

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. In multiparameter exponential families, the maximum likelihood estimate (MLE) for the vector parameter is a sufficient statistic. This is no longer true in non-exponential families. Fisher (1925, 1934) argued persuasively that in this case data sets yielding the same MLE could differ greatly in their estimation accuracy. This effect was examined in Efron (1975, 1978), where “statistical curvature” was introduced as a measure of non-exponentiality. One-parameter families were the only ones considered there.

The goal of this paper is to quantitatively examine curvature effects 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,

$$(1.1) \quad y_i \stackrel{\text{ind}}{\sim} \text{Poi}(\mu_i) \quad \text{for } i = 1, 2, \dots, 7.$$

We assume that

$$(1.2) \quad \mu_i = \alpha_1 + \alpha_2 x_i, \quad x = (-3, -2, -1, 0, 1, 2, 3).$$

The vector of observed values y_i was

$$(1.3) \quad y = (1, 1, 6, 11, 7, 14, 15).$$

Direct numerical maximization yielded

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

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

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

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

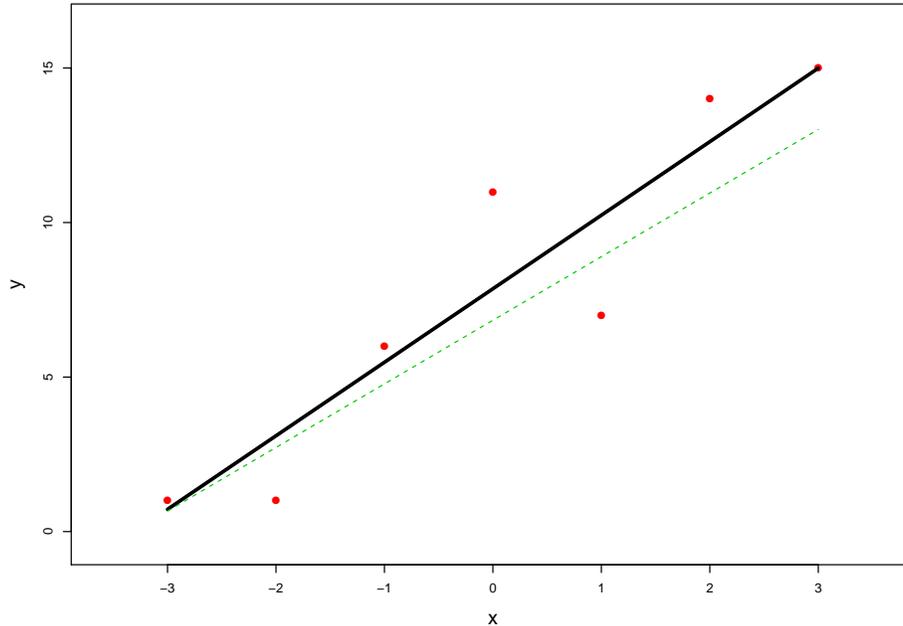


FIG 1. Toy example of a curved exponential family, (1.1)–(1.3); Poisson observation y_i (dots) are assumed to follow linear model $\mu_i = \alpha_1 + \alpha_2 x_i$ for $x_i = -3, -2, \dots, 3$; heavy line is MLE fit $\hat{\alpha} = (7.86, 2.38)$. Light dashed line is penalized MLE of Section 4.

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* $-\ddot{l}_{\hat{\alpha}}(y)$ (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 (1982) developed a full multiparametric theory of curved exponential families. His impressive results, and those of Madsen (1979), concentrate on the second order estimation accuracy of the MLE and its competitors. Their concerns are, in a sense to be made clear, orthogonal to the ones in this paper.

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

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

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

$$(2.2) \quad \boldsymbol{\mu}_{\boldsymbol{\eta}} = E_{\boldsymbol{\eta}}\{\mathbf{y}\} \quad \text{and} \quad \mathbf{V}_{\boldsymbol{\eta}} = \text{cov}_{\boldsymbol{\eta}}\{\mathbf{y}\},$$

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

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

$$(2.3) \quad \boldsymbol{\eta} = \boldsymbol{\eta}_{\alpha},$$

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

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

For simplified notation we write

$$(2.5) \quad \boldsymbol{\mu}_{\alpha} = \boldsymbol{\mu}_{\boldsymbol{\eta}_{\alpha}} \quad \text{and} \quad \mathbf{V}_{\alpha} = \mathbf{V}_{\boldsymbol{\eta}_{\alpha}}.$$

It is assumed that $\boldsymbol{\eta}_{\alpha}$ stays within the convex set of possible $\boldsymbol{\eta}$ vectors in (2.1), say $\alpha \in A$. The family

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

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 ,

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

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

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

The log likelihood function corresponding to (2.4) is

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

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

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

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

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

$\mathbf{0}$ here indicating a p -vector of zeroes.

