The Bayes deconvolution problem

Bradley Efron*†
Stanford University

Abstract

An unknown prior density $g(\theta)$ has yielded realizations $\Theta_1, \Theta_2, \ldots, \Theta_N$. They are unobservable, but each Θ_i produces an observable value X_i according to a known probability mechanism, for instance $X_i \sim \text{Poisson}(\Theta_i)$. We wish to estimate $g(\theta)$ from the observed sample $X_1, X_2, \ldots, X_N$. Traditional asymptotic calculations are discouraging, indicating very slow nonparametric rates of convergence. Here we show that parametric exponential family modeling of $g(\theta)$ can give useful estimates in moderate-sized samples. A variety of real and artificial examples illustrates the methodology. Covariate information can be incorporated into the deconvolution process, leading to a more detailed theory of Generalized Linear Mixed Models.

Keywords: g-modeling, generalized mixed models, exponential family models, Fourier deconvolution, frailty

1 Introduction

We are interested in the following situation: an *unknown* probability density $g(\theta)$ yields an *unobserved* random sample of realizations $\Theta_1, \Theta_2, \ldots, \Theta_N$,

$$\Theta_i \overset{\text{iid}}{\sim} g(\theta), \qquad i = 1, 2, \ldots, N; \tag{1.1}$$

each Θ_i independently produces an *observed* random variable X_i according to a *known* family of probability densities for X_i given Θ_i,

$$X_i \overset{\text{ind}}{\sim} p_i(X_i|\Theta_i); \tag{1.2}$$

finally, from the observed sample $\boldsymbol{X} = (X_1, X_2, \ldots, X_N)$ we wish to estimate the prior density $g(\theta)$.

In the second example of Section 4, X_i is a binomial variate, observed after n_i independent draws each of probability Θ_i,

$$X_i \sim \text{Binom}(n_i, \Theta_i), \tag{1.3}$$

so $p_i(x_i|\theta_i)$ is the corresponding binomial density function, a discrete density in this case. The n_i differ, which is why we need the extra subscript on $p_i(\cdot|\cdot)$.

There are at least two reasons to be interested in estimating the prior density $g(\theta)$. First of all, we may want to learn ensemble properties such as $E\{\Theta\}$ or $\Pr\{\Theta = 0\}$. This is the case in the first example of Section 4, where the Θ_i are effect sizes in a microarray experiment and $\Pr\{\Theta = 0\}$ is the proportion of "null genes". Empirical Bayes calculations, for instance of $\Pr\{\Theta_i = 0|X_i \geq 3\}$, provide the second reason, as emphasized in Efron (2014).

*Research supported in part by NIH grant 8R37 EB002784 and by NSF grant DMS 1208787.

†Sequoia Hall, 390 Serra Mall, Stanford University, Stanford, CA 94305-4065; brad@stat.stanford.edu

There is an impressive theoretical literature on the deconvolution problem, as in Laird (1978), Fan (1991), Hall and Meister (2007), and Butucea and Comte (2009), mostly focused on the *additive model*,

$$X_i = \Theta_i + \epsilon_i, \tag{1.4}$$

where the ϵ_i are an i.i.d. sample from a known density; typically

$$\epsilon_i \overset{\text{ind}}{\sim} \mathcal{N}(0,1) \tag{1.5}$$

so

$$X_i \sim \mathcal{N}(\Theta_i, 1). \tag{1.6}$$

The results are discouraging: asymptotic rates of convergence, of estimates $\hat{g}(\theta)$ to $g(\theta)$, are much slower than $N^{-1/2}$, as slow as $(\log N)^{-1}$ under general conditions. See for example Carroll and Hall (1988). A good part of the discouragement relates to nonparametric modeling of $g(\theta)$, allowing its fine structure to dictate convergence rates.

A more aggressive modeling approach, yielding more optimistic results, is taken here. Section 2 and Section 3 discuss low-parameter exponential family models for the prior density $g(\theta)$. Examples, both genuine and artificial, appear in Sections 2 through 6. They show the deconvolution problem as being difficult but feasible, at least in the modern "big data" context of sample sizes N in the hundreds or thousands.

Section 5 extends (1.1)–(1.2) to the situation where, in addition to X_i, the statistician observes a covariate vector u_i, the observed data being pairs

$$(X_i, u_i), \qquad i = 1, 2, \ldots, N. \tag{1.7}$$

This brings the deconvolution problem into the realm of "frailty" and generalized linear mixed models.

Let X_i denote the *marginal density* of X_i under model (1.1)–(1.2),

$$f_i(x_i) = \int p_i(X_i|\theta_i) g(\theta_i)\, d\theta_i, \tag{1.8}$$

the integral being taken over the Θ space $\mathcal{T}$. Effectively, the statistician only observes $X_i \overset{\text{ind}}{\sim} f_i(\cdot)$ for $i = 1, 2, \ldots, N$. Another approach to the deconvolution problem is to directly model the densities f_i (called "f-modeling" in Efron, 2014). The elegant Fourier deconvolution method of Stefanski and Carroll (1990), applying to the additive situation (1.4), is featured in Section 6, where efficiency comparisons are made between f-modeling and the exponential family "g-modeling" approach. Most of the derivations are deferred to Section 7, **Proofs and details**.

Our interest here is in the practical aspects of the deconvolutuion problem, where theoretical considerations — these being mainly of a standard nature in our exponential family framework — are of secondary concern. The label "Bayes deconvolution problem" is intended to emphasize the more general nature of situation (1.1)–(1.2) compared to (1.4) or (1.6). The likelihood methodology described in Section 2 accommodates a wide variety of applied problems, as the examples will show.

