$-dimensional linear subspace $\mathcal{X}^\perp$, while the second conditioning restricts the lengths of permissible vectors, so that the co-sufficient manifold is a hypersphere embedded in $\mathcal{X}^\perp$, which can also be viewed as an $(n-d)$-dimensional sphere embedded in $\mathbb{R}^n$. Depending on context, we refer to the co-sufficient manifold interchangeably as a hypersphere in $\mathcal{X}^\perp$ and an $(n-d)$-dimensional sphere in $\mathbb{R}^n$, the term *hypersphere* meaning that there is one dimension missing in the ambient space. Having restricted to $\mathcal{X}^\perp$, we can, for most purposes, think of $\mathcal{Q}(s^o)$ as being a unit hypersphere in $\mathbb{R}^{n-d_\theta}$, except that the coordinates within the larger space $\mathbb{R}^n$ are specified by the orientation of $\mathcal{X}^\perp$. The distribution on the sphere is uniform under correct specification of the model, as shown by the following calculation, analogous to those used in the invariant testing literature (e.g., Kariya, 1980; King, 1980).

The subspaces $\mathcal{X}$ and $\mathcal{X}^\perp$ are groups under addition, which implies that, for any matrix U with column span $\mathrm{Im}(U) = \mathcal{X}^\perp$, the statistic

$$q(y, U) = U^{\mathrm{T}}y/\|U^{\mathrm{T}}y\| = U^{\mathrm{T}}y/(y^{\mathrm{T}}UU^{\mathrm{T}}y)^{1/2}$$

is invariant under the group with action $g(a, b)\colon y \mapsto ay + b$, where $a > 0$, $b \in \mathcal{X}$. A specification of $q(y, U)$ of smallest dimension is obtained by taking the columns of U to be a linear basis for $\mathcal{X}^\perp$, the natural choice being the $n \times (n - d_\theta)$ matrix of orthonormal eigenvectors corresponding to the unit eigenvalues of the standard orthogonal projection $M = I_n - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$. The remaining eigenvalues are zero, and it follows that $UU^{\mathrm{T}} = M$ and $U^{\mathrm{T}}U = I_{n-d_\theta}$.

Consider $W = w(Y) = U^{\mathrm{T}}Y$, which is normally distributed with mean zero and covariance $U^{\mathrm{T}}\Sigma U = \sigma^2 I_{n-d_\theta}$ under the postulated model. In the transformation $w \mapsto (w/\|w\|, w^{\mathrm{T}}w) =: (q, r^2)$, the volume element transforms as

$$\mathrm{d}w_1 \cdots \mathrm{d}w_{n-d_\theta} = \frac{(r^2)^{(n-d_\theta)/2-1}\mathrm{d}r^2 \prod_{j=1}^{n-d_\theta-1}\mathrm{d}q_j}{2(1 - \sum_{j-1}^{n-d_\theta-1} q_j^2)^{1/2}} = \frac{1}{2}(r^2)^{(n-d_\theta)/2-1}\mathrm{d}r^2\mathrm{dVol}^{n-d}(q), \tag{3}$$

where

$$\mathrm{dVol}^{n-d}(q) = \mathrm{dVol}^{n-d_\theta-1}(q) = \frac{\prod_{j=1}^{n-d_\theta-1}\mathrm{d}q_j}{(1 - \sum_{j-1}^{n-d_\theta-1} q_j^2)^{1/2}}$$

is the natural volume element on the surface of the hypersphere in $\mathbb{R}^{n-d_\theta}$. The joint density function is therefore

$$\frac{(r^2)^{(n-d_\theta)/2-1}\exp(-r^2 q^{\mathrm{T}}q/2\sigma^2)}{2(2\pi\sigma^2)^{(n-d_\theta)/2}}\mathrm{d}r^2\mathrm{dVol}^{n-d}(q),$$

Downloaded from academic.oup.com/biomet/article/113/4/asag045/8726260 by Imperial College of Science Technology and Medicine user on 21 October 2026

from which the marginal density function of Q is obtained by integration using a change of variables from r^2 to $r^2/2\sigma^2$ as

$$\frac{\Gamma\{(n-d_\theta)/2\}(q^{\mathrm{T}}q)^{-(n-d_\theta)/2}}{2\pi^{(n-d_\theta)/2}}\mathrm{dVol}^{n-d}(q), \tag{4}$$

showing that Q is uniformly distributed on the unit hypersphere in $\mathcal{X}^\perp$. Since, under the assumed mean model $\mu \in \mathcal{X}$, $R^2 = W^{\mathrm{T}}W$ is distributed as σ^2 times a standard χ^2 random variable with $n-d_\theta$ degrees of freedom, independently of Q, the conditional density function of Y given $R^2 = r^2$ is equal to the marginal density function of Q, which is explicitly defined in this case.

This calculation can be generalized to allow for variance components of the form $\Sigma = \sigma^2\{I_n + V(\alpha)\}$, where $V(\alpha)$ is a known matrix function of a vector parameter $\alpha = (\alpha_1, \ldots, \alpha_s)^{\mathrm{T}}$ with $V(0) = 0$. In this case, the appropriate analogue of (4) is given by Equation (10) of Battey & McCullagh (2024). Alternatively, a transformation to uniformity on the unit hypersphere in $\mathcal{X}^\perp$ is $q \mapsto A_\alpha^{-1/2}q/(q^{\mathrm{T}}A_\alpha^{-1}q)^{1/2} =: Q'$, where $\sigma^2 A_\alpha = U^{\mathrm{T}}\Sigma U$.

There are some fundamental and operational difficulties in applying the model assessment strategies of § 3 directly. These are discussed in § 5 below, where we provide a more elaborate blueprint.

4.2. *Distribution of the ancillary component*

Note that R^2 is the residual sum of squares $\hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon}$, which is part of the minimal sufficient statistic S and has a distribution independent of θ. It is ancillary for θ according to the more general definitions on page 38 of Barndorff-Nielsen & Cox (1994) and is therefore also available for the assessment of model adequacy via a version of (2) allowing for nuisance parameters.

Suppose initially that σ^2 and the mean subspace $\mathcal{X}$ are known. Then R^2 carries all the information relevant for distinguishing, say, normal errors from Student-t errors. This follows from the properties of elliptically symmetric distributions, whose density functions are of the form

$$f_Y(y;\mu,\Sigma,g) \propto |\Sigma|^{-1/2}g\{(y-\mu)^{\mathrm{T}}\Sigma^{-1}(y-\mu)\},$$

with the generator $g\colon \mathbb{R}^+ \to \mathbb{R}^+$ identifying the member of the elliptically symmetric class. To provide insight, and a basis for some arguments later in the paper, we briefly consider this setting.

Let U be as defined in § 4.1. Then $W = U^{\mathrm{T}}Y$ is a member of the elliptically symmetric class of distributions in $n-d_\theta$ dimensions, with zero mean and covariance $\Sigma_W = \sigma^2 I_{n-d_\theta} = U^{\mathrm{T}}\Sigma U$ (e.g., Muirhead, 1982), and therefore has a stochastic representation (Cambanis et al., 1981, Theorem 1) $W \overset{\mathrm{D}}{=} Z\Sigma_W^{1/2}Q'$, where $\overset{\mathrm{D}}{=}$ denotes equality in distribution and Q' is uniformly distributed on the surface of the unit hypersphere in $\mathbb{R}^{n-d_\theta}$. The distribution of Z with density function f_Z determines the member of the elliptic class and is related to the generator g for the standardized variable $\Sigma_W^{-1/2}W$ as (Fang et al., 1990)

$$f_z(z) = \frac{2\pi^{(n-d_\theta)/2}}{\Gamma\{(n-d_\theta)/2\}}z^{n-d_\theta-1}g(z^2).$$

Downloaded from academic.oup.com/biomet/article/113/4/asag045/8726260 by Imperial College of Science Technology and Medicine user on 21 October 2026

In the Gaussian case of § 4.1, the random variable R satisfies $R \overset{\text{D}}{=} \sigma Z$, where Z is distributed as the square root of a χ^2 random variable with $n - d_\theta$ degrees of freedom under the null model $\mu \in \mathcal{X}$. As an example, for the Student-t family of errors with ν degrees of freedom, the distribution of Z is (Kano, 1994) $Z \overset{\text{D}}{=} (\chi^2_{n-d_\theta}/V)^{1/2}$, where $V \overset{\text{D}}{=} \chi^2_\nu/\nu$ is independent of the numerator.