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

$$(2.12) \quad \mathcal{I}_{\alpha} = \dot{\boldsymbol{\eta}}'_{\alpha} \mathbf{V}_{\alpha} \dot{\boldsymbol{\eta}}_{\alpha}.$$

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

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

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

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

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

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

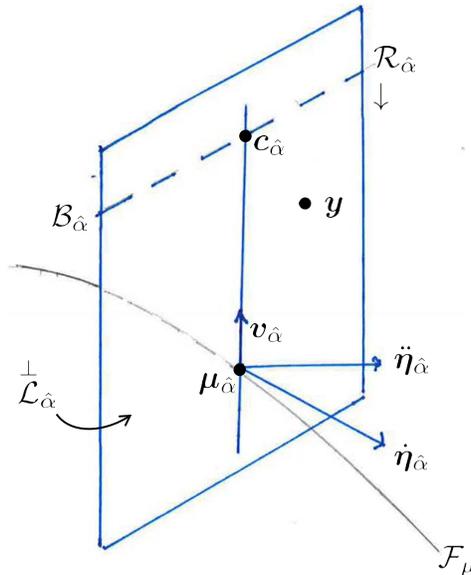


FIG 2. *Maximum likelihood estimation in a one-parameter curved exponential family* (2.16). Observed vector \mathbf{y} in $\perp\!\!\!\perp_{\hat{\alpha}}$ (2.17) gives MLE $\hat{\alpha}$; the observed Fisher information $\hat{\mathbf{I}}(\mathbf{y})$ (2.15) varies linearly in $\perp\!\!\!\perp_{\hat{\alpha}}$, equaling zero along the critical boundary $\mathcal{B}_{\hat{\alpha}}$ (2.26). Closest point $\mathbf{c}_{\hat{\alpha}}$ 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 \mathbf{y} values below $\mathcal{B}_{\hat{\alpha}}$; $\hat{\alpha}$ is a local maximum of the likelihood only for $\mathbf{y} \in \mathcal{R}_{\hat{\alpha}}$.

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

$$(2.16) \quad \mathcal{F}_\mu = \{\boldsymbol{\mu}_\alpha = E_\alpha\{\mathbf{y}\}, \alpha \in A\}.$$

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

$$(2.17) \quad \perp\!\!\!\perp(\dot{\eta}_{\hat{\alpha}}) = \{\mathbf{y} : \dot{\eta}'_{\hat{\alpha}}(\mathbf{y} - \boldsymbol{\mu}_{\hat{\alpha}}) = 0\},$$

denoted more simply as $\perp\!\!\!\perp_{\hat{\alpha}}$.

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

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

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

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

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

$$(2.20) \quad \langle \boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \rangle = \boldsymbol{\eta}_1' \mathbf{V}_{\hat{\alpha}} \boldsymbol{\eta}_2,$$

is

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

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

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

see Remark B of Section 6. It satisfies

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

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

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

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

(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* $\mathbf{c}_{\hat{\alpha}}$,

$$(2.25) \quad \mathbf{c}_{\hat{\alpha}} = \boldsymbol{\mu}_{\hat{\alpha}} + \mathbf{v}_{\hat{\alpha}}/\gamma_{\hat{\alpha}},$$

and the *critical boundary*

$$(2.26) \quad \mathcal{B}_{\hat{\alpha}} = \{\mathbf{y} = \boldsymbol{\mu}_{\hat{\alpha}} + \mathbf{c}_{\hat{\alpha}} + \mathbf{r}, \dot{\boldsymbol{\eta}}_{\hat{\alpha}}' \mathbf{r} = \ddot{\boldsymbol{\eta}}_{\hat{\alpha}}' \mathbf{r} = 0\}$$

(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 $\mathcal{B}_{\hat{\alpha}}$,

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

as the *region of stability*. Only for \mathbf{y} in $\mathcal{R}_{\hat{\alpha}}$ does the local stability equation $\dot{\boldsymbol{\eta}}_{\hat{\alpha}}'(\mathbf{y} - \boldsymbol{\mu}_{\hat{\alpha}}) = 0$ 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 $\mathbf{c}_{\hat{\alpha}}$ is infinitely far from $\boldsymbol{\mu}_{\hat{\alpha}}$ and $\mathcal{R}_{\hat{\alpha}}$ is all of $\mathcal{L}_{\hat{\alpha}}^{\perp}$. Otherwise $\gamma_{\hat{\alpha}}$ is > 0 , larger values moving $\mathcal{B}_{\hat{\alpha}}$ closer to $\boldsymbol{\mu}_{\hat{\alpha}}$.
- The point $\mathbf{y} = \boldsymbol{\mu}_{\hat{\alpha}} + b\mathbf{v}_{\hat{\alpha}}$ is Mahalanobis distance

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

from $\boldsymbol{\mu}_{\hat{\alpha}}$, which is the minimum distance for points $\boldsymbol{\mu}_{\hat{\alpha}} + b\mathbf{v}_{\hat{\alpha}} + \mathbf{r}$.

- The usual estimate for the standard deviation of $\hat{\alpha}$ is

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

Fisher suggested instead using the observed information,

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

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,

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

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

- A large value of $\gamma_{\hat{\alpha}}$ is worrisome from a frequentist point of view even if \mathbf{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 \mathbf{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 $\mathbf{a}'\mathbf{D}\mathbf{b}$: \mathbf{D} equal $\mathbf{V}_{\hat{\alpha}}$, $\mathbf{V}_{\hat{\alpha}}^{-1}$, and the identity. As discussed in the next section, we can always transform \mathcal{F} to make $\mathbf{V}_{\hat{\alpha}}$ the identity, simplifying both the derivations and their interpretation. Figure 2 assumes this to be the case, with $\mathbf{v}_{\hat{\alpha}}$ lying along $\hat{\boldsymbol{\eta}}_{\hat{\alpha}}$ (2.21), and $\mathcal{B}_{\hat{\alpha}}$ orthogonal to $\mathbf{v}_{\hat{\alpha}}$.

