INDUCEMENT OF POPULATION SPARSITY

H. S. BATTEY [1] *

Abstract. The pioneering work on parameter orthogonalization by Cox and Reid (1987) is presented as an inducement of abstract population-level sparsity. This is taken as a unifying theme for this article, in which sparsity-inducing parameterizations or data transformations are sought. Three recent examples are framed in this light: sparse parameterizations of covariance models, the construction of factorizable transformations for the elimination of nuisance parameters, and inference in high-dimensional regression. Strategies for the problem of exact or approximate sparsity inducement appear to be context specific and may entail, for instance, solving one or more partial differential equation or specifying a parameterized path through transformation or parameterization space. Open problems are emphasized.

Some key words: Nuisance parameter; parameter orthogonalization; partial likelihood; sparse parameterizations.

1. Introduction

Sparsity, the existence of many zeros or near-zeros in some domain, plays at least two roles in high-dimensional statistical theory: to aid interpretation and to restrain estimation error associated with multitudinous nuisance parameters. The ideas to be presented are motivated primarily by the latter, with a low-dimensional parameter of interest encapsulating relevant aspects of interpretation. In some contexts there is a natural and interpretable notion of sparsity, and the statistical challenge is the now rather routine task of specifying an estimator that exploits this structure to give appropriate statistical guarantees. See, e.g., Wainwright (2019) for an extensive account covering numerous examples.

The present article is barely concerned with estimators and other sample quantities. Its contribution is to explore the idea that certain forms of abstract, population-level sparsity may be systematically induced, and to seek unification of some isolated examples through this perspective. The precursor, although not framed as sparsity inducement, appears to be the paper on parameter orthogonalization by Cox and Reid (1987).

The four principal examples considered fall broadly into two categories: sparsity induced by reparameterization and sparsity induced by transformations of the data. The exposition here follows this separation, although it is plausible that the two

Date: October 4, 2022.

MSC 2020 subject classification codes: Primary 62-02, 62A01; Secondary 62B99, 00A27.

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

* Author to whom correspondence may be addressed.

approaches are connected. To avoid repetition, only the essential aspects of each case are presented, from which a synthesis is attempted.

2. Sparsity induced by reparameterization

2.1. **Parameter orthogonalization (Cox and Reid, 1987).** Direct use of the likelihood function often produces misleading estimates of parameters of interest when the dimension of the nuisance parameter vector is of a similar order of magnitude to the number of independent observations, and is typically suboptimal even for moderately many nuisance parameters. One resolution, an implicit invocation of sparsity, is parameter orthogonalization (Cox and Reid, 1987), prior to likelihood inference on the parameter of interest.

Let ψ and λ represent parameters of interest and nuisance parameters, respectively, in a parametric statistical model. Where $i_{\psi\lambda}(\psi, \lambda)$ denotes the corresponding off-diagonal block of the Fisher information matrix, ψ and λ are said to be globally orthogonal if $i_{\psi\lambda}(\psi, \lambda) = 0$ for all ψ and λ and locally orthogonal if this equality holds at particular values. The implication of parameter orthogonality is that the maximum likelihood estimator of ψ behaves "almost as if" λ were fixed at its true value in the sense that $\widehat{\psi} - \widehat{\psi}_\lambda = O_p(n^{-1})$ for λ in an $O(n^{-1/2})$-neighbourhood of the true value. Here, $\widehat{\psi}$ is the unconstrained maximum likelihood estimator and $\widehat{\psi}_\lambda$ maximizes the likelihood over the constrained parameter space at λ. The previous statements assume that the dimension of λ is fixed.

Parameter orthogonality also enables higher-order inference via a simple modification to the profile log-likelihood function without the specification of an ancillary complement to the maximum likelihood estimator (Barndorff-Nielsen, 1983; Cox and Reid, 1987).

From an initial parameterization (ψ, λ), Cox and Reid (1987) provided a way to construct an interest-respecting reparameterization $(\psi, \phi(\psi, \lambda))$ such that $i_{\psi\phi} = 0$. In other words, they proposed a sparsity-inducing reparameterization. In general, establishment of the appropriate reparameterization entails solving a set of partial differential equations, although there is a simpler route if the parameter of interest is a canonical parameter of a full exponential family, as noted by Huzurbazar (1956) for two-parameter families and more generally by Barndorff-Nielsen (1978, p. 183). A version of the more explicit derivation of Appendix A was presented by Barndorff-Nielsen and Cox (1994, p. 64).

