

# SOME PERSPECTIVES ON INFERENCE AND ASYMPTOTIC ANALYSIS IN HIGH DIMENSIONS

By

H. S. Battey

*Department of Mathematics, Imperial College London, SW7 2AZ, UK*

and

D. R. Cox

*Nuffield College, University of Oxford OX1 1NF, UK*

## SUMMARY

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 and the use of asymptotic analysis in statistical inference, we provide some new perspectives on aspects of high-dimensional statistical theory, emphasizing particularly a number of important open problems.

*Some key words:* asymptotic analysis; 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 the security of those conclusions. 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

40 according to which of  $(n, q, p)$  are fruitfully treated as large. Theoretical questions can  
 41 then be studied from at least two perspectives. For what values of  $(n, q, p)$  do specific  
 42 procedures of analysis being studied give numerical procedures of acceptable statistical  
 43 behaviour? When is further refinement needed?

## 44 2 Broad conceptual issues

### 45 2.1 Two roles of conditioning

46 Suppose that we regard as provisionally given a parametric probability model for the  
 47 data containing a parameter vector  $(\psi, \lambda)$ , where  $\psi$  is of specific subject-matter interest  
 48 and  $\lambda$  is a vector of nuisance parameters. Consideration of conditional distributions can  
 49 arise in two quite different ways.

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

55 The second quite distinct possibility is that even though there are no nuisance param-  
 56 eters, or these have in effect been eliminated from the discussion, the minimal sufficient  
 57 statistic is of higher dimension than the parameter space, the difference in dimensionality  
 58 being  $d_A$ , the dimension of a statistic  $A$ . Then  $A$ , part of the minimal sufficient statistic,  
 59 is called *ancillary* if its distribution does not depend on the parameter,  $\psi$ . The qualita-  
 60 tive interpretation is that the observed value  $a$  has bearing on the precision achievable in  
 61 the particular set of data under analysis, as contrasted with that in some larger less well  
 62 specified context.

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

### 74 2.2 Significance testing as a primary mode of inference

75 This paper concentrates on the asymptotic considerations that may well arise when a  
 76 relatively complicated model is needed, but first we give a single example illustrating dif-  
 77 ficulties in a notionally very simple situation which therefore may, possibly unexpectedly,  
 78 endanger larger problems.

79 In some forms of bioassay (Bliss, 1935), estimation of a key property, ED50, of a dose-  
 80 response function is essentially equivalent to the study of the ratio of two normal means.  
 81 Here, in an idealized form of the original problem, observed random variables  $(Y_1, Y_2)$  are  
 82 independently normally distributed with means  $(\mu_1, \mu_2)$  and unit variance, and interest  
 83 lies in  $\theta = \mu_2/\mu_1$ . Depending on the data, reasonable conclusions about  $\theta$  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. See the appendix of Bliss (1935) for a seamless application of Fisherian ideas to this problem, and the supplementary material for historical context.

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 totally 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 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 Asymptotic analysis

#### 3.1 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  $p(n)/n \rightarrow 0$ ;
- (ii)  $p(n)/n \rightarrow \kappa \in (0, 1]$ ;
- (iii)  $p(n)/n$  diverges or tends to a limit greater than one.

These categorizations belong to formal theory. For any notional limiting operation,  $\alpha \rightarrow \alpha^*$  say, applicability of procedures thus justified often extends to values of  $\alpha$  far from  $\alpha^*$ . 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  $x \rightarrow \infty$ . A natural question emerges: what are the fundamental limits of inference, in terms of the maximum permissible value of  $\kappa$  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 *et al.* (2018), Sur and Candès (2019), Fan *et al.* (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 (see section 3.2). 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, formally justified under regime (i), continue to hold under regime (ii).

### 3.2 Techniques of asymptotic analysis for statistical inference

The techniques of asymptotic discussion appearing in the literature on statistical inference differ considerably. To compare and contrast these approaches, it suffices to consider the following highly simplistic example of regime (i). With suitable reformulation, some of the ideas are applicable under regimes (ii) and (iii) as indicated in the references provided.

Let  $X_1, \dots, X_n$  be independent and identically distributed random variables with mean zero and standard deviation  $\nu$ . On defining  $S_n = \sum_{i=1}^n X_i$ , and  $Z_n = (\sqrt{n}\nu)^{-1}S_n$ , the central limit theorem holds for  $Z_n$ . Specifically

$$\sup_{z \in \mathbb{R}} |\text{pr}(Z_n \leq z) - \Phi(z)| \rightarrow 0, \quad (n \rightarrow \infty),$$

where  $\Phi(z)$  is the standard normal distribution function at  $z$ . Two refinements to this limiting result entail: retaining  $\Phi(z)$  but quantifying the error in the approximation; or replacing  $\Phi(z)$  by an approximation that is more accurate in small samples. The Berry-Esseen theorem (Berry, 1941; Esseen, 1942) provides the bound

$$\sup_{z \in \mathbb{R}} |\text{pr}(Z_n \leq z) - \Phi(z)| \leq \frac{C\mathbb{E}(|X_1|^3)}{\nu^3\sqrt{n}}, \quad (1)$$

where  $0.409 \leq C \leq 0.479$  (Esseen, 1942; Tyurin, 2010), showing that the rate of convergence in the central limit theorem is rather slow. Nevertheless, the simplicity of the normal approximation is an advantage. While independence of  $X_1, \dots, X_n$  is critical, Berry-Esseen type results for the Lindeberg version of the central limit theorem are available with the same bounds on  $C$  as above, making the result applicable in a variety of statistical contexts. For a sophisticated example under regime (iii) see Zhou *et al.* (2018, Theorem 2.2.) and for a Berry-Esseen type result establishing the rate of convergence

to a non-normally distributed random variable under regime (iii), see Fan *et al.* (2018, Theorem 3.1).

Improvements to the approximation (1) are obtained by writing the density function of  $Z_n$  in terms of its Laplace transform and approximating the resulting integral by the complex variable method of steepest descents or its localized version, the saddlepoint approximation (Cauchy, 1829; Daniels, 1954). A detailed exposition of these complex variable arguments is given in the supplementary material. Much fruitful work on statistical inference under regime (i) has emanated from Daniels (1954); see, for instance, Barndorff-Nielsen and Cox (1979; 1989, Chapter 4), Davison *et al.* (2014), Fraser *et al.* (2016), and for a large number of examples with detailed implementation, see Brazzale *et al.* (2007).