2 Likelihood and deconvolution

We will pursue a likelihood approach to the Bayes deconvolution problem (1.1)–(1.2), with the prior $g(\theta)$ modeled by an exponential family of densities on the Θ space $\mathcal{T}$. To simplify the presentation it is assumed that $\mathcal{T}$ is a finite, discrete set,

$$\mathcal{T} = \{\theta_1, \theta_2, \ldots, \theta_m\}. \tag{2.1}$$

(This is a convenience, not a necessity; see Remark A of Section 7.)

The prior $g(\theta)$ is now an m-vector $g = (g_1, g_2, \ldots, g_m)$ specifying probability g_j on θ_j,

$$g = g(\alpha) = e^{Q\alpha - \phi(\alpha)}. \tag{2.2}$$

Here α is a p-dimensional parameter vector while Q is a known $m \times p$ structure matrix, say with jth row Q_j'. Notation (2.2) indicates that the jth component of $g(\alpha)$ is

$$g_j(\alpha) = e^{Q_j'\alpha - \phi(\alpha)} \qquad \text{for } j = 1, 2, \ldots, m, \tag{2.3}$$

with function $\phi(\alpha)$ normalizing $g(\alpha)$ to sum to one,

$$\phi(\alpha) = \log \sum_{j=1}^{m} e^{Q_j'\alpha}. \tag{2.4}$$

Let

$$p_{ij} = p_i(X_i|\Theta_i = \theta_j) \tag{2.5}$$

be the probability that X_i equals its observed value if Θ_i equals θ_j, and define P_i as the m-vector of possible such probabilities for X_i,

$$P_i = (p_{i1}, p_{i2}, \ldots, p_{im})'. \tag{2.6}$$

In our discrete setting,the marginal probability (1.8) for X_i becomes

$$f_i(\alpha) = \sum_{j=1}^{m} p_{ij} g_j(\alpha) = P_i' g(\alpha). \tag{2.7}$$

The log likelihood function for parameter vector $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_p)'$ is

$$l_i(\alpha) = \log f_i(\alpha) = \log P_i' g(\alpha), \tag{2.8}$$

whose p-dimensional first derivative vector and $p \times p$-dimensional second derivative matrix,

$$\dot{l}_i(\alpha) = \left(\ldots \frac{\partial l_i(\alpha)}{\partial \alpha_h} \ldots \right)' \quad \text{and} \quad \ddot{l}_i(\alpha) = \left(\ldots \frac{\partial^2 l_i(\alpha)}{\partial \alpha_h \partial \alpha_k} \ldots \right), \tag{2.9}$$

will be needed for the maximum likelihood calculations of Section 3.

Lemma 1. *Define*

$$w_{ij}(\alpha) = g_j(\alpha)\left(p_{ij}/f_i(\alpha) - 1\right), \tag{2.10}$$

and let $W_i(\alpha)$ be the m-vector

$$W_i(\alpha) = (w_{i1}(\alpha), w_{i2}(\alpha), \ldots, w_{im}(\alpha))'. \tag{2.11}$$

Then

$$\dot{l}_i(\alpha) = Q' W_i(\alpha) \tag{2.12}$$

and

$$-\ddot{l}_i(\alpha) = Q'\left\{W_i(\alpha)W_i(\alpha)' + W_i(\alpha)g(\alpha)' + g(\alpha)W_i(\alpha)' - \operatorname{diag}\left(W_i(\alpha)\right)\right\} Q; \tag{2.13}$$

$\operatorname{diag}(v)$ *indicates a $p \times p$ diagonal matrix with diagonal elements obtained from the vector v. (Notice that the first three bracketed terms are outer products.)*

See Remark B of Section 7 for the derivations. An attractive alternate expression for $-\ddot{l}_i(\alpha)$ appears in (7.17), Remark C.

Summing over the N observations X_i, the total log likelihood $l(\alpha) = \sum_{i=1}^{N} l_i(\alpha)$ has

$$\dot{l}(\alpha) = \sum_{i=1}^{N} \dot{l}_i(\alpha) = Q'W_+(\alpha), \tag{2.14}$$

where

$$W_+(\alpha) = \sum_{i=1}^{N} W_i(\alpha). \tag{2.15}$$

Similarly,

$$-\ddot{l}(\alpha) = Q' \left\{ \sum_{i=1}^{N} W_i(\alpha)W_i(\alpha)' + W_+(\alpha)g(\alpha)' + g(\alpha)W_+(\alpha)' - \mathrm{diag}\left(W_+(\alpha)\right) \right\} Q. \tag{2.16}$$

Lemma 2. *The maximum likelihood estimate (MLE) $\hat{\alpha}$ for α satisfies*

$$Q'W_+(\hat{\alpha}) = 0, \tag{2.17}$$

while $-\ddot{l}(\hat{\alpha})$, the observed Fisher information matrix*, equals*

$$-\ddot{l}(\hat{\alpha}) = Q' \left\{ \sum_{i=1}^{N} W_i(\hat{\alpha})W_i(\hat{\alpha})' - \mathrm{diag}\left(W_+(\hat{\alpha})\right) \right\} Q. \tag{2.18}$$

The proof is immediate from (2.12), (2.13), after noting that the middle two terms in (2.16) vanish because of (2.17).

