Statistical Science
2022, Vol. 37, No. 1, 110–122
https://doi.org/10.1214/21-STS824

Some Perspectives on Inference in High Dimensions

H. S. Battey and D. R. Cox

Abstract. With very large amounts of data, important aspects of statistical analysis may appear largely descriptive in that the role of probability sometimes seems limited or totally absent. The main emphasis of the present paper lies on contexts where formulation in terms of a probabilistic model is feasible and fruitful but to be at all realistic large numbers of unknown parameters need consideration. Then many of the standard approaches to statistical analysis, for instance direct application of the method of maximum likelihood, or the use of flat priors, often encounter difficulties. After a brief discussion of broad conceptual issues, we provide some new perspectives on aspects of high-dimensional statistical theory, emphasizing a number of open problems.

Key words and phrases: Inference, likelihood, model uncertainty, nuisance parameters, parameter orthogonalization, sparsity.

1. INTRODUCTION

In broad terms, probability may be needed to describe a context in the initial planning phases of an investigation, in particular to assess measurement methods, and at least in outline to check that the proposed data are potentially capable of addressing the questions of concern. Then for the various later phases of analysis, leading to the presentation of conclusions and their consequences, an implicit or explicit probability base is typically needed, allowing incisive exposition of conclusions and some assessment of their security. The contrast and interplay between probability as a representation of empirical phenomena and as a tool for representing uncertainty of knowledge raises subtleties even with small amounts of data, only enhanced with delicate changes of emphasis when there are very large amounts of data.

In the current paper, we restrict ourselves to data for which the broad form of probabilistic model appropriate for interpretation has been temporarily agreed. We study situations with n study individuals regarded as independent and with some common underlying structure. On each individual there are q variables representing outcomes and p variables measuring features that could potentially influence the outcome if hypothetically altered, keeping all other features at their given levels. Even in this relatively simple context, whenever there is not an exact solution, a number of distinct formulations occur according to which of (n, q, p) are fruitfully treated as large. Theoretical questions can then be studied from at least two perspectives. For what values of (n, q, p) do specific procedures of analysis being studied give numerical procedures of acceptable statistical behaviour? When is further refinement needed?

H. S. Battey is Assistant Professor, Department of Mathematics, Imperial College London, SW7 2AZ, UK (e-mail: h.battey@imperial.ac.uk). D. R. Cox is Honorary Fellow, Nuffield College, University of Oxford OX1 1NF, UK.

2. BROAD CONCEPTUAL ISSUES

2.1 Two Roles of Conditioning

Suppose that we regard as provisionally given a parametric probability model for the data specified in terms of a parameter vector (ψ, λ), where ψ is of specific subject-matter interest and λ is a vector of nuisance parameters. Consideration of conditional distributions can arise in two quite different ways.

First, for arbitrary fixed ψ there may be a minimal sufficient statistic for λ, and then the distribution of the data conditionally on that statistic does not depend on λ and so is potentially available for inference about ψ. This is uncontroversial unless the conditional distribution is nearly degenerate pointing perhaps to inadequacies of the data or to too demanding a specification.

The second quite distinct possibility is that even though there are no nuisance parameters, or these have in effect been eliminated from the discussion, the minimal sufficient statistic is of higher dimension than the parameter space, the difference in dimensionality being d_A, the dimension of a statistic A. Then A, part of the minimal sufficient statistic, is called *ancillary* if its distribution does not depend on the parameter, ψ. The qualitative interpretation

is that the observed value a has bearing on the precision achievable in the particular set of data under analysis, as contrasted with that in some larger less tightly specified context.

This is of common relevance in regression type problems, sometimes even in simple linear regression. Analysis concerning the regression coefficient is based conditionally on the observed values of the explanatory variables even if these have a known probability distribution; precision is to be assessed through the values realized, not by those that would be observed in some hypothetical future realizations. In fact, for standard low-dimensional linear regression, by the form of the minimal sufficient statistic for the regression coefficients and error variance, we need condition only on the mean and sum of squares and products of the explanatory variables. The key point is to calibrate our assessments of uncertainty by behaviour in hypothetical repetitions as relevant as is feasible for the specific data under study. See Buja et al. (2019) for a contrary view, rebutted by Davison, Koch and Koh (2019).

2.2 Significance Testing as a Primary Mode of Inference

This paper concentrates on considerations that arise when a relatively complicated model is needed, but first we give a single example illustrating difficulties in a notionally very simple situation which therefore may, possibly unexpectedly, endanger larger problems.

In some forms of bioassay (Bliss, 1935), estimation of a key property, ED50, of a dose-response function is essentially equivalent to the estimation of the ratio of two normal means. Here, in an idealized form of the original problem, observed random variables (Y_1, Y_2) are independently normally distributed with means (μ_1, μ_2) and unit variance, and interest lies in $\psi = \mu_2/\mu_1$. Depending on the data, reasonable conclusions about ψ may be that it lies in the inside of an interval, the outside of an interval or the whole real line. The latter is not a vacuous statement but a strong warning about the limitations of the data. This example exposes difficulties with over-formal interpretations of the Neyman–Pearson theory, which require attainment of a prespecified unconditional coverage probability.

The situation just mentioned, and other similar anomalies, arise naturally in applied contexts. They exhibit a general formulation in which estimation by some form of confidence region may unexpectedly be misleading. For general discussion, primary emphasis should therefore be on assessing whether the data are consistent with an arbitrarily specified value of the unknown parameter of interest. Often but not always this will lead to confidence intervals or closed regions.

A central question even in low dimensional situations concerns the most fruitful and secure summarization of evidence about a parameter of interest, which may in particular be scalar. While specification by a series of nested confidence limits or regions at various levels may often seem an appealing resolution, it cannot be regarded as a satisfactory general approach because of the possibility that in some realizations, specifications by confidence intervals or limits are inappropriate. Therefore, we prefer to regard as the primary base for interpretation a series of p-values corresponding to possible values of the parameter of interest. This may indicate that all possible values are reasonably consistent with the data, that none is, that values within certain designated subsets are acceptable, and so on.

The distinction between applications of and motivations for various types of significance tests is discussed in detail by Cox (2020).

3. THREE BROAD ASYMPTOTIC REGIMES

For a preliminary discussion, we do not distinguish between parameters of direct interest and nuisance parameters needed to complete the specification but of no special concern. Nor do we differentiate between potential explanatory variables and outcomes, in effect setting q to zero. It is then useful to separate the following situations. As n increases:

(i) p is fixed or $p(n)/n \to 0$;
(ii) $p(n)/n \to \kappa \in (0, 1]$;
(iii) $p(n)/n$ diverges or tends to a limit greater than one.

These categorizations belong to formal theory. For any notional limiting operation, $\alpha \to \alpha^*$ say, applicability of procedures thus justified often extends to values of α far from α^*. This is typified by Stirling's approximation to the gamma function $\Gamma(x)$, remarkably accurate for all $x \geq 1$ in spite of the notional requirement that $x \to \infty$. A natural question emerges: what are the fundamental limits of inference, in terms of the maximum permissible value of κ or similar, beyond which inferential tools developed under regime (i) deteriorate beyond use? Partial results in specific contexts have been obtained, inter alia, by Lei, Bickel and El Karoui (2018), Sur and Candès (2019), Fan, Demirkaya and Lv (2019), Tang and Reid (2020) and Anastasiou and Reinert (2020).

Regime (i) is commonly referred to as *low dimensional*, (ii) as *high dimensional* and (iii) as *ultra-high dimensional*. The objective is to produce approximations and associated statistical recommendations that perform well for realistic values of (p, n). Thus, if a problem entails $p \geq n$, regime (ii) or (iii) is implicated by a theoretical analysis for large n.

We focus primarily on the special issues raised by regimes (ii) and (iii). Inference under regime (i) is

discussed in detail elsewhere. For instance, Barndorff-Nielsen and Cox (1994) emphasize Fisherian ideas and approaches to inference on interest parameters in the presence of nuisance parameters. Their treatment in Chapters 4–8 involves modifying likelihood-based statistics, modifying the associated limiting distributions, or both, such that the distributional approximation holds to a higher order of accuracy in n than would hold for the unmodified statistic/distribution pair. Thus, while formally justified under regime (i), the errors in the distributional approximations decay, by construction, faster than $n^{-1/2}$, the approximation error rate in the central limit theorem. The implication is that, relative to the number of nuisance parameters, a smaller sample size is needed in the adjusted procedure for an approximation with the same accuracy as the unadjusted procedure. So-called higher-order inference is therefore relevant in the types of applications for which regime (ii) could alternatively be considered. Tang and Reid (2020) provide some formal analysis of the extent to which distributional approximations for modified likelihood procedures, originally justified under regime (i), continue to hold under regime (ii).