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 $\mathcal{L}^{\perp}(\hat{\boldsymbol{\eta}}_{\hat{\alpha}})$ (2.17). Here we return to multiparameter curved families $\mathcal{F} = \{f_{\alpha}(\mathbf{y}), \alpha \in A\}$ (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 $\mathcal{L}^{\perp}(\hat{\boldsymbol{\eta}}_{\hat{\alpha}})$, 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 $g_{\boldsymbol{\eta}}(\mathbf{y})$ (2.1). Let

$$(3.1) \quad \mathbf{M} = \mathbf{V}_{\hat{\alpha}}^{1/2},$$

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

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

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

$$(3.3) \quad g_{\boldsymbol{\eta}^\dagger}(\mathbf{y}^\dagger) = e^{\boldsymbol{\eta}'\mathbf{y}^\dagger - \psi^\dagger(\boldsymbol{\eta}^\dagger)} \left[e^{\boldsymbol{\eta}'\mathbf{M}^{-1}\boldsymbol{\mu}_{\hat{\alpha}}} g_0^\dagger(\mathbf{y}^\dagger) \right]$$

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

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

$$(3.4) \quad \boldsymbol{\eta}_\alpha^\dagger = \mathbf{M}\boldsymbol{\eta}_\alpha,$$

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

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

that is, that $(\boldsymbol{\eta}, \mathbf{y})$ 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

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

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

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

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

$$(3.8) \quad \dot{\eta}_u = \dot{\boldsymbol{\eta}}_{\hat{\alpha}}u,$$

(2.7), and similarly

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

where $\ddot{\eta}_u$ is the n -vector with i th component $\sum_j \sum_k \ddot{\boldsymbol{\eta}}_{\hat{\alpha}ijk}u_ju_k$. Using (3.8), the Fisher information \mathcal{I}_u , at $\lambda = 0$ in \mathcal{F}_u , is

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

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

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

LEMMA 1.

$$(3.11) \quad \hat{I}_u(\mathbf{y}) = u'\hat{\mathbf{I}}(\mathbf{y})u.$$

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

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

the bottom line again invoking linearity. \square

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

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

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

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

as at (2.21), and curvature (2.19)

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

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

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

It points toward $\mathbf{c}_u = \mathbf{v}_u / \gamma_u$ (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 \mathbf{y} in $\overset{\perp}{\mathcal{L}}(\dot{\eta}_{\hat{\alpha}})$ lying beyond \mathcal{B}_u have $\hat{I}_u(\mathbf{y}) < 0$ and so

$$(3.17) \quad u' \hat{I}(\mathbf{y}) u < 0$$

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 $\overset{\perp}{\mathcal{L}}(\dot{\eta}_{\hat{\alpha}})$ looks like.

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

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

$$(3.18) \quad d_u = 1 / (\gamma_u \cos \theta_u),$$

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

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

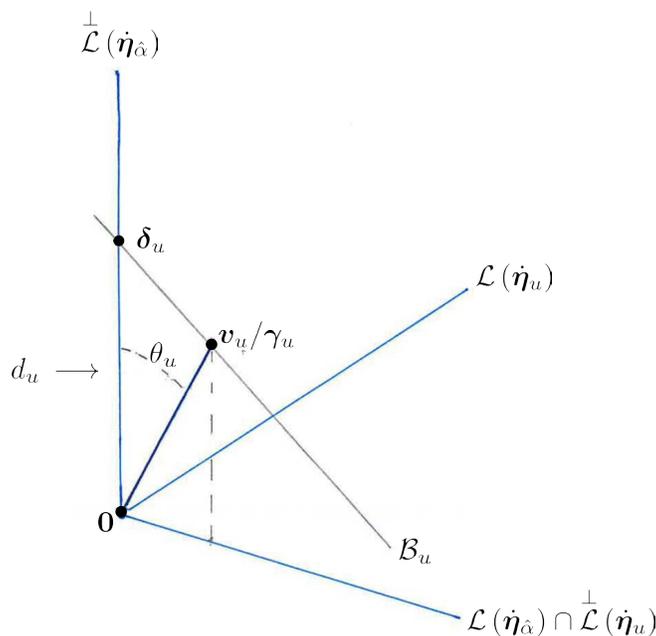


FIG 3. Curvature effects in multiparametric curved exponential families: p -dimensional unit vector u determines one-parameter curved subfamily \mathcal{F}_u and direction $\dot{\eta}_u$ (3.6)–(3.8) as well as curvature γ_u and critical vector \mathbf{v}_u/γ_u (3.15)–(3.16), tilted at angle θ_u to the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

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

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

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

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

for use in (3.18).

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

$$(3.21) \quad \mathbf{w}_u = \overset{\perp}{\mathbf{P}} \mathbf{v}_u / \cos \theta_u.$$

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

$$(3.22) \quad \mathbf{y} = b\mathbf{w}_u + \mathbf{r}$$

we have

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

which by Lemma 1 implies

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

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

Remark D in Section 6 verifies Theorem 2.

