Submitted to the Annals of Statistics

CURVATURE AND INFERENCE FOR MAXIMUM LIKELIHOOD ESTIMATES

By Bradley Efron

Stanford University

Maximum likelihood estimates are sufficient statistics in exponential families, but not in general. The theory of statistical curvature was introduced to measure the effects of MLE insufficiency in one-parameter families. Here we analyze curvature in the more realistic venue of multiparameter families—more exactly, *curved exponential families*, a broad class of smoothly defined non-exponential family models. We show that within the set of observations giving the same value for the MLE, there is a "region of stability" outside of which the MLE is no longer even a local maximum. Accuracy of the MLE is affected by the location of the observation vector within the region of stability. Our motivating example involves "g-modeling", an empirical Bayes estimation procedure.

1. Introduction. The modern theory of maximum likelihood estimation (MLE) evolved in three increasingly nuanced papers by R.A. Fisher. His 1922 paper considered MLEs to be sufficient statistics in smoothly defined probability models. This was amended in 1925 to sufficiency holding within what are now called exponential families and, moreover, to the MLE being more efficient than competitors such as minimum chi-squared even in non-exponential families. The final step, in 1934, was the most subtle: in non-exponential families, two data sets giving the same value of the MLE may nevertheless differ greatly in their estimation accuracy, making conditional estimates of accuracy necessary.

There are two distinct themes here. The first, and the most studied in subsequent work, concerns the efficiency of maximum likelihood estimation. Fisher claimed that even when not a sufficient statistic, the MLE lost less information than its competitors. Rao (1961, 1962, 1963) developed the theory of "second order efficiency" to verify Fisher's claim. The concept of "statistical curvature" was introduced in Efron (1975), justifying the Fisher–Rao theory in geometric terms.

Almost all of this work concerned one-parameter families. Amari's seminal 1982 paper used the full power of differential geometry to extend statistical curvature to multiparameter families. A multivariate version of Fisher–Rao theory was developed. Amari's paper launched the thriving field of *information geometry*: "GSI2017", the 3rd conference on geometric science of information, lists 28 separate topic areas, most of them far outside the confines of statistical inference. For statisticians, Kass and Vos' 2011 book *Geometric Foundations of Asympototic Inference* is a key reference.

This paper concerns Fisher's second main theme: that data sets having the same MLE $\hat{\theta}$ may still vary in how accurately the true value θ is being estimated. To this end, he recommended using the *observed information*, the second derivative of the log likelihood function evaluated at $\hat{\theta}$, rather than its expectation (the more familiar "expected Fisher information") to assess $\hat{\theta}$'s accuracy.

Efron (1978) showed that the relationship between observed and expected information depended directly on the statistical curvature, larger curvatures implying larger differences. Again, this applied only to one-parameter families.

MSC 2010 subject classifications: Primary 62Bxx; secondary 62Hxx

Keywords and phrases: observed information, g-modeling, region of stability, curved exponential families

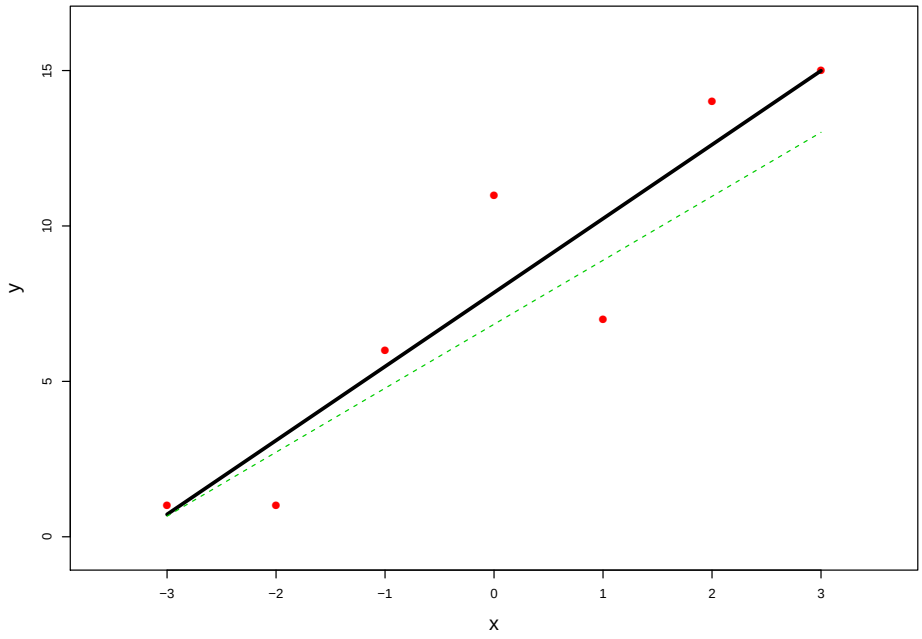

FIG 1. *Toy example of a curved exponential family,* (1.1)–(1.3)*; Poisson observation* y_i *(dots) are assumed to follow linear model* $\mu_i = \alpha_1 + \alpha_2 x_i$ *for* $x_i = -3, -2, \ldots, 3$*; heavy line is MLE fit* $\hat{\alpha} = (7.86, 2.38)$*. Light dashed line is penalized MLE of Section 4.*

The goal of this paper is to examine curvature and information in multiparameter families—more precisely, in multiparameter *curved exponential families*, as defined in Section 2. Figure 1 illustrates a toy example of our situation of interest. Independent Poisson random variables have been observed,

$$y_i \overset{\text{ind}}{\sim} \text{Poi}(\mu_i) \qquad \text{for } i = 1, 2, \ldots, 7. \tag{1.1}$$

We assume that

$$\mu_i = \alpha_1 + \alpha_2 x_i, \qquad x = (-3, -2, -1, 0, 1, 2, 3). \tag{1.2}$$

The vector of observed values y_i was

$$y = (1, 1, 6, 11, 7, 14, 15). \tag{1.3}$$

Direct numerical maximization yielded

$$\hat{\alpha} = (7.86, 2.38) \tag{1.4}$$

as the MLE of $\alpha = (\alpha_1, \alpha_2)$.

If, instead of (1.2), we had specified $\log(\mu_i) = \alpha_1 + \alpha_2 x_i$, a generalized linear model (GLM), the resulting MLE $\hat{\alpha}$ would be a sufficient statistic. Not so for model (1.1)–(1.2). In Section 3 we will see that the set of observation vectors y giving MLE $\hat{\alpha} = (7.86, 2.38)$ lies in a 5-dimensional linear subspace, passing through the point $\hat{\mu} = (\cdots \hat{\alpha}_1 + \hat{\alpha}_2 x_i \cdots)$; and that the *observed Fisher information matrix* $-\ddot{l}_{\hat{\alpha}}(y)$ (minus the second derivative matrix of the log likelihood function with respect to α) varies in a simple but impactful way as y ranges across its subspace.

The motivating example for this paper concerns empirical Bayes estimation: "g-modeling" (Efron, 2016) proposes GLM models for unseen parameters $\hat{\theta}_i$, which then yield observations X_i, say by normal, Poisson, or binomial sampling, in which case the X_i follow multiparameter curved exponential families. Our paper's second goal is to assess the stability of the ensuing maximum likelihood estimates, in the sense of Fisher's arguments.

The paper proceeds as follows. Section 2 reviews one-parameter curved exponential families. Section 3 extends the theory to multiparameter families. Some regularization may be called for in the multiparameter case, modifying our results as described in Section 4. Section 5 presents the analysis of a multiparameter g-modeling example. Some proofs and remarks are deferred to Section 6.

Our results here are obtained by considering all one-parameter subfamilies of the original multiparameter family. By contrast, Amari's 1982 work and that of Madsen (1979) use full multiparametric differential geometry to attack the problem of MLE efficiency (carried on in the information geometry literature, for example in Hayashi and Watanabe, 2016). This paper does not concern the accuracy of the MLE compared to competitors, but, rather, changes in its own accuracy as the observed data varies within the space of constant $\hat{\theta}$–what might be called conditional rather than marginal accuracy. The two concerns are, in a technical sense mentioned at the end of Section 3, orthogonal to each other.

2. One-parameter curved families. After introducing some basic notation and results, this section reviews the curvature theory for one-parameter families. We begin with a full n-parameter exponential family

$$g_{\boldsymbol{\eta}}(\boldsymbol{y}) = e^{\boldsymbol{\eta}'\boldsymbol{y}-\psi(\boldsymbol{\eta})}g_0(\boldsymbol{y}), \tag{2.1}$$

$\boldsymbol{\eta}$ and $\boldsymbol{y}$ n-vectors; $\boldsymbol{\eta}$ is the natural or canonical parameter vector, taking values in a convex set Ω, and $\boldsymbol{y}$ the sufficient data vector; $\psi(\boldsymbol{\eta})$ is the normalizing value that makes $g_{\boldsymbol{\eta}}(\boldsymbol{y})$ integrate to one with respect to the carrier $g_0(\boldsymbol{y})$. The mean vector and covariance matrix of $\boldsymbol{y}$ given $\boldsymbol{\eta}$ are

$$\boldsymbol{\mu}_{\boldsymbol{\eta}} = E_{\boldsymbol{\eta}}\{y\} \quad \text{and} \quad \boldsymbol{V}_{\boldsymbol{\eta}} = \mathrm{cov}_{\boldsymbol{\eta}}\{y\}, \tag{2.2}$$

which can be obtained by differentiating $\psi(\boldsymbol{\eta})$: $\boldsymbol{\mu}_{\boldsymbol{\eta}} = (\partial\psi/\partial\eta_i)$, $\boldsymbol{V}_{\boldsymbol{\eta}} = (\partial^2\psi/\partial\eta_i\partial\eta_j)$.

Now suppose $\boldsymbol{\eta}$ is a smoothly defined function of a p-dimensional vector α,

$$\boldsymbol{\eta} = \boldsymbol{\eta}_\alpha, \tag{2.3}$$

and define the p-parameter subfamily of densities for $\boldsymbol{y}$

$$f_\alpha(\boldsymbol{y}) = g_{\boldsymbol{\eta}_\alpha}(\boldsymbol{y}) = e^{\boldsymbol{\eta}_\alpha'\boldsymbol{y}-\psi(\boldsymbol{\eta}_\alpha)}g_0(\boldsymbol{y}). \tag{2.4}$$

For simplified notation we write

$$\boldsymbol{\mu}_\alpha = \mu_{\boldsymbol{\eta}_\alpha} \quad \text{and} \quad \boldsymbol{V}_\alpha = V_{\boldsymbol{\eta}_\alpha}. \tag{2.5}$$

It is assumed that $\boldsymbol{\eta}_\alpha$ stays within the convex set of possible $\boldsymbol{\eta}$ vectors Ω, say $\alpha \in A$. The family

$$\mathcal{F} = \{f_\alpha(\boldsymbol{y}),\ \alpha \in A\} \tag{2.6}$$

is by definition a *p-parameter curved exponential family*. In the GLM situation where $\boldsymbol{\eta}_\alpha = M\alpha$ for some given $n \times p$ structure matrix M, $\mathcal{F}$ is an (uncurved) p-parameter exponential family, not the case for family (1.1)–(1.2).

Let $\dot{\boldsymbol{\eta}}_\alpha$ denote the $n \times p$ derivative matrix of $\boldsymbol{\eta}_\alpha$ with respect to α,

$$\dot{\boldsymbol{\eta}}_\alpha = (\partial \eta_{\alpha_i}/\partial \alpha_j)\,, \tag{2.7}$$

$i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$; and $\ddot{\boldsymbol{\eta}}_\alpha$ the $n \times p \times p$ array of second derivatives,

$$\ddot{\boldsymbol{\eta}}_\alpha = \left(\partial^2 \eta_{\alpha_i}/\partial \alpha_j \partial \alpha_k\right). \tag{2.8}$$

The log likelihood function corresponding to (2.4) is

$$l_\alpha(\boldsymbol{y}) = \log\left[f_\alpha(\boldsymbol{y})\right] = \boldsymbol{\eta}_\alpha' \boldsymbol{y} - \psi(\boldsymbol{\eta}_\alpha). \tag{2.9}$$

Its derivative vector with respect to α (the "score function") is

$$\dot{l}_\alpha(\boldsymbol{y}) = \dot{\boldsymbol{\eta}}_\alpha'(\boldsymbol{y} - \boldsymbol{\mu}_\alpha). \tag{2.10}$$

The MLE equations $\dot{l}_{\hat\alpha}(\boldsymbol{y}) = \boldsymbol{0}$ for curved exponential families reduce to

$$\dot{\boldsymbol{\eta}}_{\hat\alpha}'(\boldsymbol{y} - \boldsymbol{\mu}_{\hat\alpha}) = \boldsymbol{0}, \tag{2.11}$$

$\boldsymbol{0}$ here indicating a p-vector of zeroes. For convenient discussion, solutions $\hat\alpha$ to (2.11) will sometimes be referred to as "the MLE", even though they may not be global maximums.

From (2.10) we see that the *Fisher information matrix* $\boldsymbol{\mathcal{I}}_\alpha$ for α is

$$\boldsymbol{\mathcal{I}}_\alpha = \dot{\boldsymbol{\eta}}_\alpha' \boldsymbol{V}_\alpha \dot{\boldsymbol{\eta}}_\alpha. \tag{2.12}$$

We will be particularly interested in the second derivative matrix of the log likelihood: $\ddot{l}_\alpha(\boldsymbol{y}) = (\partial^2 l_\alpha(\boldsymbol{y})/\partial\alpha_j\partial\alpha_k)$. Some calculation — see Remark A in Section 6 — gives the important result

$$-\ddot{l}_\alpha(\boldsymbol{y}) = \boldsymbol{\mathcal{I}}_\alpha - \ddot{\boldsymbol{\eta}}_\alpha'(\boldsymbol{y} - \boldsymbol{\mu}_\alpha), \tag{2.13}$$

where $\ddot{\boldsymbol{\eta}}_\alpha'(\boldsymbol{y} - \boldsymbol{\mu}_\alpha)$ is the $p \times p$ matrix having jkth element

$$\sum_{i=1}^n \frac{\partial^2 \eta_{\alpha i}}{\partial \alpha_j \partial \alpha_k}(y_i - \mu_{\alpha i}). \tag{2.14}$$

The *observed Fisher information matrix* $\hat{\boldsymbol{I}}(\boldsymbol{y})$ is defined to be $-\ddot{l}_\alpha(\boldsymbol{y})$ evaluated at $\alpha = \hat\alpha$,

$$\hat{\boldsymbol{I}}(\boldsymbol{y}) = \boldsymbol{\mathcal{I}}_{\hat\alpha} - \ddot{\boldsymbol{\eta}}_{\hat\alpha}'(\boldsymbol{y} - \boldsymbol{\mu}_{\hat\alpha}). \tag{2.15}$$

In the one-dimensional case, $p = 1$, $\dot{\boldsymbol{\eta}}_\alpha$ and $\ddot{\boldsymbol{\eta}}_\alpha$ are each vectors of length n. Figure 2 illustrates the geometry of maximum likelihood estimation: $\mathcal{F}_{\boldsymbol{\mu}}$ is the one-dimensional curve of possible expectations $\boldsymbol{\mu}_\alpha$ in R^n,

$$\mathcal{F}_{\boldsymbol{\mu}} = \{\boldsymbol{\mu}_\alpha = E_\alpha\{\boldsymbol{y}\}, \alpha \in A\}\,. \tag{2.16}$$

The set of observation vectors $\boldsymbol{y}$ that yield MLE $\hat\alpha$ (2.11) lies in the $(n-1)$-dimensional hyperplane passing through $\mu_{\hat\alpha}$ orthogonally to $\dot{\boldsymbol{\eta}}_{\hat\alpha}$,

$$\overset{\perp}{\mathcal{L}}(\dot{\boldsymbol{\eta}}_{\hat\alpha}) = \{\boldsymbol{y} : \dot{\boldsymbol{\eta}}_{\hat\alpha}'(\boldsymbol{y} - \boldsymbol{\mu}_{\hat\alpha}) = 0\}\,, \tag{2.17}$$

denoted more simply as $\overset{\perp}{\mathcal{L}}_{\hat\alpha}$.