4. INFERENCE IN HIGH DIMENSIONS

4.1 Consequences of Parameterization

When the number of nuisance parameters is appreciable relative to the number of independent observations, maximum likelihood may produce severely biased estimators of interest parameters. Bartlett (1937) gave a simple yet striking example. In a normal theory linear model the maximum likelihood estimate of the unknown variance is the residual sum of squares divided by sample size. In the extreme case of matched pairs with pair-specific means and equal variances, this would have expectation one-half the true variance. This example is sometimes referred to as a Neyman–Scott problem. In view of the connection between maximum likelihood and Bayesian inference with flat priors, together with the extensive use of Markov chain Monte Carlo to fit large Bayesian models, Bartlett's (1937) example highlights the considerable potential for highly miscalibrated procedures in high dimensions.

The analysis of the saturated $n = 2^k$ factorial experiment is a helpful further example. Normally distributed outcomes $Y_1, \ldots, Y_n$ are observed on each of the possibly large number n of treatment combinations and the corresponding vector of treatment combination means is denoted by μ. An alternative representation is $\mu = H^{-1}\tau$, where H is a $2^k \times 2^k$ Hadamard matrix and $\tau = (2^k\tau_1, 2^{k-1}\tau_2, \ldots, 2^{k-1}\tau_n)$ where $\tau_1 = \mu_{(1)}$ is the overall mean and $\tau_2, \ldots, \tau_n$ are contrasts, defined as main effects and interactions of various orders; each such is a contrast of two groups of $n/2$ observations. See Cox and Reid (2000, pp. 110–116) for a detailed account of the analysis of 2^k factorial experiments. The maximum likelihood estimator of the mean vector μ is the single observation $Y = (Y_1, \ldots, Y_n)^T$ while that of the contrasts is the average of 2^k mean differences and has variance $4\sigma^2/n$, where σ^2 is the common variance of Y_i, $i = 1, \ldots, n$. Thus while the maximum likelihood estimators of the means are unbiased but inconsistent, the estimators of the contrasts are consistent at the usual low dimensional parametric rate $n^{-1/2}$, even though there are $n - 1$ of them.

As an aside, this illustrates a version of Stein's (1956) problem. The contrast estimators $\hat{\tau}_2, \ldots, \hat{\tau}_n$ are independent by the orthogonality of the Hadamard matrix and therefore, on using Markov's inequality after exponential transformation, Jensen's inequality and the moment generating function of normal random variables,

$$\mathrm{pr}\Big(\max_{2\le j\le n} |\hat{\tau}_j - \tau_j| > t\Big) \le (n-1)\exp\{-nt^2/(32\sigma^2)\}.$$

This is $o(1)$ provided that $t \gtrsim 2\sigma\{8n^{-1}\log(n)\}^{1/2}$. Thus, the vector of contrast estimators is consistent in the maximum norm but not in the stronger Euclidean norm. However, this is not the point we wish to emphasize here, rather that some parameterizations are conducive to good behaviour while others are not.

The analysis of factorial experiments is not the only example illustrating this problem. Consider m normally distributed matched pairs (Y_{i0}, Y_{i1}) with known unit variance and means $(\lambda_i, \lambda_i + \delta)$. There are $n = 2m$ observations in all. Write $Z = MY$, where Y is $n \times 1$, Z is $m \times 1$ and $\mathbb{E}(Z) = \Delta = \delta 1_m$ so that M is the $m \times n$ matrix whose jth row has $(1, -1)$ in its $(2j-1)$th and $2j$th positions and zeros elsewhere. Suppose an initial formulation has $E(Y) = X\Delta$, that is, $X = (M^TM)^{-1}M^T$ and the standardized information matrix for Δ is $X^TX = 2I$. Now suppose we make a nonsingular linear transformation in the parameters, writing $\Theta = L\Delta$ so that $E(Y) = XL^{-1}\Theta$ and the standardized information matrix for Θ is

$$(L^{-1})^T X^T X L^{-1} = 2(L^{-1})^T L^{-1}.$$

If the errors in the original model are independent and identically distributed with finite variance and the sum of squares of the elements of each row of L^{-1} diverge as n increases, then, under standard regularity conditions, any finite subset of components of $\hat{\Theta}$ is asymptotically normal. The argument illustrates that there is nothing special about the matched pair problem in that, for any normal-theory linear problem with known variance and the number of parameters increasing proportionally to n, standard results will apply unless we choose, as is always possible, to have parameter components that are perverse. The only special feature of the matched pair problem is that the perverse components are directly interpretable.

For a more general formalization, let θ be coordinates of a parameterization, if one exists, in which the (s, s)th

entry of the associated Fisher information matrix satisfies $i_{ss}^{(\theta)}(\hat{\theta}) = O(n)$ for all $s = 1, \ldots, p$, where $\hat{\theta}$ is the maximum likelihood estimate. In other words, we suppose there exists a parameterization in which the variances of the maximum likelihood estimates decay at the usual parametric rate. Motivated by the saturated factorial experiment, consider a linear nonsingular transformation to $\psi = \psi(\theta)$ such that the elements of the transformation matrix satisfy

$$\frac{\partial \theta^s}{\partial \psi^t} = \gamma_{ts} B_s, \quad t = 1, \ldots, p \tag{1}$$

where $B_s = O(p^{-1})$ and $\gamma_{ts} \in \{-1, 1\}$ for all s. The information matrix transforms as

$$\begin{aligned} & i_{tt}^{(\psi)}\{\psi(\hat{\theta})\} \\ &\quad = \sum_{s,r} \frac{\partial \theta^s}{\partial \psi^t} i_{sr}^{(\theta)}(\hat{\theta}) \frac{\partial \theta^r}{\partial \psi^t} \\ &\quad = \sum_s \left(\frac{\partial \theta^s}{\partial \psi^t} \right)^2 i_{ss}^{(\theta)}(\hat{\theta}) + 2 \sum_s \sum_{r>s} \frac{\partial \theta^s}{\partial \psi^t} i_{sr}^{(\theta)}(\hat{\theta}) \frac{\partial \theta^r}{\partial \psi^t}. \end{aligned}$$

The first term is $\sum_s B_s^2 i_{ss}^{(\theta)}(\hat{\theta}) = O(n^2/p^2) = O(1)$ if $p \asymp n$. The second is

$$2 \sum_s \sum_{r>s} \gamma_{ts} \gamma_{tr} B_s B_r i_{sr}^{(\theta)}(\hat{\theta}),$$

which is $O\{\max_{s,r} i_{sr}^{(\theta)}(\hat{\theta})\}$ almost everywhere in reparameterization space, but can always be made to disappear by suitable choices of the γ_{ts} terms. This shows that, starting from a sensible parameterization, formal manoeuvres may lead to perverse parameterizations in which the information does not accumulate with the sample size.

The transformation (1) was chosen by working backwards from the contrasts parameterization of the $n = 2^k$ factorial experiment, but the conclusion holds more generally. On writing $\theta = \tau$, the vector comprising the population mean and factorial contrasts, and $\psi = \mu$, the vector of population treatment means, the associated transformation matrix with entries described by (1) is $(\partial\theta/\partial\psi) = C \circ H$ where $\circ$ denotes the Hadamard product, H is a Hadamard matrix with entries in $\{-1, 1\}$ and C is a $2^k \times 2^k$ matrix whose first column entries are $1/2^{k-1} = 2n^{-1} = O(p^{-1})$ and whose remaining entries are $1/2^k = n^{-1} = O(p^{-1})$.

There are two broad questions:

(a) Why and when do standard inferential procedures such as maximum likelihood fail in high dimensions? In what sense do they fail and in which directions of parameter space?

(b) What is the resolution to any such failure?

These questions remain mostly unanswered but we provide some general insights in Section 4.2 and highlight open problems in Section 5.3.

4.2 Likelihood: Two Sources of Bad Behaviour in High Dimensions

The previous section illustrates two types of problem that may arise in high dimensions. One is concentration of the likelihood function near a wrong point due to the nuisance parameters, as exemplified by the maximum likelihood estimator of the variance in Bartlett's (1937) example. The other is failure to concentrate at all, as exemplified by the maximum likelihood estimators of the pairwise means in the same example with known variance, and by that of the treatment combination means in the factorial experiment. Between these two extreme situations a combination of the two aspects operate.