The random variable Z^2 thus carries information on the distribution of the errors when the mean model is known. Conversely, $Z^2 = R^2/\sigma^2$ contains valuable information for assessment of the mean model under the assumption that the errors are Gaussian.

5. An operational blueprint for Gaussian linear regression

5.1. *Co-sufficiency and confidence sets of models: difficulties*

Sections 3 and 4 covered general principles and the corresponding calculations of the normal-theory linear model. These sections were detached from the problem of post-reduction inference for confidence sets of models introduced in § 1, to which we now return.

From a comprehensive model $\hat{\mathcal{E}}$ computed using all available data, we consider all submodels $\mathcal{E}_m$ of reasonable size, seeking to assess each one for compatibility with the data. Thus, for each submodel $\mathcal{E}_m$, we can construct one realization q^o of the random variable $Q = U^{\mathrm{T}}Y/\|U^{\mathrm{T}}Y\|$ from § 4.1. The dependence of Q on the model $\mathcal{E}_m$ is not made explicit in the notation; nor is that of the corresponding null space $\mathcal{X}^\perp$ or the embedded manifold $\mathcal{Q}(s^o)$. The entries of Q are the coordinates, in the basis U, of the residual $\hat{\varepsilon}$ after standardization to unit length. Under the correct model $\mathcal{E}^*$, Q is uniformly distributed on the unit hypersphere in $\mathbb{R}^{n-d_\theta}$, as shown in § 4.1.

The difficulty is that, in the absence of a known alternative, a test based on q^o alone has no power in an assessment of model adequacy: regardless of which mean model the true distribution belongs, a single realization of Q under the induced distribution on the unit sphere will always be deemed compatible with uniformity. Although power in certain directions could be restored by consideration of alternative models for comparison, this leads to a test statistic essentially equivalent to the F-test, i.e., the likelihood-ratio test for Gaussian models, and therefore back to the initial difficulties discussed in § 1. Note, however, that even with two replicate observations of the random variable Q, there would be a degree of power against uniformity of distribution on $\mathcal{Q}(s^o)$, this tending to increase with the number of replicate observations.

This observation points to an initial suggestion for inference based on co-sufficient information partitioning to achieve some replication. The simplest partitioning approach, while conceptually revealing, is ineffectual for reasons given in § A in the Supplementary Material. Instead, we propose in § 5.2 below a more elaborate reformulation, adapting and extending the approach of Rasines & Young (2023), introduced in that work for a different purpose.

5.2. *Co-sufficient replication through randomization*

In the vein of Rasines & Young (2023), we first generate independent pseudo-replicates of Y with common distribution $N(\mu, k\sigma^2 I_n)$. The idealized, but infeasible randomization, procedure is as follows. Let L be an $n \times (k-1)$ matrix, independent of Y, with independent

Downloaded from academic.oup.com/biomet/article/113/4/asag045/8726260 by Imperial College of Science Technology and Medicine user on 21 October 2026

$N(0,1)$ entries, and define the $n \times k$-dimensional matrix of pseudo-replicates as

$$[\tilde{Y}^{(1)} \cdots \tilde{Y}^{(k)}] = [Y\ L]\Gamma, \tag{5}$$

where Γ is a $k \times k$ matrix of coefficients tuned to obtain the desired distribution as

$$\Gamma = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ a_1 & -b_1 & -b_1 & \cdots & -b_1 \\ 0 & a_2 & -b_2 & \cdots & -b_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & a_{k-1} & -b_{k-1} \end{bmatrix}. \tag{6}$$

Here $\{a_i, b_i \colon i = 1, \ldots, k-1\}$ are deterministic coefficients that can be computed recursively as $a_i = \sigma\tilde{a}_i$ and $b_i = \sigma\tilde{b}_i$, where

$$\tilde{a}_1 = (k-1)^{1/2},\ \tilde{b}_1 = \frac{1}{\tilde{a}_1},\ \tilde{a}_{i+1} = \left(k - 1 - \sum_{j=1}^{i} \tilde{b}_j^2\right)^{1/2},\ \tilde{b}_{i+1} = \frac{1 + \sum_{j=1}^{i} \tilde{b}_j^2}{\tilde{a}_{i+1}}.$$

Proposition 1. *If* $Y \sim N(\mu, \sigma^2 I_n)$ *then* $\tilde{Y}^{(1)}, \ldots, \tilde{Y}^{(k)} \sim N(\mu, k\sigma^2 I_n)$ *independently.*

A proof is provided in § C.1 in the Supplementary Material. The above construction can be generalized to allow for any known positive-definite covariance matrix Σ in the distribution of Y by adjusting Γ appropriately.

The randomization scheme is a direct generalization of the construction introduced by Rasines & Young (2023) in a different context. In particular, Rasines & Young (2023) decomposed a random vector $Y \sim N(\mu, \sigma^2 I_n)$ into two vectors of the same size

$$Y^{(1)} = Y + \gamma\sigma L \sim N\{\mu, (1+\gamma^2)\sigma^2 I_n\},$$
$$Y^{(2)} = Y - \sigma L/\gamma \sim N\{\mu, (1+\gamma^{-2})\sigma^2 I_n\},$$

where $L \sim N(0,1)$ and $\gamma > 0$ calibrates the variance allocation. On applying this argument recursively $k-1$ times, we have, with $\gamma_1 = (k-1)^{1/2}$, a decomposition of Y into $Y^{(1)} \sim N(\mu, k\sigma^2 I_n)$ and $V_1 \sim N\{\mu, (k/k-1)\sigma^2 I_n\}$, and with $\gamma_j = (k-j)^{1/2}$, a decomposition of V_{j-1} into $Y^{(j)} \sim N(\mu, k\sigma^2 I_n)$ and $V_j \sim N\{\mu, (k/k-j)\sigma^2 I_n\}$. The matrix Γ emerges on writing $Y^{(i)}$ in terms of Y and the Gaussian errors, and collecting coefficients. The same construction can be obtained by an application of Proposition 2 of Dharamshi et al. (2025), but the explicit form of Γ that achieves a decomposition into equal-variance components is new to this work.

From pseudo-replicates $Y^{(i)}$, we obtain independent replicates of Q, constructed for any given model as $\tilde{Q}^{(i)} = U^{\mathrm{T}}\tilde{Y}^{(i)}/\|U^{\mathrm{T}}\tilde{Y}^{(i)}\|$ for $i = 1, \ldots, k$, where U is that of § 4.1. Since $U^{\mathrm{T}}Y$ are the coordinates of the projection $UU^{\mathrm{T}}Y$ onto $\mathcal{X}^{\perp}$, under correct submodel specification, $\mu = X\theta$, these k replicates are independent and uniformly distributed on the unit sphere in $\mathbb{R}^{n-d_\theta}$.

If σ^2 is known or precisely and accurately estimated, then a larger choice of k is intuitively better, as this plays the role of sample size for model assessment, determining the

Downloaded from academic.oup.com/biomet/article/113/4/asag045/8726260 by Imperial College of Science Technology and Medicine user on 21 October 2026

number of observations on the sphere. Since an actionable version of the randomization approach requires estimation of σ^2, which induces some asymptotically negligible dependence between pseudo-replicates, small k is better at small sample sizes, a point to which we return below.

The classical estimator, constructed from the residual sum of squares of the linear model including all covariates in the comprehensive set $\hat{\mathcal{E}}$, is consistent provided that $n-|\hat{\mathcal{E}}| \to \infty$. Since $\hat{\mathcal{E}}$ tends to overfit to the data in small samples, it typically underestimates σ^2 unless the sample size is large. Small-sample bias is reduced through the use of the refitted cross-fitting procedure of Fan et al. (2012), who provided statistical guarantees. Below we propose a minor adaptation of this estimator, whose distribution theory is simple modulo the necessary conditioning arguments, and convenient for the Gaussian linear model.