Efron (2014) discusses the "i.i.d. case" where $p_i(\cdot|\cdot)$ in (1.2) does not depend on i (as in (1.6)); also assuming that $\mathcal{X}$, the space of possible X values, is discrete, say

$$\mathcal{X} = \{x_1, x_2, \ldots, x_n\}. \tag{2.19}$$

We can then define the $n \times m$ matrix $\boldsymbol{P} = (p_{kj})$,

$$p_{kj} = p(X = x_k|\Theta = \theta_j), \tag{2.20}$$

with the n-vector of marginal probabilities $f_k(\alpha)$ given by

$$f(\alpha) = \boldsymbol{P}g(\alpha). \tag{2.21}$$

Let $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$ be the vector of counts,

$$y_k = \#\{X_i = x_k\}; \tag{2.22}$$

$\boldsymbol{y}$ is a sufficient statistic in the i.i.d. case, having a multinomial distribution of N draws on n categories with probability vector $f(\alpha)$,

$$\boldsymbol{y} \sim \mathrm{Mult}_n\left(N, f(\alpha)\right). \tag{2.23}$$

Now $W_+(\alpha) = \sum_1^n y_k W_k(\alpha)$, with $w_{kj}(\alpha)$ and $W_k(\alpha)$ as defined in (2.10)–(2.11), k replacing index i; the observed Fisher information (2.18) becomes

$$-\ddot{l}(\hat{\alpha}) = Q' \left\{ \sum_{k=1}^{n} W_k(\hat{\alpha})y_k W_k(\hat{\alpha})' - \mathrm{diag}\left(W_+(\hat{\alpha})\right) \right\} Q. \tag{2.24}$$

Lemma 3. *In the i.i.d. case* (2.23), *the* expected Fisher information matrix $\mathcal{I}(\alpha) = E_\alpha\{-\ddot{l}_{\boldsymbol{y}}(\alpha)\}$ *is*

$$\mathcal{I}(\alpha) = Q' \left\{ \sum_{k=1}^{n} W_k(\alpha)\,(Nf_k(\alpha))\,W_k(\alpha)' \right\} Q. \tag{2.25}$$

See Remark C for the derivation. The sufficient vector $\boldsymbol{y}$ is now explicitly denoted in $-\ddot{l}_{\boldsymbol{y}}(\alpha)$ since it is the random quantity in the frequentist calculation of $\mathcal{I}(\alpha)$.

The expectation of $\boldsymbol{y} \sim \text{Mult}_n(N, f(\alpha))$ is $Nf(\alpha)$. Comparing (2.24) with (2.25), we see that the latter is the former with each y_k replaced by $Nf_k(\alpha)$, except that the term $\text{diag}(W_+(\hat{\alpha}))$ is dropped. In fact, we have equality between the expected and observed Fisher information evaluated at $\boldsymbol{y} = Nf(\alpha)$:

Theorem 1. *In the i.i.d. case,*

$$\mathcal{I}(\hat{\alpha}) = -\ddot{l}_{Nf(\hat{\alpha})}(\hat{\alpha}) = Q' \left\{ \sum_{k=1}^{n} W_k(\hat{\alpha})\,(Nf_k(\hat{\alpha}))\,W_k(\hat{\alpha})' \right\} Q. \tag{2.26}$$

Proof. See Remark C. ■

The right-hand side of (2.26) can be thought of as a smoothed version of the observed information, where the parametric estimate $Nf(\hat{\alpha})$ is substituted for the nonparametric value $\boldsymbol{y}$. There is no obvious analogue of Theorem 1 for the non-i.i.d. case. In general however it suggests ignoring the term $\text{diag}(W_+(\hat{\alpha}))$ (which in any case has expectation zero, Remark C) in (2.18), and taking

$$-\ddot{l}(\hat{\alpha}) \doteq Q' \left\{ \sum_{i=1}^{N} W_i(\hat{\alpha}) W_i(\hat{\alpha})' \right\} Q. \tag{2.27}$$

This made little numerical difference in our example, and had the benefit that (2.27) was guaranteed to be non-negative definite.

Figure 1 illustrates an artificial deconvolution problem in which $g(\theta)$ is a mixture of one-eighth uniform over the interval $[-3, 3]$ and seven-eighths $\mathcal{N}(0, 0.5^2)$,

$$g(\theta) = \frac{1}{8}\,\frac{I_{(-3,3)}(\theta)}{6} + \frac{7}{8}\,\frac{1}{\sqrt{2\pi\sigma^2}}\,e^{-\frac{1}{2}\frac{\theta^2}{\sigma^2}} \qquad [\sigma = 0.5], \tag{2.28}$$

with normal observations

$$X_i \sim \mathcal{N}(\Theta_i, 1), \qquad i = 1, 2, \ldots, N, \tag{2.29}$$

as in (1.6) (an "i.i.d. case"). The figure graphs $g(\theta)$ and the marginal density $f(x)$, the convolution of $g(\theta)$ with $\mathcal{N}(0, 1)$. The deconvolution task is to estimate $g(\theta)$ based on a sample $X_1, X_2, \ldots, X_N$ from $f(x)$.

Model (2.1)–(2.2) was implemented taking $\mathcal{T} = (-3, -2.8, \ldots, 3)$, $m = 31$, and Q from the R function `ns(`$\mathcal{T}$`,df=5)`; that is, Q was a 31×5 matrix of natural splines over the set $\mathcal{T}$. $N = 1000$ pairs (Θ_i, X_i) were independently generated according to (2.28)–(2.29), and MLE $\hat{\alpha}$ calculated from the sample $(X_1, X_2, \ldots, X_{1000})$, giving a maximum likelihood estimate of the m-vector g (2.2),

$$\hat{g} = g(\hat{\alpha}) = e^{Q\hat{\alpha} - \psi(\hat{\alpha})}. \tag{2.30}$$

See Remark D for further details.