With ψ an interest parameter and $\lambda = (\lambda_1, \ldots, \lambda_{p-1})$ a vector of nuisance parameters, let s be a jointly minimal sufficient statistic for (ψ, λ). To allow the situation in which s does not consist of p separate sufficient statistics for each parameter, write $s = (s_\psi, s_{-\psi}) = (s_j, s_{-j})$ for $j = 1, \ldots, p-1$, where s_ψ is a minimal sufficient statistic for ψ and s_j is minimal sufficient for λ_j. Here we suppress the explicit form of randomness and refer instead to the sensitivity of $\hat{\psi}$ to perturbations in an arbitrary scalar component x of the observed data. The total derivative of $\hat{\psi}$ with respect to x is

$$\frac{d\hat{\psi}}{dx} = \frac{\partial \hat{\psi}}{\partial s_\psi} \frac{ds_\psi}{dx} + \sum_{j=1}^{p-1} \frac{\partial \hat{\psi}}{\partial \hat{\lambda}_j} \frac{\partial \hat{\lambda}_j}{\partial s_j} \frac{ds_j}{dx}, \tag{2}$$

with the obvious changes if any of the terms s_ψ or s_j $(j = 1, \ldots, p-1)$ are not scalar valued. This isolates two sources of variation. One is due to sensitivity of the sufficient statistic s_ψ to perturbations in the data. A second is due to sensitivity of sufficient statistics s_j combined with sensitivity of $\hat{\psi}$ to small variations in $\hat{\lambda}_j$. The latter situation is characterized by the geometry of the likelihood function, as illustrated below. A further aspect, strongly attributable to dimension, is that the second pair of sensitivities might be individually small but collectively large on summation of $p-1$ nuisance parameter contributions.

To understand which coordinates of parameter space correspond to large values of $|\partial\hat{\psi}/\partial\hat{\lambda}_j|$, a geometric illustration is helpful. In low dimensional contexts these geometric properties of the log likelihood function ℓ are sources of finite sample bias, which disappears asymptotically.

Large deviations in an estimate of ψ would arise due to sampling error in other estimates $\hat{\lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_p)$ if the corresponding contour $\{(\psi, \lambda) : \nabla_\psi \ell(\psi, \lambda) = 0\}$ curves in the ψ direction as one moves in one or more of the λ directions. Intuitively, if the contour is (locally) flat as a function of the nuisance parameter(s), then the maximum likelihood estimator of ψ is the same regardless of whether the log likelihood is evaluated at the true value of

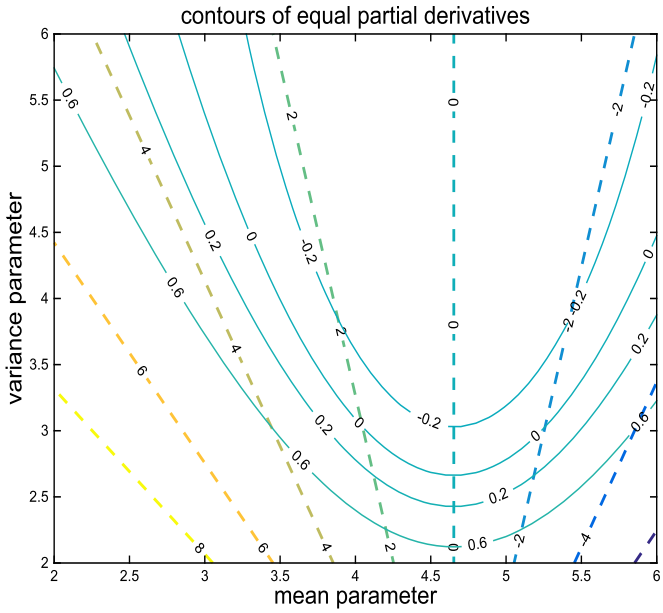

FIG. 1. *Contours* $\{(\psi,\lambda):\nabla_\psi \ell(\psi,\lambda)=c\}$ *for different values of* c *for one realization of* ℓ*, the log likelihood function corresponding to* 10 *observations from a normal distribution of mean* 4 *and variance* 4. *Dashed lines correspond to* ψ *being the mean parameter and solid lines to* ψ *being the variance parameter.*

λ or at an estimate. Figure 1 gives one of the simplest illustrations of the problem for estimation of the mean and variance of a normal distribution. This gives geometric insight into what is known from simple algebra: estimation of the mean is unbiased for any sample size, while estimation of the variance suffers small-sample bias. As is clear from equation (2), unless there are cancellations, the problem is amplified by there being a large number of such parameters, as in the normal-theory linear model with an appreciable number of regression coefficients and unknown error variance, treated as the interest parameter. For average behaviour of $\hat{\psi}$, deviations of $\hat{\lambda}_j$ in both directions are relevant, so that finite sample bias is exacerbated if the curvature is symmetric about $\hat{\lambda}_j$.

The randomness from the sampling was suppressed in equation (2), but is captured in essence by the terms ds_ψ/dx and ds_j/dx. Suppose that $X_1,\dots,X_n$ are independent or weakly dependent random variables, not identically distributed in general. The fluctuations of a statistic $s(X_1,\dots,X_n)$ around its mean $Es(X_1,\dots,X_n)$ are small provided that the function s is not too sensitive to any of the coordinates x_i. For bounded s, a two-sided generalization of McDiarmid's inequality (e.g., Vershynin, 2018, p. 40) provides the following upper bound.

LEMMA 1. *Let* $X_1,\dots,X_n$ *be independent. For all* $t>0$,

$$\operatorname{pr}\bigl(|s(X_1,\dots,X_n)-Es(X_1,\dots,X_n)|>t\bigr) \le \exp(-t^2/2v),$$

where $v=\sup_x\{\sum_{i=1}^n |D_i s(x)|^2\}$ *and*

$$D_i s(x)=\sup_z s(x_1,\dots,x_{i-1},z,x_{i+1},\dots,x_n) - \inf_z s(x_1,\dots,x_{i-1},z,x_{i+1},\dots,x_n).$$

Lemma 1 is unhelpful if s is not a uniformly bounded function of its arguments. Weaker inequalities, controlling only the size of the fluctuations around their mean rather than their distributions, are usually in terms of the expected squared Euclidean norm of the gradient of s and are called Poincaré inequalities. Another simple bound is

$$\operatorname{var}\{s(X_1,\dots,X_n)\} \le E\sum_{i=1}^n \operatorname{var}_i\{s(X_1,\dots,X_n)\}, \tag{3}$$

with equality if s is linear, where

$$\operatorname{var}_i\{s(x_1,\dots,x_n)\} = \operatorname{var}\{s(x_1,\dots,x_{i-1},X_i,x_{i+1},\dots,x_n)\},$$

is the variance of s with respect to X_i only, the other variables fixed.

Consider equation (2) for the means and contrasts parameterization of the $n=2^k$ factorial experiment, treating a single mean or contrast as the interest parameter ψ and the remaining parameters as components of λ. Neither the means nor the contrasts are affected by the geometry of the log likelihood function because the contours $\{(\psi,\lambda):\nabla_\psi\ell(\psi,\lambda)=0\}$ are flat in both parameterizations. However, the sufficient statistics for the contrasts, being an average of $n/2$ observations, concentrate in a $O(n^{-1/2})$ neighbourhood around their means as $n\to\infty$. By contrast, the sufficient statistic for the ith mean is simply X_i and has fluctuations of order $O(1)$ regardless of n.

A referee has remarked that the situation is detectable through an eigen decomposition of the observed information at the estimated parameter vector, indicating orthogonal directions in which the information is highest. If $\nabla_\theta\psi(\hat{\theta})$ has a high multiple correlation with the first few dominant eigenvectors this would suggest that the data are informative about ψ. There is likely to be a formulation in which the directions of principle curvature, important in differential geometry (Weatherburn, 1957, p. 120), play a similar role.