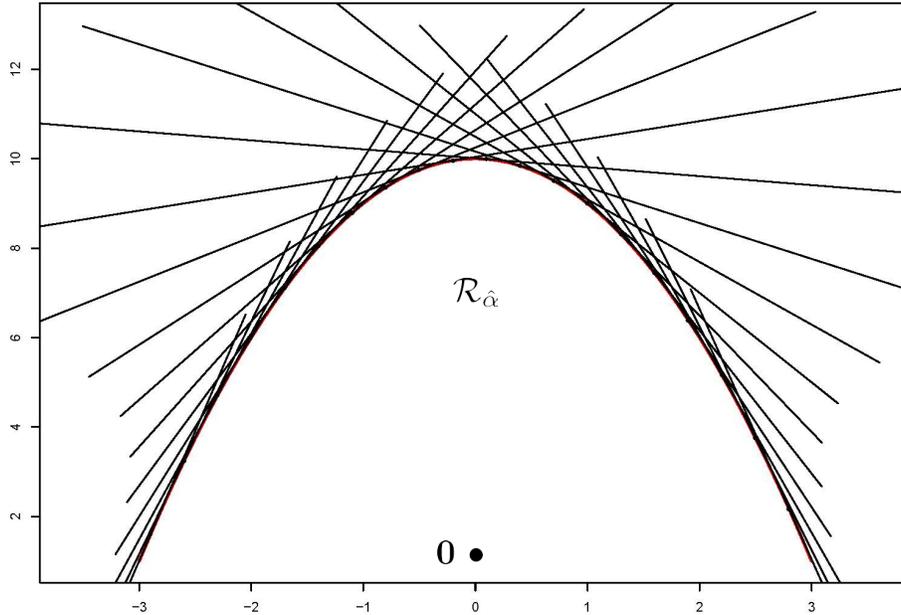


FIG 4. Schematic illustration showing construction of the region of stability $\mathcal{R}_{\hat{\alpha}}$. Each dot represents critical point δ_u for a particular choice of u , with its tangent line representing the boundary \mathcal{B}_u^\perp . The intersection of the half-spaces \mathcal{R}_u containing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

Now let \mathcal{R}_u denote the half-space of $\mathcal{L}(\hat{\eta}_{\hat{\alpha}})$ lying below \mathcal{B}_u^\perp , 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,

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

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

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

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

PROOF. If \mathbf{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 $u' \hat{\mathbf{I}}(\mathbf{y})u < 0$. However for \mathbf{y} in $\mathcal{R}_{\hat{\alpha}}$, b must be less than $1/(\gamma_u \cos \theta_u)$ for all u , implying

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

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

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

$$(3.27) \quad u(t) = (\cos t, \sin t),$$

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

$$(3.28) \quad \mathbf{w}_u = (-.390, .712, .392, .145, -.049, -.210, -.340),$$

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

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

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

The information matrix in the toy example was

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

The observed information matrix $\hat{\mathbf{I}}(b\mathbf{w}_u)$ 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 $u(t)' \hat{\mathbf{I}}(b\mathbf{w}_u) u(t)$ to zero, as in (3.24). Boundary distance 2.85 is small enough to allow substantial differences between $\hat{\mathbf{I}}(\mathbf{y})$ and $\mathcal{I}_{\hat{\alpha}}$. For example, in a Monte Carlo simulation of model (1.1)–(1.2), with $\alpha = \hat{\alpha}$, the ratio of lower right corner elements $\hat{\mathbf{I}}_{22}/\mathcal{I}_{\hat{\alpha}22}$ 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 $\mathcal{L}(\hat{\boldsymbol{\eta}}_{\hat{\alpha}})$, as seen in the covariance approximation

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

while stability refers to behavior in the orthogonal space $\mathcal{L}(\hat{\boldsymbol{\eta}}_{\hat{\alpha}})^\perp$. The full differential geometric developments in Amari (1982) and Madsen (1979) aim toward second-order accuracy expressions for assessing $\text{cov}(\hat{\alpha})$, and in this sense are orthogonal to the considerations here. Following Amari's influential 1982 paper there has been continued interest in differential geometric curved family inference, centered in the Japanese school; see for example Hayashi and Watanabe (2016).

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 $m_\alpha(\mathbf{y})$ to be

$$(4.1) \quad m_\alpha(\mathbf{y}) = l_\alpha(\mathbf{y}) - s_\alpha,$$

where $l_\alpha(\mathbf{y})$ is the usual log likelihood $\log f_\alpha(\mathbf{y})$, 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

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

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 $\dot{m}_{\hat{\alpha}}(\mathbf{y}) = 0$,

$$(4.3) \quad \hat{\alpha} : \quad \dot{m}_{\hat{\alpha}}(\mathbf{y}) = \dot{\eta}'_{\hat{\alpha}}(\mathbf{y} - \boldsymbol{\mu}_{\hat{\alpha}}) - \dot{s}_{\hat{\alpha}} = \mathbf{0},$$

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 \mathbf{y} satisfying (4.3) is an $(n - p)$ -dimensional hyperplane

$$(4.4) \quad \mathcal{M}_{\hat{\alpha}}^\perp = \{ \mathbf{y} : \dot{\eta}'_{\hat{\alpha}}(\mathbf{y} - \boldsymbol{\mu}_{\hat{\alpha}}) = \dot{s}_{\hat{\alpha}} \};$$