The estimator of Fan et al. (2012) partitions the outcome and covariate data into two disjoint subsets, $(Y^{(1)}, X^{(1)})$ and $(Y^{(2)}, X^{(2)})$, each of size $n/2$. Fan et al. (2012) proposed fitting a conservative variable selection procedure within each subset, such that the true set $\mathcal{E}^*$ is contained in both selected sets $\hat{\mathcal{E}}^{(1)}$ and $\hat{\mathcal{E}}^{(2)}$ with high probability. Let $\mathcal{X}_2^{(1)}$ be the space spanned by the columns of $X^{(1)}$ indexed by $\hat{\mathcal{E}}^{(2)}$, and let $\mathcal{X}_1^{(2)}$ be the space spanned by the columns of $X^{(2)}$ indexed by $\hat{\mathcal{E}}^{(1)}$. The projection matrices onto these two subspaces are written $P_2^{(1)}$ and $P_1^{(2)}$. The refitted cross-validation estimator $\hat{\sigma}^2_{\text{rcv}}$ of σ^2 is then given by the average of the two independent estimators

$$\hat{\sigma}_1^2 = \frac{\|(I - P_2^{(1)})Y^{(1)}\|_2^2}{n/2 - |\hat{\mathcal{E}}^{(2)}|}, \qquad \hat{\sigma}_2^2 = \frac{\|(I - P_1^{(2)})Y^{(2)}\|_2^2}{n/2 - |\hat{\mathcal{E}}^{(1)}|}.$$

Empirically, the best approach computes this estimate, or the one below, using only $[\gamma n]$ observations for some $0.5 < \gamma < 1$, so that n in the previous display is replaced by $[\gamma n]$. This further limits the dependence between $\hat{\sigma}^2_{\text{rcv}}$ and $U^{\mathrm{T}}Y$, resulting in a better calibration of the test statistic under the null distribution. In simulations, we found that $\gamma = 0.6$ was a good choice across a broad range of scenarios.

For the distribution theory of the test statistic in Theorem 1 below, it is convenient to use a slightly modified version of the refitted cross-validation estimator. Let $df_1 = n/2 - |\hat{\mathcal{E}}^{(2)}|$ and $df_2 = n/2 - |\hat{\mathcal{E}}^{(1)}|$, and define the modified refitted cross-validation estimator as

$$\hat{\sigma}^2_{\text{mrcv}} := \frac{df_1\hat{\sigma}_1^2 + df_2\hat{\sigma}_2^2}{df_1 + df_2}. \tag{7}$$

Proposition 2 below establishes a distributional approximation to the distribution of this estimator when the sizes of the larger models, $|\hat{\mathcal{E}}^{(1)}|$ and $|\hat{\mathcal{E}}^{(2)}|$, are fixed independently of the data. When $|\hat{\mathcal{E}}^{(1)}|$ and $|\hat{\mathcal{E}}^{(2)}|$ are fixed in advance, it is sensible to choose them to be equal, in which case $\hat{\sigma}^2_{\text{mrcv}}$ reduces to $\hat{\sigma}^2_{\text{rcv}}$. Unreported simulations for data dependent $|\hat{\mathcal{E}}^{(1)}|$ and $|\hat{\mathcal{E}}^{(2)}|$ show, however, that the conditional distributional approximation given in Proposition 2 is almost exact, provided that the sample size is not too small.

PROPOSITION 2. *Let $\nu = df_1 + df_2$. Let $|\hat{\mathcal{E}}^{(1)}|$ and $|\hat{\mathcal{E}}^{(2)}|$ be fixed independently of the data and satisfy $|\hat{\mathcal{E}}^{(1)}| + |\hat{\mathcal{E}}^{(2)}| = o(n)$. Assume that, for any set E such that $E \subseteq \{1, \ldots, p\}$ and $|E| = |\hat{\mathcal{E}}^{(-j)}|$, $X_E^{(j)}$ has full column rank for $j = 1, 2$, where the notation $-j$ denotes*

Downloaded from academic.oup.com/biomet/article/113/4/asag045/8726260 by Imperial College of Science Technology and Medicine user on 21 October 2026

the complementary index. Assume further that $\mathrm{pr}(\mathcal{E}^* \not\subseteq \hat{\mathcal{E}}^{(1)} \cap \hat{\mathcal{E}}^{(2)}) = O(n^{-1/2})$. *Then,*

$$\hat{\sigma}^2 \overset{\mathrm{D}}{=} \sigma^2 \frac{\chi^2_\nu}{\nu} + o_p(n^{-1/2}) \quad \text{as } n \to \infty,$$

where χ^2_ν *is a chi-square random variable with* ν *degrees of freedom.*

Let $\hat{Y}^{(1)}, \ldots, \hat{Y}^{(k)}$ be the k columns of the $n \times k$ matrix obtained by replacing Γ in (5) with $\hat{\Gamma}$, the realizable version of Γ with $\hat{\sigma}$ in place of σ. With $\hat{Q}^{(i)} = U^{\mathrm{T}}\hat{Y}^{(i)}/\|U^{\mathrm{T}}\hat{Y}^{(i)}\|$, Proposition B.1 in the Supplementary Material quantifies the maximum Euclidean distance between $\hat{Q}^{(1)}, \ldots, \hat{Q}^{(k)}$ and the idealized versions $\tilde{Q}^{(1)}, \ldots, \tilde{Q}^{(k)}$ that treat σ as known. This result, although reassuring, misses an important point: the relevant metric in which to measure differences between $\hat{Q}^{(1)}, \ldots, \hat{Q}^{(k)}$ and $\tilde{Q}^{(1)}, \ldots, \tilde{Q}^{(k)}$ depends on the way in which they are used to assess model adequacy, which is via their distribution under the null hypothesis. Since the weak topology is weaker than the norm topology, we can expect better results than those reflected by Proposition B.1 in the Supplementary Material. In particular, an intuitive reaction to the discussion of § 4.2 is that the distribution of $\tilde{Q}^{(1)}, \ldots, \tilde{Q}^{(k)}$ should be relatively unaffected by estimation of σ. The discussion to follow, and the proof of Theorem 1, verifies this intuition, and clarifies some more subtle points.

In contrast to the simple data-partitioning scheme considered in § A in the Supplementary Material, the replicates of Q obtained through randomization via (5) are approximately independent and identically distributed under mean models other than the postulated model $\mu = X\theta$. As a consequence, this decomposition is more amenable to standard testing procedures than that of § A in the Supplementary Material. While there are many tests of uniformity on a sphere, we focus on the high-dimensional Rayleigh test, as its properties are guaranteed under a high-dimensional regime that matches our requirements, and efficient implementations are available in standard software.

In the present context, the Rayleigh test is based on the statistic

$$R_n := \frac{\{2(n-d_\theta)\}^{1/2}}{k} \sum_{1 \leqslant i < j \leqslant k} \left\langle \frac{U^{\mathrm{T}}\tilde{Y}^{(i)}}{||U^{\mathrm{T}}\tilde{Y}^{(i)}||_2}, \frac{U^{\mathrm{T}}\tilde{Y}^{(j)}}{||U^{T}\tilde{Y}^{(j)}||_2} \right\rangle \xrightarrow{\mathrm{D}} N(0,1), \tag{8}$$

or, rather, its analogue with $\hat{Y}^{(i)}$ in place of $\tilde{Y}^{(i)}$. The test rejects for large values of R_n, calibrated against the asymptotic null distribution. If $\mu \notin \mathcal{X}$ then $\tilde{Q}^{(1)}, \ldots, \tilde{Q}^{(k)}$ are approximately a random sample from a projected Gaussian with nonzero mean, so small angles among them are indicative of model misspecification. The distributional convergence was ascertained by Paindaveine & Verdebout (2016) in a more general setting than needed here. Under the null hypothesis that the postulated mean model is correct, $U^{\mathrm{T}}\tilde{Y}^{(i)}$ is normally distributed with mean zero and variance $k\sigma^2 I_{n-d_\theta}$; thus, exact distribution theory for R_n can, in principle, be derived from the calculations of Fisher (1915). We have

$$\frac{U^{\mathrm{T}}\tilde{Y}^{(i)}}{||U^{\mathrm{T}}\tilde{Y}^{(i)}||_2} = \frac{\sigma^{-1}k^{-1/2}U^{\mathrm{T}}\tilde{Y}^{(i)}}{||\sigma^{-1}k^{-1/2}U^{\mathrm{T}}\tilde{Y}^{(i)}||_2},$$