Sparsity of the Fisher information matrix has emerged in recent literature on inference in high-dimensional regression (e.g., van de Geer *et al.*, 2014; Ning and Liu, 2017; Fang *et al.*, 2017). However, its presence is assumed rather than induced as in Cox and Reid (1987). Direct imposition of such sparsity is highly restrictive, as can be seen from the simple example of a normal-theory linear regression model. In that context, with ψ taken as an arbitrary regression coefficient and λ the remaining coefficients, the requirement $i_{\psi\lambda} = 0$ without preliminary manoeuvres is equivalent to assuming that all columns of the design matrix corresponding to λ are orthogonal to those corresponding to ψ. We return to this example in Section 3.2.

2.2. **Sparse parameterizations of covariance models.** Covariance matrices and their inverses are encountered throughout classical multivariate analysis, almost always as nuisance parameters to be estimated, which necessitates a sparsity assumption to extend multivariate procedures to high-dimensional settings. More precisely, the requirement is typically that an estimator of the covariance or precision matrix is consistent in the spectral matrix norm as the dimension p of the matrix tends to infinity with the effective sample size n under a suitable scaling condition. This notional asymptotic regime is a theoretical device, a means of studying the probabilistic behaviour of an estimator as a function of n when $p > n$.

Spectral-norm consistency, while achievable under a sparsity constraint, is only relevant if the assumptions made are valid to an adequate order of approximation. This motivates a search for parameterizations under which relevant covariance models are sparse. In particular, it raises the question, to which we refer henceforth as Q^*: for a given (relevant) covariance model not obviously sparse in any domain, can a sparsity-inducing parameterization be deduced? An answer would enable reparameterization to achieve maximal sparsity, and on the transformed scale, a more effective and valid estimation could be achieved by exploiting the sparsity before transforming the conclusions back to the scale of interest. Battey (2019) established statistical guarantees for such a procedure assuming the sparsity scale is known or can be reliably estimated, which does not address Q^*. The idea of parameterizing and thereby estimating the sparsity scale using a device analogous to that proposed by Box and Cox (1964) has been suggested by Peter McCullagh in unpublished communication.

The problem Q^* remains open. The proof of concept to be outlined follows Battey (2017), who gave an example of a covariance model that is unexpectedly sparse after reparameterization. The starting point was the converse formulation: to impose sparsity in unusual domains and study the structure induced on the original and inverse scales. Consider the matrix logarithm L of a covariance matrix Σ, implicitly defined through the Taylor series expansion of the matrix exponential:

$$\Sigma = \exp(L) = \textstyle\sum_{k=0}^{\infty} L^k/k!.$$

In contrast to Σ, which belongs to the open cone of positive definite matrices, L belongs to the vector space of symmetric matrices, for which there exists a natural basis, $\mathcal{B} = \{B_1, \ldots, B_{p(p+1)/2}\}$ say, consisting of $p(p+1)/2$ p-dimensional square matrices. The choice to study the matrix logarithm was made on mathematical grounds, as the properties of vector spaces make the formulation feasible and fruitful, and allow a simple characterization of sparsity through the basis expansion $L = L(\alpha) = \sum_{m=1}^{p(p+1)/2} \alpha_m B_m$. In particular, the sparsity of L corresponds to that of the basis coefficient vector $\alpha = (\alpha_1, \ldots, \alpha_{p(p+1)/2})^\top$. With the constraint $\|\alpha\|_0 = s^* \ll p$, where $\|\alpha\|_0$ is the number of nonzero entries of α, the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ and corresponding eigenvectors $(\gamma_j)_{j=1}^p$ of the resulting matrix $\Sigma(\alpha)$ inherit substantial structure, as illustrated in Figure 1. The right-hand panel corrects Figure 1 of Battey (2017), which depicted a single realization instead of a Monte Carlo average.