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

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

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

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

$$(4.6) \quad \mathbf{y}_{\hat{\alpha}} = \dot{\eta}_{\hat{\alpha}}\mathcal{I}_{\hat{\alpha}}^{-1}\dot{s}_{\hat{\alpha}} \quad \text{with} \quad \|\mathbf{y}_{\hat{\alpha}}\|^2 = \dot{s}'_{\hat{\alpha}}\mathcal{I}_{\hat{\alpha}}^{-1}\dot{s}_{\hat{\alpha}}.$$

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

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

$\ddot{s}_{\hat{\alpha}}$ being the $p \times p$ second derivative matrix $(\partial^2 s_\alpha / \partial \alpha_j \partial \alpha_k)$. $\hat{\mathbf{J}}(\mathbf{y})$ 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

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

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 $\exp\{m_\alpha(\mathbf{y})\}$ is proportional to the posterior density of α given \mathbf{y} . Applying Laplace's method, as in Tierney and Kadane (1986), yields the normal approximation

$$(4.9) \quad \alpha \mid \mathbf{y} \sim \mathcal{N}_p(\hat{\alpha}, \hat{\mathbf{J}}(\mathbf{y})^{-1}).$$

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

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

LEMMA 3.

$$(4.10) \quad \hat{J}_u(\mathbf{y}) = u' \hat{\mathbf{J}}(\mathbf{y}) u.$$

PROOF. Let

$$(4.11) \quad \dot{s}_u = u' \dot{s}_{\hat{\alpha}} \quad \text{and} \quad \ddot{s}_u = u' \ddot{s}_{\hat{\alpha}} u,$$

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),

$$(4.12) \quad \begin{aligned} \hat{J}_u(\mathbf{y}) &= \mathcal{I}_u - \ddot{\eta}'_u \mathbf{y} + \ddot{s}_u \\ &= u' \mathcal{I}_{\hat{\alpha}} u - (u' \ddot{\eta}'_{\hat{\alpha}} u)' \mathbf{y} + u' \ddot{s}_{\hat{\alpha}} u = u' \hat{\mathbf{J}}(\mathbf{y}) u. \end{aligned} \quad \square$$

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

$$(4.13) \quad \mathbf{v}_u = \dot{\eta}_u^\perp / (\mathcal{I}_u \gamma_u),$$

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

$$(4.14) \quad \dot{\mathcal{M}}_u^\perp = \{\mathbf{y} : \dot{\eta}'_u \mathbf{y} = \dot{s}_u\}$$

passing through its nearest point to the origin

$$(4.15) \quad \mathbf{y}_u = \dot{\eta}_u \mathcal{I}_u^{-1} \dot{s}_u$$

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

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

$$(4.16) \quad \Delta_u = \mathbf{y}_u - \mathbf{y}_{\hat{\alpha}} = \dot{\eta}_u \mathcal{I}_u^{-1} \dot{s}_u - \dot{\eta}_{\hat{\alpha}} \mathcal{I}_{\hat{\alpha}}^{-1} \dot{s}_{\hat{\alpha}},$$

lies in $\dot{\mathcal{L}}(\dot{\eta}_u) \cap \dot{\mathcal{L}}(\dot{\eta}_{\hat{\alpha}})$; $\Delta_u = 0$ 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 $\dot{\mathcal{M}}_u^\perp$ at which $\hat{J}_u(\mathbf{y})$ is zero:

LEMMA 4. For any vector \mathbf{r} in $\dot{\mathcal{M}}_u^\perp$ orthogonal to the unit vector \mathbf{v}_u ,

$$(4.17) \quad \hat{J}_u(\mathbf{y}_u + b\mathbf{v}_u + \mathbf{r}) = \mathcal{I}_u(1 - b\gamma_u) - \frac{\nu_{u12}}{\nu_{u11}} \dot{s}_u + \ddot{s}_u;$$

$\hat{J}_u(\mathbf{y}_u + b\mathbf{v}_u + \mathbf{r})$ equals 0 for $b = c_u$,

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

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

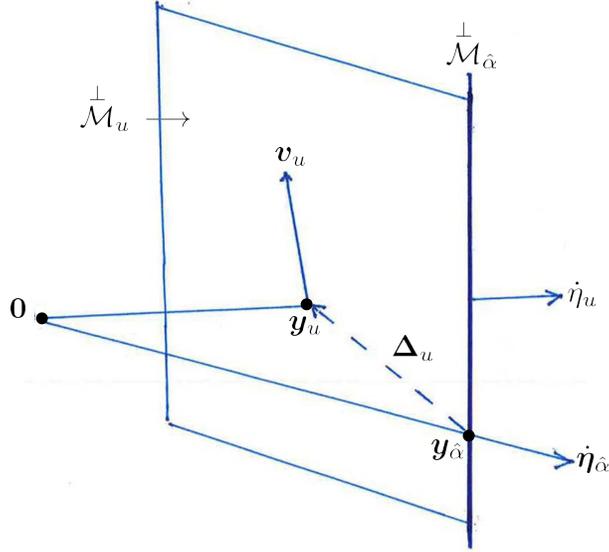


FIG 5. Penalized maximum likelihood estimation: $\mathcal{M}_{\hat{\alpha}}^{\perp}$ is the $(n-p)$ -dimensional hyperplane containing those observation vectors \mathbf{y} having pMLE = $\hat{\alpha}$ (4.3); it is orthogonal to $\hat{\eta}_{\hat{\alpha}}$, passing through $\mathbf{y}_{\hat{\alpha}}$,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