Downloaded from academic.oup.com/biomet/article/113/4/asag045/8726260 by Imperial College of Science Technology and Medicine user on 21 October 2026

and the standardized quantity in the numerator is standard normally distributed in $n - d_\theta$ dimensions. It follows that the distribution of

$$\left\langle \frac{U^{\mathrm{T}}\tilde{Y}^{(i)}}{||U^{\mathrm{T}}\tilde{Y}^{(i)}||_2}, \frac{U^{\mathrm{T}}\tilde{Y}^{(j)}}{||U^{T}\tilde{Y}^{(j)}||_2} \right\rangle = \cos\left\{ \angle\left(\frac{U^{\mathrm{T}}\tilde{Y}^{(i)}}{k^{1/2}\sigma}, \frac{U^{\mathrm{T}}\tilde{Y}^{(j)}}{k^{1/2}\sigma} \right) \right\} =: \bar{R}$$

is that of the sample correlation coefficient between two standard normally distributed random variables from $n - d_\theta$ observation points. In the above display, $\angle(a, b)$ denotes the angle between vectors a and b. The density function of the sample correlation coefficient $\bar{R}$ is (Fisher, 1915)

$$f_{\bar{R}}(r) = \frac{\Gamma\{(n - d_\theta)/2\}}{\sqrt{\pi}\Gamma\{(n - d_\theta - 1)/2\}}(1 - r^2)^{(n-d_\theta-3)/2}, \qquad -1 < r < 1. \tag{9}$$

That the distribution of R_n in (8) is asymptotically standard normal follows essentially by the central limit theorem on noting that $\mathbb{E}(\bar{R}) = 0$ and

$$\begin{aligned} \mathrm{var}(\bar{R}) = & \frac{\Gamma\{(n - d_\theta)/2\}}{\sqrt{\pi}\Gamma\{(n - d_\theta - 1)/2\}} \int_0^1 s^{1/2}(1 - s)^{(n-d_\theta-3)/2}\,\mathrm{d}s \\ = & \frac{\Gamma\{(n - d_\theta)/2\}}{2\Gamma\{(n - d_\theta)/2 + 1\}} \\ \simeq & \frac{1}{n - d_\theta} \quad \text{as } n - d_\theta \to \infty, \end{aligned}$$

where we have used the gamma recursion $\Gamma(3/2) = \sqrt{\pi}/2$, and the asymptotic form comes from Stirling's approximation $\Gamma(x+a)/\Gamma(x) \simeq x^a$ for fixed a and large x. For the Gaussian case, the asymptotic distribution is essentially exact for k as small as 2, because in (9)

$$(1 - r^2)^{(n-d_\theta-3)/2} \simeq \mathrm{e}^{-(n-d_\theta)r^2/2} \quad \text{as } r \to 0, \tag{10}$$

which is the non-normalized density function of a Gaussian random variable with mean zero and variance $(n - d_\theta)^{-1}$.

Theorem 1 below shows how estimation of Γ from (6) using the modified refitted cross-validation estimator (7) affects the theoretical guarantee (8).

THEOREM 1. *Under the conditions of Proposition 2,*

$$\begin{aligned} \hat{R}_n := & \frac{\{2(n - d_\theta)\}^{1/2}}{k} \sum_{1 \leqslant i < j \leqslant k} \left\langle \frac{U^{\mathrm{T}}\hat{Y}^{(i)}}{||U^{\mathrm{T}}\hat{Y}^{(i)}||_2}, \frac{U^{\mathrm{T}}\hat{Y}^{(j)}}{||U^{\mathrm{T}}\hat{Y}^{(j)}||_2} \right\rangle \\ \stackrel{\mathrm{D}}{=} & R_n + O_p(\max\{k\nu^{-3/2}, n^{-1/2}\}) \quad \text{as } n, \nu \to \infty, \end{aligned}$$

where $\nu = df_1 + df_2 \asymp n$. *It follows from* (8) *that* $\hat{R}_n \stackrel{\mathrm{D}}{\to} N(0, 1)$ *provided that* $k \lesssim n^{1/2}$.

Downloaded from academic.oup.com/biomet/article/113/4/asag045/8726260 by Imperial College of Science Technology and Medicine user on 21 October 2026

Remark 1. Although the estimator of error variance looks to play a minimal role in Theorem 1, it is important for the proof strategy that the modified cross-validation estimator $\hat{\sigma}^2_{\mathrm{mrcv}}$ is used.

The error term is minimized with $k = 2$ and, in view of (10), we expect this choice to provide a close distributional approximation under the null model. There may be reasons to prefer larger k, particularly if this provides appreciable improvements in power to detect departures from the null model for only a small loss in coverage. Conditioning on $|\hat{\mathcal{E}}^{(1)}|$ and $|\hat{\mathcal{E}}^{(2)}|$ in the statement of Theorem 1 breaks the dependence between $\hat{Q}^{(1)}, \ldots, \hat{Q}^{(k)}$. In the simulations of § 6 below, where this aspect is implicitly averaged over in the simulation replicates, some mild dependence is introduced, which may contribute to the deterioration in the distribution theory for increasing k.

It may be questioned, in view of the discussion preceding Theorem 1, why the randomization of (5) is needed at all, as rotational invariance of the Haar measure implies that any vector v on a hypersphere in $\mathbb{R}^{n-d_\theta}$ gives rise to an angle $\langle Q, v\rangle$ that follows the distribution in (9). Such an approach, like that of § A in the Supplementary Material, ignores the orientation of the embedding manifold $\mathcal{X}^\perp$ in $\mathbb{R}^n$. Thus, while the null distribution theory is correct in an unconditional sense, the analysis is not relevant for detecting departures from the null model.

5.3. *A test based on the ancillary component*

As discussed in § 4.1 and § 4.2, $R^2 = \hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon}$ has a $\sigma^2\chi^2_{n-d_\theta}$ distribution and is ancillary for θ under the postulated mean model. To exploit this for model assessment, a consistent estimator of σ^2 is needed, as in § 5.2. Proposition 3 below establishes the error incurred by using the $\chi^2_{n-d_\theta}$ quantiles of the random variable $V := R^2/\sigma^2$ to approximate those of $\hat{V} := R^2/\hat{\sigma}^2$, on the basis of the arbitrary consistent estimator $\hat{\sigma}^2$, which needs to be assumed independent of R^2 for the theoretical derivations. The two estimators $\hat{\sigma}^2_{\mathrm{rcv}}$ and $\hat{\sigma}^2_{\mathrm{mrcv}}$ described in § 5.2 clearly do not satisfy this independence assumption, and it does not seem possible to define an estimator with this property from the same sample. Nevertheless, our numerical investigations found that $\hat{\sigma}^2_{\mathrm{rcv}}$ and $\hat{\sigma}^2_{\mathrm{mrcv}}$, computed on a proportion $\gamma = 0.6$ of the data, have small enough dependence to yield reasonably good inferential performance, although slightly miscalibrated in small samples.