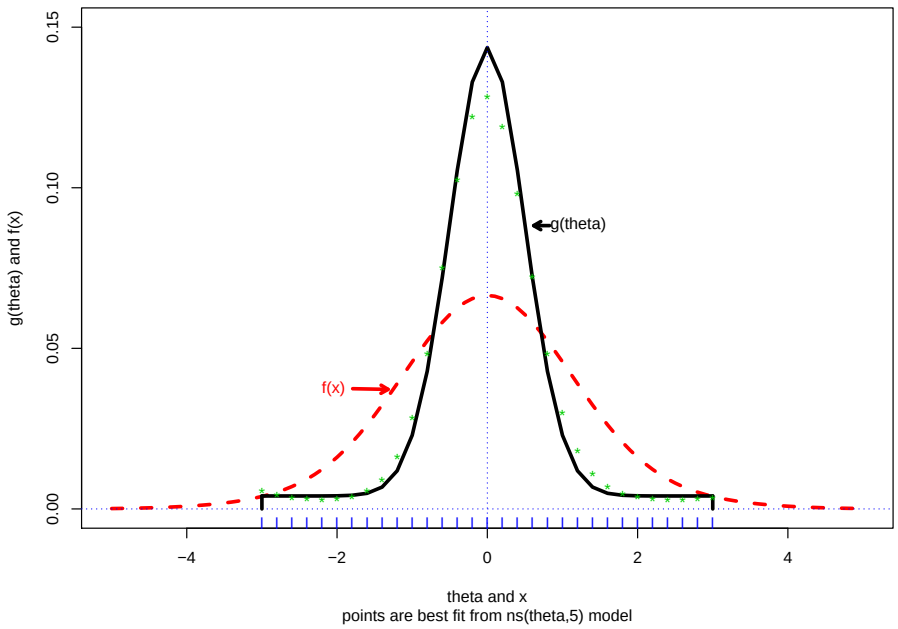

Figure 1: Prior $g(\theta)$ (solid curve) and marginal density $f(x)$ (dashed) for situation (2.28)–(2.29). Deconvolution methods aim to estimate $g(\theta)$ based on observations from $f(x)$. Dots indicated closest curve to g (2.28) in 5-parameter exponential family.

All of this was independently repeated 500 times, yielding estimates $\hat{g}^{(1)}, \hat{g}^{(2)}, \ldots, \hat{g}^{(500)}$, with means and standard deviations of the component values $\hat{g}_j$,

$$\bar{g}_j = \sum_{b=1}^{500} \hat{g}_j^{(b)}/500 \quad \text{and} \quad \overline{\text{sd}}_j = \left[\sum_{b=1}^{500} \left(\hat{g}_j^{(b)} - \bar{g}_j\right)^2 \Big/ 499\right]^{1/2}. \tag{2.31}$$

The vertical bars in Figure 2 plot

$$\bar{g}_j \pm \overline{\text{sd}}_j \tag{2.32}$$

versus θ_j for $j = 1, 2, \ldots, 31$. We see that the $\hat{g}$'s were reasonably accurate in estimating g, successfully capturing its long flat tails. However, some bias is apparent. It comes from the fact that our five-parameter exponential family of priors does *not* contain the true prior $g(\theta)$ (2.28). The starred points in Figure 1 show the closest possible member of the five-parameter family to $g(\theta)$, say $\tilde{g}(\theta)$; see Remark D. In Figure 2 the dashed curve graphing the means $\bar{g}_j$ closely approximates $\tilde{g}_j$. This kind of *definitional bias* is a price we pay for employing low-dimensional parametric families, the payoff being reduced variability. Figure 9 in Section 6 indicates a substantial payoff in this case.

3 Regularization and accuracy

The accuracy of a deconvolution estimate obtained from exponential family model (2.2) can be greatly improved by regularization of the maximum likelihood algorithm. Rather than maximizing $l(\alpha) = \sum \log f_i(\alpha)$ (2.8), we maximize a *penalized* log likelihood

$$m(\alpha) = l(\alpha) - s(\alpha). \tag{3.1}$$

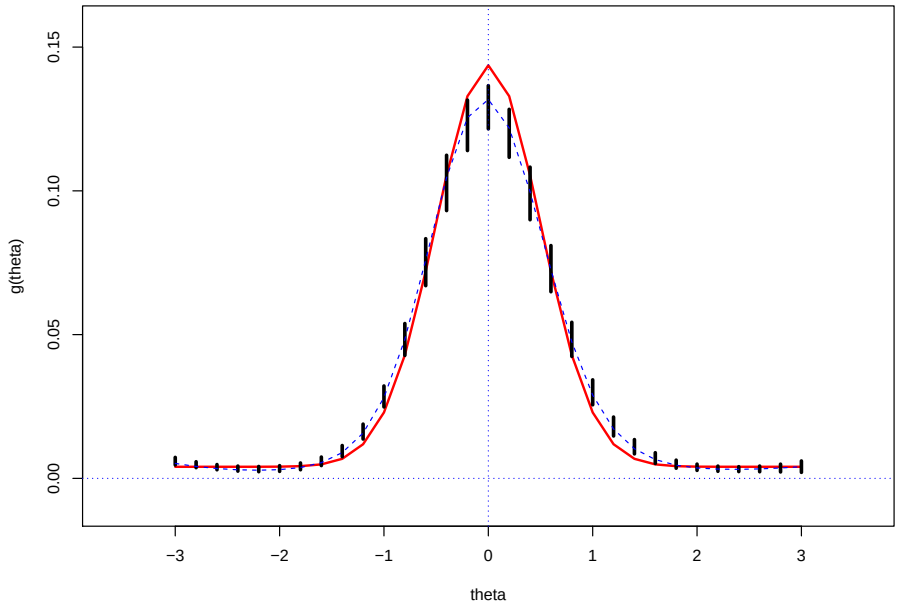

Figure 2: Vertical bars are $\bar{g}_j \pm \overline{\mathrm{sd}}_j$ (2.32), from 500 MLE estimates $\hat{g}$, each based on 1000 observations X_i (2.28)–(2.29). Solid curve shows true g_j values. Dashed line through $\bar{g}_j$ values closely follows natural spline best possible fit curve indicated by points in Figure 1.