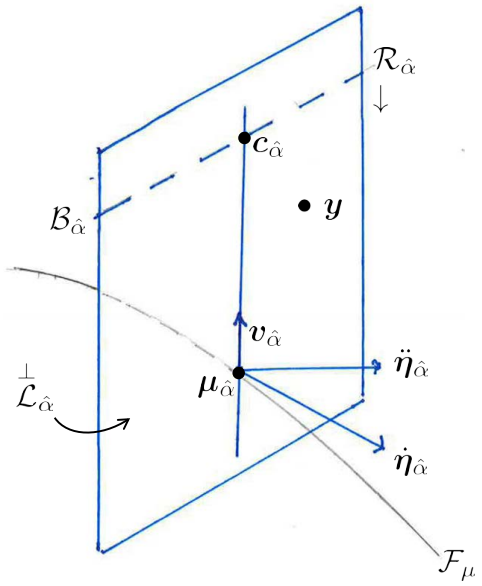

FIG 2. *Maximum likelihood estimation in a one-parameter curved exponential family* (2.16). *Observed vector* $\boldsymbol{y}$ *in* $\overset{\perp}{\mathcal{L}}_{\hat\alpha}$ (2.17) *gives MLE* $\hat\alpha$*; the observed Fisher information* $\hat{\boldsymbol{I}}(\boldsymbol{y})$ (2.15) *varies linearly in* $\overset{\perp}{\mathcal{L}}_{\hat\alpha}$*, equaling zero along the critical boundary* $\mathcal{B}_{\hat\alpha}$ (2.26). *Closest point* $\boldsymbol{c}_{\hat\alpha}$ *is Mahalanobis distance* $1/\gamma_{\hat\alpha}$ *from* $\boldsymbol{\mu}_{\hat\alpha}$*. Region of stability* $\mathcal{R}_{\hat\alpha}$ (2.27) *is the set of* $\boldsymbol{y}$ *values below* $\mathcal{B}_{\hat\alpha}$*;* $\hat\alpha$ *is a local maximum of the likelihood only for* $\boldsymbol{y} \in \mathcal{R}_{\hat\alpha}$.

We see from (2.15) that the observed Fisher information $\hat{\boldsymbol{I}}(\boldsymbol{y})$ is a linear function of $\boldsymbol{y}$, equaling $\boldsymbol{\mathcal{I}}_{\hat\alpha}$ at $\boldsymbol{y} = \boldsymbol{\mu}_{\hat\alpha}$. A quantitative description of this function is provided by the curvature theory of Efron (1975, 1978). Let

$$\nu_{11} = \dot{\boldsymbol{\eta}}_{\hat\alpha}' \boldsymbol{V}_{\hat\alpha} \dot{\boldsymbol{\eta}}_{\hat\alpha}, \quad \nu_{12} = \dot{\boldsymbol{\eta}}_{\hat\alpha}' \boldsymbol{V}_{\hat\alpha} \ddot{\boldsymbol{\eta}}_{\hat\alpha} \quad \text{and} \quad \nu_{22} = \ddot{\boldsymbol{\eta}}_{\hat\alpha}' \boldsymbol{V}_{\hat\alpha} \ddot{\boldsymbol{\eta}}_{\hat\alpha}. \tag{2.18}$$

(Notice that $\nu_{11} = \boldsymbol{\mathcal{I}}_{\hat\alpha}$ (2.12).) The *statistical curvature* of $\mathcal{F}$ at $\alpha = \hat\alpha$ is then defined to be

$$\gamma_{\hat\alpha} = (\nu_{22} - \nu_{12}^2/\nu_{11})^{1/2}/\nu_{11}. \tag{2.19}$$

The residual of $\ddot{\boldsymbol{\eta}}_{\hat\alpha}$ after linear regression on $\dot{\boldsymbol{\eta}}_{\hat\alpha}$, working in the $\boldsymbol{V}_{\hat\alpha}$ inner product,

$$\langle \boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \rangle = \boldsymbol{\eta}_1' \boldsymbol{V}_{\hat\alpha} \boldsymbol{\eta}_2, \tag{2.20}$$

is

$$\overset{\perp}{\boldsymbol{\eta}}_{\hat\alpha} = \ddot{\boldsymbol{\eta}}_{\hat\alpha} - \frac{\nu_{12}}{\nu_{11}} \dot{\boldsymbol{\eta}}_{\hat\alpha}. \tag{2.21}$$

The direction vector $\boldsymbol{v}_{\hat\alpha}$ in Figure 2 is

$$\boldsymbol{v}_{\hat\alpha} = \boldsymbol{V}_{\hat\alpha} \overset{\perp}{\boldsymbol{\eta}}_{\hat\alpha} / (\mathcal{I}_{\hat\alpha} \gamma_{\hat\alpha}); \tag{2.22}$$

see Remark B of Section 6. It satisfies

$$\dot{\boldsymbol{\eta}}_{\hat\alpha}' \boldsymbol{v}_{\hat\alpha} = 0 \quad \text{and} \quad \boldsymbol{v}_{\hat\alpha}' \boldsymbol{V}_{\hat\alpha}^{-1} \boldsymbol{v}_{\hat\alpha} = 1; \tag{2.23}$$

that is, $\boldsymbol{\mu}_{\hat\alpha} + \boldsymbol{v}_{\hat\alpha}$ lies in $\overset{\perp}{\mathcal{L}}_{\hat\alpha}$, at Mahalanobis distance 1 from $\boldsymbol{\mu}_{\hat\alpha}$.

THEOREM 1. *Let $\boldsymbol{r}$ be any vector such that $\dot{\boldsymbol{\eta}}'_{\hat\alpha}\boldsymbol{r} = \ddot{\boldsymbol{\eta}}'_{\hat\alpha}\boldsymbol{r} = 0$. Then the observed Fisher information at $\boldsymbol{y} = \boldsymbol{\mu}_{\hat\alpha} + b\boldsymbol{v}_{\hat\alpha} + \boldsymbol{r}$ is*

$$\hat{I}(\boldsymbol{\mu}_{\hat\alpha} + b\boldsymbol{v}_{\hat\alpha} + \boldsymbol{r}) = \mathcal{I}_{\hat\alpha}(1 - b\gamma_{\hat\alpha}) \tag{2.24}$$

(dropping the boldface notation for $\hat{I}$ and $\mathcal{I}_{\hat\alpha}$ in the one-parameter case).

Theorem 1 is a slight extension of Theorem 2 in Efron (1978), and a special case of multiparametric result (3.23) in the next section.

Define the *critical point* $\boldsymbol{c}_{\hat\alpha}$,

$$\boldsymbol{c}_{\hat\alpha} = \boldsymbol{\mu}_{\hat\alpha} + \boldsymbol{v}_{\hat\alpha}/\gamma_{\hat\alpha}, \tag{2.25}$$

and the *critical boundary*

$$\mathcal{B}_{\hat\alpha} = \{\boldsymbol{y} = \boldsymbol{c}_{\hat\alpha} + \boldsymbol{r},\, \dot{\boldsymbol{\eta}}'_{\hat\alpha}\boldsymbol{r} = \ddot{\boldsymbol{\eta}}'_{\hat\alpha}\boldsymbol{r} = 0\} \tag{2.26}$$

(indicated by the dashed line in Figure 2). Above $\mathcal{B}_{\hat\alpha}$, $\hat\alpha$ *is a local minimum of the likelihood* rather than a local maximum. We define the region below $\mathcal{B}_{\hat\alpha}$,

$$\mathcal{R}_{\hat\alpha} = \{\boldsymbol{y} = \boldsymbol{\mu}_{\hat\alpha} + b\boldsymbol{v} + \boldsymbol{r},\, b < 1/\gamma_{\hat\alpha}\} \tag{2.27}$$

as the *region of stability*. Only for $\boldsymbol{y}$ in $\mathcal{R}_{\hat\alpha}$ does the local stability equation $\dot{\boldsymbol{\eta}}'_{\hat\alpha}(\boldsymbol{y} - \boldsymbol{\mu}_{\hat\alpha}) = 0$ yield a local maximum.

Some comments on Theorem 1:

- If $\mathcal{F}$ is a genuine (uncurved) exponential family then $\gamma_{\hat\alpha}$ is zero, in which case $\boldsymbol{c}_{\hat\alpha}$ is infinitely far from $\boldsymbol{\mu}_{\hat\alpha}$ and $\mathcal{R}_{\hat\alpha}$ is all of $\overset{\perp}{\mathcal{L}}_{\hat\alpha}$. Otherwise $\gamma_{\hat\alpha}$ is > 0, larger values moving $\mathcal{B}_{\hat\alpha}$ closer to $\boldsymbol{\mu}_{\hat\alpha}$.
- The point $\boldsymbol{y} = \boldsymbol{\mu}_{\hat\alpha} + b\boldsymbol{v}_{\hat\alpha}$ is Mahalanobis distance

  $$\left[(\boldsymbol{y} - \boldsymbol{\mu}_{\hat\alpha})'\boldsymbol{V}_{\hat\alpha}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_{\hat\alpha})\right]^{1/2} = b \tag{2.28}$$

  from $\boldsymbol{\mu}_{\hat\alpha}$, which is the minimum distance for points $\boldsymbol{\mu}_{\hat\alpha} + b\boldsymbol{v}_{\hat\alpha} + \boldsymbol{r}$.
- The usual estimate for the standard deviation of $\hat\alpha$ is

  $$\mathrm{sd}(\hat\alpha) \doteq 1/\mathcal{I}_{\hat\alpha}^{1/2}. \tag{2.29}$$

  Fisher suggested instead using the observed information,

  $$\mathrm{sd}(\hat\alpha) \doteq 1/\hat{I}^{1/2}; \tag{2.30}$$

  see Efron and Hinkley (1978) for some justification; (2.30) is smaller than (2.29) for b negative in (2.24), and larger for b positive, approaching infinity — total instability — as b approaches $1/\gamma_{\hat\alpha}$.
- Asymptotically, as $\mathcal{I}_{\hat\alpha}$ goes to infinity,

  $$\hat{I}(\boldsymbol{y})/\mathcal{I}_{\hat\alpha} \longrightarrow \mathcal{N}(1, \gamma_\alpha^2); \tag{2.31}$$

  see Remark 2 of Efron (1978, Sect. 5). A large value of the curvature implies possibly large differences between $\hat{I}(\boldsymbol{y})$ and $\mathcal{I}_{\hat\alpha}$.

- A large value of $\gamma_{\hat\alpha}$ is worrisome from a frequentist point of view even if $\boldsymbol{y}$ is in $\mathcal{R}_{\hat\alpha}$. It suggests a substantial probability of global instability, with observations on the far side of $\mathcal{B}_{\hat\alpha}$ producing wild MLE values, undermining (2.29).
- For preobservational planning, before $\boldsymbol{y}$ is seen, large values of γ_α over a relevant range of α warn of eccentric behavior of the MLE. *Penalized* maximum likelihood estimation, Sections 4 and 5, can provide substantially improved performance.
- Theorem 1 involves three different inner products $\boldsymbol{a}'\boldsymbol{D}\boldsymbol{b}$: $\boldsymbol{D}$ equal $\boldsymbol{V}_{\hat\alpha}$, $\boldsymbol{V}_{\hat\alpha}^{-1}$, and the identity. As discussed in the next section, we can always transform $\mathcal{F}$ to make $\boldsymbol{V}_{\hat\alpha}$ the identity, simplifying both the derivations and their interpretation. Figure 2 assumes this to be the case, with $\boldsymbol{v}_{\hat\alpha}$ lying along $\overset{\perp}{\dot{\boldsymbol{\eta}}}_{\hat\alpha}$ (2.21), and $\mathcal{B}_{\hat\alpha}$ orthogonal to $\boldsymbol{v}_{\hat\alpha}$.

3. Regions of stability for multiparameter families. Figure 2 pictures the region of stability $\mathcal{R}_{\hat\alpha}$ for the MLE $\hat\alpha$ in a one-parameter curved exponential family (2.28) as a half-space of the hyperplane $\overset{\perp}{\dot{\mathcal{L}}}(\dot{\boldsymbol{\eta}}_\alpha)$ (2.17). Here we return to multiparameter curved families $\mathcal{F} = \{f_\alpha(\boldsymbol{y}), \alpha \in A\}$ (2.6) where α has dimension $p > 1$. Now the region of stability $\mathcal{R}_{\hat\alpha}$, naturally defined, will turn out to be a convex subset of $\overset{\perp}{\dot{\mathcal{L}}}(\dot{\boldsymbol{\eta}}_\alpha)$, though not usually a half-space.