Proposition 3. *Let $\hat{\sigma}^2$ be a consistent estimator of σ^2, independent of R^2. Then,*

$$\sup_{v>0} |\,\mathrm{pr}(\hat{V} \leqslant v) - \mathrm{pr}(V \leqslant v)| = o(1).$$

Proof. The result is a direct application of Polya's theorem. An alternative derivation with possibly very conservative rates is given in § E in the Supplementary Material via a more explicit statement about the rate of convergence of $\hat{\sigma}^2$. □

As expected in view of the previous observations, we found in simulations that confidence sets of models based on the ancillary component R^2 have favourable properties over those based on the co-sufficient pseudo-replicates $\tilde{Q}^{(1)}, \ldots, \tilde{Q}^{(k)}$ if σ^2 is known, but that calibration is compromised in small samples once σ^2 is estimated. For sufficiently large sample size, the approach based on the ancillary information with σ^2 estimated is well calibrated. By calibrating $\hat{V}/(n - d_\theta)$ against the F distribution on $(n - d_\theta, \nu)$ degrees of freedom, with ν from Proposition 2 rather than the $\chi^2_{n-d_\theta}$ distribution, the ancillary test produces larger

Downloaded from academic.oup.com/biomet/article/113/4/asag045/8726260 by Imperial College of Science Technology and Medicine user on 21 October 2026

confidence sets with coverage close to 1. The explanation is that the F distribution assumes independence as in Proposition 3, an assumption that is violated in practice. Since R^2 and $\hat{\sigma}^2$ are positively correlated, their ratio has less variability than under independence, so it is more difficult to reject the test.

That the test based on R^2 is sensitive to estimation of σ^2 is not unexpected: R^2 estimates σ^2 when the submodel being tested is correct, so it is impossible to completely break the dependence between R^2 and $\hat{\sigma}^2_{\mathrm{rcv}}$ or $\hat{\sigma}^2_{\mathrm{mrcv}}$, as required in the idealized Proposition 3. On the other hand, since Q and R^2 are independent, $\tilde{Q}^{(1)}, \ldots, \tilde{Q}^{(k)}$ and $\hat{\sigma}^2_{\mathrm{mrcv}}$ are expected to be asymptotically independent. This can also be seen on noting that $\mathbb{E}(W)$ and σ^2 are orthogonal parameters in the sense of Huzurbazar (1950).

5.4. *Relation of co-sufficient and ancillary constructions to sample splitting*

Suppose, hypothetically, that the error variance was known. Then $\hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon}$ is no longer part of the minimal sufficient statistic, and all co-sufficient information would be contained in $W = U^{\mathrm{T}}Y$. A succinct summary is the scalar statistic $W^{\mathrm{T}}W = \hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon}$, whose distribution is known. From this perspective, if a precise and accurate external estimate of σ^2 was available, the problem is essentially solved and sample splitting can be viewed as an approximation to that. The perspective does not extend to most other types of regression context, where sample splitting is more problematic.

Another approximation to the above is the ancillary test of § 5.3, where $W^{\mathrm{T}}W$ is computed on the full sample. Violation of the independence assumption in Proposition 3 is inevitable unless $\hat{\sigma}^2$ is an external estimate, which is analogous to sample splitting in a larger sample.

6. Numerical comparisons

6.1. *Experimental design*

This section explores the numerical performance of the approaches based on the co-sufficient and ancillary separations relative to the simpler approach based on sample splitting, which inevitably discards relevant information. We focus on the Gaussian linear model of § 4.1. Conditional on all signal variables surviving the reduction phase, all three approaches have the correct theoretical coverage probability for sufficiently large sample size, but the approach based on the ancillary separation is expected to be more sensitive to estimation of the error variance in small samples, as discussed at the end of the previous section. The extent of this effect is uncovered through simulation in idealized settings. The approach based on sample splitting discards information at both the reduction phase and the model assessment phase. In small samples, it therefore discards signal variables with higher probability in the initial phase for a set $\hat{\mathcal{E}}$ of given size, or, conversely, includes more variables in $\hat{\mathcal{E}}$ for a fixed probability $\mathrm{pr}(\mathcal{E}^* \subseteq \hat{\mathcal{E}})$.

In 500 simulation replicates, n independent rows of X were generated from a normal distribution of dimension p with zero mean and covariance matrix Σ, where Σ is the identity matrix with the top-left 10×10 block replaced by an equi-correlation matrix with correlation $\rho \in (0, 1)$. The outcome Y was generated as $Y \sim N(X\theta, \sigma^2 I_n)$, where $\sigma^2 = 1$ and $\theta = (t, t, t, 0, \ldots, 0)^{\mathrm{T}}$; in other words, the true model is $\mathcal{E}^* = \{1, 2, 3\}$. The three signal variables were therefore correlated with each other and with seven irrelevant variables.

Since the emphasis of this paper is on the model assessment phase, we simplify the reduction phase by choosing $p = 400$, so that a single-stage version of Cox reduction can

Downloaded from academic.oup.com/biomet/article/113/4/asag045/8726260 by Imperial College of Science Technology and Medicine user on 21 October 2026

be used rather than the more general two-stage reductions that would be required with a considerably larger number of variables. If p is larger, the reduction phase is more difficult, and the true set of signal variables is less likely in a reduced data set to survive reduction, strengthening the argument for a version that uses all data for both phases of analysis. Detailed analyses of Cox reduction, with extensive numerical comparisons for higher dimensions and different modelling frameworks, can be found in Battey & Cox (2018), Hoeltgebaum & Battey (2019) and Lewis & Battey (2025). The latter paper provides insights into how the procedure used at the reduction phase interacts with the model assessment phase. Cox reduction for $p = 400$ proceeds by arranging the variable indices at random in a 20×20 grid and traversing the rows and columns, fitting a linear regression to each set of variables. Thus, each variable appears in two regressions alongside different companion variables. Variables that are not significant at a given significance level in either analysis are discarded, with the resulting set of variables forming the comprehensive model $\hat{\mathcal{E}}$. In a real application, the significance level in each 20-dimensional regression would ideally be chosen to give an acceptable level of stability for $\hat{\mathcal{E}}$ upon re-randomization of the variable indices in the grid. For simplicity here, we start with a significance level of 0.05, reducing it in increments of 0.001 until the model $\hat{\mathcal{E}}$ contains at most 15 variables. In very high-dimensional problems, a provisional reduction to 15 variables is likely to be too severe; the number can, in principle, be any value smaller than n, the main difficulty being the computational burden associated with testing a large number of submodels. This, however, is less problematic in a single application than in the many-replicate simulation considered here.

For comparison, we also use, for reduction, an undertuned lasso, with the parameter chosen such that approximately the same number of variables is retained in $\hat{\mathcal{E}}$ as were retained by Cox reduction in each simulation replicate. Specifically, the lasso was implemented with the smallest penalty that ensured selection of at most 15 predictors.

As both the co-sufficiency-based and ancillary-based tests require a consistent estimator of σ^2, prior to model assessment, we computed $\hat{\sigma}^2_{\text{mrcv}}$ as defined in (7), using the first $[0.6n]$ observations in order to reduce the dependence between $\hat{\sigma}^2_{\text{mrcv}}$ and the test statistics in the idealized known-variance setting. Estimation with the full sample led to invalid tests under the null, whereas the 60% version worked well over many scenarios. The conservative variable selection procedure employed for screening in the definition of $\hat{\sigma}^2_{\text{mrcv}}$ was the undertuned lasso procedure described before. The lasso, having excellent predictive properties, delivers good estimates of variance even when signal variables are discarded in favour of other variables.

For the construction of the confidence set of models $\mathcal{M}$, we compared the following testing procedures, assessing all subsets of $\hat{\mathcal{E}}$ of size 5 or smaller for their compatibility with the data. Those tests not rejected at the $\alpha = 0.05$ level constitute $\mathcal{M}$, a 5% confidence set for $\mathcal{E}^*$.

Co-sufficient information: k replicates $\hat{Q}^{(1)}, \ldots, \hat{Q}^{(k)}$ of Q were computed as described in § 5.2, and were tested for uniformity on the hypersphere in $\mathbb{R}^{n-d_\theta}$ using the Rayleigh test.

Ancillary information: the observed test statistic $\hat{V} = R^2 / \hat{\sigma}^2_{\text{mrcv}}$ was calibrated against its asymptotic null distribution $\chi^2_{n-d_\theta}$ and, after rescaling by $n - d_\theta$, against the F distribution on $(n - d_\theta, \nu)$ degrees of freedom, with ν from Proposition 2, as discussed in § 5.3. The two calibrations are labelled *ancillary 1* and *ancillary 2*, respectively.