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

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

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

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

But

$$(4.21) \quad \ddot{\eta}'_u(b\mathbf{v}_u + \mathbf{r}) = b\dot{\eta}'_u \mathbf{v}_u = b\mathcal{I}_u \gamma_u$$

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

Each choice of u produces a bounding hyperplane \mathcal{B}_u^{\perp} in $\mathcal{M}_{\hat{\alpha}}^{\perp}$, the boundary being the intersection of \mathcal{B}_u with $\mathcal{M}_{\hat{\alpha}}^{\perp}$ (as in Lemma 2). Each \mathcal{B}_u^{\perp} 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 \mathcal{B}_u^{\perp} is obtained as an extension of Theorem 2. Let $\mathbf{w}_u = \frac{\perp}{\mathbf{P}} \mathbf{v}_u / \cos \theta_u$ as in (3.21), the unit projection of \mathbf{v}_u into $\mathcal{M}_{\hat{\alpha}}^{\perp}$.

THEOREM 4. For

$$(4.22) \quad \mathbf{y} = \mathbf{y}_{\hat{\alpha}} + b\mathbf{w}_u + \mathbf{r},$$

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

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

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

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

PROOF.

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

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

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

and

$$(4.27) \quad \ddot{\eta}'_u b\mathbf{w}_u = b \frac{\perp}{\eta}'_u \mathbf{w}_u = b \|\frac{\perp}{\eta}_u\| \cos \theta_u = b \mathcal{I}_u \gamma_u \cos \theta_u.$$

Also

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

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

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

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

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 $-\dot{l}_{\hat{\alpha}}(\mathbf{y}) = \hat{\mathbf{I}}(\mathbf{y})$ rather than $\hat{\mathbf{J}}(\mathbf{y})$. The only change needed is to remove the \dot{s}/ν_{u11} term in the definition of c_u (4.18), after which $\hat{I}_u(\mathbf{y})$ can replace $\hat{J}_u(\mathbf{y})$ 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 $\Theta_1, \Theta_2, \dots, \Theta_N$ of unobserved parameters sampled from an unknown density function $g(\theta)$,

$$(5.1) \quad \Theta_i \stackrel{\text{ind}}{\sim} g(\theta), \quad \text{for } i = 1, 2, \dots, N.$$

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

$$(5.2) \quad X_i | \Theta_i \stackrel{\text{ind}}{\sim} p(X_i | \Theta_i),$$

for example,

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

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

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

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

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

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

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

$$(5.6) \quad X_i \stackrel{\text{ind}}{\sim} f_\alpha(x_i), \quad \text{for } i = 1, 2, \dots, N,$$

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 $\hat{\mathbf{I}}(y)$ and $\mathcal{I}_{\hat{\alpha}}$ can be strikingly different—or, worse, where \mathbf{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,