Since the standard deviation,  $\nu$ , is typically unknown, introduce an estimator,  $V_n$ , of  $\nu$ , independent of  $S_n$ . Let  $D_n := V_n/\nu$  with density function  $f_{D_n}(\delta)$  at  $\delta$ , ideally approximable by a simpler density function  $f_D(\delta)$ . Then with  $T_n := (\sqrt{n}V_n)^{-1}S_n$ ,

$$\begin{aligned} \text{pr}(T_n \leq t) &= \int_0^\infty \Phi(t\delta) f_D(\delta) d\delta + \int_0^\infty \Phi(t\delta) (f_{D_n} - f_D)(\delta) d\delta \\ &+ \int_0^\infty \{\text{pr}(Z_n \leq t\delta) - \Phi(t\delta)\} f_{D_n}(\delta) d\delta. \end{aligned} \quad (2)$$

Provided that  $f_D(\delta)$  is sufficiently simple, there is hope that the first of the above integrals is approximable by Laplace's method or by treating  $\delta$  as a complex variable and applying the method of steepest descents (supplementary material). Suppose such an approximation,  $A(t)$  say, has been found, with remainder  $R_n(t)$ , that is

$$\int_0^\infty \Phi(t\delta) f_D(\delta) d\delta = A(t) + R_n(t). \quad (3)$$

The function  $A(t)$  approximates the distribution function of a leptokurtic random variable (Beale and Mallows, 1959). By the triangle inequality and Hölder's inequality,

$$|\text{pr}(T_n \leq t) - A(t)| \leq \sup_{z \in \mathbb{R}} |\text{pr}(Z_n \leq z) - \Phi(z)| + |R_n(t)| + \frac{1}{2} \|f_{D_n} - f_D\|_{\mathbb{L}_1}. \quad (4)$$

The first term on the right hand side is controlled by (1).

Laplace's method approximates  $I_L(\lambda; \varepsilon)$  for large  $\lambda$  in the decomposition of a general integral of the form

$$\int_a^b g(x) e^{\lambda u(x)} dx = \left( \int_a^{c-\varepsilon} + \int_{c-\varepsilon}^{c+\varepsilon} + \int_{c+\varepsilon}^b \right) g(x) e^{\lambda u(x)} dx,$$

where  $\varepsilon$  is small and  $c$  is the point at which the function  $u(x)$  achieves its maximum. In the present example, which falls under regime (i), a change of variables would be chosen such that  $\lambda$  is an increasing function of  $n \rightarrow \infty$ . For an application of Laplace approximation in a high-dimensional context, see Battey and Cox (2018). The supplementary material details approximations to the dominant middle term.

To illustrate, in simplified form, the non-asymptotic probability inequalities commonly encountered in the modern statistical literature, we present some simple ways of controlling the remainders in a Laplace approximation. By Hölder's inequality again,

$$\int_b^\infty \Phi(tw) f_W(w) dw \leq \sup_{x \in \mathbb{R}} \Phi(x) \text{pr}(W > b).$$

197 For all increasing functions  $h$ , Markov's inequality gives

$$\text{pr}(W > b) = \text{pr}\{h(W) > h(b)\} \leq \frac{\mathbb{E}h(W)}{h(b)}, \quad (5)$$

198 and careful choices of  $h$  deliver useful bounds. For instance, if the moment generating  
199 function of  $W$  satisfies  $\mathbb{E}(e^{\eta W}) \leq \exp(\eta^2 \sigma^2 / 2)$  for some  $\sigma$ ,  $W$  is said to be sub-Gaussian  
200 of parameter  $\sigma$  and taking  $h(b) = e^{\eta b}$  yields

$$\text{pr}(W > b) \leq e^{-\eta b} \mathbb{E}(e^{\eta W}) \leq \exp\{-\eta b + \eta^2 \sigma^2 / 2\}.$$

201 The choice  $\eta = b/\sigma^2$  minimizes the term in the exponential, giving  $\text{pr}(W > b) \leq$   
202  $\exp(-b^2/2\sigma^2)$ . Again, to apply these ideas, typically a change of variables is needed  
203 to ensure that either  $\sigma \rightarrow 0$  or  $b \rightarrow \infty$  or both, as  $n \rightarrow \infty$ . See section 4.1 for example  
204 usage in a slightly different context, Boucheron *et al.* (2013) for advanced coverage of non-  
205 asymptotic probability inequalities and Wainwright (2019) for numerous high-dimensional  
206 statistical procedures studied using non-asymptotic probability inequalities.

207 Suppose that  $D^2$  is exactly distributed as a  $\chi_n^2$  random variable, rescaled by  $n$  to  
208 have unit mean. The left hand side of (3) can then be exactly solved in terms of special  
209 functions showing that  $T_n$  is approximately distributed as a Student  $t$  random variable  
210 with  $n$  degrees of freedom, whose density function is

$$f_T(t) = \frac{\Gamma\{(n+1)/2\}}{\Gamma(n/2)\sqrt{\pi n}} (1 + t^2/n)^{-(n+1)/2}.$$

211 By Stirling's formula,

$$\frac{\Gamma(\frac{n}{2} + \frac{1}{2})}{\Gamma(\frac{n}{2})} \simeq \left(\frac{n}{2}\right)^{1/2} \quad (n \rightarrow \infty).$$

212 so that the constant term in  $f_T(t)$  is approximately  $(2\pi)^{1/2}$  for large  $n$ . On taking the  
213 logarithm of  $(1 + t^2/n)^{-(n+1)/2}$  and expanding for small  $t^2/n$ , Fisher (1925) noted that

$$-\frac{(n+1)}{2} \log\left(1 + \frac{t^2}{n}\right) = -\frac{t^2}{2} + n^{-1} \left(\frac{t^4}{4} - \frac{t^2}{2}\right) + n^{-1} \left(\frac{t^4}{4} - \frac{t^6}{6}\right) + O(n^{-3}).$$

214 Thus

$$f_T(t) = \phi(t) \left\{ 1 + n^{-1} \left(\frac{t^4}{4} - \frac{t^2}{2}\right) + n^{-1} \left(\frac{t^4}{4} - \frac{t^6}{6}\right) \right\} + O(n^{-3}),$$

215 where  $\phi(t)$  is the standard normal density function at  $t$ , suggesting a version of (4) in  
216 which  $A(t) = \Phi(t)$ . It is easily checked that this is also the Laplace approximation to the  
217 integral in (3).

218 Normal approximations of this nature for studentized statistics can be established  
219 more generally whenever  $T_n$  is expressible as a self-normalized sum  $S_n/Q_n$  where  $Q_n^2 =$   
220  $\sum_{i=1}^n X_i^2$ . Specifically, if  $\nu$  is estimated by the sample standard deviation of  $X_1, \dots, X_n$ ,  
221  $T_n$  can be written as (Efron, 1969; Logan *et al.*, 1973)

$$T_n = \frac{S_n/Q_n}{[n - (S_n/Q_n)^2/(n-1)]^{1/2}},$$

222 from which we obtain

$$\text{pr}(T_n \geq t) = \text{pr}\left\{ \frac{S_n}{Q_n} \geq \left( \frac{nt^2}{n-1+t^2} \right)^{1/2} \right\}, \quad t \geq 0.$$

For an extensive survey of self-normalised processes and their usage in statistical inference, see Shao and Zhou (2017). Alternative distributional approximations based on self-normalization arguments characterize the relative error rather than the absolute error and are of the form

$$\left| \frac{\text{pr}(T_n \geq t)}{1 - \Phi(t)} - 1 \right| \leq C(t)\varepsilon_n,$$

where  $\varepsilon_n$  is a positive decreasing sequence, depending on  $n$  but not on  $t$ . Such inequalities are called large or moderate deviation bounds depending on the relative magnitude of  $t$  to  $n$ .