Here $s(\alpha)$ is a *penalty function* that smoothly increases as α moves farther away from the origin. In our examples,

$$s(\alpha) = c_0\|\alpha\| = c_0\left(\sum_{h=1}^{p}\alpha_h^2\right)^{1/2}, \tag{3.2}$$

with c_0 equal 1 or 2. (The calculations in Figure 2 used $c_0 = 2$.) The effect of this kind of regularization is to bias $\hat{\alpha}$ toward the origin and pull $g(\hat{\alpha})$ toward a flat prior over the Θ space. Regularization tamps down excursions of the exponent $Q\alpha - \phi(\alpha)$ in (2.2), decreasing the variability of $\hat{\alpha}$, at the possible expense of increased definitional bias.

The first derivative vector and second derivative matrix of $m(\alpha)$ are

$$\dot{m}(\alpha) = \dot{l}(\alpha) - \dot{s}(\alpha) \quad \text{and} \quad \ddot{m}(\alpha) = \ddot{l}(\alpha) - \ddot{s}(\alpha). \tag{3.3}$$

Let α_0 represent the true value of α. Following the usual MLE derivation, the maximizer $\hat{\alpha}$ of $m(\alpha)$, that is, the *penalized likelihood estimate*, satisfies

$$\begin{aligned} 0 = \dot{m}(\hat{\alpha}) &\doteq \dot{m}(\alpha_0) + \ddot{m}(\alpha_0)(\hat{\alpha} - \alpha_0) \\ &= \left(\dot{l}(\alpha_0) - \dot{s}(\alpha_0)\right) + \left(\ddot{l}(\alpha_0) - \ddot{s}(\alpha_0)\right)(\hat{\alpha} - \alpha_0), \end{aligned} \tag{3.4}$$

giving

$$\hat{\alpha} - \alpha_0 \doteq \left(-\ddot{l}(\alpha_0) + \ddot{s}(\alpha_0)\right)^{-1}\left(\dot{l}(\alpha_0) - \dot{s}(\alpha_0)\right). \tag{3.5}$$

But $\dot{l}(\alpha_0)$ has mean vector and covariance matrix $(0, \mathcal{I}(\alpha_0))$, and $-\ddot{l}(\alpha_0) \doteq \mathcal{I}(\alpha_0)$, so (3.5) leads to familiar expressions for the mean and covariance of $\hat{\alpha}$:

Theorem 2. *If α_0 is the true value of α, the penalized maximum likelihood estimator $\hat{\alpha}$ has approximate mean vector and covariance matrix*

$$\hat{\alpha} - \alpha_0 \doteadd \left[-\left(\mathcal{I}(\alpha_0) + \ddot{s}(\alpha_0)\right)^{-1} \dot{s}(\alpha_0),\, \left(\mathcal{I}(\alpha_0) + \ddot{s}(\alpha_0)\right)^{-1} \mathcal{I}(\alpha_0) \left(\mathcal{I}(\alpha_0) + \ddot{s}(\alpha_0)\right)^{-1} \right]; \tag{3.6}$$

where $\mathcal{I}(\alpha_0)$ is the Fisher information matrix for $\hat{\alpha}$.

Note: For $s(\alpha) = c_0\|\alpha\|$ (3.2),

$$\dot{s}(\alpha) = c_0 \frac{\alpha}{\|\alpha\|} \quad \text{and} \quad \ddot{s}(\alpha) = \frac{c_0}{\|\alpha\|}\left(I - \frac{\alpha\alpha'}{\|\alpha\|^2}\right), \tag{3.7}$$

I the $p \times p$ identity matrix.

Writing (3.6) as

$$\hat{\alpha} - \alpha_0 \doteadd \left(\text{Bias}(\alpha_0), \text{Cov}(\alpha_0)\right), \tag{3.8}$$

we obtain convenient approximations for the bias and covariance of $\hat{g} = g(\hat{\alpha})$:

Corollary 1. *In notation* (3.6)–(3.8)*,*

$$g(\hat{\alpha}) - g(\alpha_0) \doteadd \left(D(\alpha_0)Q\,\text{Bias}(\alpha_0), D(\alpha_0)Q\,\text{Cov}(\alpha_0)Q'D(\alpha_0)\right), \tag{3.9}$$

where

$$D(\alpha_0) = \text{diag}\left(g(\alpha_0)\right) - g(\alpha_0)g(\alpha_0)'. \tag{3.10}$$

The corollary follows from (3.8) and the differential relationship

$$dg(\alpha)/d\alpha = Q'D(\alpha); \tag{3.11}$$

see Remark B. In practice, α_0 would be replaced by $\hat{\alpha}$ in (3.6)–(3.9), and $\mathcal{I}(\alpha_0)$ replaced by $-\ddot{l}(\hat{\alpha})$ (2.18) (perhaps dropping the term $\text{diag}(W_+(\hat{\alpha}))$, as suggested by (2.26), as was done in all the following examples). *Note:* $D(\alpha_0)Q\,\text{Bias}(\alpha_0)$ does not include possible definitional bias.