Downloaded from academic.oup.com/biomet/article/113/4/asag045/8726260 by Imperial College of Science Technology and Medicine user on 21 October 2026

Table 1. $n = 100$, $t = 0.5$, $\rho = 0.1$

	$\mathbb{P}(\mathcal{E}^* \subseteq \mathcal{M})$	$\mathbb{P}(\mathcal{E}^* \subseteq \hat{\mathcal{E}})$	$\mathbb{E}(\lvert\mathcal{M}\rvert)$
Cox + co-sufficient ($k = 2$)	0.95 (0.01)	0.96 (0.01)	3029 (68)
Cox + co-sufficient ($k = 8$)	0.92 (0.01)	0.96 (0.01)	2463 (73)
Cox + ancillary 1	0.92 (0.01)	0.96 (0.01)	2399 (72)
Cox + ancillary 2	0.96 (0.01)	0.96 (0.01)	3221 (65)
Cox + F-test	0.02 (0.01)	0.96 (0.01)	5 (1)
Lasso + co-sufficient ($k = 2$)	0.98 (0.01)	0.99 (<0.01)	3263 (66)
Lasso + co-sufficient ($k = 8$)	0.95 (0.01)	0.99 (<0.01)	2696 (69)
Lasso + ancillary 1	0.94 (0.01)	0.99 (<0.01)	2605 (69)
Lasso + ancillary 2	0.99 (<0.01)	0.99 (<0.01)	3436 (61)
Lasso + F-test	0.00 (0.00)	0.99 (<0.01)	0 (0)
Split Cox + split F-test	0.52 (0.02)	0.55 (0.02)	1347 (60)
Split Lasso + split F-test	0.81 (0.02)	0.84 (0.02)	1522 (60)

No adjustment for selection: a naïve approach uses the same data for reduction and for model assessment based on a likelihood-ratio test of each submodel against the comprehensive model. Since we are in a Gaussian setting, the likelihood-ratio statistic reduces to the F-statistic. Under each null model $\mathcal{E}_m$ to be assessed, this F-statistic is calibrated against an F distribution with $|\hat{\mathcal{E}}| - |\mathcal{E}_m|$ and $n - |\hat{\mathcal{E}}|$ degrees of freedom, which would be the exact distribution under the null model had $\hat{\mathcal{E}}$ not been selected in the light of the data.
Sample splitting: the same F-test was performed as described above, except that the first 60% of observations were used for reduction and the remaining 40% for model assessment. This ratio is in the range suggested by Lewis & Battey (2025).

We did not attempt the Markov chain Monte Carlo approaches noted in the introduction, as there were 4943 submodels of $\hat{\mathcal{E}}$ to be assessed in each simulation run when restricting to models of maximal size 5; Markov chain Monte Carlo methods seem poorly suited for the confidence sets of models problem, in view of the computational requirements.

6.2. *Coverage and expected size of the confidence sets*

The quantities of interest are reported in Tables 1–4 for values of n, t and ρ each taken at one of two levels in combinations corresponding to a half-replicate of a factorial design: (low, low, low), (high, high, low), (high, low, high), (low, high, high). The reported outcome variables are the simulated probability that $\mathcal{E}^*$ is retained through the initial reduction; the simulated coverage probability of the confidence set $\mathcal{M}$ at nominal coverage 0.95 and the simulation average of the number of models in the confidence set. Numbers in parentheses indicate empirical standard deviations of the simulation averages.

The half-replicate of Tables 1–4 allows estimation of the main effects of increasing each of n, t and ρ from its low to its high level. Estimates are presented in Table 5, where the reported effects are multiplicative on the odds scale for probabilities and additive on the log scale for expected size, with values close to 1 and 0, respectively, suggesting a null effect.

The tables reveal several aspects that were anticipated qualitatively. Use of the F-test without sample splitting leads to extremely poor calibration of the confidence set of models, which is restored through sample splitting provided that the true model survives the reduction phase, this being more likely when n and t are large. Conditional on survival, the

Downloaded from academic.oup.com/biomet/article/113/4/asag045/8726260 by Imperial College of Science Technology and Medicine user on 21 October 2026

Table 2. $n = 120$, $t = 1$, $\rho = 0.1$

	$\mathbb{P}(\mathcal{E}^* \subseteq \mathcal{M})$	$\mathbb{P}(\mathcal{E}^* \subseteq \hat{\mathcal{E}})$	$\mathbb{E}(\lvert\mathcal{M}\rvert)$
Cox + co-sufficient ($k = 2$)	0.96 (0.01)	1.00 (0.00)	332 (24)
Cox + co-sufficient ($k = 8$)	0.88 (0.01)	1.00 (0.00)	137 (13)
Cox + ancillary 1	0.86 (0.02)	1.00 (0.00)	124 (12)
Cox + ancillary 2	0.98 (0.01)	1.00 (0.00)	311 (23)
Cox + F-test	0.41 (0.02)	1.00 (0.00)	31 (1)
Lasso + co-sufficient ($k = 2$)	0.96 (0.01)	1.00 (0.00)	402 (29)
Lasso + co-sufficient ($k = 8$)	0.88 (0.01)	1.00 (0.00)	158 (14)
Lasso + ancillary 1	0.86 (0.02)	1.00 (0.00)	146 (13)
Lasso + ancillary 2	0.98 (0.01)	1.00 (0.00)	364 (24)
Lasso + F-test	0.00 (0.00)	1.00 (0.00)	0 (0)
Split Cox + split F-test	0.94 (0.01)	0.99 ($<$0.01)	77 (3)
Split Lasso + split F-test	0.95 (0.01)	1.00 (0.00)	84 (3)

Table 3. $n = 120$, $t = 0.5$, $\rho = 0.5$

	$\mathbb{P}(\mathcal{E}^* \subseteq \mathcal{M})$	$\mathbb{P}(\mathcal{E}^* \subseteq \hat{\mathcal{E}})$	$\mathbb{E}(\lvert\mathcal{M}\rvert)$
Cox + co-sufficient ($k = 2$)	0.97 (0.01)	1.00 (0.00)	2863 (67)
Cox + co-sufficient ($k = 8$)	0.92 (0.01)	1.00 (0.00)	1872 (65)
Cox + ancillary 1	0.90 (0.01)	1.00 (0.00)	1747 (64)
Cox + ancillary 2	0.99 (0.00)	1.00 (0.00)	3010 (64)
Cox + F-test	0.37 (0.02)	1.00 (0.00)	54 (3)
Lasso + co-sufficient ($k = 2$)	0.98 (0.01)	1.00 (0.00)	2597 (55)
Lasso + co-sufficient ($k = 8$)	0.91 (0.01)	1.00 (0.00)	1841 (54)
Lasso + ancillary 1	0.90 (0.01)	1.00 (0.00)	1745 (53)
Lasso + ancillary 2	0.99 ($<$0.01)	1.00 (0.00)	2682 (53)
Lasso + F-test	0.00 (0.00)	1.00 (0.00)	0 (0)
Split Cox + split F-test	0.93 (0.01)	0.98 (0.01)	1200 (44)
Split Lasso + split F-test	0.91 (0.01)	0.95 (0.01)	1121 (37)

co-sufficient and ancillary tests are more conservative in settings where t is low or ρ is high. Additional simulations, not reported, found that the co-sufficient and ancillary tests were exactly calibrated at their nominal levels across all settings when the sample size was sufficiently large or when the true error variance was used instead of $\hat{\sigma}^2_{\text{mrcv}}$, suggesting that the small-sample over-coverage in low-information settings is due to variance estimation. When the variance is known or the sample is sufficiently large that the sampling variability of $\hat{\sigma}^2_{\text{mrcv}}$ is negligible, taking k as large as 20 gives coverage close to nominal, and produces smaller confidence sets than the co-sufficient test with smaller values of k, verifying intuition. These high-information settings are, however, cases where the split F-test dominates.

Unreported simulations with the error term following a Student-t distribution gave qualitatively similar conclusions. In that case, the distribution theory for the oracle version of the co-sufficient test remains valid, but that for the F-test is invalidated. The survival probabilities of the full set of signal variables are lowered in this case, and the distribution theory for $\hat{\sigma}^2_{\text{mrcv}}$ established in Proposition 2 is no longer valid, affecting Theorem B.1 in the Supplementary Material.