The definition and computation of $\mathcal{R}_{\hat\alpha}$ is the subject of this section. All of this is simplified by transforming coordinates in the full family $g_{\boldsymbol{\eta}}(\boldsymbol{y})$ (2.1). Let

$$\boldsymbol{M} = \boldsymbol{V}_{\hat\alpha}^{1/2}, \tag{3.1}$$

a symmetric square root of the $n \times n$ covariance matrix $\boldsymbol{V}_{\hat\alpha}$ at $\alpha = \hat\alpha$ (2.5), assumed to be of full rank, and define

$$\boldsymbol{\eta}^\dagger = \boldsymbol{M}\boldsymbol{\eta} \quad \text{and} \quad \boldsymbol{y}^\dagger = \boldsymbol{M}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_{\hat\alpha}). \tag{3.2}$$

With $\hat\alpha$ considered fixed at its observed value, $g_{\boldsymbol{\eta}}(\boldsymbol{y})$ transforms into the n-parameter exponential family

$$g^\dagger_{\boldsymbol{\eta}^\dagger}(\boldsymbol{y}^\dagger) = e^{\boldsymbol{\eta}^{\dagger\prime}\boldsymbol{y}^\dagger - \psi^\dagger(\boldsymbol{\eta}^\dagger)}\left[e^{\boldsymbol{\eta}^{\dagger\prime}\boldsymbol{M}^{-1}\boldsymbol{\mu}_{\hat\alpha}} g_0^\dagger(\boldsymbol{y}^\dagger)\right] \tag{3.3}$$

where $\psi^\dagger(\boldsymbol{\eta}^\dagger) = \psi(\boldsymbol{M}^{-1}\boldsymbol{\eta}^\dagger)$.

The curved family $\mathcal{F}$ (2.6) can just as well be defined by

$$\boldsymbol{\eta}^\dagger_\alpha = \boldsymbol{M}\boldsymbol{\eta}_\alpha, \tag{3.4}$$

with the advantage that $\boldsymbol{y}$ has mean vector $\boldsymbol{0}$ and covariance matrix the identity $\boldsymbol{I}_n$ at $\alpha = \hat\alpha$. In what follows it will be assumed that the mean vector and covariance matrix are

$$\boldsymbol{\mu}_{\hat\alpha} = \boldsymbol{0} \quad \text{and} \quad \boldsymbol{V}_{\hat\alpha} = \boldsymbol{I}_n; \tag{3.5}$$

that is, that $(\boldsymbol{\eta}, \boldsymbol{y})$ have already been transformed into the convenient form (3.5).

We will employ one-parameter subfamilies of $\mathcal{F}$ to calculate the p-parameter stable region $\mathcal{R}_{\hat\alpha}$. Let u be a p-dimensional unit vector. It determines the one-parameter subfamily

$$\mathcal{F}_u = \{f_{\hat\alpha + u\lambda}, \lambda \in \Lambda\}, \tag{3.6}$$

where Λ is an interval of the real line containing 0 as an interior point.

Looking at (2.4), $\mathcal{F}_u$ is a one-parameter curved exponential family having natural parameter vector

$$\eta_\lambda = \boldsymbol{\eta}_{\hat{\alpha}+u\lambda}, \tag{3.7}$$

with MLE $\lambda = 0$. We will denote $\eta_{\lambda=0}$ by η_u in what follows, and likewise $\dot{\eta}_u$ and $\ddot{\eta}_u$ for the derivatives of η_λ at $\lambda = 0$, thus

$$\dot{\eta}_u = \dot{\boldsymbol{\eta}}_{\hat{\alpha}} u, \tag{3.8}$$

(2.7), and similarly

$$\ddot{\eta}_u = u' \ddot{\boldsymbol{\eta}}_{\hat{\alpha}} u, \tag{3.9}$$

$\ddot{\eta}_u$ having ith component $\sum_j \sum_k \ddot{\boldsymbol{\eta}}_{\hat{\alpha} ijk} u_j u_k$. Using (3.8), the Fisher information $\mathcal{I}_u$, at $\lambda = 0$ in $\mathcal{F}_u$, is

$$\mathcal{I}_u = u' \dot{\boldsymbol{\eta}}'_{\hat{\alpha}} \dot{\boldsymbol{\eta}}_{\hat{\alpha}} u = u' \boldsymbol{\mathcal{I}}_{\hat{\alpha}} u \tag{3.10}$$

(from (2.12), remembering that $\boldsymbol{V}_{\hat{\alpha}} = \boldsymbol{I}_n$).

There is a similar expression for the observed Fisher information $\hat{I}_u(\boldsymbol{y})$ in $\mathcal{F}_u$:

Lemma 1.

$$\hat{I}_u(\boldsymbol{y}) = u' \hat{\boldsymbol{I}}(\boldsymbol{y}) u. \tag{3.11}$$

Proof. Applying (2.15) to $\mathcal{F}_u$,

$$\begin{aligned} \hat{I}_u(\boldsymbol{y}) &= \mathcal{I}_u - \ddot{\eta}'_u \boldsymbol{y} = u' \boldsymbol{\mathcal{I}}_{\hat{\alpha}} u - (u' \ddot{\boldsymbol{\eta}}_{\hat{\alpha}} u)' \boldsymbol{y} \\ &= u' (\boldsymbol{\mathcal{I}}_\alpha - \ddot{\boldsymbol{\eta}}'_{\hat{\alpha}} \boldsymbol{y}) u = u' \hat{\boldsymbol{I}}(\boldsymbol{y}) u, \end{aligned} \tag{3.12}$$

the bottom line again invoking linearity. □

We can apply the one-parameter curvature theory of Section 2 to $\mathcal{F}_u$: with $\boldsymbol{V}_{\hat{\alpha}} = \boldsymbol{I}_n$ in (2.18),

$$\nu_{u11} = \dot{\eta}'_u \dot{\eta}_u, \quad \nu_{u12} = \dot{\eta}'_u \ddot{\eta}_u, \quad \text{and} \quad \nu_{u22} = \ddot{\eta}'_u \ddot{\eta}_u, \tag{3.13}$$

$\nu_{u11} = \mathcal{I}_u$ (3.10), giving

$$\overset{\perp}{\eta}_u = \ddot{\eta}_u - \frac{\nu_{u12}}{\nu_{u11}} \dot{\eta}_u \tag{3.14}$$

as at (2.21), and curvature (2.19)

$$\gamma_u = (\nu_{u22} - \nu_{u12}^2/\nu_{u11})^{1/2} / \nu_{u11}. \tag{3.15}$$

The direction vector $\boldsymbol{v}_{\hat{\alpha}}$ (2.22) in Figure 2 is

$$\boldsymbol{v}_u = \overset{\perp}{\eta}_u / (\mathcal{I}_u \gamma_u). \tag{3.16}$$

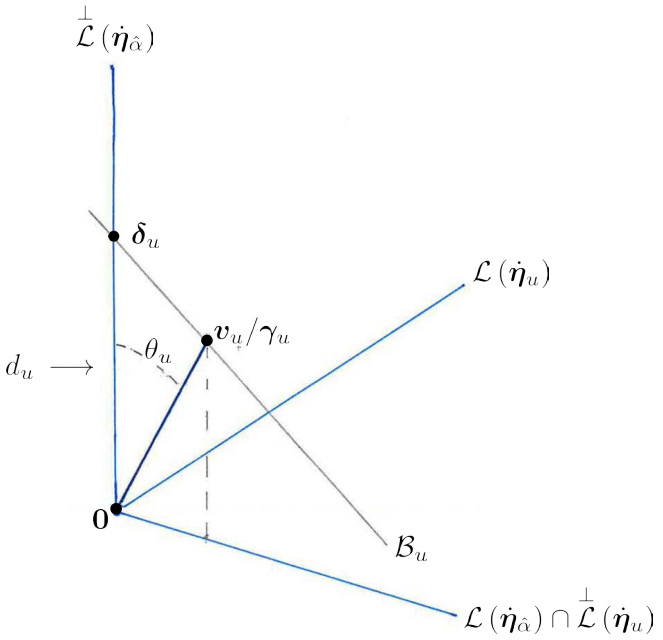

FIG 3. *Curvature effects in multiparametric curved exponential families: p-dimensional unit vector u determines one-parameter curved subfamily $\mathcal{F}_u$ and direction $\dot{\eta}_u$ (3.6)–(3.8) as well as curvature γ_u and critical vector $\boldsymbol{v}_u/\gamma_u$ (3.15)–(3.16), tilted at angle θ_u to the nearest point in $\overset{\perp}{\mathcal{L}}(\dot{\boldsymbol{\eta}}_{\hat{\alpha}})$; critical boundary $\mathcal{B}_u$, corresponding to $\mathcal{B}_{\hat{\alpha}}$ in Figure 2, intersects $\overset{\perp}{\mathcal{L}}(\dot{\boldsymbol{\eta}}_{\hat{\alpha}})$ in hyperplane $\overset{\perp}{\mathcal{B}}_u$ at distance $d_u = \|\boldsymbol{\delta}_u\|$ from $\mathbf{0}$; see Lemma 2. $\mathcal{R}_u$ is the half-space of $\overset{\perp}{\mathcal{L}}(\dot{\boldsymbol{\eta}}_{\hat{\alpha}})$ below $\overset{\perp}{\mathcal{B}}_u$. Intersection of all possible $\mathcal{R}_u$'s gives $\mathcal{R}_{\hat{\alpha}}$, the region of stability.*

It points toward $\boldsymbol{c}_u = \boldsymbol{v}_u/\gamma_u$ (2.25), lying on the boundary $\mathcal{B}_u$ at distance $1/\gamma_u$ from the origin ("distance" now being ordinary Euclidean distance).

Observation vectors $\boldsymbol{y}$ in $\overset{\perp}{\mathcal{L}}(\dot{\boldsymbol{\eta}}_{\hat{\alpha}})$ lying beyond $\mathcal{B}_u$ have $\hat{I}_u(\boldsymbol{y}) < 0$ and so

$$u'\hat{\boldsymbol{I}}(\boldsymbol{y})u < 0 \tag{3.17}$$

according to Lemma 1. Such points will be excluded from our definition of the multiparameter stable region $\mathcal{R}_{\hat{\alpha}}$. It remains to compute what the excluded region of $\overset{\perp}{\mathcal{L}}(\dot{\boldsymbol{\eta}}_{\hat{\alpha}})$ looks like.

Figure 3 illustrates the geometry. Three orthogonal linear subspaces are pictured: $\mathcal{L}(\dot{\eta}_u)$, dimension 1; $\mathcal{L}(\dot{\boldsymbol{\eta}}_{\hat{\alpha}}) \cap \overset{\perp}{\mathcal{L}}(\dot{\eta}_u)$, the (p-1)-dimensional subspace of $\mathcal{L}(\dot{\boldsymbol{\eta}}_\alpha)$ orthogonal to $\mathcal{L}(\dot{\eta}_u)$; and $\overset{\perp}{\mathcal{L}}(\dot{\boldsymbol{\eta}}_{\hat{\alpha}})$, the $(n-p)$-dimensional space of $\boldsymbol{y}$ vectors giving MLE $\hat{\alpha}$. The critical point $\boldsymbol{v}_u/\gamma_u$, corresponding to $\boldsymbol{c}_{\hat{\alpha}}$ in Figure 2, is shown at angle θ_u to $\overset{\perp}{\mathcal{L}}(\dot{\boldsymbol{\eta}}_{\hat{\alpha}})$, this being the angle between $\boldsymbol{v}_u$ and its projection into $\overset{\perp}{\mathcal{L}}(\dot{\boldsymbol{\eta}}_{\hat{\alpha}})$, i.e., the smallest possible angle between $\boldsymbol{v}_u$ and a vector in $\overset{\perp}{\mathcal{L}}(\dot{\boldsymbol{\eta}}_{\hat{\alpha}})$.

LEMMA 2. *$\mathcal{B}_u$ intersects $\overset{\perp}{\mathcal{L}}(\dot{\boldsymbol{\eta}}_{\hat{\alpha}})$ in a $(n-p-1)$-dimensional hyperplane, say $\overset{\perp}{\mathcal{B}}_u$; $\boldsymbol{\delta}_u$, the nearest point to the origin in $\overset{\perp}{\mathcal{B}}_u$, is at distance*

$$d_u = 1/(\gamma_u \cos\theta_u), \tag{3.18}$$

and lies along the projection of $\boldsymbol{v}_u$ into $\overset{\perp}{\mathcal{L}}(\dot{\boldsymbol{\eta}}_{\hat{\alpha}})$.

The geometry of Figure 2 provides heuristic support for Lemma 2. A more analytic justification appears in Remark C of Section 6.

Let $\overset{\perp}{\boldsymbol{P}}$ be the projection matrix into $\overset{\perp}{\mathcal{L}}(\dot{\boldsymbol{\eta}}_{\hat\alpha})$,

$$\overset{\perp}{\boldsymbol{P}} = \boldsymbol{I}_n - \dot{\boldsymbol{\eta}}_{\hat\alpha}\boldsymbol{\mathcal{I}}_{\hat\alpha}^{-1}\dot{\boldsymbol{\eta}}_{\hat\alpha}' \qquad (\boldsymbol{\mathcal{I}}_{\hat\alpha} = \dot{\boldsymbol{\eta}}_{\hat\alpha}'\dot{\boldsymbol{\eta}}_{\hat\alpha}). \tag{3.19}$$

Since $\boldsymbol{v}_u$ is a unit vector we obtain

$$\cos\theta_u = \left(\boldsymbol{v}_u'\overset{\perp}{\boldsymbol{P}}\boldsymbol{v}_u\right)^{1/2} \tag{3.20}$$

for use in (3.18).

A version of Theorem 1 applies here. Let $\boldsymbol{w}_u$ be the unit projection of $\boldsymbol{v}_u$ into $\overset{\perp}{\mathcal{L}}_{\dot{\boldsymbol{\eta}}_{\hat\alpha}}$,

$$\boldsymbol{w}_u = \overset{\perp}{\boldsymbol{P}}\boldsymbol{v}_u/\cos\theta_u. \tag{3.21}$$

THEOREM 2. *Let $\boldsymbol{r}$ be any vector in $\overset{\perp}{\mathcal{L}}(\dot{\boldsymbol{\eta}}_{\hat\alpha})$ orthogonal to $\boldsymbol{w}_u$. Then for*

$$\boldsymbol{y} = b\boldsymbol{w}_u + \boldsymbol{r} \tag{3.22}$$

we have

$$\hat{I}_u(\boldsymbol{y}) = \mathcal{I}_u(1 - b\gamma_u\cos\theta_u), \tag{3.23}$$

which by Lemma 1 implies

$$u'\hat{\boldsymbol{I}}(\boldsymbol{y})u = u'\boldsymbol{\mathcal{I}}_{\hat\alpha}u(1 - b\gamma_u\cos\theta_u). \tag{3.24}$$

Choosing $b = d_u = 1/(\gamma_u\cos\theta_u)$ as in Lemma 2 gives $u'\hat{\boldsymbol{I}}(\boldsymbol{y})u = 0$.