## 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 the 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.

Fisher's (1945) analysis of the saturated  $n = 2^k$  factorial experiment is a helpful further example. Normally distributed outcomes  $Y_1, \dots, Y_n$  are observed on each of the possibly large number  $n$  of treatment combinations and the corresponding vector of treatment combination means is denoted by  $\mu$ . An alternative representation is  $\mu = H^{-1}\tau$ , where  $H$  is a  $2^k \times 2^k$  Hadamard matrix and  $\tau = (2^k\tau_1, 2^{k-1}\tau_2, \dots, 2^{k-1}\tau_n)$  where  $\tau_1 = \mu_{(1)}$  is the overall mean and  $\tau_2, \dots, \tau_n$  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  $\mu$  is the single observation  $Y = (Y_1, \dots, Y_n)^T$  while that of the contrasts is the average of  $2^k$  mean differences and has variance  $4\sigma^2/n$ , where  $\sigma^2$  is the common variance of  $Y_i$ ,  $i = 1, \dots, n$ . 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 Stein's (1956) problem. The contrast estimators  $\hat{\tau}_2, \dots, \hat{\tau}_n$  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,

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

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  $\mathbb{E}(Z) = \Delta = \delta \mathbf{1}_m$  so that  $M$  is the  $m \times n$  matrix whose  $j$ th 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,  $X = (M^T M)^{-1} M^T$  and the standardized information matrix for  $\Delta$  is  $X^T X = 2I$ . 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  $\Theta$  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  $\theta$  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  $s = 1, \dots, p$ , where  $\hat{\theta}$  is the maximum likelihood solution. 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 Fisher's (1945) saturated factorial experiment, consider a linear nonsingular transformation to  $\psi = \psi(\theta)$  such that the elements of the transformation matrix satisfy

$$\frac{\partial \theta^s}{\partial \psi^t} = \gamma_{ts} B_s, \quad t = 1, \dots, p \quad (6)$$

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

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

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  $\gamma_{ts}$  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 (6) 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 (6) 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  $\psi$  an interest parameter and  $\lambda = (\lambda_1, \dots, \lambda_{p-1})$  a vector of nuisance parameters, let  $s$  be a jointly minimal sufficient statistic for  $(\psi, \lambda)$ . 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  $j = 1, \dots, p-1$ , where  $s_\psi$  is a minimal sufficient statistic for  $\psi$  and  $s_j$  is minimal sufficient for  $\lambda_j$ . 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

$$\frac{d\hat{\psi}}{dx} = \frac{\partial\hat{\psi}}{\partial s_\psi} \frac{ds_\psi}{dx} + \sum_{j=1}^{p-1} \frac{\partial\hat{\psi}}{\partial \hat{\lambda}_j} \frac{\partial \hat{\lambda}_j}{\partial s_j} \frac{ds_j}{dx}. \quad (7)$$

This isolates two sources of variation. One is due to sensitivity of the sufficient statistic  $s_\psi$  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  $\ell$  are sources of finite sample bias, which disappears asymptotically.

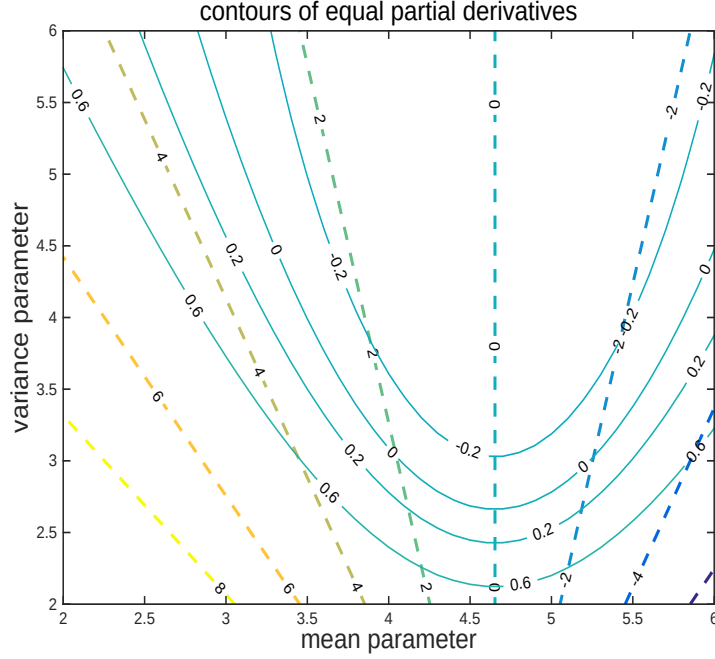


Figure 1: contours  $\{(\psi, \lambda) : \nabla_{\psi} \ell(\psi, \lambda) = c\}$  for different values of  $c$  for one realization of  $\ell$ , the log likelihood function corresponding to 10 observations from a normal distribution of mean 4 and variance 4. Dashed lines correspond to  $\psi$  being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

Large deviations in an estimate of  $\psi$  would arise due to sampling error in other estimates  $\hat{\lambda} = (\hat{\lambda}_1, \dots, \hat{\lambda}_p)$  if the corresponding contour  $\{(\psi, \lambda) : \nabla_{\psi} \ell(\psi, \lambda) = 0\}$  curves in the  $\psi$  direction as one moves in one or more of the  $\lambda$  directions. Intuitively, if the contour is (locally) flat as a function of the nuisance parameter(s), then the maximum likelihood estimator of  $\psi$  is the same regardless of whether the log likelihood is evaluated at the true value of  $\lambda$  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 (7), 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 (7), but is captured in essence by the terms  $ds_{\psi}/dx$  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, p40) provides the following upper bound.

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

$$\Pr(|s(X_1, \dots, X_n) - Es(X_1, \dots, X_n)| > t) \leq \exp(-t^2/2v),$$

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

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

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

$$\text{var}\{s(X_1, \dots, X_n)\} \leq E \sum_{i=1}^n \text{var}_i\{s(X_1, \dots, X_n)\}, \quad (8)$$

365 with equality if  $s$  is linear, where

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

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

367 Consider equation (7) for the means and contrasts parameterization of the  $n = 2^k$   
 368 factorial experiment, treating a single mean or contrast as the interest parameter  $\psi$   
 369 and the remaining parameters as components of  $\lambda$ . Neither the means nor the con-  
 370 trasts are affected by the geometry of the log likelihood function because the contours  
 371  $\{(\psi, \lambda) : \nabla_\psi \ell(\psi, \lambda) = 0\}$  are flat in both parameterizations. However, the sufficient statis-  
 372 tics for the contrasts, being an average of  $n/2$  observations, concentrate in a  $O(n^{-1/2})$   
 373 neighbourhood around their means as  $n \rightarrow \infty$ . By contrast, the sufficient statistic for the  
 374  $i$ th mean is simply  $X_i$  and has fluctuations of order  $O(1)$  regardless of  $n$ .