Figure 1 was obtained by generating 100 realizations of a random, sparse L by taking the support of α to be random samples of size s^* from the index set $\{1, \ldots, p(p+$

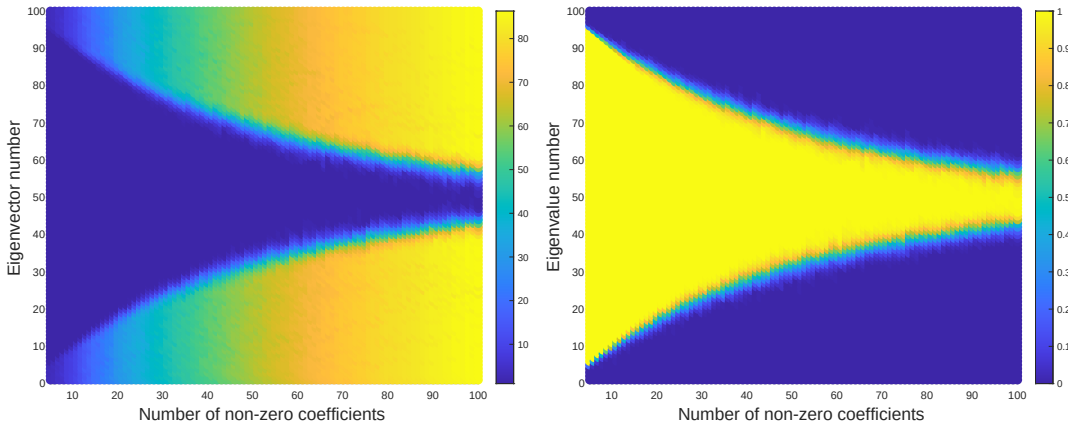

FIGURE 1. Simulation averages of $\|\gamma_j\|_0$ (left) and $\mathbb{1}\{\lambda_j = 1\}$ (right) for 100 random logarithmically s^*-sparse covariance matrices, plotted against the index j of ordered eigenvalues (y-axis) and $s^* \in \{1, \ldots, p\}$ (x-axis) for $p = 100$.

1)/2}. This was done for different values of s^* as indicated in Figure 1. The values of the nonzero basis coefficients were then drawn from a standard normal distribution, although the latter aspect is irrelevant as far as the induced structure on Σ is concerned: any other distribution could have been used instead.

Figure 1 indicates that a priori unexpected structure is present in the eigenvectors and eigenvalues of the covariance matrix, which translates to structure on the covariance matrix and its inverse. Specifically, there exists a permutation matrix P such that $\Sigma = PWP^\top$, where W consists of a potentially large, dense block and is otherwise diagonal. The dimension of the dense block is provided by theoretical analysis and can be expressed in terms of sparsity s^* and dimension p through a random-matrix perspective. The key conclusion is that L can be appreciably sparser than Σ.

This idea was extended to another class of examples by Rybak and Battey (2021), but the scope for further progress seems substantial. Most notably, the earlier work only contains a brief discussion of how one may traverse paths through a parameterization space in search of a sparse representation. These paths were chosen, following McCullagh's proposal, to pass through the covariance and inverse covariance parameterizations, as well as the matrix logarithmic parameterization. In the Box and Cox (1964) analysis of transformations, a path through a model space was viewed as a technically convenient way of assessing a discrete set of plausible models for compatibility with data. Any values of their key transformation parameter that yielded nonphysical models were not used in the final analysis. The situation in the present context is different as the covariance matrix is almost always a nuisance parameter and the goal of the reparameterization is to aid inference on the parameters of interest — a line of argument similar to that of Cox and Reid (1987).

3. Sparsity induced by transformations of data

3.1. **Construction of factorizable transformations in matched-comparison problems.** Suppose that $X_1, \ldots, X_b$ are responses on b blocks of individuals, where each X_i is a vector of random variables. The simplest realistic example has $(X_i)_{i=1}^b = (T_i, C_i)_{i=1}^b$, an outcome variable on b pairs of homozygotic twins, where one twin from each pair has been chosen at random to receive a treatment and their outcome denoted by T_i, while the other twin is the untreated control with outcome C_i. The goal is inference on the treatment effect ψ in the presence of pair-specific nuisance parameters $\lambda_1, \ldots, \lambda_b$, the latter arising from an inability or unwillingness to model the generating process in detail. These may, for instance, represent genetic differences between the b twin pairs.