Remark D in Section 6 verifies Theorem 2.

Now let $\mathcal{R}_u$ denote the half-space of $\overset{\perp}{\mathcal{L}}(\dot{\boldsymbol{\eta}}_{\hat\alpha})$ lying below $\overset{\perp}{\mathcal{B}}_u$, that is, containing the origin. We define the region of stability for the multiparameter MLE $\hat\alpha$ to be the intersection of all such regions $\mathcal{R}_u$, a convex set,

$$\mathcal{R}_{\hat\alpha} = \bigcap_{u\in\mathcal{S}^p}\mathcal{R}_u, \tag{3.25}$$

$\mathcal{S}^p$ the unit sphere in p dimensions. The construction is illustrated in Figure 4.

The $p\times p$ information matrix $\boldsymbol{\mathcal{I}}_{\hat\alpha} = \dot{\boldsymbol{\eta}}_{\hat\alpha}'\dot{\boldsymbol{\eta}}_{\hat\alpha}$ is positive definite if $\dot{\boldsymbol{\eta}}_{\hat\alpha}$ is of rank p, now assumed to be the case.

THEOREM 3. *For $\boldsymbol{y}$ in $\overset{\perp}{\mathcal{L}}_{\hat\alpha}$, the observed information matrix $-\ddot{l}_{\hat\alpha}(\boldsymbol{y}) = \hat{\boldsymbol{I}}(\boldsymbol{y})$ (2.15) is positive definite if and only if $\boldsymbol{y}\in\mathcal{R}_{\hat\alpha}$, the region of stability.*

PROOF. If $\boldsymbol{y}$ is not in some $\mathcal{R}_u$ then b in (3.22) must exceed $1/(\gamma_u\cos\theta_u)$, in which case (3.24) implies $u'\hat{\boldsymbol{I}}(\boldsymbol{y})u < 0$. However for y in $\mathcal{R}_{\hat\alpha}$, b must be less than $1/(\gamma_u\cos\theta_u)$ for *all* u, implying

$$u'\hat{\boldsymbol{I}}(\boldsymbol{y})u > 0 \qquad \text{for all } u\in\mathcal{S}^p, \tag{3.26}$$

verifying the positive definiteness of $\hat{\boldsymbol{I}}(\boldsymbol{y})$. □

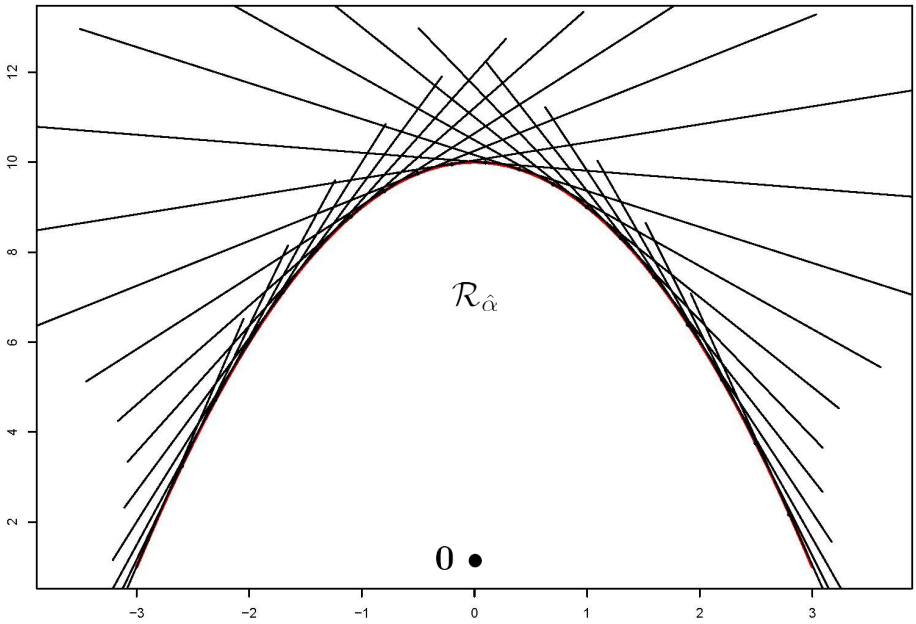

FIG 4. *Schematic illustration showing construction of the region of stability $\mathcal{R}_{\hat\alpha}$. Each dot represents critical point $\boldsymbol{\delta}_u$ for a particular choice of u, with its tangent line representing the boundary $\overset{\perp}{\mathcal{B}}_u$. The intersection of the half-spaces $\mathcal{R}_u$ containing $\mathbf{0}$ determines $\mathcal{R}_{\hat\alpha}$.*

In the toy example of Figure 1, the equivalent of Figure 4 (now in dimension $n - p = 5$) was computed for u in (3.6) equaling

$$u(t) = (\cos t, \sin t), \tag{3.27}$$

$t = k\pi/100$ for $k = 0, 1, \ldots, 100$. See Remark E in Section 6. In this special situation the unit direction vector $\boldsymbol{w}_u$ in (3.22) was always the same,

$$\boldsymbol{w}_u = (-.390, .712, .392, .145, -.049, -.210, -.347), \tag{3.28}$$

though the distance to the boundary d_u (3.18) varied. Its minimum was

$$d_u = 2.85 \text{ at } u = (0, 1). \tag{3.29}$$

The stable region $\mathcal{R}_{\hat\alpha}$ was a half-space of the 5-dimensional hyperplane $\overset{\perp}{\mathcal{L}}(\dot{\boldsymbol{\eta}}_{\hat\alpha})$, the boundary $\mathcal{B}_{\hat\alpha}$ having minimum distance 2.85 from $\mathbf{0}$.

The constancy of $\boldsymbol{w}_u$ (3.29) was perhaps surprising given that $\boldsymbol{v}_u$ (3.16) varied with u. Constancy is not the case in the more elaborate example of Section 5, though even there the $\boldsymbol{w}_u$ vary only slightly–again allowing the region of stability $\mathcal{R}_{\hat\alpha}$ to extend infinitely in certain directions, as suggested by Figure 4. The author has not found an example of bounded $\mathcal{R}_{\hat\alpha}$.

The information matrix in the toy example was

$$\boldsymbol{\mathcal{I}}_{\hat\alpha} = \begin{pmatrix} 2.26 & -4.54 \\ -4.54 & 14.98 \end{pmatrix}. \tag{3.30}$$

The observed information matrix $\hat{\boldsymbol{I}}(b\boldsymbol{w}_u)$ decreases toward singularity as b increases; it becomes singular at $b = 2.85$, at which point its lower right corner equals zero. Further increases of b reduce

other quadratic forms $u(t)'\hat{\boldsymbol{I}}(b\boldsymbol{w}_u)u(t)$ to zero, as in (3.24). Boundary distance 2.85 is small enough to allow substantial differences between $\hat{\boldsymbol{I}}(\boldsymbol{y})$ and $\boldsymbol{\mathcal{I}}_{\hat\alpha}$. For example, in a Monte Carlo simulation of model (1.1)–(1.2), with $\alpha = \hat\alpha$, the ratio of lower right corner elements $\hat{\boldsymbol{I}}_{22}/\boldsymbol{\mathcal{I}}_{\hat\alpha 22}$ averaged near 1.0 (as they should) but with standard deviation 0.98.

Stability theory can be thought of as a complement to more familiar accuracy calculations for the MLE $\hat\alpha$. The latter depend primarily on $\mathcal{L}(\dot{\boldsymbol{\eta}}_{\hat\alpha})$, as seen in the covariance approximation

$$\operatorname{cov}(\hat\alpha) \doteq \boldsymbol{\mathcal{I}}_{\hat\alpha}^{-1} = (\dot{\boldsymbol{\eta}}_{\hat\alpha}'\dot{\boldsymbol{\eta}}_{\hat\alpha})^{-1}, \tag{3.31}$$

while stability refers to behavior in the orthogonal space $\overset{\perp}{\mathcal{L}}(\dot{\boldsymbol{\eta}}_{\hat\alpha})$. The full differential geometric developments in Amari (1982) and Madsen (1979) aim toward second-order accuracy expressions for assessing $\operatorname{cov}(\hat\alpha)$, and in this sense are orthogonal to the considerations here.

4. Penalized maximum likelihood. The performance of maximum likelihood estimation can often be improved by regularization, that is by adding a penalty term to the log likelihood so as to tamp down volatile behavior of $\hat\alpha$. This is the case for empirical Bayes "g-modeling" (Efron, 2016), our motivating example discussed in Section 5.

We define the penalized log likelihood function $m_\alpha(\boldsymbol{y})$ to be

$$m_\alpha(\boldsymbol{y}) = l_\alpha(\boldsymbol{y}) - s_\alpha, \tag{4.1}$$

where $l_\alpha(\boldsymbol{y})$ is the usual log likelihood $\log f_\alpha(\boldsymbol{y})$, and s_α is a nonnegative *penalty function* that penalizes undesirable aspects of α. The idea goes back at least to Good and Gaskins (1971), where s_α penalized roughness in density estimates. Ridge regression (Hoerl and Kennard, 1970) takes $s_\alpha = c\|\alpha\|^2$ in the context of ordinary least squares estimation, while the lasso (Tibshirani, 1996) employs $c\|\alpha\|_1$. Here we will use

$$s_\alpha = c\left[\sum_1^p \alpha_j^2\right]^{1/2} \tag{4.2}$$

for our example, as in Efron (2016), though the development does not depend on this choice.

The penalized maximum likelihood estimate (pMLE) is a solution to the local maximizing equations $\dot m_{\hat\alpha}(\boldsymbol{y}) = 0$,

$$\hat\alpha: \quad \dot m_{\hat\alpha}(\boldsymbol{y}) = \dot{\boldsymbol{\eta}}_{\hat\alpha}'(\boldsymbol{y} - \boldsymbol{\mu}_{\hat\alpha}) - \dot s_{\hat\alpha} = \boldsymbol{0}, \tag{4.3}$$

where $\dot s_{\hat\alpha}$ is the p-dimensional gradient vector $(\partial s_\alpha/\partial\alpha_j)$. For a given value of $\hat\alpha$, the set of observation vectors $\boldsymbol{y}$ satisfying (4.3) is an $(n-p)$-dimensional hyperplane

$$\overset{\perp}{\mathcal{M}}_{\hat\alpha} = \{\boldsymbol{y} : \dot{\boldsymbol{\eta}}_{\hat\alpha}'(\boldsymbol{y} - \boldsymbol{\mu}_{\hat\alpha}) = \dot s_{\hat\alpha}\}; \tag{4.4}$$

$\overset{\perp}{\mathcal{M}}_{\hat\alpha}$ lies parallel to $\overset{\perp}{\mathcal{L}}_{\hat\alpha} = \overset{\perp}{\mathcal{L}}(\boldsymbol{\eta}_{\hat\alpha})$ in Figure 2, but offset from $\boldsymbol{\mu}_{\hat\alpha}$.

The nearest point to $\boldsymbol{\mu}_{\hat\alpha}$ in $\overset{\perp}{\mathcal{M}}_{\hat\alpha}$, say $\boldsymbol{y}_{\hat\alpha}$, is calculated to be

$$\boldsymbol{y}_{\hat\alpha} = \boldsymbol{\mu}_{\hat\alpha} + \dot{\boldsymbol{\eta}}_{\hat\alpha}(\dot{\boldsymbol{\eta}}_{\hat\alpha}'\dot{\boldsymbol{\eta}}_{\hat\alpha})^{-1}\dot s_{\hat\alpha}, \tag{4.5}$$

at squared distance $\|\boldsymbol{y}_{\hat\alpha} - \boldsymbol{\mu}_{\hat\alpha}\|^2 = \dot s_{\hat\alpha}'(\dot{\boldsymbol{\eta}}_{\hat\alpha}'\dot{\boldsymbol{\eta}}_{\hat\alpha})^{-1}\dot s_{\hat\alpha}$. From now on we will revert to the *transformed coordinates* (3.5) having $\boldsymbol{\mu}_{\hat\alpha} = 0$ and $\boldsymbol{V}_{\hat\alpha} = \boldsymbol{I}_n$, for which $\dot{\boldsymbol{\eta}}_{\hat\alpha}'\dot{\boldsymbol{\eta}}_{\hat\alpha}$ equals the Fisher information matrix $\boldsymbol{\mathcal{I}}_{\hat\alpha}$ (2.12), and

$$\boldsymbol{y}_{\hat\alpha} = \dot{\boldsymbol{\eta}}_{\hat\alpha}\boldsymbol{\mathcal{I}}_{\hat\alpha}^{-1}\dot s_{\hat\alpha} \qquad \text{with } \|\boldsymbol{y}_{\hat\alpha}\|^2 = \dot s_{\hat\alpha}'\boldsymbol{\mathcal{I}}_{\hat\alpha}^{-1}\dot s_{\hat\alpha}. \tag{4.6}$$

By analogy with the observed information matrix $\hat{\boldsymbol{I}}(\boldsymbol{y}) = -\ddot{l}_{\hat\alpha}(\boldsymbol{y})$ (2.15), we define

$$\begin{aligned}\hat{\boldsymbol{J}}(\boldsymbol{y}) &= -\ddot{m}_{\hat\alpha}(\boldsymbol{y}) = \hat{\boldsymbol{I}}(\boldsymbol{y}) + \ddot{s}_{\hat\alpha} \\ &= \boldsymbol{\mathcal{I}}_{\hat\alpha} - \ddot{\boldsymbol{\eta}}'_{\hat\alpha}\boldsymbol{y} + \ddot{s}_{\hat\alpha},\end{aligned} \tag{4.7}$$

$\ddot{s}_{\hat\alpha}$ being the $p \times p$ second derivative matrix $(\partial^2 s_\alpha/\partial\alpha_j\partial\alpha_k)$. $\hat{\boldsymbol{J}}(\boldsymbol{y})$ plays a key role in the accuracy and stability of the pMLE. For instance the *influence function* of $\hat\alpha$, the $p \times n$ matrix $(\partial\hat\alpha_j/\partial y_k)$, is

$$\frac{d\hat\alpha}{d\boldsymbol{y}} = \hat{\boldsymbol{J}}(\boldsymbol{y})^{-1}\dot{\boldsymbol{\eta}}'_{\hat\alpha}; \tag{4.8}$$

see Remark F in Section 6.