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

381 The geometric ideas summarized by Figure 1 relate to a single realization of the log  
 382 likelihood function. Under hypothetical replication, provided that  $n$  is large enough and  
 383  $s(X_1, \dots, X_n)$  concentrates reasonably about its mean, the position and curvature of  
 384 the contour  $\{(\psi, \lambda) : \nabla_\psi \ell(\psi, \lambda) = 0\}$  will be relatively stable. A more formal algebraic  
 385 treatment is based on an expression for finite sample bias of the maximum likelihood  
 386 estimator under regime (i). By Taylor series expansion of a generic component  $\ell_r$   
 387 of the score vector around the true value  $\theta$  and evaluation at  $\hat{\theta}$ , Barndorff-Nielsen and Cox  
 388 (1994, p.150) gave an expansion for the  $r$ th entry  $\hat{\theta}_r$  of a  $p$ -dimensional vector of maximum  
 389 likelihood estimates. This is of the form (Barndorff-Nielsen and Cox, 1994, equation 5.25)

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

390 where  $i^{rs}$  is the  $(r, s)$ th entry of the inverse expected information (here we suppress the  
 391 earlier notation indicating which coordinate system is being used),  $\nu^{rst} = i^{rq} i^{sv} i^{tz} \nu_{qvz}$ ,  
 392 where  $\nu_{qvz} = \partial_{qvz}^3 \ell$  and  $\nabla$  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. See also Weatherburn (1957, section 22) for a discussion of covariant and contravariant tensor calculations and notation. Such results are valid only under ordinary repeated sampling with a fixed number of parameters.

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

$$E(\hat{\theta}_r - \theta_r) = \frac{1}{2} i^{rs} i^{tu} (\nu_{stu} + 2\nu_{st,u}) + \dots \quad (9)$$

where  $\nu_{stu} = E(\partial_{stu}^3 \ell)$ ,  $\nu_{st,u} = E(\partial_{st}^2 \ell \partial_u \ell)$  and evaluation of all quantities is at the true  $\theta$ . 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 (9) 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 (9) as

$$E(\hat{\theta}_r - \theta_r) = -\frac{1}{2} i^{rs} i^{tu} (-^1\Gamma_{stu}) - \frac{1}{2} i^{rs} i^{tu} (\nu_{tu,s} + \nu_{su,t} - \nu_{st,u}) + i^{rs} i^{tu} \nu_{st,u} + \dots \quad (10)$$

where  $^{-1}\Gamma_{stu}$  are the coefficients of  $\alpha$ -connection with  $\alpha = -1$ . That the bias depends on the  $-1$ -connections is perhaps not surprising as this has been documented elsewhere in curved exponential families under dimension constraints (see Barndorff-Nielsen *et al.*, 1986, p. 91). The second term on the right hand side of (10) is  $i^{rs} \delta_q^t \Upsilon_{st}^q$  where  $\Upsilon_{st}^q = \frac{1}{2} i^{qu} (\nu_{tu,s} + \nu_{su,t} - \nu_{st,u})$  is somewhat similar in form to the  $\Gamma_{st}^q$  coefficients of Riemannian connection:

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

### 4.3 Special structure in the likelihood function

As shown in section 4.1 and elsewhere (e.g. Sur and Candès, 2019; Fan *et al.*, 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  $\psi = (\psi_1, \dots, \psi_{p-p_N})$  in the presence of nuisance parameters  $\lambda = (\lambda_1, \dots, \lambda_{p_N})$  hinges on a reformulation in which the nuisance parameters play a reduced role. This is most successful when  $\lambda$  is evaded completely without loss of information on  $\psi$ , 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  $\psi$  and  $\lambda$ . Factorization (a) is a case for marginalization with  $U$  sufficient for  $\psi$ , (b) is a case for conditioning on  $U$ , which is, here, the sufficient statistic for  $\lambda$ . In (c) the jointly sufficient statistic is two individually sufficient statistics so that conditioning reduces to marginalization. Weaker factorizations

427 are

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

428 Marginalization is applicable in (d) and conditioning in (e). However, information on  $\psi$   
429 may be lost in either case because neither  $U$  nor  $V$  are individually sufficient for  $\psi$  or  $\lambda$ .

430 Bartlett (1937) proposed conditional likelihood as a means of avoiding the bias in-  
431 curred by standard maximum likelihood estimation, exemplified by estimation of the  
432 variance in the Gaussian pairs example (see section 4.1). Modified profile likelihood  
433 (Barndorff-Nielsen, 1983, 1988) arises as a higher order approximation to a marginal or  
434 conditional likelihood function for the interest parameter  $\psi$  when one of these is available  
435 (see Barndorff-Nielsen, 1994). However, its specification relies on the ability to identify an  
436 ancillary complement to the maximum likelihood estimator unless the parameterization  
437 is orthogonal for the interest parameter, or can be made such by an interest-respecting  
438 transformation (see Cox and Reid, 1987). Barndorff-Nielsen (1984) discusses a construc-  
439 tive approach for identifying such ancillary complements in curved exponential models.  
440 In the specification of the modified directed likelihood root  $r^*$ , used for inference on an in-  
441 terest parameter, conditioning on an approximate ancillary statistic is somewhat implicit;  
442 see e.g. Brazzale *et al.* (2007, chapter 8.4.3). See Tang and Reid (2020) for a discussion  
443 of the extent to which the properties of  $r^*$ , established under asymptotic regime (i), are  
444 valid under regime (ii).

445 There are, in principle, three types of systematic approaches for eliminating nuisance  
446 parameters, perhaps approximately, for a scalar interest parameter. One is to solve the  
447 system of differential equations specified by Cox and Reid (1987). A second, suitable if  
448 one of factorization (a)–(c) holds, is to specify a constructive approach for finding ancil-  
449 lary statistics for the interest or nuisance parameters. A third strategy is to develop a  
450 systematic procedure for identifying statistics  $U$ ,  $V$  or both such that one of factoriza-  
451 tions (a)–(e) holds. See Battey and Cox (2020, section 7.2) for a formulation and some  
452 suggested test cases. A referee has indicated that the third goal is achieved implicitly by  
453 the tangent exponential model of Fraser and Reid (1995) and Fraser *et al.* (2010). See  
454 also the appendix of Fraser *et al.* (2016). The various approaches are likely to be related.