The corollary provides quick assessments for the bias and covariance of $g(\hat{\alpha})$. Figure 3 concerns a Poisson example in which the Θ_i were drawn from a chi-square density with 10 degrees of freedom and $X_i|\Theta_i$ was Poisson with expectation Θ_i,

$$\Theta_i \sim \chi^2_{10} \quad \text{and} \quad X_i|\Theta_i \sim \text{Poi}(\Theta_i). \tag{3.12}$$

One thousand simulations were carried out, each with $N = 1000$ observations, as in (1.1)–(1.2); $\mathcal{T} = \{1, 2, \ldots, 32\}$, Q the R language natural spline matrix `ns(`$\mathcal{T}$`,df=5)`, and $c_0 = 1$ in (3.2).

Each simulated data set $X_1, X_2, \ldots, X_{1000}$ yielded a penalized $\hat{\alpha}$, obtained by maximizing (3.1) using the R function nlm. Figure 3 compares the empirical standard deviations and biases of $g(\hat{\alpha})$ with approximation (3.9), $\alpha = \hat{\alpha}$. The approximation is excellent for the standard deviations, and a little too small for the biases. Not all of our results were this good. Figure 9 shows the corollary somewhat underestimating standard deviations for the example of Figure 1.

Table 1 reports on some of the simulation results. $N = 1000$ observations per simulation was not excessive: the coefficient of variation of $\hat{g}(\theta)$ is still large, at least in the tails of the prior $g(\theta)$. Nevertheless, g-modeling consistently yielded useful, if not perfect, inferences for $g(\theta)$.

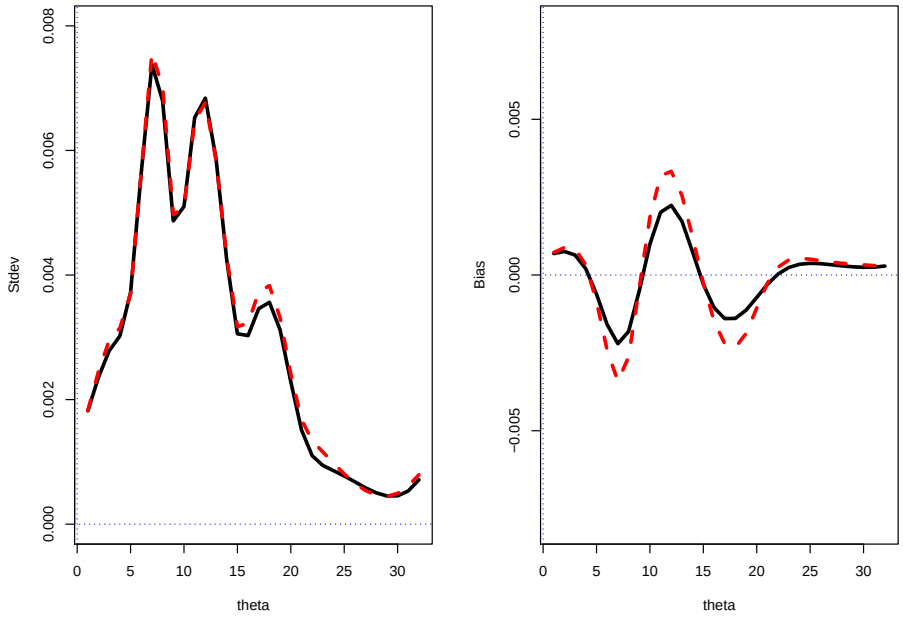

Figure 3: Standard deviations (left panel) and biases (right) for the Poisson example (3.12), $N = 1000$ observations; solid curves from formula (3.9) with $\alpha_0 = \hat{\alpha}$, dashed curve from simulations. (The vertical scales have the same spacing in both panels.)

Table 1: Simulation results for the Poisson example. (Entries for the middle four columns multiplied by 100.)

θ	$g(\theta)$	Mean	Stdev	Bias	CoefVar
5	5.50	5.59	.37	−.08	.07
10	9.42	9.24	.51	.19	.05
15	3.31	3.31	.32	−.03	.10
20	1.07	1.10	.24	−.11	.23
25	.15	.13	.08	.05	.54

The choice of c_0 can be motivated from (3.6), where $\ddot{s}(\alpha_0)$ is added to the Fisher information matrix $\mathcal{I}(\alpha_0)$. In this sense, the penalty function $s(\alpha)$ is artificially adding $\ddot{s}(\alpha_0)$ amount of information (that $g(\alpha)$ is flat). Let R_α equal the ratio of traces,

$$R(\alpha) = \operatorname{tr}\left(\ddot{s}(\alpha)\right) / \operatorname{tr}\left(\mathcal{I}(\alpha)\right). \tag{3.13}$$

From (2.18) and (3.7) we obtain an approximation for the ratio of artificial to genuine information,

$$R(\hat{\alpha}) = c_0(p-1)/\|\hat{\alpha}\| \cdot \operatorname{tr}\left\{ Q' \sum_{i=1}^{N} W_i(\hat{\alpha}) W_i(\hat{\alpha})' Q' \right\}, \tag{3.14}$$

again ignoring the $\operatorname{diag}(W_+(\hat{\alpha}))$ term; $R(\hat{\alpha}) = 0.01$ for the Poisson example, suggesting that $c_0 = 1$ was a modest choice for the regularizing constant.

The parametric bootstrap offers a direct, though more laborious, alternative to formula (3.14). Bootstrap realizations $\hat{\Theta}_i^*$, $i = 1, 2, \ldots, N$, are sampled from $\hat{g} = g(\hat{\alpha})$. Each Θ_i^* gives an X_i^*, as in

(1.2), and then $\hat{\alpha}^*$ is obtained as the penalized MLE based on $X_1^*, X_2^*, \ldots, X_N^*$. (It helps to start the nlm search for each $\hat{\alpha}^*$ from $\hat{\alpha}$.) Finally, $\hat{g}^* = g(\hat{\alpha}^*)$ (2.2), after which the bootstrap covariance and bias estimates are calculated in the usual way.