We can think of s_α in (4.1) as the log of a Bayesian prior density for α, in which case $\exp\{m_\alpha(\boldsymbol{y})\}$ is proportional to the posterior density of α given $\boldsymbol{y}$. Applying Laplace's method, as in Tierney and Kadane (1986), yields the normal approximation

$$\alpha \,\Big|\, \boldsymbol{y} \doteadot \mathcal{N}_p\left(\hat\alpha, \hat{\boldsymbol{J}}(\boldsymbol{y})^{-1}\right). \tag{4.9}$$

This supports the Fisherian covariance approximation $\operatorname{cov}(\hat\alpha) \doteq \hat{\boldsymbol{J}}(\boldsymbol{y})^{-1/2}$, similar to (2.30).

Determination of the region of stability $\mathcal{R}_{\hat\alpha}$ — now defined as those vectors $\boldsymbol{y}$ in $\overset{\perp}{\mathcal{M}}_{\hat\alpha}$ having $\hat{\boldsymbol{J}}(\boldsymbol{y})$ positive definite — proceeds as in Section 3. For a one-parameter subfamily $\mathcal{F}_u$ (3.6), the observed penalized information $\hat J_u(\boldsymbol{y}) = -\partial^2 m(\hat\alpha + u\lambda)/\partial\lambda^2|_0$ obeys the analogue of Lemma 1:

Lemma 3.

$$\hat J_u(\boldsymbol{y}) = u'\hat{\boldsymbol{J}}(\boldsymbol{y})u. \tag{4.10}$$

Proof. Let

$$\dot s_u = u'\dot s_{\hat\alpha} \quad \text{and} \quad \ddot s_u = u'\ddot s_{\hat\alpha}u, \tag{4.11}$$

so $\dot s_u = \partial s(\hat\alpha + u\lambda)/\partial\lambda|_0$ and $\ddot s_u = \partial^2 s(\hat\alpha + u\lambda)/\partial\lambda^2|_0$. Then, applying (3.10), (3.11), (4.11), and (4.7),

$$\begin{aligned}\hat J_u(\boldsymbol{y}) &= \mathcal{I}_u - \ddot{\boldsymbol{\eta}}'_u\boldsymbol{y} + \ddot s_u \\ &= u'\boldsymbol{\mathcal{I}}_{\hat\alpha}u - (u'\ddot{\boldsymbol{\eta}}_{\hat\alpha}u)'\boldsymbol{y} + u'\ddot s_{\hat\alpha}u = u'\hat{\boldsymbol{J}}(\boldsymbol{y})u.\end{aligned} \tag{4.12}$$

□

Most of the definitions in Section 3 remain applicable as stated: $\dot\eta_u$ (3.8), ν_{u11}, etc. (3.13), $\overset{\perp}{\eta}_u$ (3.14), curvature γ_u (3.15), information $\mathcal{I}_u$ (3.10), and

$$\boldsymbol{v}_u = \overset{\perp}{\eta}_u/(\mathcal{I}_u\gamma_u), \tag{4.13}$$

the unit vector whose direction determines the critical boundary $\mathcal{B}_u$. The set of vectors $\boldsymbol{y}$ giving $\hat\alpha$ as the pMLE in family $\mathcal{F}_u$ lies in the $(n-1)$-dimensional hyperplane

$$\overset{\perp}{\mathcal{M}}_u = \{\boldsymbol{y} : \dot{\boldsymbol{\eta}}'_u\boldsymbol{y} = \dot s_u\} \tag{4.14}$$

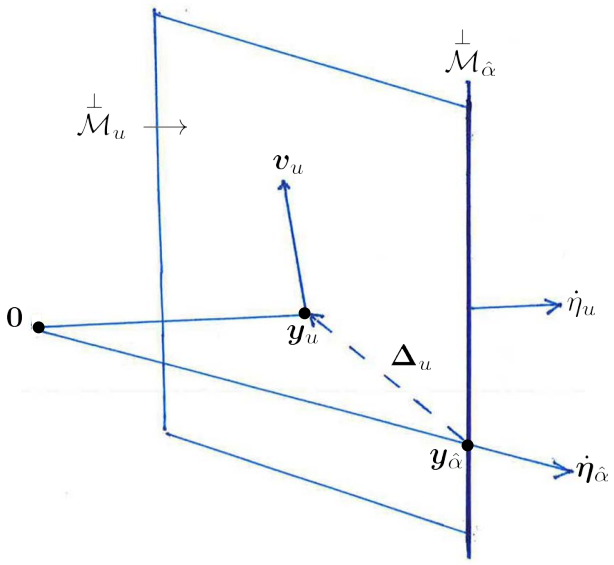

FIG 5. *Penalized maximum likelihood estimation:* $\overset{\perp}{\mathcal{M}}_{\hat\alpha}$ *is the* $(n-p)$*-dimensional hyperplane containing those observation vectors* $\boldsymbol{y}$ *having pMLE* $= \hat\alpha$ (4.3)*; it is orthogonal to* $\dot{\boldsymbol{\eta}}_{\hat\alpha}$*, passing through* $\boldsymbol{y}_{\hat\alpha}$*, its closest point to the origin;* $\overset{\perp}{\mathcal{M}}_u$ *is the* $(n-1)$*-dimensional hyperplane of pMLE solutions for a one-parameter subfamily* $\mathcal{F}_u$ (3.6)*, orthogonal to* $\dot\eta_u$*, passing through closest point* $\boldsymbol{y}_u$*; difference* $\boldsymbol{\Delta}_u = \boldsymbol{y}_u - \boldsymbol{y}_{\hat\alpha}$ *is in* $\overset{\perp}{\mathcal{L}}(\dot\eta_u) \cap \mathcal{L}(\dot{\boldsymbol{\eta}}_{\hat\alpha})$*; direction vector* $\boldsymbol{v}_u$ (4.13) *determines the boundary* $\mathcal{B}_u$ *as in Figure 3.*

passing through its nearest point to the origin

$$\boldsymbol{y}_u = \dot\eta_u \mathcal{I}_u^{-1} \dot s_u \tag{4.15}$$

at squared distance $\dot s_u' \mathcal{I}_u^{-1} \dot s_u$, (4.5)–(4.6).

Since $\dot m_{\hat\alpha}(\boldsymbol{y}) = 0$ implies $\dot m_u(\boldsymbol{y}) = 0$ (because $\dot m_u(\boldsymbol{y}) = u' \dot m_{\hat\alpha}(\boldsymbol{y})$) the $(n-p)$ space $\overset{\perp}{\mathcal{M}}_{\hat\alpha}$ (4.4) is contained in the $(n-1)$ space $\overset{\perp}{\mathcal{M}}_u$. Figure 5 illustrates the relationship. The vector $\boldsymbol{\Delta}$ connecting the two respective *nearest* points,

$$\boldsymbol{\Delta}_u = \boldsymbol{y}_u - \boldsymbol{y}_{\hat\alpha} = \dot\eta_u \mathcal{I}_u^{-1} \dot s_u - \dot{\boldsymbol{\eta}}_{\hat\alpha} \mathcal{I}_{\hat\alpha}^{-1} \dot s_{\hat\alpha}, \tag{4.16}$$

lies in $\overset{\perp}{\mathcal{L}}(\dot\eta_u) \cap \mathcal{L}(\dot{\boldsymbol{\eta}}_{\hat\alpha})$; $\boldsymbol{\Delta}_u = 0$ for an unpenalized MLE, but plays a role in determining the stable region for the pMLE.

As in Figure 2, there is a linear boundary $\mathcal{B}_u$ in $\overset{\perp}{\mathcal{M}}_u$ at which $\hat J_u(\boldsymbol{y})$ is zero:

LEMMA 4. *For any vector* $\boldsymbol{r}$ *in* $\overset{\perp}{\mathcal{M}}_u$ *orthogonal to the unit vector* $\boldsymbol{v}_u$*,*

$$\hat J_u(\boldsymbol{y}_u + b\boldsymbol{v}_u + \boldsymbol{r}) = \mathcal{I}_u(1 - b\gamma_u) - \frac{\nu_{u12}}{\nu_{u11}} \dot s_u + \ddot s_u; \tag{4.17}$$

$\hat J_u(\boldsymbol{y}_u + b\boldsymbol{v}_u + \boldsymbol{r})$ *equals 0 for* $b = c_u$*,*

$$c_u = \frac{1}{\gamma_u} \left(1 - \frac{\nu_{u12}}{\nu_{u11}^2} \dot s_u + \frac{\ddot s_u}{\nu_{u11}} \right). \tag{4.18}$$

The boundary $\mathcal{B}_u$ of vectors $\boldsymbol{y}$ having $\hat{J}_u(\boldsymbol{y}) = 0$ is an $(n-2)$-dimensional hyperplane in $\overset{\perp}{\mathcal{M}}_u$ passing through $c_u \boldsymbol{v}_u$ orthogonally to v_u.

PROOF. From (4.7), applied to $\mathcal{F}_u$, and (4.15), we get

$$\begin{aligned} \hat{J}_u(\boldsymbol{y}_u) &= \mathcal{I}_u - \ddot{\eta}_u' \boldsymbol{y}_u + \ddot{s}_u = \mathcal{I}_u - \ddot{\eta}_u' \dot{\eta}_u \mathcal{I}_u^{-1} \dot{s}_u + \ddot{s}_u \\ &= \mathcal{I}_u - \frac{\nu_{u12}}{\nu_{u11}} \dot{s}_u + \ddot{s}_u \end{aligned} \tag{4.19}$$

(remembering that $\mathcal{I}_u = \nu_{u11}$). Also

$$\hat{J}_u(\boldsymbol{y}_u + b\boldsymbol{v}_u + \boldsymbol{r}) = \hat{J}_u(\boldsymbol{y}_u) - \ddot{\eta}_u'(b\boldsymbol{v}_u + \boldsymbol{r}). \tag{4.20}$$

But

$$\ddot{\eta}_u'(b\boldsymbol{v}_u + \boldsymbol{r}) = b\overset{\perp}{\ddot{\eta}}{}_u' \boldsymbol{v}_u = b\mathcal{I}_u \gamma_u \tag{4.21}$$

using (4.13). Then (4.19) and (4.20) yield (4.17), the remainder of Lemma 4 following directly. □

Each choice of u produces a bounding hyperplane $\overset{\perp}{\mathcal{B}}_u$ in $\overset{\perp}{\mathcal{M}}_{\hat{\alpha}}$, the boundary being the intersection of $\mathcal{B}_u$ with $\overset{\perp}{\mathcal{M}}_{\hat{\alpha}}$ (as in Lemma 2). Each $\overset{\perp}{\mathcal{B}}_u$ defines a stable half-space $\mathcal{R}_u$, and their intersection $\cap \mathcal{R}_u$ defines the stable region $\mathcal{R}_{\hat{\alpha}}$ for the pMLE ((3.25) and Figure 4). The location of $\overset{\perp}{\mathcal{B}}_u$ is obtained as an extension of Theorem 2. Let $\boldsymbol{w}_u = \overset{\perp}{\boldsymbol{P}} \boldsymbol{v}_u / \cos\theta_u$ as in (3.21), the unit projection of $\boldsymbol{v}_u$ into $\overset{\perp}{\mathcal{M}}_{\hat{\alpha}}$.

THEOREM 4. *For*

$$\boldsymbol{y} = \boldsymbol{y}_{\hat{\alpha}} + b\boldsymbol{w}_u + \boldsymbol{r}, \tag{4.22}$$

where $\boldsymbol{r}$ is any vector in $\overset{\perp}{\mathcal{M}}_{\hat{\alpha}}$ orthogonal to $\boldsymbol{w}_u$, we have

$$\hat{J}_u(\boldsymbol{y}) = \mathcal{I}_u \gamma_u (c_u - b\cos\theta_u + \boldsymbol{\Delta}_u' \boldsymbol{v}_u), \tag{4.23}$$

c_u from (4.18); $\hat{J}_u(\boldsymbol{y})$ equals 0 for $b = d_u$,

$$d_u = (c_u + \boldsymbol{\Delta}_u' \boldsymbol{v}_u)/\cos\theta_u. \tag{4.24}$$

PROOF.

$$\hat{J}_u(\boldsymbol{y}_{\hat{\alpha}} + b\boldsymbol{w}_u + \boldsymbol{r}) = \hat{J}_u(\boldsymbol{y}_u) - \ddot{\eta}_u'(b\boldsymbol{w}_u + \boldsymbol{\Delta}_u + \boldsymbol{r}) \tag{4.25}$$

as at (4.20). Since $\boldsymbol{w}_u$, $\boldsymbol{\Delta}_u$, and $\boldsymbol{r}$ are orthogonal to $\dot{\eta}_u$, we can substitute $\overset{\perp}{\ddot{\eta}}_u = (\mathcal{I}_u \gamma_u)\boldsymbol{v}_u$ for $\ddot{\eta}_u$ in (4.25). Then

$$\ddot{\eta}_u' \boldsymbol{\Delta}_u = (\mathcal{I}_u \gamma_u) \boldsymbol{v}_u' \boldsymbol{\Delta}_u \tag{4.26}$$

and

$$\ddot{\eta}_u' b\boldsymbol{w}_u = b\overset{\perp}{\ddot{\eta}}{}_u' \boldsymbol{w}_u = b\|\overset{\perp}{\ddot{\eta}}_u\| \cos\theta_u = b\mathcal{I}_u \gamma_u \cos\theta_u. \tag{4.27}$$

Also

$$\hat{J}_u(\boldsymbol{y}_u) = \mathcal{I}_u - \frac{\nu_{u12}}{\nu_{u11}} \dot{s}_u + \ddot{s}_u = \mathcal{I}_u \gamma_u c_u. \tag{4.28}$$

Putting (4.24)–(4.28) together verifies (4.23), and solving for $\hat{J}_u(\boldsymbol{y}) = 0$ in (4.23) gives (4.24). □

THEOREM 5. *Assume that $\boldsymbol{\mathcal{I}}_{\hat\alpha}$ is positive definite. Then $-\ddot m_{\hat\alpha}(\boldsymbol{y}) = \hat{\boldsymbol{J}}(\boldsymbol{y})$ is positive definite if and only if $\boldsymbol{y}$ is in the region of stability*

$$\mathcal{R}_{\hat\alpha} = \bigcap_{u\in\mathcal{S}^p} \mathcal{R}_u, \tag{4.29}$$

which is (3.25).

Proof is the same as for Theorem 3.

Suppose that even though we are employing penalized maximum likelihood we remain interested in $-\ddot l_{\hat\alpha}(\boldsymbol{y}) = \hat{\boldsymbol{I}}(\boldsymbol{y})$ rather than $\hat{\boldsymbol{J}}(\boldsymbol{y})$. The only change needed is to remove the $\ddot s/\nu_{u11}$ term in the definition of c_u (4.18), after which $\hat I_u(\boldsymbol{y})$ can replace $\hat J_u(\boldsymbol{y})$ in (4.23), with an appropriate version of Theorem 5 following. The boundary distance d_u (4.24) will then be reduced from its previous value.