The geometric ideas summarized by Fig. 1 relate to a single realization of the log likelihood function. Under hypothetical replication, provided that n is large enough and $s(X_1,\dots,X_n)$ concentrates reasonably about its mean, the position and curvature of the contour $\{(\psi,\lambda):\nabla_\psi\ell(\psi,\lambda)=0\}$ will be relatively stable. A more formal algebraic treatment is based on an expression for finite sample bias of the maximum likelihood estimator under regime (i). By Taylor series expansion of a generic component ℓ_r of the score vector around the true value θ and evaluation at $\hat{\theta}$, Barndorff-Nielsen and Cox (1994, p. 150)

gave an expansion for the rth entry $\hat{\theta}_r$ of a p-dimensional vector of maximum likelihood estimates, where p was treated as fixed. This is of the form (Barndorff-Nielsen and Cox, 1994, equation 5.25)

$$\hat{\theta}_r - \theta_r = i^{rs}\ell_s \blacktriangledown + \frac{1}{2}\nu^{rst}\ell_s\ell_t + i^{rs}i^{tu}H_{st}\ell_u \blacktriangledown + \cdots,$$

where i^{rs} is the (r,s)th entry of the inverse expected information (here we suppress the earlier notation indicating which coordinate system is being used), $\nu^{rst} = i^{rq}i^{sv}i^{tz}\nu_{qvz}$, where $\nu_{qvz} = \partial^3_{qvz}\ell$ and $\blacktriangledown$ indicates a drop in asymptotic magnitude of $n^{-1/2}$. We use the convention that symbols appearing both as subscripts and superscripts in the same product are summed. Such results are valid only under ordinary repeated sampling with a fixed number of parameters.

We now show the largest conceivable magnitude of p as a function of n such that asymptotic unbiasedness is still assured without imposing and exploiting sparsity (discussed in Section 4.4). The bias is

$$E(\hat{\theta}_r - \theta_r) = \frac{1}{2}i^{rs}i^{tu}(\nu_{stu} + 2\nu_{st,u}) + \cdots, \tag{4}$$

where $\nu_{stu} = E(\partial^3_{stu}\ell)$, $\nu_{st,u} = E(\partial^2_{st}\ell\partial_u\ell)$ and evaluation of all quantities is at the true θ. Suppose, optimistically, that all elements of the inverse information matrix converge at rate n^{-1}. Then the remainder is of smaller order than the leading term. The leading term in (4) is a sum over approximately $p^3/3!$ combinations of things by Stirling's approximation, each of which converges at rate n^{-1}. A sufficient condition for the estimator to be asymptotically unbiased in this setting is then $p = o(n^{1/3})$.

For an interpretation in terms of known geometric quantities, use the Bartlett identities (Barndorff-Nielsen and Cox, 1994, equation 5.7) to write equation (4) as

$$\begin{aligned} E(\hat{\theta}_r - \theta_r) = &-\frac{1}{2}i^{rs}i^{tu}({}^{-1}\Gamma_{stu}) \\ &-\frac{1}{2}i^{rs}i^{tu}(\nu_{tu,s} + \nu_{su,t} - \nu_{st,u}) \\ &+ i^{rs}i^{tu}\nu_{st,u} + \cdots, \end{aligned} \tag{5}$$

where ${}^{-1}\Gamma_{stu}$ are the coefficients of α-connection with $\alpha = -1$. The second term on the right-hand side of (5) is $i^{rs}\delta^t_q\Upsilon^q_{st}$ where $\Upsilon^q_{st} = \frac{1}{2}i^{qu}(\nu_{tu,s} + \nu_{su,t} - \nu_{st,u})$ is somewhat similar in form to the Γ^q_{st} coefficients of Riemannian connection:

$$\Gamma^q_{st} = \frac{1}{2}i^{qu}\left(\frac{\partial i_{tu}}{\partial\theta_s} + \frac{\partial i_{su}}{\partial\theta_t} - \frac{\partial i_{st}}{\partial\theta_u}\right).$$

4.3 Special Structure in the Likelihood Function

As shown in Section 4.1 and elsewhere (e.g., Sur and Candès, 2019; Fan, Demirkaya and Lv, 2019), with n independent observations and p parameters there may be failure of standard likelihood-based procedures if p increases proportionally to n in the notional limit as n tends to infinity. Sensible estimation of interest parameters $\psi = (\psi_1, \ldots, \psi_{p-p_N})$ in the presence of nuisance parameters $\lambda = (\lambda_1, \ldots, \lambda_{p_N})$ hinges on a reformulation in which the nuisance parameters play a reduced role. This is most successful when λ is evaded completely without loss of information on ψ, which is possible when the joint density from which the data are drawn admits one of the following factorizations:

$$\begin{aligned} &\text{(a)} \quad f_Y(y;\psi,\lambda) = f_{V|U}(v|u;\lambda)f_U(u;\psi), \\ &\text{(b)} \quad f_Y(y;\psi,\lambda) = f_{V|U}(v|u;\psi)f_U(u;\lambda), \\ &\text{(c)} \quad f_Y(y;\psi,\lambda) = f_V(v;\lambda)f_U(u;\psi), \end{aligned}$$

where U and V are jointly sufficient statistics for ψ and λ. Factorization (a) is a case for marginalization with U sufficient for ψ, (b) is a case for conditioning on U, which is, here, the sufficient statistic for λ. In (c) the jointly sufficient statistic is two individually sufficient statistics so that conditioning reduces to marginalization. Weaker factorizations are:

$$\begin{aligned} &\text{(d)} \quad f_Y(y;\psi,\lambda) = f_{V|U}(v|u;\lambda,\psi)f_U(u;\psi), \\ &\text{(e)} \quad f_Y(y;\psi,\lambda) = f_{V|U}(v|u;\psi)f_U(u;\lambda,\psi). \end{aligned}$$

Marginalization is applicable in (d) and conditioning in (e). However, information on ψ may be lost in either case because neither U nor V are individually sufficient for ψ or λ.

Bartlett (1937) proposed conditional likelihood as a means of avoiding the bias incurred by standard maximum likelihood estimation, exemplified by estimation of the variance in the Gaussian pairs example (see Section 4.1). Modified profile likelihood (Barndorff-Nielsen, 1983, 1988) arises as a higher order approximation to a marginal or conditional likelihood function for the interest parameter ψ when one of these is available (see Barndorff-Nielsen, 1994). However, its specification relies on the ability to identify an ancillary complement to the maximum likelihood estimator unless the parameterization is orthogonal for the interest parameter, or can be made such by an interest-respecting transformation (see Cox and Reid, 1987). Barndorff-Nielsen (1984) discusses a constructive approach for identifying such ancillary complements in curved exponential models. In the specification of the modified directed likelihood root r^*, used for inference on an interest parameter, conditioning on an approximate ancillary statistic is somewhat implicit; see, for example, Brazalle et al. (2007, Chapter 8.4.3). See Tang and Reid (2020) for a discussion of the extent to which the properties of r^*, established under asymptotic regime (i), are valid under regime (ii).

There are, in principle, three types of systematic approaches for eliminating nuisance parameters, perhaps

approximately, for a scalar interest parameter. One is to solve the system of differential equations specified by Cox and Reid (1987). A second, suitable if one of factorization (a)–(c) holds, is to specify a constructive approach for finding ancillary statistics for the interest or nuisance parameters. A third strategy is to develop a systematic procedure for identifying statistics U, V or both such that one of factorizations (a)–(e) holds. See Battey and Cox (2020, Section 7.2) for a formulation and some suggested test cases. To some extent the third goal is achieved implicitly by the tangent exponential model of Fraser and Reid (1995). See Davison and Reid (2021) for a geometric account with full bibliographical details. The various approaches are likely to be related.

4.4 Sparsity

A concept of theoretical and subject-matter importance in very high dimensional inference problems is sparsity. This forms part of the statistical model and entails the existence of many zeros or near-zeros in an unknown interest or nuisance parameter, often after suitable reparameterization. McCullagh and Polson (2018) have formulated a rigorous definition.

There are three roles of sparsity appropriate in different settings and its implications are contingent on this. A first role is to encourage fruitful interpretation when there are a large number of parameters, initially on an equal footing and potentially of interest, and yet it is likely that few represent securely established effects of importance. The second role is to ensure that so-called plug-in estimators of interest parameters are sensible when they rely on preliminary estimates of p nuisance parameters when p/n is large. A third role is to ensure stable predictions when p/n is large. For prediction, all parameters used in the specification of the model can be viewed as nuisance parameters and the single interest parameter is the conditional expectation estimated as the prediction. Thus, the third role is treated here as a special case of the second.

4.4.1 *Sparsity for interpretation*: *Post-selection inference*. In some current areas, notably genomics, the broad form of probabilistic model may be agreed, but with appreciable uncertainty over details. This leads to an all-embracing model being specified, inevitably containing a large number p of regression parameters, many of which are, in fact, negligible. In other words, the comprehensive model embraces $\sum_{k=1}^{s^*} p!/\{(p-k)!k!\}$ submodels of size at most s^* by requiring most of the parameters to be zero.