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

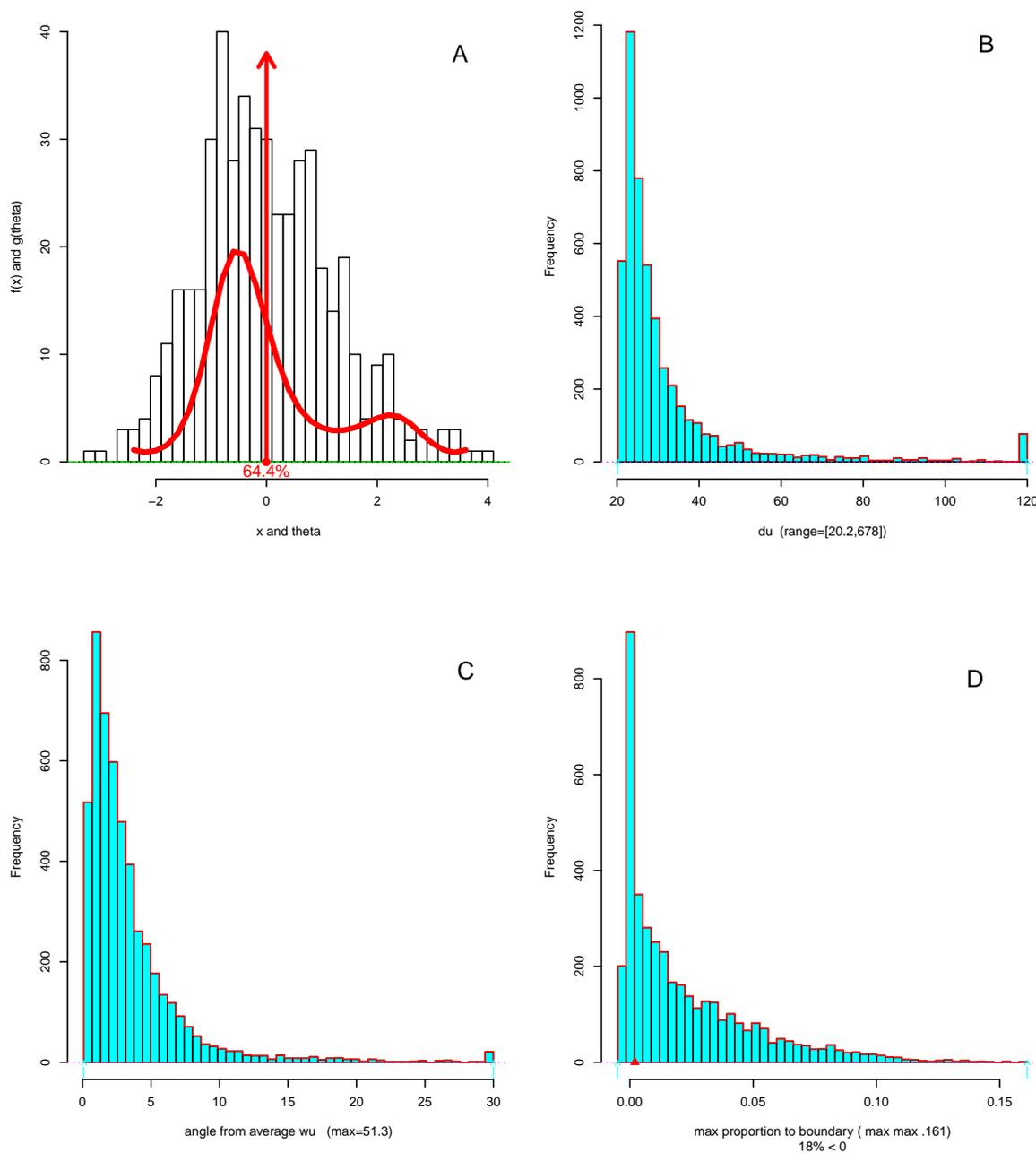


FIG 6. **Panel A:** Histogram of the 477 observations X_i , and the estimated prior density $\mathbf{g}_{\hat{\alpha}}$ based on spike and slab prior (5.16); it estimates $\Pr\{\Theta = 0\} = 0.644$. **Panel B:** Histogram of critical distances d_u (4.24) for 5000 randomly selected vectors u (5.17).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

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

$$(5.8) \quad \mathbf{x} = (-3.2, -3.0, \dots, 4.0).$$

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

$$(5.9) \quad \mathbf{y} = (y_1, y_2, \dots, y_n),$$

gives the heights of the histogram bars in Panel A. We will work with data vector \mathbf{y} rather than \mathbf{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 \mathbf{g} ,

$$(5.10) \quad \mathbf{g} = (g_1, g_2, \dots, g_m),$$

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

$$(5.11) \quad g_{\alpha k} = e^{Q'_k \alpha} / C_\alpha,$$

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 \mathbf{Q} , having k th row Q'_k , determines the exponential family of possible priors (5.4).

Define

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

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

$$(5.13) \quad \mathbf{P} = (p_{kj}, k = 1, 2, \dots, n \text{ and } j = 1, 2, \dots, m).$$

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

$$(5.14) \quad \mathbf{f}_\alpha = \mathbf{P}\mathbf{g}_\alpha.$$

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

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

the last indicating a multinomial distribution on n categories, sample size N , probability vector \mathbf{f}_α . (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,

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

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 $\mathbf{g}_{\hat{\alpha}}$ 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 $\mathbf{g}_{\hat{\alpha}}$, 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 \mathbf{y} ?

Exponential family model (5.11) leads to simple expression for $\boldsymbol{\eta}_\alpha$ and $\boldsymbol{\eta}''_\alpha$, (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 $\mathbf{y} \sim \text{Mult}_n(N, \mathbf{f}_\alpha)$, 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 $\mathbf{g}_{\hat{\alpha}}$.

The calculation for $\mathcal{R}_{\hat{\alpha}}$ were done after transformation to standardized coordinates having $\boldsymbol{\mu}_{\hat{\alpha}} = \mathbf{0}$ and $\mathbf{V}_{\hat{\alpha}} = \mathbf{I}_n$ (3.5), this being assumed from now on. The construction of $\mathcal{R}_{\hat{\alpha}}$ pictured in Figure 4 is 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 from \mathcal{S}^8 , the surface of the unit sphere in eight dimensions,

$$(5.17) \quad u_1, u_2, \dots, u_{5000}.$$

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

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

$$(5.18) \quad 20.2 \leq d_u \leq 678.$$

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 $\bar{\mathbf{w}} = \sum \mathbf{w}_{uh}/5000$ be the average \mathbf{w}_u direction vector. Panel C is a histogram of the angle in degrees between \mathbf{w}_{uh} and $\bar{\mathbf{w}}$. We see a close clustering around $\bar{\mathbf{w}}$, 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 \mathbf{Y}^* were generated,

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

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

$$(5.20) \quad m_i = \max_h \{\mathbf{y}_i^{*'} \mathbf{w}_{uh} / d_{uh}, h = 1, 2, \dots, 5000\},$$

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

In fact, Panel D of Figure 6 shows $m_i \leq 0.161$ for all i . For the actual observation vector \mathbf{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 \mathbf{y} , and would be unlikely to vary much for other possible observations \mathbf{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.

Panel D shows that 18% of the 4000 \mathbf{y}_i^* vectors gave m_i less than zero. That is, \mathbf{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 degrees — 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 $\dot{l}_\alpha(\mathbf{y}) = \dot{\eta}'_\alpha(\mathbf{y} - \boldsymbol{\mu}_\alpha)$ (2.10),

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

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

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

which is (2.13).