The toy example of Figure 1 was rerun now with penalty function (4.2) $c = 1$. This gave pMLE $= (6.84, 2.06)$ and the dashed regression line in Figure 1, rather than the MLE $\hat\alpha = (7.86, 2.38)$. Again the stable region was a half-space, minimum distance 2.95 compared with 2.85 at (3.28). There is no particular reason for regularization here, but it is essential in the g-modeling example of Section 5.

5. An empirical Bayes example. Familiar empirical Bayes estimation problems begin with a collection $\Theta_1, \Theta_2, \ldots, \Theta_N$ of unobserved parameters sampled from an unknown density function $g(\theta)$,

$$\Theta_i \overset{\text{ind}}{\sim} g(\theta), \qquad \text{for } i = 1, 2, \ldots, N. \tag{5.1}$$

Each Θ_i independently produces an observation X_i according to a known probability density kernel $p(x \mid \theta)$,

$$X_i \mid \Theta_i \overset{\text{ind}}{\sim} p(X_i \mid \Theta_i), \tag{5.2}$$

for example,

$$X_i \sim \mathcal{N}(\Theta_i, 1). \tag{5.3}$$

Having observed $\boldsymbol{X} = (X_1, X_2, \ldots, X_N)$, the statistician wishes to estimate all of the Θ_i values.

If $g(\cdot)$ were known then the Bayes posterior distribution $g(\Theta_i \mid X_i)$ would provide ideal inferences. Empirical Bayes methods attempt to approximate Bayesian results using only the observed sample $\boldsymbol{X}$. Efron (2016) suggested "g-modeling" for this purpose: a multiparameter exponential family of possible prior densities $g(\cdot)$ is hypothesized,

$$\mathcal{G} = \{g_\alpha(\theta), \alpha \in A\}. \tag{5.4}$$

It induces a family of marginal densities $f_\alpha(x)$ for the X_i,

$$\mathcal{F} = \left\{ f_\alpha(x) = \int p(x \mid \theta) g_\alpha(\theta)\, d\theta, \alpha \in A \right\}, \tag{5.5}$$

the integral taken over the sample space of Θ. The marginal model

$$X_i \overset{\text{ind}}{\sim} f_\alpha(x_i), \qquad \text{for } i = 1, 2, \ldots, N, \tag{5.6}$$

yields an estimate $\hat{\alpha}$ by maximum likelihood. This gives $g_{\hat{\alpha}}(\theta)$ as an estimate of the unknown prior, which can then be plugged into Bayes formula for inferential purposes.

Except in the simplest of situations, the convolution step (5.5) spoils exponential family structure, making $\mathcal{F}$ into a multiparameter *curved* exponential family. G-modeling was the motivating example for this paper. Does its application lead to large regions of stability $\mathcal{R}_{\hat{\alpha}}$, or to dangerously small ones where $\hat{\boldsymbol{I}}(y)$ and $\boldsymbol{\mathcal{I}}_{\hat{\alpha}}$ can be strikingly different–or, worse, where $\boldsymbol{y}$ may be prone to falling outside of $\mathcal{R}_{\hat{\alpha}}$? Here we present only a single example rather than a comprehensive analysis.

A diffusion tensor imaging study (DTI) compared six dyslexic children with six normal controls at 15,443 brain voxels (Schwartzman, Dougherty and Taylor, 2005; see also Efron, 2010, Sect. 2.5). Each voxel produced a statistic X_i comparing dyslexics with normals. Model (5.3), $X_i \sim \mathcal{N}(\Theta_i, 1)$, is reasonable here, the Θ_i being the true voxel-wise effect sizes we wish to estimate.

For our example we consider only the $N = 477$ voxels from the extreme back of the brain. Smaller sample size exacerbates curvature effects, making them easier to examine; see Remark G in Section 6. A histogram of the N X_i's appears in Panel A of Figure 6. Superimposed is an estimate of the prior density $g(\theta)$ (5.1), including a large atom of probability at $\Theta = 0$, as explained below.

Description of the g-modeling algorithm is simplified by discretizing both Θ and X. We assume that Θ_i can take on m possible values,

$$\boldsymbol{\theta} = (\theta_{(1)}, \theta_{(2)}, \ldots, \theta_{(m)}); \tag{5.7}$$

for the DTI example $\boldsymbol{\theta} = (-2.4, -2.2, \ldots, 3.6)$ with $m = 31$. The X_i were discretized by placement in $n = 37$ bins of width 0.2, having center points

$$\boldsymbol{x} = (-3.2, -3.0, \ldots, 4.0). \tag{5.8}$$

Define y_k to be the number of X_i's in bin k, so that the count vector $\boldsymbol{y}$,

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_n), \tag{5.9}$$

gives the heights of the histogram bars in Panel A. We will work with data vector $\boldsymbol{y}$ rather than $\boldsymbol{X}$, ignoring the slight loss of information from binning.

In the discrete formulation (5.7) the unknown prior $g(\theta)$ is described by a vector $\boldsymbol{g}$,

$$\boldsymbol{g} = (g_1, g_2, \ldots, g_m), \tag{5.10}$$

with $g_k = \Pr\{\Theta_i = \theta_{(k)}\}$ for $k = 1, 2, \ldots, m$. Our exponential family model $\mathcal{G}$ defines the components $\boldsymbol{g}_\alpha$ by

$$g_{\alpha k} = e^{Q_k'\alpha}/C_\alpha, \tag{5.11}$$

where Q_k is a given p-dimensional vector and α is an unknown p-dimensional parameter vector; $C_\alpha = \sum_1^m \exp\{Q_k'\alpha\}$. The $m \times p$ *structure matrix* $\boldsymbol{Q}$, having kth row Q_k', determines the exponential family of possible priors (5.4).

Define

$$p_{kj} = \Pr\{X_i \in \text{bin}_k \mid \Theta_i = \theta_{(j)}\}, \tag{5.12}$$

and $\boldsymbol{P}$ as the $n \times m$ matrix

$$\boldsymbol{P} = (p_{kj}, k = 1, 2, \ldots, n \text{ and } j = 1, 2, \ldots, m). \tag{5.13}$$

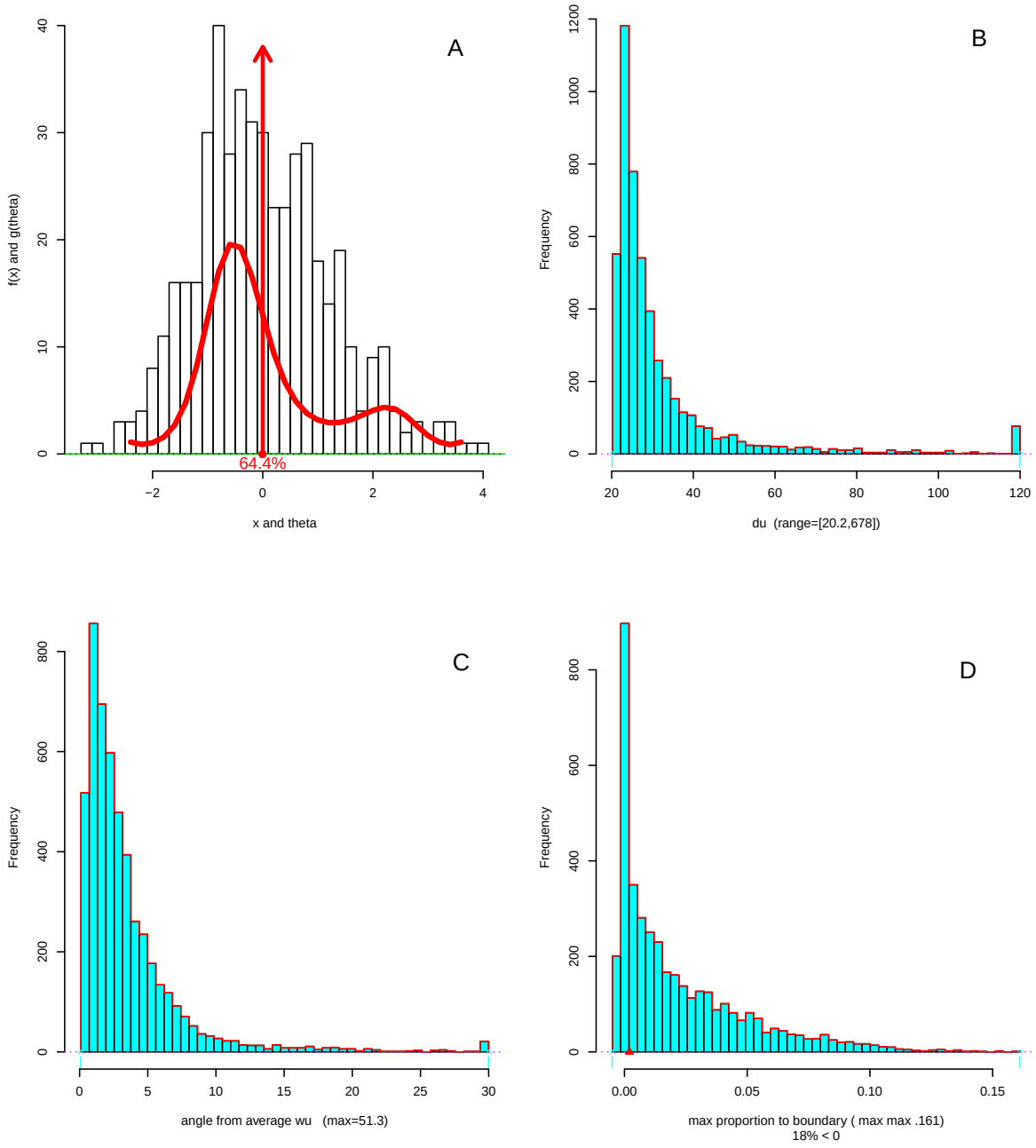

FIG 6. ***Panel A****: Histogram of the 477 observations X_i, and the estimated prior density $\boldsymbol{g}_{\hat{\alpha}}$ based on spike and slab prior* (5.16)*; it estimates* $\Pr\{\Theta = 0\} = 0.644$. ***Panel B****: Histogram of critical distances d_u* (4.24) *for 5000 randomly selected vectors u* (5.17). ***Panel C****: Angular distances in degrees of the 5000 direction vectors $\boldsymbol{w}_u$* (3.21) *from their average vector $\bar{\boldsymbol{w}}$. **Panel D**: Maximum proportional distance to the boundary of stable region $\mathcal{R}_{\hat{\alpha}}$ for 4000 bootstrap observation vectors $\boldsymbol{y}^*$; triangular point indicates maximum proportional distance for actual observation $\boldsymbol{y}$.*

The marginal density $f_\alpha(x)$ in (5.5) is given by the n-vector $\boldsymbol{f}_\alpha$,

$$\boldsymbol{f}_\alpha = \boldsymbol{P}\boldsymbol{g}_\alpha. \tag{5.14}$$

A flow chart of empirical Bayes g-modeling goes as follows:

$$\alpha \longrightarrow \boldsymbol{g}_\alpha = e^{\boldsymbol{Q}'\alpha}/C_\alpha \longrightarrow \boldsymbol{f}_\alpha = \boldsymbol{P}\boldsymbol{g}_\alpha \longrightarrow \boldsymbol{y} \sim \text{Mult}_n(N, \boldsymbol{f}_\alpha), \tag{5.15}$$

the last indicating a multinomial distribution on n categories, sample size N, probability vector $\boldsymbol{f}_\alpha$. (This assumes independence as in (5.1) and (5.2), not actually the case for the DTI data; see Remark H in Section 6.)

The estimate of $g(\theta)$ shown in Panel A was based on a p equals 8-dimensional "spike and slab" prior,

$$\boldsymbol{Q} = (I_0, \text{poly}(\boldsymbol{\theta}, 7))\,; \tag{5.16}$$

here I_0 represents a delta function at $\Theta = 0$ (vector $(\dots 0, 1, 0, \dots)$, 1 in the 13th place in (5.7)) while $\text{poly}(\boldsymbol{\theta}, 7)$ was the $m \times 7$ matrix provided by the R function `poly`. Model (5.16) allows for a spike of "null voxels" at $\Theta = 0$, and a smooth polynomial distribution for the non-null cases.

For this data set, the MLE estimate $\boldsymbol{g}_{\hat\alpha}$ put probability 0.644 on $\Theta = 0$; the remaining 0.356 probability was distributed bimodally–most of the non-null mass was close to 0, but with a small mode of possibly interesting voxels around $\Theta = 2$. See Panel A. Efron (2016) gives simple formulas for the standard errors of $\boldsymbol{g}_{\hat\alpha}$, but our interest here is in questions of stability: what does the region of stability $\mathcal{R}_{\hat\alpha}$ look like, and how close to or far from its boundary is the observed data vector $\boldsymbol{y}$?

Exponential family model (5.11) leads to simple expressions for $\dot{\boldsymbol{\eta}}_\alpha$ and $\ddot{\boldsymbol{\eta}}_\alpha$, (2.7)–(2.8), the necessary ingredients for calculating $\hat\alpha$ and $\mathcal{R}_{\hat\alpha}$, the stable region. See Remark I in Section 6. G-modeling was carried out based on $\boldsymbol{y} \sim \text{Mult}_n(N, \boldsymbol{f}_\alpha)$, as in (5.15), giving pMLE $\hat\alpha$ (with $c = 1$ in (4.1)–(4.2)). Panel A of Figure 6 shows the estimated prior $\boldsymbol{g}_{\hat\alpha}$.

The calculations for $\mathcal{R}_{\hat\alpha}$ were done after transformation to standardized coordinates having $\boldsymbol{\mu}_{\hat\alpha} = \boldsymbol{0}$ and $\boldsymbol{V}_{\hat\alpha} = \boldsymbol{I}_n$ (3.5), this being assumed from now on. The construction of $\mathcal{R}_{\hat\alpha}$ pictured in Figure 4 was carried out, here in 29 dimensions ($n - p = 37 - 8$). This brings up the problem of choosing the one-parameter bounding families $\mathcal{F}_u$ (3.6), with u 8-dimensional rather than the two dimensions of the toy example (3.27).

Five thousand u vectors were chosen randomly uniformly from $\mathcal{S}^8$, the surface of the unit sphere in eight dimensions,

$$u_1, u_2, \dots, u_{5000}. \tag{5.17}$$

Each u yielded a direction vector $\boldsymbol{w}_u$ (3.21) in the 29-dimensional space $\overset{\perp}{\mathcal{M}}_{\hat\alpha}$ (4.4), and a distance d_u to the critical boundary (4.24). The points $d_u\boldsymbol{w}_u$ are the high-dimensional equivalent of the dots in Figure 4. Here the origin $\boldsymbol{0}$ is $\boldsymbol{y}_{\hat\alpha}$ (4.6).