In the sparse regression setting, if one fails to account for the first use of the data for selection of the model, then the resulting inferential statements are typically invalid. For instance, in a hypothesis testing context the false discovery rate, in hypothetical replication, of an α-level significance test is typically not α. The issue was exposed in a simple setting by Cox (1975b) and the problem has experienced renewed interest due to its relevance for genomics and other modern scientific areas.

An important point is that if the data are relatively complicated, there is no difficulty, in principle, in using them several or many times to answer different questions, as is usual in factorial experiments. Serious issues arise when one asks essentially the same physical question in different guises. Thus, when a variable selection procedure like the lasso (Tibshirani, 1996) is used to indicate a model, and inference is subsequently performed on the associated parameters, the inference is miscalibrated unless standard low-dimensional distribution theory is adjusted for the selection. See Berk et al. (2013), Lockhart et al. (2014), Lee et al. (2016), Tibshirani et al. (2016) for further detailed discussion and some suggested resolutions. Among these are to condition on the selection event, or in other words on the region of the sample space that would have led to the same model being selected. Such conditioning is typically difficult without strong assumptions. A simple but potentially wasteful solution is to split the sample, using part of the data to decide upon a model for further analysis and the rest for inference on the parameters of that model. This has been shown in some contexts to be dominated by an approach based on a randomized response. See Tian and Taylor (2018) and Rasines and Young (2021).

In effect, the approach of implicitly or explicitly conditioning on selection, prescribes singleton sets at zero as the confidence sets for the unselected variables. An alternative viewpoint, somewhat but not completely aligned with our comments in Section 4.4.2, specifies confidence sets for all parameters, including those that are not selected by the variable selection procedure, thereby acknowledging imperfections in the initial phase. Some references are Zhang and Zhang (2014), Javanmard and Montanari (2014), van de Geer et al. (2014) and Cai and Guo (2017). Our view on the appropriate formulation is summarized in Section 4.4.2.

4.4.2 *Sparsity for interpretation*: *Confidence sets of models*. In the context of the genomics example, formal inference on the values of parameters of an estimated model is arguably less relevant than inference on the model itself. To report the single model returned by a single-optimization procedure may be, for substantive interpretation, rather misleading if there are several or indeed many models that are essentially equally compatible with the data.

Assessment of model adequacy must always be a concern, and while this aspect is exacerbated by the vast number of possible models in sparse regression problems with large numbers of potential explanatory variables, classical ideas due to Fisher (1922) provide insights on the issues. For an appreciable but not excessive number of variables ($p < 30$, say, due to current computational limitations), the situation was discussed by Cox and Snell (1974, 1989)

and one resolution is Fisherian (Barndorff-Nielsen and Cox, 1994, p. 29). On letting z denote a realization of $Z = (Y_i, x_i)_{i=1}^n$ one specifies, for each sparse model indexed by m, the minimal sufficient statistic S_m for the parameter vector $\theta(m)$ of nonzero components. All models compatible with the data in the sense that z is not extreme when calibrated against the distribution of $Z \mid S_m = s_m$ should be reported as a confidence set of models alongside the associated confidence statements for $\theta(m)$. These ideas also underpinned the recommendations of Cox and Battey (2017) who advocate confidence sets of models in high-dimensional regression problems. When p is of the order of tens of thousands as in the genomics examples, it is practically infeasible to check all low-dimensional models for their compatibility with the data and some preliminary reduction is needed. Cox and Battey (2017) suggest a version of backwards reduction based on partially balanced incomplete block arrangements (Yates, 1936). This discards variables for which there is little or no evidence for having a real effect. See Battey and Cox (2018) for detailed discussion of the aspects raised by this two-stage procedure.

Our preference for confidence sets of models is not wholly incompatible with the references given in the last paragraph of Section 4.4.1, although the conceptual differences are substantial.

4.4.3 *Sparsity as a mathematical aspect.* As indicated in Section 4.3, nuisance parameters will ideally be eliminated by problem-specific manoeuvres without appreciable information loss. There is, however, no generally applicable strategy for (exactly or approximately) eliminating nuisance parameters in high dimensions. It is widespread practice to replace them by estimates. Depending on the objective, this sometimes necessitates a sparsity assumption in asymptotic regimes (ii) and (iii).

The following simple example is somewhat artificial in that it is designed to illustrate one particular point, ignoring other aspects such as those indicated in Section 2.1.

One treatment of linear regression in observational studies considers the data on outcome Y and covariates $X = (X_1, \ldots, X_p)$ as realizations of random couples (Y, X), so that both variables are initially treated as random. One then conditions on X, and if X and Y are jointly normally distributed we obtain

$$E(Y \mid X = x) = \mu_Y + \Sigma_{XY}^T \Sigma_{XX}^{-1}(x - \mu_X),$$

where μ_X and μ_Y are the means of X and Y, Σ_{XX} is the covariance matrix of the X variables and Σ_{XY} is the column of the full covariance matrix of X and Y corresponding to the Y variable. Thus, the linear regression coefficient β is equal by definition to

$$\frac{\partial E(Y \mid X = x)}{\partial x^T} = \Sigma_{XX}^{-1}\Sigma_{XY}.$$

If the relationship is not multivariate normal then one seeks the best linear approximation to $E(Y \mid X = x)$, where the approximation error is measured by the expected squared distance of $E(Y \mid X = x)$ from Y. This leads to the same expression $\beta = \Sigma_{XX}^{-1}\Sigma_{XY}$ regardless of assumptions on the joint distribution of Y and X (see Cox and Wermuth, 1996, pp. 63–64) but its relevance is small if the relationship between Y and X is highly nonlinear.

Let β_j, the jth element of the vector β, be an interest parameter. Let $\Omega = \Sigma_{XX}^{-1}$. A natural estimator of $\beta_j = (\Sigma_{XX}^{-1}\Sigma_{XY})_j$ replaces Ω and Σ_{XY} by estimates, $\hat{\Omega}$ and $\hat{\Sigma}_{XY}$ say, producing an estimator $\hat{\beta}_j$ of β_j. On writing

$$\begin{aligned}\hat{\beta}_j - \beta_j = {} & \|\Omega_{j\cdot}\|_2 \|\hat{\Sigma}_{XY} - \Sigma_{XY}\|_2 t_1 \\ & + \|(\hat{\Omega} - \Omega)_{j\cdot}\|_2 \|\Sigma_{XY}\|_2 t_2,\end{aligned} \tag{6}$$

where $\Omega_{j\cdot}$ denotes the jth row of Ω, $t_1 \in (-1, 1)$ is the cosine of the angle between $\Omega_{j\cdot}$ and $\hat{\Sigma}_{XY} - \Sigma_{XY}$ and $t_2 \in (-1, 1)$ is the cosine of the angle between $(\hat{\Omega} - \Omega)_{j\cdot}$ and Σ_{XY}, we see that a sparsity assumption is in general needed to prevent excessive accumulation of estimation error from matrix multiplication when p grows with n.

To illustrate the role of sparsity and the scaling conditions that arise in high dimensional problems, we will consider only the term $\|\hat{\Sigma}_{XY} - \Sigma_{XY}\|_2$ in equation (6). The permissible scaling of p with n in high dimensional problems with sparsity is typically $\log p = o(n)$. This derives from the use of nonasymptotic deviation bounds such as Hoeffding's inequality for weighted sums of sub-Gaussian random variables or Bernstein's inequality for weighted sums of subexponential random variables.

Bernstein's inequality for mean-zero subexponential random variables $\tilde{Z}_1, \ldots, \tilde{Z}_n$ is

$$\begin{aligned}&\operatorname{pr}\left(\left|\sum_{i=1}^n a_i \tilde{Z}_i\right| \geq t\right) \\ &\quad \leq 2\exp\{-c\min(t^2/K^2\|a\|_2^2, t/K\|a\|_\infty)\},\end{aligned} \tag{7}$$

where K is a constant related to the subexponential tail behaviour of $\tilde{Z}_i$. Let $(\hat{\Sigma}_{XY})_k = n^{-1}\sum_{i=1}^n X_{ik}Y_i = n^{-1}\sum_{i=1}^n Z_i$, say. Suppose initially that one estimates the kth element of Σ_{XY} by $(\hat{\Sigma}_{XY})_k$. Provided that the distributions of X_{ik} and Y_i have sub-Gaussian tails, the distribution of $Z_i - E(Z_i)$ has subexponential tails (see, e.g., Vershynin, 2018, Lemma 2.7.7). Thus by equation (7), we have, on setting $\tilde{Z}_i = Z_i - E(Z_i)$ and $a = (n^{-1}, \ldots, n^{-1})$,

$$\begin{aligned}&\operatorname{pr}\{|(\hat{\Sigma}_{XY} - \Sigma_{XY})_k| \geq t\} \\ &\quad \leq 2\exp\{-c\min(nt^2/K^2, nt/K)\},\end{aligned}$$

so that

$$\begin{aligned}&\operatorname{pr}\Big\{\max_k |(\hat{\Sigma}_{XY} - \Sigma_{XY})_k| \geq t\Big\} \\ &\quad \leq 2p\exp\{-c\min(nt^2/K^2, nt/K)\}.\end{aligned} \tag{8}$$

The requirement for the right-hand side of (8) to tend is zero as n tends to infinity is that $t \gtrsim \max[\{(\log p)/n\}^{1/2}, (\log p)/n]$ and convergence of (8) to zero is assured by the scaling $(\log p)/n \to 0$.