In the Gaussian context, because the F-test is exact, the coverage probability is affected only by the reduced survival probability at the first stage and not by the small sample size at the second stage, and the efficacy of sample splitting is high, the directionality of the

Downloaded from academic.oup.com/biomet/article/113/4/asag045/8726260 by Imperial College of Science Technology and Medicine user on 21 October 2026

Table 4. $n = 100$, $t = 1$, $\rho = 0.5$

	$\mathbb{P}(\mathcal{E}^* \subseteq \mathcal{M})$	$\mathbb{P}(\mathcal{E}^* \subseteq \hat{\mathcal{E}})$	$\mathbb{E}(\lvert\mathcal{M}\rvert)$
Cox + co-sufficient ($k = 2$)	0.94 (0.01)	1.00 (0.00)	620 (29)
Cox + co-sufficient ($k = 8$)	0.86 (0.02)	1.00 (0.00)	228 (13)
Cox + ancillary 1	0.83 (0.02)	1.00 (0.00)	200 (11)
Cox + ancillary 2	0.98 (0.01)	1.00 (0.00)	691 (27)
Cox + F-test	0.74 (0.02)	1.00 (0.00)	53 (1)
Lasso + co-sufficient ($k = 2$)	0.95 (0.01)	1.00 (0.00)	642 (27)
Lasso + co-sufficient ($k = 8$)	0.85 (0.02)	1.00 (0.00)	305 (16)
Lasso + ancillary 1	0.83 (0.02)	1.00 (0.00)	265 (14)
Lasso + ancillary 2	0.98 (0.01)	1.00 (0.00)	712 (25)
Lasso + F-test	0.00 (0.00)	1.00 (0.00)	0 (0)
Split Cox + split F-test	0.96 (0.01)	1.00 (0.00)	248 (11)
Split Lasso + split F-test	0.96 (0.01)	1.00 (0.00)	237 (10)

Table 5. *Point estimates of the effects of n, t and ρ from the half-replicate of the factorial design (Tables 1–4): multiplicative on the odds scale for $\mathbb{P}(\mathcal{E}^* \subseteq \mathcal{M})$ and additive for $\log \mathbb{E}(\lvert\mathcal{M}\rvert)$*

	$\mathbb{P}(\mathcal{E}^* \subseteq \mathcal{M})$			$\log \mathbb{E}(\lvert\mathcal{M}\rvert)$		
	n	t	ρ	n	t	ρ
Cox + co-sufficient ($k = 2$)	1.61	0.78	1.05	–0.34	–1.87	0.28
Cox + co-sufficient ($k = 8$)	1.09	0.58	0.92	–0.39	–2.50	0.12
Cox + ancillary 1	0.99	0.54	0.79	–0.40	–2.56	0.08
Cox + ancillary 2	2.03	1.01	2.03	–0.43	–1.90	0.37
Cox + F-test	2.65	12.8	10.9	0.92	0.90	1.46
Lasso + co-sufficient ($k = 2$)	1.12	0.44	0.89	–0.35	–1.75	0.12
Lasso + co-sufficient ($k = 8$)	0.83	0.47	0.64	–0.52	–2.32	0.14
Lasso + ancillary 1	0.85	0.50	0.68	–0.50	–2.38	0.10
Lasso + ancillary 2	1.00	0.49	1.00	–0.46	–1.79	0.21
Lasso + F-test	–	–	–	–	–	–
Split Cox + split F-test	2.83	5.11	4.33	–0.64	–2.22	0.53
Split Lasso + split F-test	1.37	3.25	1.73	–0.67	–2.23	0.37

F-test compensating to a large extent for the smaller sample size, provided that all signal variables survive reduction. In broad terms, therefore, the main aspect influencing the relative advantage of the co-sufficient and ancillary tests over the split-sample tests is the difficulty of the reduction phase. In nonlinear regression models, where a likelihood-ratio test has to be used in place of the F-test, the coverage probability of $\mathcal{M}$ is also affected by the smaller sample size in the second stage, due to the χ^2 approximation to the likelihood-ratio statistic, although this too could be improved by Bartlett correction. We discuss briefly in § F in the Supplementary Material the extension of co-sufficient model assessment to settings other than the Gaussian linear model.

Intermediate choices of k interpolate between the reported values of Tables 1–4 in terms of coverage probability and the expected size of the confidence set, with small k achieving higher coverage at the expense of larger confidence sets.

In all correctly calibrated cases, we see that the size of the confidence set of models is large, particularly for the co-sufficient and F-calibrated ancillary tests. Small errors of calibration seem worth the appreciable reduction in the number of models retained from splitting the

Downloaded from academic.oup.com/biomet/article/113/4/asag045/8726260 by Imperial College of Science Technology and Medicine user on 21 October 2026

sample. The large errors of coverage, as reported in Table 1, are more severe, making the co-sufficient approach more relevant at small sample sizes. The maximal size of the confidence set obtained by assessing models up to size 5 from a reduced set with $|\hat{\mathcal{E}}| = 15$ is 4943. Thus, all approaches have power to reject false models in these low-information settings. Cox & Battey (2017) discussed how to extract compact messages from large confidence sets of models.

7. Closing discussion

There is a strong contrast with the widely studied post-selection inference problems concerned with performing inference on a parameter following a data-based selection of effects on which formal inference is sought. That problem involves almost the same scientific question in different guises: a selection phase asking whether the effect is zero, and an inference phase asking what values of strongly suggested effects are compatible with the data. The context of the present paper is conceptually different in that the object of inference $\mathcal{E}^*$ is unambiguously defined. There is, furthermore, a clearer separation of information in the data for the different phases of analysis.

In the confidence sets of models formulation, the most relevant hypotheses, being small submodels of $\hat{\mathcal{E}}$, are stable under hypothetical replication from the same generating process. Provided that $\hat{\mathcal{E}}$ reliably includes $\mathcal{E}^*$, the double use of the data is only problematic if the overfitted $\hat{\mathcal{E}}$ is used as a basis for comparison of fit, as in the likelihood-ratio test. We avoided this comparison through co-sufficiency and ancillary separations. If the distribution of an ancillary statistic for the interest component depends on nuisance parameters, then the small-sample coverage properties of the ancillary-based assessment may be compromised through estimation of these. Numerical work showed that the efficiency of sample splitting is high for the Gaussian model due to the exactness of the second-stage test. There are modest advantages of co-sufficient randomized inference in settings where the reduction phase is difficult, for instance when the sample size is small relative to dimension or the signal strength is low.

A natural reaction to the tables in § 6 is that the number of models in the confidence set is inconveniently large, with the smallest correctly calibrated set containing 77 models. This is an honest reflection of the limitations of the data, and emphasizes the dangers of reporting a single model when many are compatible with the data at the chosen significance level. Statistical analysis can go no further, and any choice between models in the confidence set requires either additional data or subject-matter expertise. Cox & Battey (2017) discussed how one might extract compact messages from large confidence sets of models, as discussed in § 1.

If prediction is the sole objective, an arbitrary well-fitting model, such as that selected by the lasso, will typically suffice, and a confidence set of models is unnecessary. Confidence sets may nevertheless be used to form so-called confidence distributions for predictions, as discussed in § S3 of the supplement of Lewis & Battey (2025).

Although the ideas of § 3.1 and § 3.2 extend in principle to any parametric regression model admitting a sufficient, or approximately sufficient, statistic of dimension smaller than n, the details of implementation for non-Gaussian models need to be worked out on a case-by-case basis and may well involve considerable methodological innovation beyond the ideas presented in § 5. An exception is the use of the lognormal distribution for modelling positive

Downloaded from academic.oup.com/biomet/article/113/4/asag045/8726260 by Imperial College of Science Technology and Medicine user on 21 October 2026

outcomes, for which the ideas of § 5 apply directly after logarithmic transformation. The log-normal distribution is one of several popular parametric distributions for survival outcomes with covariate-dependence usually modelled in traditional accelerated-life form.