Panel B of Figure 6 is a histogram of the 5000 d_u values,

$$20.2 \le d_u \le 678. \tag{5.18}$$

The minimum of d_u over all of $\mathcal{S}_8$ was 20.015, found by Newton–Raphson minimization (starting from any point on $\mathcal{S}_8$).

Let $\bar{\boldsymbol{w}} = \sum \boldsymbol{w}_{uh}/5000$ be the average $\boldsymbol{w}_u$ direction vector. Panel C is a histogram of the angle in degrees between $\boldsymbol{w}_{uh}$ and $\bar{\boldsymbol{w}}$. We see a close clustering around $\bar{\boldsymbol{w}}$, the mean angular difference being only 3.5 degrees.

The stable region $\mathcal{R}_{\hat\alpha}$ (3.25) has its boundary more than 20 Mahalanobis distance units away from the origin. Is this sufficiently far to rule out unstable behavior? As a check, 4000 parametric bootstrap observation vectors $\boldsymbol{Y}^*$ were generated,

$$\boldsymbol{Y}_i^* \sim \text{Mult}_{37}(477, \boldsymbol{f}_{\hat\alpha}) \qquad (i = 1, 2, \dots, 4000) \tag{5.19}$$

and then standardized and projected into vectors $\boldsymbol{y}_i^*$ in the 29-dimensional space $\overset{\perp}{\mathcal{M}}_{\hat{\alpha}}$ (4.4); see Remark J in Section 6. For each $\boldsymbol{y}_i^*$ we define

$$m_i = \max_h\{\boldsymbol{y}_i^{*'}\boldsymbol{w}_{uh}/d_{uh}, h = 1, 2, \ldots, 5000\}, \tag{5.20}$$

this being the maximum proportional distance of $\boldsymbol{y}_i^*$ to the linear boundary $\overset{\perp}{\mathcal{B}}_{uh}$ (the tangent lines in Figure 4); $m_i > 1$ would indicate $\boldsymbol{y}_i^*$ beyond the boundary of $\mathcal{R}_{\hat{\alpha}}$.

In fact, Panel D of Figure 6 shows $m_i \le 0.161$ for all i. For the actual observation vector $\boldsymbol{y}$, m equaled 0.002. In this case we need not worry about stability problems. Observed and expected Fisher information are almost the same for $\boldsymbol{y}$, and would be unlikely to vary much for other possible observations $\boldsymbol{y}^*$. And there is almost no possibility of an observation falling outside of the region of stability $\mathcal{R}_{\hat{\alpha}}$. G-modeling looks to be on stable ground in this example. (But see the cautionary note in Remark M.)

Panel D shows that 18% of the 4000 $\boldsymbol{y}_i^*$ vectors gave m_i less than zero. That is, $\boldsymbol{y}_i^*$ had negative correlation with all 5000 w_{uh} direction vectors (it was in their "polar cone"). This implies that $\mathcal{R}_{\hat{\alpha}}$ is open, as in Figure 4. A circular polar cone that included 18% of the unit sphere in 29-dimensional space would have angular radius 73.9 degreees — see Remark L in Section 6 — so the polar opening is substantial.

6. Remarks. This section presents comments, details, and proofs relating to the previous material.

Remark A. *Formula* (2.13) Result (2.13) is obtained by differentiating $\dot{l}_\alpha(\boldsymbol{y}) = \dot{\boldsymbol{\eta}}_\alpha'(\boldsymbol{y} - \boldsymbol{\mu}_\alpha)$ (2.10),

$$\ddot{l}_\alpha(\boldsymbol{y}) = \ddot{\boldsymbol{\eta}}_\alpha'(\boldsymbol{y} - \boldsymbol{\mu}_\alpha) - \dot{\boldsymbol{\eta}}_\alpha'\frac{d\boldsymbol{\mu}_\alpha}{d\alpha}. \tag{6.1}$$

Since $\boldsymbol{\mu}_{\boldsymbol{\eta}} = d\psi(\boldsymbol{\eta})/d\boldsymbol{\eta}$ and $\boldsymbol{V}_{\boldsymbol{\eta}} = d^2\psi(\boldsymbol{\eta})/d\boldsymbol{\eta}^2$ give $d\boldsymbol{\mu}_{\boldsymbol{\eta}}/d\boldsymbol{\eta} = \boldsymbol{V}_{\boldsymbol{\eta}}$, we get $d\boldsymbol{\mu}_\alpha/d\alpha = \boldsymbol{V}_\alpha\dot{\boldsymbol{\eta}}_\alpha$ and

$$\ddot{l}_\alpha(\boldsymbol{y}) = \ddot{\boldsymbol{\eta}}_\alpha'(\boldsymbol{y} - \boldsymbol{\mu}_\alpha) - \dot{\eta}_\alpha'\boldsymbol{V}_\alpha\dot{\eta}_\alpha = \ddot{\boldsymbol{\eta}}_\alpha'(\boldsymbol{y} - \boldsymbol{\mu}_\alpha) - \boldsymbol{\mathcal{I}}_\alpha, \tag{6.2}$$

which is (2.13).

Remark B. *Formula* (2.22) Suppose first that we have transformed to standardized coordinates (3.2) where $\boldsymbol{\mu}_{\hat{\alpha}} = \boldsymbol{0}$ and $\boldsymbol{V}_{\hat{\alpha}} = \boldsymbol{I}_n$ (3.5). Then the projection of $\ddot{\boldsymbol{\eta}}_{\hat{\alpha}}^\dagger$ into $\overset{\perp}{\mathcal{L}}{}_{\hat{\alpha}}^\dagger$ in Figure 2 is, by the usual OLS calculations,

$$\overset{\perp}{\boldsymbol{\eta}}{}_{\hat{\alpha}}^\dagger = \ddot{\eta}_{\hat{\alpha}}^\dagger - \frac{\nu_{12}}{\nu_{11}}\dot{\eta}_{\hat{\alpha}}^\dagger, \tag{6.3}$$

with length

$$\|\overset{\perp}{\boldsymbol{\eta}}{}_{\hat{\alpha}}^\dagger\| = \nu_{22} - \nu_{12}^2/\nu_{11} = \mathcal{I}_{\hat{\alpha}}\gamma_{\hat{\alpha}}, \tag{6.4}$$

so $\boldsymbol{v}_{\hat{\alpha}}^\dagger = \overset{\perp}{\boldsymbol{\eta}}{}_{\hat{\alpha}}^\dagger/\mathcal{I}_{\hat{\alpha}}\gamma_{\hat{\alpha}}$ has unit length.

Notice that $\mathcal{I}_{\hat{\alpha}}$, $\gamma_{\hat{\alpha}}$, ν_{11}, ν_{12}, ν_{22} are all invariant under the transformations (3.2) as is the observed information,

$$-\ddot{l}_{\hat{\alpha}}(\boldsymbol{y}) = \mathcal{I}_{\hat{\alpha}} - \ddot{\boldsymbol{\eta}}_{\hat{\alpha}}'(\boldsymbol{y} - \boldsymbol{\mu}_{\hat{\alpha}}) = \mathcal{I}_{\hat{\alpha}} - \ddot{\boldsymbol{\eta}}_{\hat{\alpha}}'^\dagger\boldsymbol{y}^\dagger. \tag{6.5}$$

Also

$$\boldsymbol{v}_{\hat\alpha} = \boldsymbol{V}_{\hat\alpha}\overset{\perp}{\boldsymbol{\eta}}_{\hat\alpha}/\mathcal{I}_{\hat\alpha}\gamma_{\hat\alpha} = \boldsymbol{M}\boldsymbol{v}_{\hat\alpha}^{\dagger} \tag{6.6}$$

satisfies

$$\boldsymbol{v}'_{\hat\alpha}(\boldsymbol{y}-\boldsymbol{\mu}_{\hat\alpha}) = \boldsymbol{v}_{\hat\alpha}^{\prime\dagger}\boldsymbol{y}^{\dagger}. \tag{6.7}$$

We can rewrite (6.5) in terms of $\boldsymbol{v}_{\hat\alpha}$ (2.22):

$$-\ddot{l}_{\hat\alpha}(\boldsymbol{y}) = \mathcal{I}_{\hat\alpha} - \mathcal{I}_{\hat\alpha}\gamma_{\hat\alpha}\boldsymbol{v}'_{\hat\alpha}(\boldsymbol{y}-\boldsymbol{\mu}_{\hat\alpha}) = \mathcal{I}_{\hat\alpha} - \mathcal{I}_{\hat\alpha}\gamma_{\hat\alpha}\boldsymbol{v}_{\hat\alpha}^{\prime\dagger}\boldsymbol{y}^{\dagger}. \tag{6.8}$$

This justifies the use of $\boldsymbol{v}_{\hat\alpha}$ in (2.24), and quickly leads to verification of Theorem 1.

Remark C. *Lemma 2* By rotations

$$\boldsymbol{\eta} \longrightarrow \boldsymbol{\Gamma}\boldsymbol{\eta} \quad \text{and} \quad \boldsymbol{y} \longrightarrow \boldsymbol{\Gamma}'\boldsymbol{y}, \tag{6.9}$$

where $\boldsymbol{\Gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_n)$ is an $n \times n$ orthogonal matrix, we can simplify calculations relating to Figure 3: select γ_1 to lie along $\mathcal{L}(\dot\eta_u)$ and $\gamma_2, \gamma_3, \ldots, \gamma_p$ to span $\mathcal{L}(\dot{\boldsymbol{\eta}}_{\hat\alpha}) \cap \overset{\perp}{\mathcal{L}}(\dot\eta_u)$. (Notice that $\boldsymbol{y}$ still has mean $\boldsymbol{0}$ and covariance $\boldsymbol{I}_n$.) For any n-vector $\boldsymbol{z}$, write

$$\boldsymbol{z} = (z_1, z_2, z_3), \tag{6.10}$$

where z_1 is the first coordinate, z_2 coordinates $2, 3, \ldots, p$, and z_3 coordinates $p+1$ through n. Then

$$\begin{aligned} \boldsymbol{v}_u &= (0, v_{u2}, v_{u3}) \\ \text{and } \boldsymbol{y} &= (0, 0, y_3), \end{aligned} \tag{6.11}$$

the zeros following from $\boldsymbol{v}_u \in \overset{\perp}{\mathcal{L}}(\dot\eta_u)$ and $\boldsymbol{y} \in \overset{\perp}{\mathcal{L}}(\dot{\boldsymbol{\eta}}_{\hat\alpha})$.

The projection $\overset{\perp}{\boldsymbol{P}}v_u$ of $\boldsymbol{y}_u$ into $\overset{\perp}{\mathcal{L}}(\dot{\boldsymbol{\eta}}_{\hat\alpha})$ must equal $(0, 0, v_{u3})$, giving

$$\begin{aligned} \cos\theta_u &= \|v_{u3}\|/\|\boldsymbol{v}_u\| \\ &= \|v_{u3}\|. \end{aligned} \tag{6.12}$$

Also $\boldsymbol{w}_u$ (3.21) in $\overset{\perp}{\mathcal{L}}(\dot{\boldsymbol{\eta}}_{\hat\alpha})$ equals

$$\boldsymbol{w}_u = (0, 0, w_{u3}) = (0, 0, v_{u3}/\cos\theta_u), \tag{6.13}$$

$\boldsymbol{w}_u$ being the unit projection of $\boldsymbol{v}_u$ into $\overset{\perp}{\mathcal{L}}(\dot{\boldsymbol{\eta}}_{\hat\alpha})$.

The vector $\boldsymbol{\delta}_u = \boldsymbol{w}_u/(\mathcal{I}_u\gamma_u\cos\theta_u)$ lies on the hyperplane $\mathcal{B}_u$, which is defined by

$$\boldsymbol{\delta}'_u\boldsymbol{v}_u = 1/\gamma_u, \tag{6.14}$$

since $\boldsymbol{w}'_u\boldsymbol{v}_u = \|v_{u3}\|^2/\cos\theta_u = \cos\theta_u$, and it has length $\|\delta_u\| = d_u = 1/(\gamma_u\cos\theta_u)$ (3.18). Suppose $\boldsymbol{\delta}_u + \boldsymbol{r}$ is any vector in $\overset{\perp}{\mathcal{L}}(\dot{\boldsymbol{\eta}}_{\hat\alpha})$, $\boldsymbol{r} = (0, 0, r_3)$, that is also in $\mathcal{B}_u$. Then $(\boldsymbol{\delta}_u + \boldsymbol{r})'\boldsymbol{v}_u = 1/\gamma_u$ implies $\boldsymbol{r}'\boldsymbol{v}_u = 0$ and so, from (6.13), $\boldsymbol{r}'\boldsymbol{\delta}_u = 0$. This verifies that $\boldsymbol{\delta}_u$ is the nearest point in $\overset{\perp}{\mathcal{B}}_u$ to $\boldsymbol{0}$ as claimed in Lemma 2.

Remark D. *Theorem 2* For $\boldsymbol{y} = b\boldsymbol{w}_u + \boldsymbol{r}$,

$$\begin{aligned}\ddot{\eta}'_u\boldsymbol{y} = \overset{\perp}{\ddot{\eta}}{}'_u\boldsymbol{y} = \mathcal{I}_u\gamma_u\boldsymbol{v}'_u\boldsymbol{y} = \\ = \mathcal{I}_u\gamma_u\boldsymbol{v}'_u(b\boldsymbol{w}_u + \boldsymbol{r}) = \mathcal{I}_u\gamma_u\cos\theta_u \cdot b.\end{aligned} \tag{6.15}$$

Then (3.23) follows from $\hat{I}_u(\boldsymbol{y}) = \mathcal{I}_u - \ddot{\eta}'_u\boldsymbol{y}$.