Maximum likelihood estimation without preliminary manoeuvres typically produces misleading inference for ψ. It is, however, sometimes possible to eliminate b nuisance parameters from the analysis by exploiting the natural separation in the data due to the matching, and transforming the observations in such a way that the resulting likelihood function, $L(\psi, \lambda; x)$ say, fruitfully factorizes. Such a factorization is of the form

$$L(\psi, \lambda; x) = L_{\mathrm{pa}}(\psi; x) L_{\mathrm{r}}(\psi, \lambda; x), \tag{3.1}$$

where the factor $L_{\mathrm{pa}}(\psi; x)$ is called the partial likelihood (Cox, 1975). Ideally, little or no information for inference on ψ is lost through relinquishment of the remainder likelihood $L_{\mathrm{r}}(\psi, \lambda; x)$.

Conditional (Bartlett, 1936, 1937) and marginal (Fraser, 1968) likelihood are special cases of partial likelihood in which $L_{\mathrm{pa}}(\psi; x)$ is replaced by the product of appropriate marginal and conditional probability functions, respectively, evaluated at the data. A different construction leading to the encompassing form in Equation (3.1) was given by Cox (1972) to evade estimation of the baseline hazard function in the proportional hazards model.

In a matched-comparison setting, a suitable marginal likelihood is found by making a transformation $s(x)$ such that the probability density or mass function f_S of $S_i = s(X_i)$ is free of λ_i, so that $L_{\mathrm{pa}}(\psi, x)$ can be taken as $\prod_i f_S(s_i; \psi)$. Such factorizations, with a partial-likelihood component that is free of nuisance parameters, need not exist, which raises the question of whether useful approximate versions are available.

The following example (Lindsay, 1980) solidifies ideas. Similar examples based on different distributions appear in Cox (1958), Cox and Hinkley (1974) and Barndorff-Nielsen and Cox (1994).

Example 1. Suppose that T_i and C_i are exponentially distributed of rates $\lambda_i \psi$ and λ_i / ψ, respectively. The marginal density of $S_i = T_i / C_i$ at s is $f_S(s) = \psi^2 / (1 + \psi^2 s)^2$, which does not depend on λ_i. Thus $(S_i)_{i=1}^b$ are independent and identically distributed and can be used for likelihood-based inference on ψ.

The connection to sparsity is that the partial derivative of f_S with respect to λ is identically zero when the distribution of S_i is free of λ. Thus, search for a transformation $S_i = s(X_i)$ that produces sparsity of $\nabla_\lambda f_S$ at any set of evaluation points corresponds to a search for a factorizable transformation of X_i. Battey, Cox and

Lee (2022) used this sparsity as the basis for systematically deducing a factorizable transformation. This was achieved via integro-differential equations constructed from the Laplace transform of f_S and convertible to standard forms of partial differential equations in some contexts. The approach can also be viewed as inducing sparsity on the (ψ, λ) off-diagonal blocks of the Fisher information matrix, as in Cox and Reid (1987).

3.2. **Inference in high-dimensional linear regression.** The work to be outlined (Battey and Reid, 2022) stemmed from an attempt to formulate the high-dimensional linear regression problem in a way that would evade nuisance parameters as in Section 3.1. Sparsity is induced through linear transformations of the data, which exploits the linearity of the regression model.

For an arbitrary component β_v of the p-dimensional regression coefficient vector β, write the linear regression model for the n observations in matrix form as

$$Y = X\beta + \varepsilon = x_v\beta_v + X_{-v}\beta_{-v} + \varepsilon,$$

where ε is an error term with a mean of zero and a variance of τ, β_{-v} is the nuisance component of β, and X_{-v} is the matrix X of observations on explanatory variables after removing the column corresponding to β_v. Although $p \gg n$, we assume β is sparse in the sense that $\|\beta\|_0 = s \ll p$.

Each coefficient is treated in turn as the parameter of interest and an interest-respecting transformation is sought that produces sparsity (the existence of many zeros or near-zeros) in the (β_v, β_{-v}) off-diagonal block of the conditional Fisher information matrix or, rather, the notional conditional Fisher information matrix that would have emerged had a normality assumption been made.