Under an assumption that only a small number s^* of the entries of Σ_{XY} are nonzero, one can define an estimator $\tilde{\Sigma}_{XY}$ that exploits the sparsity and retains the same elementwise convergence rate as $\hat{\Sigma}_{XY}$, so that

$$\begin{aligned}&\mathrm{pr}\{\|\tilde{\Sigma}_{XY} - \Sigma_{XY}\|_2 \geq t\}\\ &\quad\leq \mathrm{pr}\Big\{\max_k |(\hat{\Sigma}_{XY} - \Sigma_{XY})_k| \geq t/s^*\Big\}\\ &\quad\leq 2p\exp[-c\min\{n(t/s^*)^2/K^2, n(t/s^*)/K\}].\end{aligned}$$

Thus consistency is assured provided that $(s^* \log p)/n \to 0$. We emphasize that these are upper bounds. Suitable lower bounds would be needed to show that a better scaling is impossible.

4.4.4 *Sparsity-inducing reparameterizations*. Under asymptotic regime (i), Cox and Reid (1987) specified the set of differential equations whose solution yields an interest-respecting orthogonal reparameterization. This is a parameterization in which a single interest parameter ψ is orthogonal to a vector of nuisance parameters λ. Orthogonality is meant in the sense of setting the off-diagonal entries of the Fisher information matrix to zero in the row corresponding to entry ψ, identically in ψ and λ or in a weaker sense.

In regime (i), this form of sparsity is sufficient to deliver strong properties of the maximum likelihood estimator. In high dimensions, errors introduced into the interest parameter estimate by estimation of nuisance parameters, although substantially reduced by parameter orthogonalization, accumulate as $p \to \infty$ because maximum likelihood estimation does not exploit the sparsity.

The sparsity induced by parameter orthogonalization is an essential condition in some recent papers operating under regime (iii). See Ning and Liu (2017) and Fang, Ning and Liu (2017). Parameter orthogonalization is not explicitly mentioned as a prerequisite, although without this, plausibility of the conditions is small. The idea of parameter orthogonality has also resurfaced indirectly in work aimed at achieving so-called double robustness in high dimensions; see Chernozhukov et al. (2017).

Ning and Liu's (2017) adjustment to the score function to accommodate a high-dimensional nuisance parameter assumes that there is an estimator of the latter which is consistent in ℓ_1 norm at rate $s^*(\log p)^{1/2}n^{-1/2}$ where n is the sample size, p is dimension and s^* is the number of nonzero entries of the parameter vector. Regression settings are considered. Thus, in the notation of Section 4.4.3, let $\psi = \beta_v$ be the interest parameter, let $\lambda = \beta_{-v}$ be a $(p-1)$-dimensional nuisance parameter and let ℓ denote the log likelihood function, now divided by n. Ning and Liu's (2017) *decorrelated score function*, evaluated at the true values of the parameters is

$$\begin{aligned}S(\psi^*, \lambda^*) &= \nabla_v \ell(\beta_v^*, \beta_{-v}^*) - w^{*T}\nabla_{-v}\ell(\beta_v^*, \beta_{-v}^*)\\ &\text{where } w^{*T} = J_{v,-v}^* J_{-v,-v}^{*-1}.\end{aligned} \tag{9}$$

Asterisks denote true values and the J^* quantities are partitions of the information matrix $J^* = J(\beta^*)$:

$$\begin{aligned}J(\beta) &= \begin{pmatrix} J_{v,v} & J_{v,-v} \\ J_{-v,v} & J_{-v,-v}\end{pmatrix}\\ &= \begin{pmatrix} \mathbb{E}\nabla_{v,v}^2\ell(\beta) & \mathbb{E}\nabla_{v,-v}^2\ell(\beta) \\ \mathbb{E}\nabla_{-v,v}^2\ell(\beta) & \mathbb{E}\nabla_{-v,-v}^2\ell(\beta)\end{pmatrix}.\end{aligned} \tag{10}$$

For an estimator $\hat{w}$ of w^*, not discussed here, a mean value expansion of

$$\hat{S}(\beta_v^*, \beta_{-v}) \stackrel{\text{def}}{=} \nabla_v \ell(\beta_v^*, \beta_{-v}) - \hat{w}^T \nabla_{-v}\ell(\beta_v^*, \beta_{-v}) \tag{11}$$

around the true value, β_{-v}^* of β_{-v} followed by evaluation at its estimated value gives, in slightly misleading notation explained immediately below,

$$\begin{aligned}&\hat{S}(\beta_v^*, \hat{\beta}_{-v})\\ &\quad= \nabla_v \ell(\beta_v^*, \beta_{-v}^*) - \hat{w}^T \nabla_{-v}\ell(\beta_v^*, \beta_{-v}^*)\\ &\qquad+ [\nabla_{v,-v}^2\ell(\beta_v^*, \beta_{-v,\alpha})\\ &\qquad- \hat{w}^T\nabla_{-v,-v}^2\ell(\beta_v^*, \beta_{-v,\alpha})](\hat{\beta}_{-v} - \beta_{-v}^*).\end{aligned} \tag{12}$$

The notation is inaccurate because $\beta_{-v,\alpha} = \alpha\hat{\beta}_{-v} + (1-\alpha)\beta_{-v}^*$ for $\alpha \in (0,1)$ has a different α for each component of the $1 \times (p-1)$ dimensional vector $\nabla_{v,-v}^2\ell$. The key to understanding how the decorrelated score remedies the problems faced by the classical score function in high dimensions is the observation that

$$\begin{aligned}&[\nabla_{v,-v}^2\ell(\beta_v^*, \beta_{-v,\alpha}) - \hat{w}^T\nabla_{-v,-v}^2\ell(\beta_v^*, \beta_{-v,\alpha})]\\ &\quad\approx \mathbb{E}[\nabla_{v,-v}^2\ell(\beta_v^*, \beta_{-v}^*) - w^{*T}\nabla_{-v,-v}^2\ell(\beta_v^*, \beta_{-v}^*)]\\ &\quad= J_{v,-v}^* - J_{v,-v}^* J_{-v,-v}^{*-1} J_{-v,-v}^* = 0,\end{aligned} \tag{13}$$

where $w^{*T} = J_{v,-v}^* J_{-v,-v}^{*-1}$. Hence, provided that w^* is sufficiently sparse to avoid excessive noise accumulation, rate acceleration is possible in equation (12) because two quantities, both tending to zero with the sample size, are multiplied together, ultimately giving rise to a tractable limit distribution of a suitable rescaling of $\hat{S}(\beta_v^*, \hat{\beta}_v)$. To apply these ideas, β_v^* is replaced by a hypothesized value.

In the linear regression setting, the assumption of sparsity of w^* is essentially that most of the nuisance covariates are uncorrelated with signal variables.

4.5 Supersaturated Designs

Our discussion in this paper applies primarily to data coming from observational studies. The traditional literature on the design of experiments has extensive treatment

of situations in which the number of unknown parameters, p, is less than but quite close to the number, n, of independent individuals. For the so-called supersaturated case, $p > n$, Satterthwaite (1959) suggested randomization of factor levels so that in studying the effect of one factor, the others could be treated as random, an argument that led to much criticism. Booth and Cox (1962) suggested systematic supersaturated designs with suitable properties; there is no record of application. Exact orthogonality of all columns of the design matrix simultaneously is infeasible when $p > n$. The implication is that certain treatment effects are aliased meaning, for them, that statistical analysis is unable to attribute certain significant effects to one variable over another; thus both should be reported as possible explanations. More recent work has built on heavily fractionated factorial designs.