455 If exact elimination of nuisance parameters is achievable, or global parameter orthog-  
456 onalization is possible based on the observed Fisher information function, high dimen-  
457 sionality poses no additional difficulty. However, approaches that rely on approximations  
458 to these situations in low dimensions are likely to fail when  $p$  is allowed to grow with  $n$ .

## 459 4.4 Sparsity

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

465 There are three roles of sparsity appropriate in different settings and its implications  
466 are contingent on this. A first role is to encourage fruitful interpretation when there are a  
467 large number of parameters, initially on an equal footing and potentially of interest, and  
468 yet it is likely that few represent securely established effects of importance. The second  
469 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 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.

Consider a regression problem in which a response variable  $Y_i$  is observed on study individuals  $i = 1, \dots, n$  alongside a vector of  $p$  potential explanatory variables,  $x_i$ , where  $p$  is large, typically tens of thousands in a genomics context. The assumption that there is an unknown low-dimensional submodel consisting of just a few important features seems vital for communication of results.

While the design of any study, observational or experimental, needs attention to the feasibility of analysis capable of addressing the questions at issue, new challenges may arise when the data themselves are obtained. Fisher (1922) distinguished between issues of specification and those of analysis. Indeed in many contexts the above formulation in which a large number of simple representations comprise a single overparameterized model, renders the overarching model partially specified (specified up to the unknown positions of the zero entries), leading to an important conceptual distinction. This is between the phase of analysis in which a specification is chosen for detailed analysis, leading to a second phase that is interpretive. The initial model here is a device for choosing a base for the primary interpretation. Indeed, a model is often chosen partly in the light of the data before conducting inference on the associated parameters.

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  $\alpha$ -level significance test is typically not  $\alpha$ . 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 applications.

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 single-optimization procedure like the lasso (Tibshirani, 1996) is used to identify a model, and inference is subsequently performed on the associated parameters, the inference is highly misleading 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. 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.

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 *et al.* (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 the lasso 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) 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 non-zero 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 sections 2.1 and

4.6.

One treatment of linear regression in observational studies (not the recommended one; see section 4.6) considers the data on outcome  $Y$  and covariates  $X = (X_1, \dots, X_p)$  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

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

where  $\mu_X$  and  $\mu_Y$  are the means of  $X$  and  $Y$ ,  $\Sigma_{XX}$  is the covariance matrix of the  $X$  variables and  $\Sigma_{XY}$  is the column of the full covariance matrix of  $X$  and  $Y$  corresponding to the  $Y$  variable. Thus the linear regression coefficient  $\beta$  is equal by definition to

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

If the relationship is not multivariate normal then one seeks the best linear approximation to  $E(Y | X = x)$ , where the approximation error is measured by the expected squared distance of  $E(Y | X = x)$  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, p. 63–4) but its relevance is small if the relationship between  $Y$  and  $X$  is highly non-linear.

Let  $\beta_j$ , the  $j$ th element of the vector  $\beta$ , be an interest parameter. Let  $\Omega = \Sigma_{XX}^{-1}$ . A natural estimator of  $\beta_j = (\Sigma_{XX}^{-1} \Sigma_{XY})_j$  replaces  $\Omega$  and  $\Sigma_{XY}$  by estimates,  $\hat{\Omega}$  and  $\hat{\Sigma}_{XY}$  say, producing a so called plug in estimator  $\hat{\beta}_j$  of  $\beta_j$ . On writing

$$\hat{\beta}_j - \beta_j = \|\Omega_{j\cdot}\|_2 \|\hat{\Sigma}_{XY} - \Sigma_{XY}\|_2 t_1 + \|(\hat{\Omega} - \Omega)_{j\cdot}\|_2 \|\Sigma_{XY}\|_2 t_2, \quad (11)$$

where  $\Omega_{j\cdot}$  denotes the  $j$ th row of  $\Omega$ ,  $t_1 \in (-1, 1)$  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  $\Sigma_{XY}$ , 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 (11). 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 non-asymptotic deviation bounds such as Hoeffding's inequality for weighted sums of sub-Gaussian random variables or Bernstein's inequality for weighted sums of sub-exponential random variables.

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

$$\text{pr}(|\sum_{i=1}^n a_i \tilde{Z}_i| \geq t) \leq 2 \exp\{-c \min(t^2/K^2 \|a\|_2^2, t/K \|a\|_\infty)\}, \quad (12)$$

where  $K$  is a constant related to the sub-exponential 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  $k$ th element of  $\Sigma_{XY}$  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 sub-exponential tails (see e.g. Vershynin, 2018, Lemma 2.7.7). Thus by equation (12), we have, on setting  $\tilde{Z}_i = Z_i - E(Z_i)$  and  $a = (n^{-1}, \dots, n^{-1})$ ,

$$\text{pr}\{|\hat{\Sigma}_{XY} - \Sigma_{XY}|_k \geq t\} \leq 2 \exp\{-c \min(nt^2/K^2, nt/K)\},$$

so that

$$\text{pr}\{\max_k |(\hat{\Sigma}_{XY} - \Sigma_{XY})_k| \geq t\} \leq 2p \exp\{-c \min(nt^2/K^2, nt/K)\}. \quad (13)$$

The requirement for the right hand side of (13) to tend to zero as  $n$  tends to infinity is that  $t \gtrsim \max[\{(\log p)/n\}^{1/2}, (\log p)/n]$  and convergence of (13) to zero is assured by the scaling  $(\log p)/n \rightarrow 0$ . Similar arguments for the maxima of sums of random variables underpin much of the high dimensional statistical literature.

A crude bound for  $\|\hat{\Sigma}_{XY} - \Sigma_{XY}\|_2$  takes the form

$$\|\hat{\Sigma}_{XY} - \Sigma_{XY}\|_2 \leq p \max_k |(\hat{\Sigma}_{XY} - \Sigma_{XY})_k|.$$

The right hand side is unbounded with probability 1 as  $p \rightarrow \infty$ , but under an assumption that only a small number  $s^*$  of the entries of  $\Sigma_{XY}$  are non-zero, 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

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

Thus consistency is assured provided that  $(s^* \log p)/n \rightarrow 0$ . 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  $\psi$  is orthogonal to a vector of nuisance parameters  $\lambda$ . 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  $\psi$ , identically in  $\psi$  and  $\lambda$  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  $p \rightarrow \infty$  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 *et al.* (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's 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  $\ell_1$  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 non-zero 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  $\ell$  denote the log likelihood function, now divided by  $n$ . Ning's and Liu's (2017) so called decorrelated score function, evaluated at the true values of the parameters is

$$S(\psi^*, \lambda^*) = \nabla_v \ell(\beta_v^*, \beta_{-v}^*) - w^{*T} \nabla_{-v} \ell(\beta_v^*, \beta_{-v}^*), \text{ where } w^{*T} = J_{v,-v}^* J_{-v,-v}^{*-1}. \quad (14)$$