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

$$(6.3) \quad \dot{\eta}_{\hat{\alpha}}^{\perp\dagger} = \dot{\eta}_{\hat{\alpha}}^\dagger - \frac{\nu_{12}}{\nu_{11}} \eta_{\hat{\alpha}}^\dagger,$$

with length

$$(6.4) \quad \|\dot{\eta}_{\hat{\alpha}}^{\perp\dagger}\| = \nu_{22} - \nu_{12}^2/\nu_{11} = \mathcal{I}_{\hat{\alpha}} \gamma_{\hat{\alpha}},$$

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

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

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

Also

$$(6.6) \quad \mathbf{v}_{\hat{\alpha}} = \mathbf{V}_{\hat{\alpha}}^\perp \dot{\eta}_{\hat{\alpha}}^\dagger / \mathcal{I}_{\hat{\alpha}} \gamma_{\hat{\alpha}} = \mathbf{M} \mathbf{v}_{\hat{\alpha}}^\dagger$$

satisfies

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

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

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

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

Remark C. *Lemma 2* By rotations

$$(6.9) \quad \boldsymbol{\eta} \longrightarrow \boldsymbol{\Gamma} \boldsymbol{\eta} \quad \text{and} \quad \mathbf{y} \longrightarrow \boldsymbol{\Gamma}' \mathbf{y},$$

where $\boldsymbol{\Gamma} = (\gamma_1, \gamma_2, \dots, \gamma_n)$ 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 $\gamma_2, \gamma_3, \dots, \gamma_p$ to span $\mathcal{L}(\dot{\eta}_{\hat{\alpha}}) \cap \mathcal{L}(\dot{\eta}_u)^\perp$. (Notice that \mathbf{y} still has mean $\mathbf{0}$ and covariance \mathbf{I}_n .) For any n -vector \mathbf{z} , write

$$(6.10) \quad \mathbf{z} = (z_1, z_2, z_3),$$

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

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

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

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

$$(6.12) \quad \begin{aligned} \cos \theta_u &= \|v_{u3}\| / \|\mathbf{v}_u\| \\ &= \|v_{u3}\|. \end{aligned}$$

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

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

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

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

$$(6.14) \quad \boldsymbol{\delta}'_u \mathbf{v}_u = 1 / \gamma_u,$$

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

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

$$(6.15) \quad \begin{aligned} \ddot{\eta}'_u \mathbf{y} &= \dot{\eta}'_u \mathbf{y} = \mathcal{I}_u \gamma_u \mathbf{v}'_u \mathbf{y} = \\ &= \mathcal{I}_u \gamma_u \mathbf{v}'_u (b\mathbf{w}_u + \mathbf{r}) = \mathcal{I}_u \gamma_u \cos \theta_u \cdot b. \end{aligned}$$

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

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

$$(6.16) \quad \ddot{\eta}_{\hat{\alpha}i} = -\frac{1}{\mu_{\hat{\alpha}i}} \begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \end{pmatrix}.$$

Remark F. *Influence function* (4.8) From

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

(4.3), we get the local relationship

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

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

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

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

The linear expansion (4.8) suggests the covariance approximation

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

for the pMLE (Efron, 2016, Thm. 2), in contrast with the Bayesian covariance estimate $\hat{\mathbf{J}}(\mathbf{y})^{-1}$ 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 \mathbf{y} with $2\mathbf{y}$) 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 \mathbf{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. $\dot{\eta}_{\hat{\alpha}}$ and $\ddot{\eta}_{\hat{\alpha}}$ for g -models Section 2 of Efron (2016) calculates $\dot{\eta}_{\hat{\alpha}}$ and $\ddot{\eta}_{\hat{\alpha}}$ for model (5.15): define

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

giving the $m \times n$ matrix $\mathbf{W}(\alpha) = (w_{kj}(\alpha))$, having k th column $\mathbf{W}_k(\alpha) = (w_{k1}(\alpha), \dots, w_{km}(\alpha))'$. Then

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

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

$$(6.23) \quad \mathbf{Q}' [\text{diag } \mathbf{W}_k(\alpha) - \mathbf{W}_k(\alpha) \mathbf{W}_k(\alpha)' - \mathbf{W}_k(\alpha) \mathbf{g}_{\alpha} - \mathbf{g}_{\alpha} \mathbf{W}_k(\alpha)'] \mathbf{Q},$$

$\text{diag } \mathbf{W}_k(\alpha)$ 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

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

(see Remark J), changing the previous expressions to

$$(6.25) \quad \dot{\eta}_{\hat{\alpha}}^{\dagger} = \mathbf{M} \dot{\eta}_{\hat{\alpha}} \quad \text{and} \quad \ddot{\eta}_{\hat{\alpha}}^{\dagger} = \mathbf{M} \ddot{\eta}_{\hat{\alpha}};$$

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

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

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

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

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

which represents identity covariance matrix in the $(n-1)$ -dimensional linear space of \mathbf{y}_i^* 's variability, justifying (6.24).

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

$$(6.28) \quad y_k \stackrel{\text{ind}}{\sim} \text{Poi}(\mu_{\alpha k}) \quad \text{for } k = 1, 2, \dots, n,$$

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 $\mathbf{M} = \text{diag}(N\mathbf{f}_{\hat{\alpha}})^{1/2}$ 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 $\nu_{u11} = \dot{\eta}'_u \mathbf{V}_{\hat{\alpha}} \dot{\eta}_u$, 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

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

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

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,

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

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. (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](http://www.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.
- 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.
- 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