4 Two examples

This section pursues applications of g-modeling methods to two biomedical data sets. These are meant to illustrate deconvolution in action as a practical data-analytic tool.

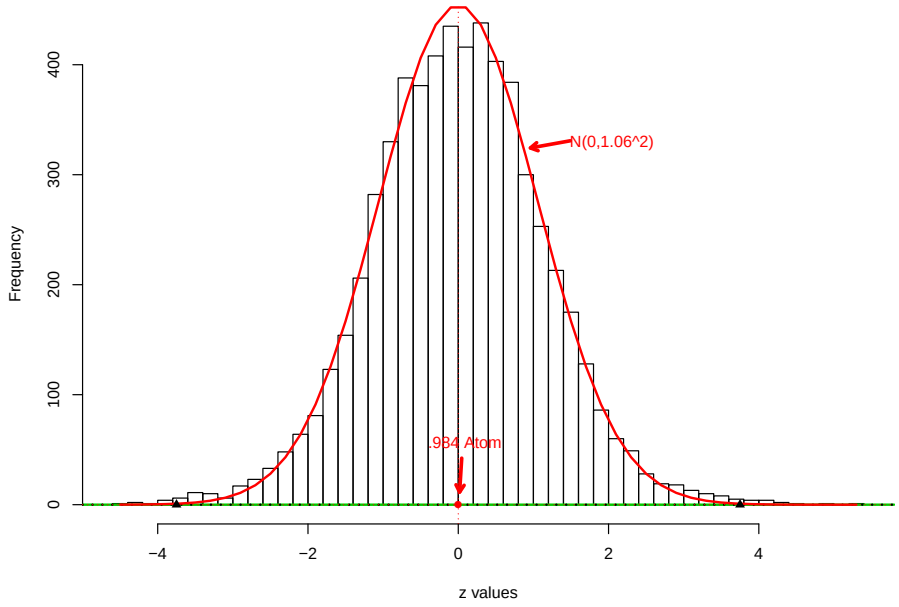

Figure 4: *Prostate data*; histogram shows test statistics X_i for $N = 6033$ genes. Local false discovery analysis using `locfdr`, an f-modeling algorithm, estimated that 98.4% of the genes were null (showing no difference between cancer and control subjects) and that the null genes had $X_i \sim \mathcal{N}(0, 1.06^2)$. The 44 genes having $|X_i| > 3.75$ had local fdr ≤ 0.2, and were flagged as non-null.

Singh et al. (2002) report on a microarray study comparing 52 prostate cancer patients with 50 healthy controls. Genetic expression levels were measured on $N = 6033$ genes. A two-sample test between patients and controls then yielded test statistic X_i for gene$_i$, $i = 1, 2, \ldots, N$, with the histogram in Figure 4 displaying the 6033 X_i values.

Efron (2010) discusses this data set, the *prostate data*: there, `locfdr` assessed 98.4% of the genes as "null", that is as having the same distribution in patients and controls, and estimated that the null genes followed the empirical null distribution

$$X_i \sim \mathcal{N}(0, 1.06^2). \tag{4.1}$$

The 44 genes having $|X_i| > 3.75$ were flagged as probably non-null. (`Locfdr` is a "f-modeling" algorithm, related to the methods discussed in Section 6.)

We wish to compare the `locfdr` results with those obtained from our g-modeling theory. A reasonable choice for the deconvolution model is

$$X_i \sim \mathcal{N}(\Theta_i, 1.06^2). \tag{4.2}$$

Here Θ_i is the *effect size* for gene$_i$; $\Theta_i = 0$ for the null genes, while of course Singh et al. were interested in spotting non-null genes, those with large values of $|\Theta_i|$. For comparison with `locfdr`, the variance in (4.2) was chosen to match (4.1). Section 7.4 of Efron (2010) justifies the normal translation model.

The deconvolution analysis used model (2.1)–(2.2), with $\mathcal{T} = (-3.6, -3.4, \ldots, 3.6)$, $m = 37$, and $Q = (I_0, Q_1)$, where I_0 was a delta function at zero (a m-vector with 1 at $\theta_{19} = 0$ and 0 elsewhere); Q_1 was the 37×5 R natural spline matrix `ns(`$\mathcal{T}$`,df=5)`, standardized so that each column had mean zero and sum of squares one. The R nonlinear maximization function `nlm` was used to find $\hat{\alpha}$, the penalized MLE (3.1), taking $s(\alpha) = c_0\|\alpha\|$, $c_0 = 1$. See Remark D.