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

$$J(\beta) = \begin{pmatrix} J_{v,v} & J_{v,-v} \\ J_{-v,v} & J_{-v,-v} \end{pmatrix} = \begin{pmatrix} \mathbb{E}\nabla_{v,v}^2\ell(\beta) & \mathbb{E}\nabla_{v,-v}^2\ell(\beta) \\ \mathbb{E}\nabla_{-v,v}^2\ell(\beta) & \mathbb{E}\nabla_{-v,-v}^2\ell(\beta) \end{pmatrix}. \quad (15)$$

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

$$\hat{S}(\beta_v^*, \beta_{-v}) \stackrel{\text{def}}{=} \nabla_v \ell(\beta_v^*, \beta_{-v}) - \hat{w}^T \nabla_{-v} \ell(\beta_v^*, \beta_{-v}) \quad (16)$$

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

$$\begin{aligned} \hat{S}(\beta_v^*, \hat{\beta}_{-v}) &= \nabla_v \ell(\beta_v^*, \beta_{-v}^*) - \hat{w}^T \nabla_{-v} \ell(\beta_v^*, \beta_{-v}^*) \\ &\quad + \left[ \nabla_{v,-v}^2 \ell(\beta_v^*, \beta_{-v,\alpha}) - \hat{w}^T \nabla_{-v,-v}^2 \ell(\beta_v^*, \beta_{-v,\alpha}) \right] (\hat{\beta}_{-v} - \beta_{-v}^*) \end{aligned} \quad (17)$$

The notation is inaccurate because  $\beta_{-v,\alpha} = \alpha \hat{\beta}_{-v} + (1 - \alpha) \beta_{-v}^*$  for  $\alpha \in (0, 1)$  has a different  $\alpha$  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

$$\begin{aligned} &\left[ \nabla_{v,-v}^2 \ell(\beta_v^*, \beta_{-v,\alpha}) - \hat{w}^T \nabla_{-v,-v}^2 \ell(\beta_v^*, \beta_{-v,\alpha}) \right] \\ &\approx \mathbb{E} \left[ \nabla_{v,-v}^2 \ell(\beta_v^*, \beta_{-v}^*) - w^{*T} \nabla_{-v,-v}^2 \ell(\beta_v^*, \beta_{-v}^*) \right] = J_{v,-v}^* - J_{v,-v}^* J_{-v,-v}^{*-1} J_{-v,-v}^* = 0, \end{aligned} \quad (18)$$

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 (17) 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,  $\beta_v^*$  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.

## 4.6 Conditioning in regression of large systems

Section 2.1 emphasized the need to condition on an ancillary statistic. The need for conditioning arises in a different guise in parts of the high dimensional literature, notably that concerned with regression where it is widespread practice in theoretical work to treat the matrix  $X$  of covariate information as random. In a low dimensional setting the ordinary least squares estimator enjoys numerous favourable properties. For instance, it is unbiased for the vector of regression coefficients both conditionally on realizations of  $X$  and also unconditionally. The former property is that which should be sought rather than the latter. To give a counterexample, an estimator whose expectation conditionally on  $X$  is far from the true value of the quantity being estimated is wildly unreasonable, even if this estimator is unbiased when averaged over the distribution of  $X$ .

## 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 non-orthogonal 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's and Reid's (1987) analysis to the misspecified situation is likely to be valuable. Further details are given by Battey and Cox (2020).

### 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 totally 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 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, e.g. 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. A related set of questions is motivated by the various examples of matched pair designs with different responses.
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, i.e. 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? A test case is exponential matched pairs with a difference in mean of  $2\psi$  rather than a difference in rate of  $2\psi$  as discussed by Battey and Cox (2020). This does not fit any of the five factorizations exactly.

- 747 9. In partial likelihood the nuisance parameter,  $\lambda$ , is a whole (baseline hazard) func-  
748 tion. Are there other situations where functional nuisance parameters can be han-  
749 dled by factorizations (a)–(e)?
- 750 10. How does one find such a factorization when the likelihood function is unavailable  
751 analytically? See Patel *et al.* (2019) for an example of such a situation motivated  
752 by a problem in biophysics, and see Shlomovich *et al.* (2020) for a more general  
753 class of examples arising in the analysis of doubly stochastic point processes.
- 754 11. When there are nuisance parameters two approaches are to transform the data and  
755 to marginalize or condition based on factorizations (a)–(e) above, or to find an  
756 interest-respecting orthogonal transformation as in Cox and Reid (1987). Is there  
757 a connection between the two and if so, can it be characterized geometrically?
- 758 12. Consider an arbitrary problem in which there are  $n$  observations and  $p$  parameters,  
759 of which one is an interest parameter and the rest are nuisance parameters. Princi-  
760 ples of inference suggest reducing the dimension  $n$  of the data to  $p$  (sufficiency) and  
761 reducing the dimension of the sufficient statistic,  $s$ , to that of the interest parameter  
762 by constraining  $s$  to the subspace of its sample space where the maximum likelihood  
763 estimate of the nuisance parameter is fixed, a version of ancillarity. Framed in this  
764 way, the situation in which  $p > n$  is precluded. But if a transformation could be  
765 applied to make the problem depend on the interest parameter and a small set of  
766 one-dimensional summaries of the original nuisance parameters, then the  $p > n$   
767 problem would have been reduced to the usual  $p < n$  situation for each interest  
768 parameter. Sparsity might still be needed for the generating process to make sense.  
769 Is it possible to construct a theory of inference that is Fisherian when  $p < n$  and  
770 can be made Fisherian by suitable transformation when  $p > n$  under a sparsity  
771 constraint?

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

- 779 13. Suppose there are blocks of individuals with different types of relation between and  
780 within blocks. If the structure is not initially clear, how would such anomalies be  
781 detected? The following situations exemplify these issues.
- 782 • Data are collected on a suitable sample of adults arranged in family groups  
783 for convenience of data collection. The family grouping may be of no direct  
784 interest but intra family dependences are likely to inflate errors of estimation  
785 as compared with random sampling.
  - 786 • There may be similar situations in which the analogue of family structure is  
787 present but not observed. That is, there are a large number, possibly individ-  
788 ually small, local structures tending to inflate variances.
  - 789 • There may be time series or spatial structure in the data, perhaps of intrinsic  
790 interest but, regardless, likely to affect standard errors. Notably in large sets

791 of data there may be multiple sources of local correlation, individually possibly  
792 small but having appreciable cumulative effect.

- 793 • A further class of problems arise when the dependence is totally unstructured,  
794 for instance it may arise in residuals due to estimation of a coefficient vector,  
795 perhaps under a misspecified model.

796 14. Are there ways of estimating the standard errors of estimators reasonably in the  
797 presence of dependent observations without without explicit estimation of the de-  
798 pendency structure? Can this be framed as a nuisance parameter problem and  
799 simplified via factorizable structure as in section 4.3?