An exact or approximate analytic approach for non-Gaussian models is likely to lead to more marked improvements over sample splitting relative to the Gaussian case, as the F-test for model adequacy is exact, conditional on the true model $\mathcal{E}^*$ having survived the first stage, while the loglikelihood test needed for more general models is based on an asymptotic approximation. We have not considered such extensions in the present work beyond the brief comments collected in § F in the Supplementary Material, emphasizing the structure in canonical exponential-family regression models and accelerated-life models that simplifies the required calculations. It remains unclear whether a randomization approach paralleling § 5.2 would be necessary or fruitful.

Acknowledgement

We thank the two reviewers and the Associate Editor for their detailed reading and for indicating an oversight in the proof of Proposition 2. EPSRC Fellowship EP/T01864X/1 is gratefully acknowledged.

Supplementary material

The Supplementary Material includes additional results, proofs and technical details, and some discussion of the considerations involved in extending the work beyond the Gaussian setting.

References

Barber, R. & Janson, L. (2022). Testing goodness of fit and conditional independence with approximate cosufficient subsampling. *Ann. Statist*. **50**, 2514–44.
Barndorff-Nielsen, O. E. & Cox, D. R. (1994). *Inference and Asymptotics*. London: Chapman and Hall.
Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proc. R. Soc. Lond. A* **106**, 268–82.
Battey, H. S. (2022). Heather Battey's contribution to the discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes. *J. R. Statist. Soc. B* **84**, 696–8.
Battey, H. S. (2024). Maximal co-ancillarity and maximal co-sufficiency. *Inf. Geom*. **7**, 355–69.
Battey, H. S. & Cox, D. R. (2018). Large numbers of explanatory variables: a probabilistic assessment. *Proc. R. Soc. Lond. A* **474**, 20170631.
Battey, H. S. & Cox, D. R. (2022). Some perspectives on inference in high dimensions. *Statist. Sci.* **37**, 110–22.
Battey, H. S. & McCullagh, P. (2024). An anomaly arising in the analysis of processes with more than one source of variability. *Biometrika* **111**, 677–89.
Battey, H. S. & Reid, N. (2023). On inference in high-dimensional regression. *J. R. Statist. Soc. B* **85**, 149–75.
Battey, H. S. & Reid, N. (2026). Induced replication and the assessment of models. *arXiv:* 2603.27718v2.
Breiman, L. (2001). Statistical modeling: the two cultures. *Statist. Sci*. **16**, 199–231.
Cambanis, S., Huang, S. & Simons, G. (1981). On the theory of elliptically contoured distributions. *J. Multi. Anal*. **11**, 368–85.
Cox, D. R. (1968). Notes on some aspects of regression analysis (with discussion). *J. R. Statist. Soc. A* **131**, 265–79.
Cox, D. R. (1975). A note on data-splitting for the evaluation of significance levels. *Biometrika* **62**, 441–4.
Cox, D. R. (1977). The role of significance tests (with discussion). *Scand. J. Statist*. **4**, 49–70.
Cox, D. R. & Battey, H. S. (2017). Large numbers of explanatory variables, a semi-descriptive analysis. *Proc. Nat. Acad. Sci*. **114**, 8592–5.
Cox, D. R. & Snell, E. J. (1974). The choice of variables in observational studies. *Appl. Statist.* **23**, 51–9.
Cox, D. R. & Snell, E. J. (1989). *Analysis of Binary Data*, 2nd edn. London: Chapman and Hall.
Dharamshi, A., Neufeld, A., Gao, L. L., Bien, J. & Witten, D. (2025). Decomposing Gaussians with unknown covariance. *Biometrika* **113**, asaf057.

Downloaded from academic.oup.com/biomet/article/113/4/asag045/8726260 by Imperial College of Science Technology and Medicine user on 21 October 2026

Engen, S. & Lillegård, M. (1997). Stochastic simulations conditioned on sufficient statistics. *Biometrika* **84**, 235–40.
Fan, J., Guo, S. & Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. R. Statist. Soc. B* **74**, 37–65.
Fang, K.-T., Kotz, S. & Ng, K. (1990). *Symmetric Multivariate and Related Distributions*. London: Chapman and Hall.
Ferrari, D. & Yang, Y. (2015). Confidence sets for model selection by F-testing. *Statist. Sinica* **25**, 1637–58.
Fithian, W., Sun, D. L. & Taylor, J. E. (2017). Optimal inference after model selection. *arXiv:* 1410.2597v4.
Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, **10**, 507–21.
Fisher, A., Rudin, C. & Dominici, F. (2019). All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.* **20**, 1–81.
Hansen, P. R., Lund, A. & Nason, J. M. (2011). The model confidence set. *Econometrica* **79**, 453–97.
Hoeltgebaum, H. H. & Battey, H. S. (2019). HCmodelSets: an R package for specifying sets of well-fitting models in high dimensions. *R J.* **11**, 370–9.
Huzurbazar, V. S. (1950). Probability distributions and orthogonal parameters. *Proc. Camb. Phil. Soc*. **46**, 281–4.
Guan, L. & Tibshirani, R. (2020). Post model-fitting exploration via a 'next-door' analysis. *Can. J. Statist*. **48**, 447–70.
Kano, Y. (1994). Consistency property of elliptic probability density functions. *J. Mult. Anal*. **51**, 139–47.
Kariya, T. (1980). Locally robust tests for serial correlation in least squares regression. *Ann. Statist.* **8**, 1065–70.
Kartsonaki, C., Pang, Y., Millwood, I., Yang, L., Guo, Y., Walters, R., Lv, J., Hill, M., Yu, C., Chen, Y. et al. (2022). Circulating proteins and risk of pancreatic cancer: a case-subcohort study among Chinese adults. *Int. J. Epidemiol*. **51**, 817–29.
Kim, I. & Ramdas, A. (2025). Locally minimax optimal confidence sets for the best model. *arXiv:* 2503.21639v4.
King, M. L. (1980). Robust tests for spherical symmetry and their application to least squares regression. *Ann. Statist.* **8**, 1265–71.
Lee, J. D., Sun, D. L., Sun, Y. & Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44**, 907–27.
Lei, J. (2020). Cross-validation with confidence. *J. Am. Statist. Assoc*. **115**, 1978–97.
Lewis, R. M. & Battey, H. S. (2025). Cox reduction and confidence sets of models: a theoretical elucidation. *Statist. Sci.* **40**, 313–28.
Lindqvist, B. H. & Taraldsen, G. (2005). Monte Carlo conditioning on a sufficient statistic. *Biometrika* **92**, 451–64.
Lockhart, R. A., O'Reilly, F. J. & Stephens, M. A. (2007). Use of the Gibbs sampler to obtain conditional tests, with applications. *Biometrika* **94**, 992–8.
Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. New York: John Wiley and Sons.
Paindaveine, D. & Verdebout, T. (2016). On high-dimensional sign tests. *Bernoulli* **22**, 1745–69.
Prosser, N. S., Greem, L. E. & Green, M. J. (2026). Role of stability and triangulation-based methods to improve identification of risk factors for lameness in ewes. *Prev. Vet. Med*. **247**, 106767.
Rasines, D. G. & Young, G. A. (2023). Splitting strategies for post-selection inference. *Biometrika* **110**, 597–614.
Su, W. (2018). When is the first spurious variable selected by sequential regression procedures? *Biometrika* **105**, 517–27.
Su, W., Bogdan, M. & Candès, E. (2017). False discoveries occur early on the lasso path. *Ann. Statist.* **45**, 2133–50.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B* **58**, 267–88.
Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. (2020). A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Statist. Soc. B* **82**, 1273–300.
Wang, P., Xie, M-X. & Zhang, L. (2026). Finite- and large sample inference for model and coefficients in high-dimensional linear regression with repro samples. *Ann. Statist.* **54**, 834–60.
Zhang, T., Lee, H. & Lei, J. (2026). Winners with confidence: discrete argmin inference with an application to model selection. *To appear in J. R. Statist. Soc. B*.

[*Received on* 11 *July* 2025. *Accepted on* 25 *June* 2026]

Downloaded from academic.oup.com/biomet/article/113/4/asag045/8726260 by Imperial College of Science Technology and Medicine user on 21 October 2026