5. DISCUSSION AND OPEN PROBLEMS

5.1 Valid Inference Under Model Misspecification

We have only discussed situations in which the probabilistic model is regarded as correctly specified. To be meaningful, a specification should be broadly correct for an interest parameter but could be vaguely specified, or perhaps even misspecified for the nuisance part of the model. An example is inference on the interest parameters of the proportional hazards model (Cox, 1972, 1975a). The nuisance parameter, the hazard function, requires specification as a fixed function of time although its detailed functional form is left arbitrary and is conveniently evaded by partial likelihood.

An example is given by Battey and Cox (2020), to some extent a reprise of an earlier result due to Lindsay (1985), in which an estimator of an interest parameter is consistent in spite of an arbitrary degree of misspecification in the nuisance part of the model. Efficiency loss over marginal likelihood based on factorization (d) was nevertheless shown to be potentially severe. The latter issues would affect conclusions of hypothesis tests that do not implicitly account for the misspecification.

The extent to which these results generalize is unclear. Analysis of misspecification is complicated by interest and nuisance parameters being nonorthogonal under arbitrary misspecification. Orthogonality of misspecified nuisance parameters to interest parameters may indeed be a requirement for consistency of the interest parameter estimator. If so, an extension of Cox and Reid's (1987) analysis to the misspecified situation is likely to be valuable.

5.2 Aspects Not Represented by a Model

Our discussion presupposes that the data are attributable to a probabilistic model. Some features of data are not always fruitfully represented in this way. It is, for instance, common in observational data for some observations on explanatory variables to be missing for reasons that are unclear. Battey, Cox and Jackson (2019) recommended a sensitivity analysis to the missingness by replacing missing entries of each affected variable by one of two relatively extreme assignments in a full 2^m factorial structure, where m is the number of variables with at least one missing entry. The analysis is performed for these 2^m combinations and the results reported for each. If answers are stable, strong assurance is provided over the conclusions but otherwise a range should be reported. This approach is in contrast to multiple imputation (e.g., Little and Rubin, 2019) in which the missingness mechanism is modelled. This entails the untestable assumption that observations are missing at random, not to be confused with missing completely at random, an even stronger assumption.

5.3 Open Problems

We close with a list of open problems motivated by this discussion.

1. Given a new statistical inference problem specified by its likelihood in some special parameterization, how does one check for anomalies that suggest some or all of the parameters are better redefined?
2. How does one deduce, from observed quantities alone, in which directions maximum likelihood will be formally sensible and in which directions inference will be seriously misleading? Can the geometric ideas of Section 4.2 be operationalized?
3. Suppose that Bartlett's (1937) matched pairs problem described in Section 4.1 is now specified in terms of a $(n+1)$-dimensional parameter, an arbitrary transformation of the original parameters, for example, the canonical parameterization of the exponential family. All parameters are wrongly estimated by maximum likelihood. Can the situation be detected in an arbitrary parameterization by analysis only of the observed likelihood function? Can it be deduced that the variance is responsible for the bad behaviour?
4. Can anything be learnt from situations with multiple anomalous parameters, as possibly exemplified by regression at various levels in a standard split-split plot experiment?

The first of the above questions relates to reparameterization while the remaining ones relate to problems that reparameterization does not solve.

5. Nuisance parameters can sometimes be evaded by problem specific manoeuvres. Is there a general theory, able to guide us towards taking ratios in one formulation of the exponential matched pairs problem and conditioning on the pairwise sums in another formulation? See Battey and Cox (2020) for a description of these two formulations and the considerations involved in a general resolution.

6. There are two types of problem associated with the previous point. One is to recognize a standard type of situation when it is presented in disguised form. The other, possibly more important, is to study problems or classes of problems where standard tricks are approximately, that is, only asymptotically, applicable.

7. Is there a geometric representation of conditioning to evade nuisance parameters, and if so, how is this different geometrically from conditioning to ensure relevance?

8. If none of factorizations (a)–(e) holds, what would be an appropriate notion of approximate factorizability?

9. How could one seek a factorization of the form (a)–(e) when the likelihood function is unavailable analytically? See Patel et al. (2019) for an example of such a situation motivated by a problem in biophysics, and Shlomovich et al. (2020) for a more general class of examples arising in the analysis of doubly stochastic point processes.

10. When there are nuisance parameters two approaches are to transform the data and to marginalize or condition based on factorizations (a)–(e) above, or to find an interest-respecting orthogonal transformation as in Cox and Reid (1987). Is there a connection between the two and if so, can it be characterized geometrically?

11. Consider an arbitrary problem in which there are n observations and p parameters, of which one is an interest parameter and the rest are nuisance parameters. Principles of inference suggest reducing the dimension n of the data to p (sufficiency) and reducing the dimension of the sufficient statistic, s, to that of the interest parameter by constraining s to the subspace of its sample space where the maximum likelihood estimate of the nuisance parameter is fixed, a version of ancillarity. Framed in this way, the situation in which $p > n$ is precluded. But if a transformation could be applied to make the problem depend on the interest parameter and a small set of one-dimensional summaries of the original nuisance parameters, then the $p > n$ problem would have been reduced to the usual $p < n$ situation for each interest parameter. Sparsity might still be needed for the generating process to make sense. Is it possible to construct a theory of inference that is Fisherian when $p < n$ and can be made Fisherian by suitable transformation when $p > n$ under a sparsity constraint?

The previous discussion has throughout assumed data are available on study individuals which share a common structure but in which the stochastic components for different individuals are mutually independent. Various complications arise if, as may be particularly likely for observational studies, some aspects of the variation are interrelated. In the simplest situations this may leave estimates based on independence assumptions being reasonable but with changed, typically inflated, variance. Connected to this point are the following important questions, relevant to both low and high dimensional problems.

12. Suppose there are blocks of individuals with different types of relation between and within blocks. If the structure is not initially clear, how would such anomalies be detected? The following situations exemplify these issues.

- Data are collected on a suitable sample of adults arranged in family groups for convenience of data collection. The family grouping may be of no direct interest but intra family dependences are, at the least, likely to inflate errors of estimation as compared with random sampling.
- There may be similar situations in which the analogue of family structure is present but not observed. That is, there are a large number, possibly individually small, local structures to inflate errors of estimation. In particular, in large sets of data there may be multiple sources of local correlation, individually possibly small but having appreciable cumulative effect.
- A further class of problems arise when the dependence is totally unstructured, for instance it may arise in residuals due to estimation of a coefficient vector, perhaps under a misspecified model.

13. Are there ways of estimating the standard errors of estimators reasonably in the presence of dependent observations without explicit estimation of the dependency structure, perhaps by a nuisance parameter formulation?

A referee has pointed out that question 12 is connected to community detection in the statistical analysis of networks literature.

We have largely assumed that a suitable probabilistic model is given, although discussion of this aspect appears in Sections 4.4.1, 4.4.2 and 5.1.

14. To what extent is inference on interest parameters valid when the nuisance part of the model is misspecified? How does this depend on the model structure and inferential procedures involved?

15. Are there classes of models and procedures that are robust to misspecification of nuisance aspects? In what sense is inference robust? For instance, consistency may be achievable but efficiency lost.

ACKNOWLEDGEMENTS

We are grateful to five anonymous referees and the Associate Editor for references and detailed constructive criticism.

FUNDING

The work was supported by a UK Engineering and Physical Sciences Research Fellowship (to HSB).

REFERENCES

ANASTASIOU, A. and REINERT, G. (2020). Bounds for the asymptotic distribution of the likelihood ratio. *Ann. Appl. Probab.* **30** 608–643. MR4108117 https://doi.org/10.1214/19-AAP1510

BARNDORFF-NIELSEN, O. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika* **70** 343–365. MR0712023 https://doi.org/10.1093/biomet/70.2.343

BARNDORFF-NIELSEN, O. E. (1984). On a formula for the distribution of the maximum likelihood estimator. *Scand. J. Stat.* **11** 157–170.

BARNDORFF-NIELSEN, O. E. (1988). *Parametric Statistical Models and Likelihood. Lecture Notes in Statistics* **50**. Springer, New York. MR0971982 https://doi.org/10.1007/978-1-4612-3934-5

BARNDORFF-NIELSEN, O. E. (1994). Adjusted versions of profile likelihood and directed likelihood, and extended likelihood. *J. Roy. Statist. Soc. Ser. B* **56** 125–140. MR1257801

BARNDORFF-NIELSEN, O. E. and COX, D. R. (1994). *Inference and Asymptotics. Monographs on Statistics and Applied Probability* **52**. CRC Press, London. MR1317097 https://doi.org/10.1007/978-1-4899-3210-5

BARTLETT, M. S. (1937). Properties of sufficiency and statistical tests. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **160** 268–282.