800 A referee has pointed out that question 13 is connected to community detection in  
801 the statistical analysis of networks literature.

802 We have largely assumed that a suitable probabilistic model is given, although dis-  
803 cussion of this aspect appears in sections 4.4.1, 4.4.2 and 5.1.

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

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

810 **Acknowledgements.** The work was supported by a UK Engineering and Physical Sci-  
811 ences Research Fellowship (to HSB). We are grateful to five anonymous referees and the  
812 associate editor for references and detailed constructive criticism.

## 813 REFERENCES

814 Anastasiou, A. and Reinert, G. (2020). Bounds for the asymptotic distribution of the  
815 likelihood ratio. *Ann. Appl. Probab.*, 30, 608–643.

816 Barndorff-Nielsen, O. E. (1983) On a formula for the distribution of the maximum  
817 likelihood estimator. *Biometrika*, 70, 343–365.

818 Barndorff-Nielsen, O. E. (1984). On a formula for the distribution of the maximum  
819 likelihood estimator. *Scand. J. Statist*, 11, 157–170.

820 Barndorff-Nielsen, O. E. (1988). *Parametric Statistical Models and Likelihood*. Springer,  
821 Heidelberg.

822 Barndorff-Nielsen, O. E. (1994). Adjusted versions of profile likelihood and directed  
823 likelihood, and extended likelihood. *J. R. Statist. Soc. B.*, 56, 125–140.

828

829 Barndorff-Nielsen, O. E. and Cox, D. R. (1979). Edgeworth and saddlepoint approxi-  
830 mations with statistical applications (with discussion). *J. R. Statist. Soc. B.*, 41, 279–312.  
831

832 Barndorff-Nielsen, O. E., Cox, D. R. and Reid, N. M. (1986). The role of differential  
833 geometry in statistical theory. *International Statistical Review*, 54, 83–96.  
834

835 Barndorff-Nielsen, O. E. and Cox, D. R. (1989). *Asymptotic Techniques for Use in*  
836 *Statistics*. Chapman and Hall, London.  
837

838 Barndorff-Nielsen, O. E. and Cox, D. R. (1994). *Inference and Asymptotics*. Chap-  
839 man and Hall, London.  
840

841 Bartlett, M.S. (1937). Properties of sufficiency and statistical tests. *Proc. R. Soc. Lond. A*,  
842 160, 268–282.  
843

844 Battey, H. S. and Cox, D. R. (2018), Large numbers of explanatory variables: a prob-  
845 abilistic assessment. *Proc. R. Soc. Lond. A.*, 474, 20170631.  
846

847 Battey, H. S., Cox, D. R. and Jackson, M. V. (2019). On the linear in probability  
848 model for binary data. *Royal Society Open Science*, 6, 190067.  
849

850 Battey, H. S. and Cox, D. R. (2020), High dimensional nuisance parameters: an ex-  
851 ample from parametric survival analysis. *Information Geometry*, to appear.  
852

853 Beale, E. M. L., Mallows, C. L. (1959). Scale mixing of symmetric distributions with  
854 zero means. *Ann. Math. Statist.* 30, 1145–1151.  
855

856 Berk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. (2013). Valid post-selection  
857 inference, *Ann. Stat.* 41, 802–837.  
858

859 Berry, A. C. (1941). The accuracy of the Gaussian approximation to the sum of in-  
860 dependent variates. *Trans. Amer. Math. Soc.*, 49, 122–136.  
861

862 Bliss, C. I. (1935). The calculation of the dosage-mortality curve. *Ann. Appl. Biology*,  
863 22, 135–167.  
864

865 Booth, K. H. V. and Cox, D. R. (1962). Some systematic supersaturated designs.  
866 *Technometrics* 4, 489–495.  
867

868 Boucheron, S., Lugosi, G., Massart, P. (2013). *Concentration inequalities. A nonasymptotic*  
869 *theory of independence*. Oxford University Press, Oxford.  
870

871 Brazzale, A., Davison, A. C. and Reid, N. M. (2007). *Applied Asymptotics*. Cam-  
872 bridge University Press.  
873

874 Cai, T. T. and Guo, Z. (2017). Confidence intervals for high-dimensional linear re-  
875 gression: minimax rates and adaptivity. *Ann. Statist.*, 45, 615–646.  
876

- 877 Cauchy, A.L. (1829). Mémoire sur divers points d'analyse. *Mémoires de l'Académie*  
878 *de France*, 8, 130–138; reprinted in *Oeuvres Complètes d'Augustin Cauchy*, série 1, 2,  
879 52–58, 1908, Gauthier-Villars, Paris.
- 880
- 881 Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey,  
882 W. and Robins, J. (2017). Double/debiased machine learning for treatment and struc-  
883 tural parameters. *The Econometrics Journal*, 21, 1–68.
- 884
- 885 Cox, D. R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.*,  
886 29, 357–372.
- 887
- 888 Cox, D. R. and Snell, E. J. (1974). The choice of variables in observational studies.  
889 *J. R. Statist. Soc. C.*, 23, 51–59.
- 890
- 891 Cox, D. R. (1972). Regression models and life-tables (with discussion). *J. R. Statist. Soc. B.*,  
892 34, 187–220.
- 893
- 894 Cox, D. R. (1975a). Partial likelihood. *Biometrika*, 62, 269–276.
- 895
- 896 Cox, D. R. (197b). A note on data-splitting for the evaluation of significance levels.  
897 *Biometrika*, 62, 441–444.
- 898
- 899 Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate condi-  
900 tional inference (with discussion). *J. R. Statist. Soc. B.*, 49, 1–39.
- 901
- 902 Cox, D. R. and Wermuth, N. (1996). *Multivariate Dependencies*. Chapman and Hall,  
903 London.
- 904
- 905 Cox, D. R. and Battey, H. S. (2017). Large numbers of explanatory variables, a semi-  
906 descriptive analysis. *Proc. Nat. Acad. Sci.*, 114, 8592–8595.
- 907
- 908 Cox, D. R. (2020). Statistical significance. *Annu. Rev. Stat. Appl.*, 7, 1–10.
- 909
- 910 Daniels, H. E. (1954). Saddlepoint approximations in statistics, *Ann. Math. Statist.*,  
911 25, 631–650.
- 912
- 913 Davison, A. C., Fraser, D. A. S., Reid, N., Sartori, N. (2014). Accurate directional  
914 inference for vector parameters in linear exponential families. *J. Amer. Statist. Assoc.*,  
915 109, 302–314.
- 916
- 917 Dawid, A. P. (1991). Fisherian inference in likelihood and prequential frames of ref-  
918 erence (with discussion). *J. R. Statist. Soc., B*, 53, 79–109.
- 919
- 920 Efron, B. (1969). Student's t-test under symmetry conditions. *J. Amer. Statist. As-*  
921 *soc.*, 64, 1278–1302.
- 922
- 923 Esseen, C-G. (1942). On the Liapunoff limit of error in the theory of probability.  
924 *Arkiv för Matematik, Astronomi och Fysik*, A28, 1–19.