Remark E. *The toy model* For model (1.1)–(1.2) it is easy to show that $\dot{\boldsymbol{\eta}}_{\hat{\alpha}}$ has ith row $(1, x_i)/\mu_{\hat{\alpha}i}$, and $\ddot{\boldsymbol{\eta}}_{\hat{\alpha}}$ has ith matrix

$$\ddot{\boldsymbol{\eta}}_{\hat{\alpha}i} = -\frac{1}{\mu^2_{\hat{\alpha}i}}\begin{pmatrix}1 & x_i\\ x_i & x_i^2\end{pmatrix}. \tag{6.16}$$

Remark F. *Influence function* (4.8) From

$$\dot{\boldsymbol{\eta}}'_{\hat{\alpha}+d\alpha}(\boldsymbol{y} - \boldsymbol{\mu}_{\hat{\alpha}+d\alpha}) = \dot{s}_{\hat{\alpha}+d\alpha} \tag{6.17}$$

(4.3), we get the local relationship

$$(\dot{\boldsymbol{\eta}}_{\hat{\alpha}} + \ddot{\boldsymbol{\eta}}_{\hat{\alpha}}d\alpha)'(\boldsymbol{y} - \dot{\boldsymbol{\eta}}_{\hat{\alpha}}d\alpha - d\boldsymbol{y}) = \dot{s}_{\hat{\alpha}} + \ddot{s}_{\hat{\alpha}}d\alpha, \tag{6.18}$$

where we have used $\boldsymbol{\mu}_{\hat{\alpha}} = \boldsymbol{0}$, $\boldsymbol{V}_{\hat{\alpha}} = \boldsymbol{I}_n$, and $d\boldsymbol{\mu}/d\boldsymbol{\eta} = \boldsymbol{V}_{\boldsymbol{\eta}}$. This reduces to

$$(\ddot{\boldsymbol{\eta}}'_{\hat{\alpha}}\boldsymbol{y} - \boldsymbol{\mathcal{I}}_{\hat{\alpha}} - \ddot{s}_{\hat{\alpha}})d\alpha = \dot{\boldsymbol{\eta}}'_{\hat{\alpha}}d\boldsymbol{y} \tag{6.19}$$

(using $\ddot{\eta}_{\hat{\alpha}ijk} = \ddot{\eta}_{\hat{\alpha}ikj}$), which yields (4.8).

The linear expansion (4.8) suggests the covariance approximation

$$\begin{aligned}\operatorname{cov}(\hat{\alpha}) &\doteq \hat{\boldsymbol{J}}(\boldsymbol{\mu}_{\hat{\alpha}})^{-1}\dot{\eta}'_{\hat{\alpha}}\boldsymbol{V}_{\hat{\alpha}}\dot{\eta}_{\hat{\alpha}}\hat{\boldsymbol{J}}(\boldsymbol{\mu}_{\hat{\alpha}})^{-1}\\ &= (\boldsymbol{\mathcal{I}}_{\hat{\alpha}} + \ddot{s}_{\hat{\alpha}})^{-1}\mathcal{I}_{\hat{\alpha}}(\boldsymbol{\mathcal{I}}_{\alpha} + \ddot{s}_{\hat{\alpha}})^{-1}\end{aligned} \tag{6.20}$$

for the pMLE (Efron, 2016, Thm. 2), in contrast with the Bayesian covariance estimate $\hat{\boldsymbol{J}}(\boldsymbol{y})^{-1}$ in (4.9).

Remark G. *Sample size effects* Curvature γ_u decreases at order $O(N^{-1/2})$ as sample size N increases (Efron, 1975). This suggests that the distance d_u to the boundary of $\mathcal{R}_{\hat{\alpha}}$ should increase as $O(N^{1/2})$, (3.18) and (4.24). Doubling the sample size in the DTI example (by replacing the count vector $\boldsymbol{y}$ with $2\boldsymbol{y}$) increased the minimum distance from 20.02 to 30.3; doubling again gave 47.6, increasing somewhat faster than predicted.

Remark H. *Correlation* The X_i observations for the DTI study of Section 5 suffer from *local correlation*, nearby brain voxels being highly correlated, as illustrated in Section 2.5 of Efron (2010) and discussed at length in Chapters 7 and 8 of that work. This doesn't bias g-modeling estimates $\hat{\alpha}$, but does increase variability of the count vectors $\boldsymbol{y}$. The effect is usually small for local correlation models — as opposed to the kinds of global correlations endemic to microarray studies — and can sometimes be calculated by the methods of Efron (2010). In any case, correlation has been ignored here for the sake of presenting an example of the stability calculations.

Remark I. $\dot{\boldsymbol{\eta}}_{\hat{\alpha}}$ *and* $\ddot{\boldsymbol{\eta}}_{\hat{\alpha}}$ *for g-models* Section 2 of Efron (2016) calculates $\dot{\boldsymbol{\eta}}_{\hat{\alpha}}$ and $\ddot{\boldsymbol{\eta}}_{\hat{\alpha}}$ for model (5.15): define

$$w_{kj}(\alpha) = g_{\alpha j}\left(\frac{p_{kj}}{f_{\alpha k}} - 1\right), \tag{6.21}$$

giving the $m \times n$ matrix $\boldsymbol{W}(\alpha) = (w_{kj}(\alpha))$, having kth column $\boldsymbol{W}_k(\alpha) = (w_{k1}(\alpha), \ldots, w_{km}(\alpha))'$. Then

$$\dot{\boldsymbol{\eta}}_{\hat{\alpha}} = \boldsymbol{W}(\alpha)'\boldsymbol{Q}, \tag{6.22}$$

where $\boldsymbol{Q}$ is the $m \times p$ g-modeling structure matrix. The $n \times p \times p$ array $\ddot{\boldsymbol{\eta}}_{\hat{\alpha}}$ has kth $p \times p$ matrix

$$\boldsymbol{Q}' \left[\operatorname{diag} \boldsymbol{W}_k(\alpha) - \boldsymbol{W}_k(\alpha)\boldsymbol{W}_k(\alpha)' - \boldsymbol{W}_k(\alpha)\boldsymbol{g}_\alpha - \boldsymbol{g}_\alpha \boldsymbol{W}_k(\alpha)'\right] \boldsymbol{Q}, \tag{6.23}$$

$\operatorname{diag} \boldsymbol{W}_k(\alpha)$ being the diagonal matrix with diagonal element $w_{kj}(\alpha)$.

These formulas apply to the original untransformed coordinates. Transformations (3.21) to standardized form employ

$$\boldsymbol{\mu}_{\hat{\alpha}} = N\boldsymbol{f}_{\hat{\alpha}} \quad \text{and} \quad \boldsymbol{M} = \operatorname{diag}(N\boldsymbol{f}_{\hat{\alpha}})^{1/2} \tag{6.24}$$

(see Remark J), changing the previous expressions to

$$\dot{\boldsymbol{\eta}}^{\dagger}_{\hat{\alpha}} = \boldsymbol{M}\dot{\boldsymbol{\eta}}_{\hat{\alpha}} \quad \text{and} \quad \ddot{\boldsymbol{\eta}}^{\dagger}_{\hat{\alpha}} = \boldsymbol{M}\ddot{\boldsymbol{\eta}}_{\hat{\alpha}}; \tag{6.25}$$

$\boldsymbol{M}\ddot{\boldsymbol{\eta}}_{\hat{\alpha}}$ indicates the multiplication of each n-vector ($\ddot{\boldsymbol{\eta}}_{\hat{\alpha}kjh}$, $k = 1, 2, \ldots, n$) by $\boldsymbol{M}$.

Remark J. *Multinomial standardization* Transformation (5.19)–(5.20) was

$$y_i^* = \operatorname{diag}(N\boldsymbol{f}_{\hat{\alpha}})^{-1/2}(Y_i^* - \hat{\boldsymbol{\mu}}_{\hat{\alpha}}), \tag{6.26}$$

$\operatorname{diag}(N\boldsymbol{f}_{\hat{\alpha}})^{-1}$ being a pseudo-inverse of the singular multinomial covariance matrix $N[\operatorname{diag}(\boldsymbol{f}_{\hat{\alpha}}) - \boldsymbol{f}_{\hat{\alpha}}\boldsymbol{f}'_{\hat{\alpha}}]$. This gives

$$\operatorname{cov}(\boldsymbol{y}_i^*) = \boldsymbol{I} - \sqrt{\frac{\boldsymbol{f}_{\hat{\alpha}}}{N}}\sqrt{\frac{\boldsymbol{f}'_{\hat{\alpha}}}{N}}, \tag{6.27}$$

which represents identity covariance matrix in the $(n-1)$-dimensional linear space of y_i^*'s variability, justifying (6.24).

The multinomial sampling model at the end of (5.15), $\boldsymbol{y} \sim \operatorname{Mult}_n(N, \boldsymbol{f}_\alpha)$ can be replaced by a Poisson model

$$y_k \overset{\text{ind}}{\sim} \operatorname{Poi}(\mu_{\alpha k}) \qquad \text{for } k = 1, 2, \ldots, n, \tag{6.28}$$

where $\mu_{\alpha k} = \beta f_{\alpha k}$, with β a free parameter. This gives maximum likelihood $\hat{\beta} = N$ and $\hat{\alpha}$ the same as before (an application of "Lindsey's method", Lindsey, 1974). The choice $\boldsymbol{M} = \operatorname{diag}(N\boldsymbol{f}_{\hat{\alpha}})^{1/2}$ in (6.24) is obviously correct for the Poisson model.

Remark K. *Original coordinates* By inverting transformations (3.2) we can express our results directly in terms of the original coordinates of Section 2. Various forms may be more or less convenient. For instance in (3.18), $d_u = 1/(\gamma_u \cos\theta_u)$, γ_u still follows expression (3.15) but now having $\nu_{u11} = \dot{\eta}'_u \boldsymbol{V}_{\hat{\alpha}} \dot{\eta}_u$, and likewise for ν_{u12} and ν_{u22}, and with $\dot{\eta}_u$ and $\ddot{\eta}_u$ still given by (3.8)–(3.9); $\cos\theta_u$ can be computed from

$$\sin^2\theta_u = 1 - \cos^2\theta_u = \frac{(\overset{\perp}{\dot{\eta}}{}'_u \boldsymbol{V}_{\hat{\alpha}} \dot{\boldsymbol{\eta}}_{\hat{\alpha}}) \boldsymbol{\mathcal{I}}_{\hat{\alpha}}^{-1} (\dot{\boldsymbol{\eta}}'_{\hat{\alpha}} \boldsymbol{V}_{\hat{\alpha}} \overset{\perp}{\dot{\eta}}_u)}{(\mathcal{I}_u \gamma_u)^2}, \tag{6.29}$$

$\mathcal{I}_u = \nu_{u11}$.

Remark L. *Areas on the sphere* A spherical cap of radius ρ radians on the surface of a unit sphere in R^d has $(d-1)$-dimensional area, relative to the full sphere,

$$c_d \int_0^\rho \sin(r)^{d-2}\, dr \qquad \left(c_d = \frac{1}{\sqrt{\pi}}\,\frac{\Gamma(d/2)}{\Gamma[(d-1)/2]}\right). \tag{6.30}$$

Remark M. All of our results concerning the observed likelihood $\hat{\boldsymbol{I}}(\boldsymbol{y})$, or $\hat{\boldsymbol{J}}(\boldsymbol{y})$, are exact. What isn't exact are the probability consequences of statements like "the boundary distance is at Mahalanobis distance 2.85." Such statements relate, at least formally, to conditional distributions within $\overset{\perp}{\mathcal{L}}(\dot{\boldsymbol{\eta}}_{\hat{\alpha}})$ or $\overset{\perp}{\mathcal{M}}(\dot{\boldsymbol{\eta}}_{\hat{\alpha}})$, and can be delicate; see Figure 5 in Efron (1975), and Barndorff-Nielsen's discussion following Efron and Hinkley (1978). Notice that the development here at (5.19)–(5.20) avoids conditioning by using the simpler approach of projection.

Acknowledgement. The author's research is supported in part by National Science Foundation award DMS 1608182.

References.

Amari, S.-i. (1982). Differential geometry of curved exponential families—curvatures and information loss. *Ann. Statist.* **10** 357–385. MR653513

Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. Statist.* **3** 1189–1242. With discussion by C.R. Rao, D.A. Pierce, D.R. Cox, D.V. Lindley, L. LeCam, J.K. Ghosh, J. Pfanzagl, N. Keiding, A.P. Dawid, and J. Reeds, and a reply by the author. MR0428531

Efron, B. (1978). The geometry of exponential families. *Ann. Statist.* **6** 362–376. MR0471152

Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. Institute of Mathematical Statistics Monographs* **1**. Cambridge University Press, Cambridge. MR2724758

Efron, B. (2016). Empirical Bayes deconvolution estimates. *Biometrika* **103** 1–20. doi: 10.1093/biomet/asv068.

Efron, B. and Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika* **65** 457–487. With comments by O. Barndorff-Nielsen, A.T. James, G.K. Robinson and D.A. Sprott, and a reply by the authors. MR521817

Fisher, R. A. (1922). On the Mathematical Foundations of Theoretical Statistics. *Phil. Trans. Roy. Soc. London Ser. A* **222** 309–368. doi: 10.2307/91208.

Fisher, R. A. (1925). Theory of statistical estimation. *Math. Proc. Cambridge Phil. Soc.* **22** 700–725. doi: 10.1017/S0305004100009580.

Fisher, R. A. (1934). Two new properties of mathematical likelihood. *Proc. Roy. Soc. London Ser. A* **144** 285–307. jstor.org/stable/2935559.

Good, I. J. and Gaskins, R. A. (1971). Nonparametric roughness penalties for probability densities. *Biometrika* **58** 255–277. MR0319314

Hayashi, M. and Watanabe, S. (2016). Information geometry approach to parameter estimation in Markov chains. *Ann. Statist.* **44** 1495–1535. MR3519931

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67. doi: 10.1080/00401706.1970.10488634.

Kass, R. E. and Vos, P. W. (2011). *Geometrical Foundations of Asymptotic Inference. Wiley Series in Probability and Statistics*. John Wiley & Sons, Inc.

Lindsey, J. K. (1974). Construction and comparison of statistical models. *J. Roy. Statist. Soc. Ser. B* **36** 418–425. MR0365794

Madsen, L. T. (1979). The geometry of statistical model—a generalization of curvature. Technical Report, Danish Medical Research Council. Statistical Research Unit Report 79-1.

Rao, C. R. (1961). Asymptotic efficiency and limiting information. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I* 531–545. Univ. California Press, Berkeley, Calif. MR0133192

Rao, C. R. (1962). Efficient estimates and optimum inference procedures in large samples. *J. Roy. Statist. Soc. Ser. B* **24** 46–72. MR0293766

Rao, C. R. (1963). Criteria of estimation in large samples. *Sankhyā Ser. A* **25** 189–206. MR0175225

Schwartzman, A., Dougherty, R. F. and Taylor, J. E. (2005). Cross-subject comparison of principal diffusion direction maps. *Magn. Reson. Med.* **53** 1423–1431. doi: 10.1002/mrm.20503.

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

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **81** 82–86. MR830567

Department of Statistics
Sequoia Hall, 390 Serra Mall
Stanford, CA 94305
E-mail: brad@stat.stanford.edu