Sparsity is achieved approximately by premultiplication of the regression equation by an $n \times n$ matrix A^v:

$$\begin{aligned} A^vY &= A^vX\beta + A^v\varepsilon \\ \widetilde{Y}^v &= \widetilde{X}^v\beta + \widetilde{\varepsilon}^v = \widetilde{x}^v_v\beta_v + \widetilde{X}^v_{-v}\beta_{-v} + \widetilde{\varepsilon}^v, \end{aligned}$$

with A^v such that $\widetilde{x}^v_v$ is as orthogonal as achievable to the $p-1$ columns of $\widetilde{X}^v_{-v}$. Orthogonality means that direct regression of $\widetilde{Y}^v$ on $\widetilde{x}^v_v$ estimates β_v without bias, as in a factorial experiment. In practice, exact orthogonalization is not achievable from this premultiplication strategy, so some bias is expected from the simple least squares regression of Y^v on the single column $\widetilde{x}^v_v$. Let $\widetilde{\beta}_v = \widetilde{x}^v_v\widetilde{Y}^v/(\widetilde{x}^{v\mathrm{T}}_v\widetilde{x}^v_v)$ denote such a marginal least-squares estimator of β_v. Choose A^v to minimize an observable upper bound on the mean squared error of $\widetilde{\beta}_v$. This leads, after a re-expression in terms of $q_v = A^{v\mathrm{T}}A^vx_v$, to a simple, unconstrained optimization problem with an exact analytic solution. The estimator of β_v in terms of q_v is

$$\widetilde{\beta}_v = (\widetilde{x}^{v\mathrm{T}}_v\widetilde{x}^v_v)^{-1}\widetilde{x}^v_v\widetilde{Y}^v = (x_v^{\mathrm{T}}A^{v\mathrm{T}}A^vx_v)^{-1}x_v^{\mathrm{T}}A^{v\mathrm{T}}A^vY = (q_v^{\mathrm{T}}x_v)^{-1}q_v^{\mathrm{T}}Y.$$

Using the optimized value $q_v = a(I + X_{-v}X_{-v}^{\mathrm{T}})^{-1}x_v$ ensures that the observable upper bound on the mean squared error of $\widetilde{\beta}_v$ is minimized, where a is any nonzero real number and can be taken to be one without loss of generality.

Let b_v be the bias associated with $\widetilde{\beta}_v$. This decays at a rate of $O(s/n)$, where n is the sample size and s the sparsity of β. The implication is that, by ignoring b_v, Wald-based inference for each β_v is accurate to an order of $O(s/\sqrt{n})$.

There are strong connections to earlier work on the problem of inference in high-dimensional regression, notably to Zhang and Zhang (2014) and van de Geer *et al.* (2014). A key conceptual distinction, that aligns these works with the theme of this present article, is that the construction outlined above induces sparsity where earlier contributions assume it.

Battey and Reid (2022) proposed using the confidence intervals for each β_v as part of a broader inferential framework concerned with uncertainty over the sparse regression model. The need for such confidence sets of models was emphasized in Cox (1968, 1995), Cox and Snell (1974, 1989) and Cox and Battey (2017).

4. Discussion and open problems

This short article is an attempt to synthesize four ideas through a unifying theme of sparsity inducement: a search for parameterizations or data transformations that yield zeros or near-zeros as components of key population-level objects. This is in contrast to a large body of work in which any sparsity on population-level objects is assumed and subsequently imposed on sample objects through penalization or thresholding.

A precise mathematical characterization in full generality seems a formidable challenge. The following list of open problems may be useful endeavours in this goal.

(1) In Section 2.2, for a given covariance model, not obviously sparse in any domain, can a sparsity-inducing parameterization be deduced? What is the appropriate formalization of sparsity in such contexts?
(2) If the interpretability of nuisance parameters is accepted as immaterial in Section 2.2, then a broader search over parameterization space is suggested than that permitted through simple parameterized paths. How can this be operationalized?
(3) What is the most appropriate approximate formulation of the factorizable transformation problem highlighted in Section 3.1? Such a formulation would need to allow the probability density function of the transformed random variable to depend weakly on the nuisance parameter. How should the adequacy of such transformations be assessed?
(4) Are there systematic routes to deducing fruitful partial likelihood factorizations more generally, beyond the matched comparison problems of Section 3.1? This question was one of five posed by Cox (1975) and has remained open since, except for the modest progress by Battey, Cox and Lee (2022).
(5) The transformations of Section 3.2 used the composite of observed data, and the sparsity-inducing transformations relied on the special structure of the linear regression model. Is there a formulation broad enough to encompass the transformations of both Sections 3.1 and 3.2?
(6) In relation to points (4) and (5): a key example with a marginal likelihood component that does not appear to be recoverable through direct application