BATTEY, H. S. and COX, D. R. (2018). Large numbers of explanatory variables: A probabilistic assessment. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **474** 20170631.

BATTEY, H. S. and COX, D. R. (2020). High dimensional nuisance parameters: An example from parametric survival analysis. *Inf. Geom.* **3** 119–148. MR4182888 https://doi.org/10.1007/s41884-020-00030-6

BATTEY, H. S., COX, D. R. and JACKSON, M. V. (2019). On the linear in probability model for binary data. *R. Soc. Open Sci.* **6** 190067, 7. MR3980023 https://doi.org/10.1098/rsos.190067

BERK, R., BROWN, L., BUJA, A., ZHANG, K. and ZHAO, L. (2013). Valid post-selection inference. *Ann. Statist.* **41** 802–837. MR3099122 https://doi.org/10.1214/12-AOS1077

BLISS, C. I. (1935). The calculation of the dosage-mortality curve. *Ann. Appl. Biol.* **22** 135–167.

BOOTH, K. H. V. and COX, D. R. (1962). Some systematic supersaturated designs. *Technometrics* **4** 489–495. MR0184369 https://doi.org/10.2307/1266285

BRAZALLE, A., DAVISON, A. C. and REID, N. (2000). *Applied Asymptotics*. Cambridge Univ. Press, London.

BUJA, A., BROWN, L., BERK, R., GEORGE, E., PITKIN, E., TRASKIN, M., ZHANG, K. and ZHAO, L. (2019). Models as approximations I: Consequences illustrated with linear regression. *Statist. Sci.* **34** 523–544. MR4048582 https://doi.org/10.1214/18-STS693

CAI, T. T. and GUO, Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *Ann. Statist.* **45** 615–646. MR3650395 https://doi.org/10.1214/16-AOS1461

CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWEY, W. and ROBINS, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econom. J.* **21** C1–C68. MR3769544 https://doi.org/10.1111/ectj.12097

COX, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* **34** 187–220. MR0341758

COX, D. R. (1975a). Partial likelihood. *Biometrika* **62** 269–276. MR0400509 https://doi.org/10.1093/biomet/62.2.269

COX, D. R. (1975b). A note on data-splitting for the evaluation of significance levels. *Biometrika* **62** 441–444. MR0378189 https://doi.org/10.1093/biomet/62.2.441

COX, D. R. (2020). Statistical significance. *Annu. Rev. Stat. Appl.* **7** 1–10. MR4104183 https://doi.org/10.1146/annurev-statistics-031219-041051

COX, D. R. and BATTEY, H. S. (2017). Large numbers of explanatory variables, a semi-descriptive analysis. *Proc. Natl. Acad. Sci. USA* **114** 8592–8595. https://doi.org/10.1073/pnas.1703764114

COX, D. R. and REID, N. (1987). Parameter orthogonality and approximate conditional inference. *J. Roy. Statist. Soc. Ser. B* **49** 1–39. With a discussion. MR0893334

COX, D. R. and REID, N. (2000). *The Theory of the Design of Experiments. Monographs on Statistics and Applied Probability* **67**. CRC Press, London. MR1456990

COX, D. R. and SNELL, E. J. (1974). The choice of variables in observational studies. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **23** 51–59.

COX, D. R. and SNELL, E. J. (1989). *The Analysis of Binary Data. Monographs on Statistics and Applied Probability* **32**. CRC Press, London.

COX, D. R. and WERMUTH, N. (1996). *Multivariate Dependencies: Models, Analysis and Interpretation. Monographs on Statistics and Applied Probability* **67**. CRC Press, London. MR1456990

DAVISON, A. C., KOCH, E. and KOH, J. (2019). Comment: Models are approximations! [MR4048582; MR4048583]. *Statist. Sci.* **34** 584–590. MR4048589 https://doi.org/10.1214/19-STS746

DAVISON, A. C., REID, N. (2021). The tangent exponential model. arXiv:2106.10496.

FAN, Y., DEMIRKAYA, E. and LV, J. (2019). Nonuniformity of p-values can occur early in diverging dimensions. *J. Mach. Learn. Res.* **20** Paper No. 77, 33. MR3960931

FANG, E. X., NING, Y. and LIU, H. (2017). Testing and confidence intervals for high dimensional proportional hazards models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1415–1437. MR3731669 https://doi.org/10.1111/rssb.12224

FISHER, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc. Lond. A* **222** 309–368.

FRASER, D. A. S. and REID, N. (1995). Ancillaries and third order significance. *Util. Math.* **47** 33–53. MR1330888

JAVANMARD, A. and MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15** 2869–2909. MR3277152

LEE, J. D., SUN, D. L., SUN, Y. and TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44** 907–927. MR3485948 https://doi.org/10.1214/15-AOS1371

LEI, L., BICKEL, P. J. and EL KAROUI, N. (2018). Asymptotics for high dimensional regression M-estimates: Fixed design results. *Probab. Theory Related Fields* **172** 983–1079. MR3877551 https://doi.org/10.1007/s00440-017-0824-7

LINDSAY, B. G. (1985). Using empirical partially Bayes inference for increased efficiency. *Ann. Statist.* **13** 914–931. MR0803748 https://doi.org/10.1214/aos/1176349646

LITTLE, R. J. A. and RUBIN, D. B. (1987). *Statistical Analysis with Missing Data. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*. Wiley, New York. MR0890519

LOCKHART, R., TAYLOR, J., TIBSHIRANI, R. J. and TIBSHIRANI, R. (2014). A significance test for the lasso. *Ann. Statist.* **42** 413–468. MR3210970 https://doi.org/10.1214/13-AOS1175

MCCULLAGH, P. and POLSON, N. G. (2018). Statistical sparsity. *Biometrika* **105** 797–814. MR3877866 https://doi.org/10.1093/biomet/asy051

NING, Y. and LIU, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Ann. Statist.* **45** 158–195. MR3611489 https://doi.org/10.1214/16-AOS1448

PATEL, L., GUSTAFSSON, N., LIN, Y., OBER, R., HENRIQUES, R. and COHEN, E. (2019). A hidden Markov model approach to characterizing the photo-switching behavior of fluorophores. *Ann. Appl. Stat.* **13** 1397–1429. MR4019144 https://doi.org/10.1214/19-AOAS1240

RASINES, D. G. and YOUNG, G. A. (2021). Splitting strategies for post-selection inference. arXiv:2102.02159.

SATTERTHWAITE, F. E. (1959). Random balance experimentation. *Technometrics* **1** 111–137. MR0130765 https://doi.org/10.2307/1266466

SHLOMOVICH, L., COHEN, E. A. K., ADAMS, N. and PATEL, L. (2020). A Monte Carlo EM algorithm for the parameter estimation of aggregated Hawkes processes. arXiv:2001.07160.

STEIN, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, *Vol. I* 197–206. Univ. California Press, Berkeley and Los Angeles. MR0084922

SUR, P. and CANDÈS, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proc. Natl. Acad. Sci. USA* **116** 14516–14525. MR3984492 https://doi.org/10.1073/pnas.1810420116

TANG, Y. and REID, N. (2020). Modified likelihood root in high dimensions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 1349–1369. MR4176346

TIAN, X. and TAYLOR, J. E. (2018). Selective inference with a randomized response. *Ann. Statist.* **46** 679–710. https://doi.org/10.1214/17-AOS1564

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TIBSHIRANI, R. J., TAYLOR, J., LOCKHART, R. and TIBSHIRANI, R. (2016). Exact post-selection inference for sequential regression procedures. *J. Amer. Statist. Assoc.* **111** 600–620. MR3538689 https://doi.org/10.1080/01621459.2015.1108848

VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285 https://doi.org/10.1214/14-AOS1221

VERSHYNIN, R. (2018). *High-Dimensional Probability*: *An Introduction with Applications in Data Science*. *Cambridge Series in Statistical and Probabilistic Mathematics* **47**. Cambridge Univ. Press, Cambridge. With a foreword by Sara van de Geer. MR3837109 https://doi.org/10.1017/9781108231596

WEATHERBURN, C. E. (1957). *An Introduction to Riemannian Geometry and the Tensor Calculus*. Cambridge Univ. Press, Cambridge, UK.

YATES, F. (1936). A new method of arranging variety trials involving a large number of varieties. *J. Agric. Sci.* **26** 424–455.

ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940 https://doi.org/10.1111/rssb.12026