925

926 Fan, J., Shao, Q-M. and Zhou, W-X. (2018). Are discoveries spurious? Distributions  
927 of maximum spurious correlations and their applications. *Ann. Statist.*, 46, 989–1017.

928

929 Fan, Y., Demirkaya, E. and Lv, J. (2019) Nonuniformity of  $p$ -values can occur early  
930 in diverging dimensions. *J. Mach. Learn. Res.*, 20, 77–101.

931

932 Fang, E. X., Ning, Y. and Liu, H. (2017). Testing and confidence intervals for high  
933 dimensional proportional hazards model. *J. R. Statist. Soc. B*, 5, 1415–1437.

934

935 Fieller, E. C. (1954). Some problems in interval estimation. *J. R. Statist. Soc. B.*,  
936 16, 175–185.

937

938 Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A*, 222, 309–368.

940

941 Fisher, R. A. (1925). Expansion of “Student’s” integral in powers of  $n^{-1}$ . *Metron*, 5,  
942 109–120.

943

944 Fisher, R. A. (1945). A system of confounding for factors with more than two alterna-  
945 tives, giving completely orthogonal cubes and higher powers. *Ann. Eugenics*, 12, 283–290.

946

947 Fraser, D. A. S., Fraser, A. M. and Staicu, A-M. (2010). The second order ancillary:  
948 a differential view with continuity. *Bernoulli*, 16, 1208–1223.

949

950 Fraser, D. A. S. and Reid, N. (1995). Ancillaries and third order significance. *Utilitas*  
951 *Math.*, 47, 33–53.

952

953 Fraser, D. A. S., Reid, N. and Sartori, N. (2016). Accurate directional inference for  
954 vector parameters. *Biometrika*, 103, 625–639.

955

956 Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing  
957 for high-dimensional regression. *J. Mach. Learning Res.*, 15, 2869–2909.

958

959 Kolmogorov, A. (1933). *Foundations of the theory of probability*. Chelsea Publishing  
960 Company, New York.

961

962 Lee, J. D., Sun, D. L., Sun, Y. and Taylor, J. E. (2016), Exact post-selection inference,  
963 with application to the lasso, *Ann. Statis.*, 44, 907–927.

964

965 Lei, L., Bickel, P. J. and El Karoui, N. (2018). Asymptotics for high dimensional  
966 regression M-estimates: fixed design results. *Probability Theory and Related Fields*, 172,  
967 983–1079

968

969 Lindsay, B. G. (1985). Using empirical partially Bayes inference for increased effi-  
970 ciency. *Ann. Statist.*, 13, 914–31.

971

972 Little, R. and Rubin, D. B. (2019). *Statistical Analysis with Missing Data*. Wiley,

973 New York.

974

975 Lockhart, R., Taylor, J. E., Tibshirani, R. J. and Tibshirani, R. (2014) A significance  
976 test for the lasso. *Ann. Statist.*, 42, 413–468.

977

978 Logan, B. F., Mallows, C. L., Rice, S. O., Shepp, L. A. (1973). Limit distributions of  
979 self-normalized sums. *Ann. Probab.*, 1, 788–809.

980

981 McCullagh, P. and Polson, N. G. (2018). Statistical sparsity. *Biometrika*, 105, 797–  
982 814.

983

984 Ning, Y. and Liu, H. (2017). A general theory of hypothesis tests and confidence  
985 regions for sparse high dimensional models. *Ann. Statist.*, 1, 158–195

986

987 Patel, L., Gustafsson, N., Lin, Y., Ober, R., Henriques, R. and Cohen, E. A. K. (2019).  
988 A hidden Markov model approach to characterizing the photo-switching behavior of flu-  
989 orophores. *Ann. Appl. Stat.*, 13, 1397–1429.

990

991 Renyi, A. (1954). New axiomatic construction of probability. *Magyar Tud. Akad.*  
992 *Mat. Fiz. Oszt. Közl.*, 4, 369–427.

993

994 Satterthwaite, R. (1959). Random balance experimentation. *Technometrics*, 1, 111–  
995 159.

996

997 Shao, Q-M. and Zhou, W-X. (2017). Self-normalization: taming a wild population in  
998 a heavy-tailed world. *Appl. Math. J. Chinese Univ.*, 32, 253–269.

999

1000 Shlomovich, L., Cohen, E. A. K., Adams, N. and Patel, L. (2020). A Monte Carlo  
1001 EM Algorithm for the Parameter Estimation of Aggregated Hawkes Processes. *arXiv:*  
1002 *2001.07160*.

1003

1004 Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate  
1005 distribution. *Proc. Third Berkeley Symposium on Mathematical Statistics and Probabil-*  
1006 *ity.*, 197–206.

1007

1008 Sur, P. and Candès, E. J. (2019). A modern maximum-likelihood theory for high-  
1009 dimensional logistic regression. *Proc. Natl. Acad. Sci. USA*, 116, 14516–14525.

1010

1011 Tang, Y. and Reid, N. M. (2020). Modified likelihood root in high-dimensions.  
1012 *J. R. Statist. Soc. B*, to appear.

1013

1014 Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B.*,  
1015 58, 267–288.

1016

1017 Tibshirani, R. J., Taylor, J. E., Lockhart, R. and Tibshirani, R. (2016) Exact Post-  
1018 Selection Inference for Sequential Regression Procedures, *J. Amer. Statist. Assoc.*, 111,  
1019 600–620.

1020

1021 Tyurin, I. S. (2010). Refinement of the upper bounds of the constants in Lyapunov's  
1022 theorem (English translation). *Russian Mathematical Surveys*, 65, 586–588.

1023

1024 Wainwright, M. J. (2019) *High-dimensional statistics. A non-asymptotic viewpoint*.  
1025 Cambridge University Press, Cambridge.

1026

1027 van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014). On asymptoti-  
1028 cally optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42,  
1029 1166–1202.

1030

1031 Vershynin, R. (2018). *High Dimensional Probability*. Cambridge University Press.

1032

1033 Weatherburn, C. E. (1957). *An Introduction to Riemannian Geometry and the Tensor*  
1034 *Calculus*. Cambridge University Press, Cambridge, UK.

1035

1036 Yates, F. (1936). A new method of arranging variety trials involving a large number  
1037 of varieties. *J. Agric. Sci.*, 26, 424–455.

1038

1039 Zhang, C-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional pa-  
1040 rameters in high dimensional linear models. *J. R. Statist. Soc. B.*, 217–242.

1041

1042 Zhou, W-X., Bose, K., Fan, J., Liu, H. (2018). A new perspective on robust M-  
1043 estimation: finite sample theory and applications to dependence-adjusted multiple test-  
1044 ing. *Ann. Statist.*, 46, 1904–1931.