of the ideas of Battey, Cox and Lee (2022) is a normal-theory linear regression model with a coefficient vector β and an unknown error variance σ^2. The minimal sufficient statistic is $(\hat{\beta}, S^2)$, where $\hat{\beta}$ is the least squares estimator and S^2 is the residual sum of squares divided by the residual degrees of freedom. Direct use of the likelihood function produces an estimate of error variance that is too small, particularly when the number of covariates is appreciable relative to the number of independent observations. By the minimal sufficiency and independence of $\hat{\beta}$ and S^2, the joint density function of the responses factorizes as $f(y; \beta, \sigma^2) = f(y \mid \hat{\beta}, s^2) f(\hat{\beta}; \beta, \sigma^2) f(s^2; \sigma^2)$ and reliable inference for σ^2 can be obtained by using the final component to construct a partial likelihood. The point of considering simple examples is not to recover the correct answer, but rather to do so through a seamless application of theory, which can then be applied to more challenging situations. See Fraser, Reid and Lin (2018) for another discussion in this vein.

(7) Likelihood theory has, associated with it, an important and enlightening differential geometric interpretation. How does partial likelihood (including marginal and conditional likelihood as special cases) fit into this discussion?

(8) Can a connection be established between data-based transformations for the elimination of nuisance parameters via marginal or conditional likelihood and the interest-respecting reparameterisations of Cox and Reid (1987)? Some informal remarks in Battey, Cox and Lee (2022), which were based on Fraser (1964), made a connection between sample and parameter spaces through the notion of a local location model.

(9) McCullagh and Polson (2018) devised an approach to quantifying sparsity using ideas from extreme value theory. The connection to the present discussion is unclear.

A broad goal is a unified theory of inference that is Fisherian when the dimension of the unknown parameter is smaller than the sample size, yet is also operational when the converse is true. A notion of sparsity seems inevitable, and this may take unusual forms as exemplified in the present article.

Acknowledgements

I am grateful to two anonymous referees for their helpful feedback and to the UK Engineering and Physical Sciences Research Council for their support (EP/T01864X/1).

References

1. Barndorff-Nielsen, O. E. (1978). *Information and Exponential Families in Statistical Theory* (2013 reprint of the 1978 edition). John Wiley & Sons, Chichester.
2. Barndorff-Nielsen, O. E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, 70, 343–365.
3. Barndorff-Nielsen, O. E. and Cox, D. R. (1994). *Inference and Asymptotics*. Chapman and Hall, London.
4. Bartlett, M. S. (1936). Statistical information and properties of sufficiency. *Proc. Roy. Soc. A*, 154, 124–137.

5. Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proc. Roy. Soc. A*, 160, 268–282.
6. Battey, H. S. (2017). Eigen structure of a new class of structured covariance and inverse covariance matrices. *Bernoulli*, 23, 3166–3177.
7. Battey, H. S. (2019). On sparsity scales and covariance matrix transformations. *Biometrika*, 106, 605–617.
8. Battey, H. S. and Reid, N. (2022). On inference in high-dimensional regression. *Preprint*.
9. Battey, H. S., Cox, D. R. and Lee, S. H.(2022). On partial likelihood and the construction of factorisable transformations. *Information Geometry*, to appear.
10. Box, G. E. and Cox, D. R. (1964). An analysis of transformations (with discussion). *J. Roy. Statist. Soc. B*, 26, 211–252.
11. Cox, D. R. (1958). Two further applications of a model for binary regression. *Biometrika*, 45, 562–565.
12. Cox, D. R. (1968). Notes on some aspects of regression analysis (with discussion). *J. Roy. Statist. Soc. A*, 131, 265–279.
13. Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.
14. Cox, D. R. and Snell, E. J. (1974). The choice of variables in observational studies. *J. Roy. Statist. Soc. C*, 23, 51–59.
15. Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. Roy. Statist. Soc. B*, 49, 1–39.
16. Cox, D. R. and Snell, E. J. (1989). *Analysis of Binary Data*, 2nd Edition. Chapman and Hall, London.
17. Cox, D. R. (1995). Discussion of a paper by Chatfield. *J. Roy. Statist. Soc. A*, 158, 455–456.
18. Cox, D. R. and Battey, H. S. (2017). Large numbers of explanatory variables, a semi-descriptive analysis. *Proc. Natl. Acad. Aci., USA*, 114, 8592–8595.
19. Fang, X. E., Ning, Y. and Liu, H. (2017). Testing and confidence intervals for high dimensional proportional hazards models. *J. Roy. Statist. Soc. B*, 79, 1415–1437.
20. Fraser, D. A. S. (1964). Local conditional sufficiency. *J. Roy. Statist. Soc. Ser. B*, 26, 52–62.
21. Fraser, D. A. S. (1968). *The Structure of Inference*. Wiley, New York.
22. Fraser, D. A. S., Reid, N. and Lin, W. (2018). When should modes of inference disagree? Some simple but challenging examples. *Ann. Appl. Statist.*, 12, 750–770.
23. Huzurbazar, V. S. (1956). Sufficient Statistics and Orthogonal Parameters. *Sankhya*, 17, 217–220.
24. Lindsay, B. G. (1980). Nuisance parameters, mixture models and the efficiency of partial likelihood estimators. *Phil. Trans. Roy. Soc. London*, 196, 639–665.
25. McCullagh, P. and Polson, N. G. (2018). Statistical sparsity. *Biometrika*, 105, 797–814.
26. 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.
27. Rybak, J. and Battey, H. S. (2021). Sparsity induced by covariance transformation: some deterministic and probabilistic results. *Proc. Roy. Soc. London, A.*, 477, 20200756.
28. 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.
29. Wainwright, M. (2019). *High-dimensional statistics: a non-asymptotic viewpoint*. Cambridge University Press.
30. Zhang, C-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Statist. Soc. B*, 76, 217–242.

Appendix A. Orthogonality of the mixed parameterization

Write the log-likelihood function of a full exponential family with canonical parameters ψ and λ, and canonical statistics S and T as

$$\ell(\psi, \lambda) = s\psi + t\lambda - K(\psi, \lambda).$$

Let

$$\mu_T = \mathbb{E}(T) = K_\lambda(\psi, \lambda) = \frac{\partial}{\partial \lambda} K(\psi, \lambda),$$

and

$$\mu_S = \mathbb{E}(S) = K_\psi(\psi, \lambda) = \frac{\partial}{\partial \psi} K(\psi, \lambda).$$

Starting with the canonical parameterization $\theta = (\psi, \lambda)$, transform to the interest-respecting mixed parameterization $\phi = (\psi, \mu_T)$. The information matrix in the canonical parameterization,

$$i^\theta(\theta) = i^\theta(\psi, \lambda) = \begin{pmatrix} K_{\psi\psi} & K_{\psi\lambda} \\ K_{\lambda\psi} & K_{\lambda\lambda} \end{pmatrix},$$

transforms as

$$i^\phi_{uv}\{\psi, \phi(\theta)\} = \sum_{w,x} \frac{\partial \theta_w}{\partial \phi_u} i^\theta_{wx}(\theta) \frac{\partial \theta_x}{\partial \phi_v}, \tag{A.1}$$

where

$$\left(\frac{\partial \theta_w}{\partial \phi_u}\right) = \left(\frac{\partial \phi_u}{\partial \theta_w}\right)^{-1} = \begin{pmatrix} \partial\psi/\partial\psi & \partial\psi/\partial\lambda \\ \partial\mu_T/\partial\psi & \partial\mu_T/\partial\lambda \end{pmatrix}^{-1}$$

is the matrix with arbitrary entry $\partial\theta_w/\partial\phi_u$. If ψ and λ are one-dimensional, then

$$\left(\frac{\partial \theta_w}{\partial \phi_u}\right) = \begin{pmatrix} 1 & 0 \\ K_{\lambda\psi} & K_{\lambda\lambda} \end{pmatrix}^{-1} = K_{\lambda\lambda}^{-1} \begin{pmatrix} K_{\lambda\lambda} & 0 \\ -K_{\lambda\psi} & 1 \end{pmatrix}$$

and Equation (A.1) gives

$$i^\phi\{\psi, \phi(\theta)\} = \begin{pmatrix} K_{\psi\psi} - K_{\psi\lambda}K_{\lambda\lambda}^{-1}K_{\lambda\psi} & 0 \\ 0 & K_{\lambda\lambda}^{-1} \end{pmatrix}.$$

This holds in arbitrary dimensions, as can be shown using the expression for the inverse of a block-partitioned matrix:

$$\left(\frac{\partial \theta_w}{\partial \phi_u}\right) = \begin{pmatrix} I & 0 \\ -K_{\lambda\lambda}^{-1}K_{\lambda\psi} & K_{\lambda\lambda}^{-1} \end{pmatrix}.$$