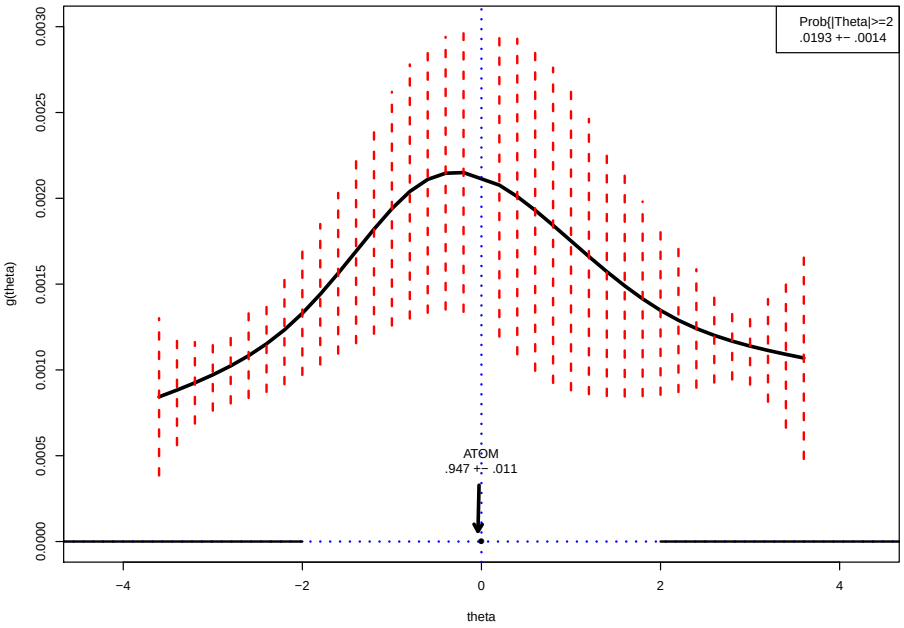

Figure 5: g-modeling estimate of prior $g(\theta)$ for the prostate data; probability of null gene $\Theta = 0$ estimated as 0.947 ± 0.011. Solid curve is estimated density $\hat{g}(\theta)$ for $\Theta \neq 0$; dashed vertical lines indicate $\pm$ one standard deviation.

The penalized MLE estimate of the prior g, $\hat{g} = g(\hat{\alpha})$, puts probability 0.947 ± 0.011 on $\Theta = 0$ (i.e., on $\theta_{19} = 0$), with the standard deviation computed from formula (3.9). Figure 5 graphs $\hat{g}(\theta)$ for $\theta \neq 0$, the vertical bars indicating $\pm$ one standard deviation. Amidst considerable noise, the curve indicates that non-null genes, those with $\Theta \neq 0$, have higher density near 0 than farther away.

Accuracy is better for larger regions of the Θ-space, for instance for

$$A = \{|\Theta| > 2\}, \tag{4.3}$$

corresponding to the subset $I_A = \{|\theta_j| > 2\}$ of $\mathcal{T}$. It has estimated probability

$$\widehat{\Pr}\{|\Theta| > 2\} = \sum_{\theta_j \in I_A} \hat{g}_j = 0.0193 \pm 0.0014, \tag{4.4}$$

with the standard deviation calculated from (3.9) according to

$$\left(v_A' D(\hat{\alpha}) Q \,\mathrm{Cov}(\hat{\alpha}) Q' D(\hat{\alpha}) v_A\right)^{1/2}, \tag{4.5}$$

where v_A is the m-vector having 1 in I_A and 0 elsewhere.

The results in Figure 4 and Figure 5 make for an interesting comparison: g-modeling puts probability 0.947 rather than 0.984 on the null case $\Theta = 0$, but indicates a substantial population of "low Θ" cases,

$$\widehat{\Pr}\{|\Theta| \le 2\} = 0.982. \tag{4.6}$$

If "interesting" genes are those with $|\Theta| > 2$, there are about two percent of them according to (4.4). This does not mean they are easy to identify, as discussed next.

Suppose the statistician is interested in the posterior expectation of some function $t(\Theta_i)$ given X_i. In our discrete setting (2.1)–(2.5), $t(\Theta)$ is represented by a vector

$$\boldsymbol{t} = (t_i, t_2, \ldots, t_m)' \qquad \text{with } t_j = t(\theta_j).$$

Bayes rule estimates the posterior expectation as

$$\hat{E}_i = \hat{E}\{t(\Theta_i)|X_i\} = \sum_{j=1}^{m} t_j p_{ij} \hat{g}_j \Big/ \sum_{i=1}^{m} p_{ij} \hat{g}_j, \tag{4.7}$$

p_{ij} as in (2.5).

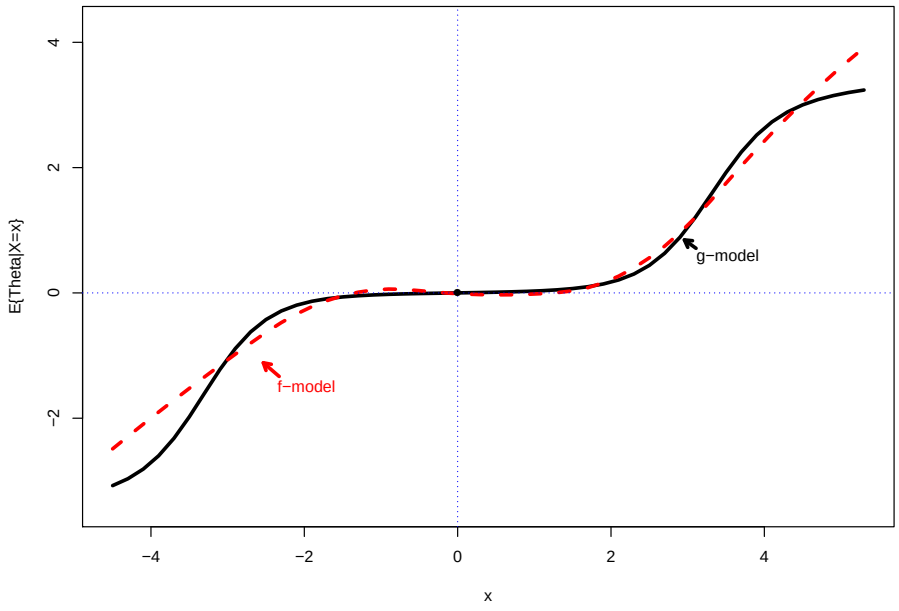

Figure 6: Solid curve is g-modeling estimate of $E\{\Theta_i|X_i = x\}$ (4.7) for the prostate data. Dashed curve is Tweedie's estimate (4.8), an f-modeling estimate.

The solid curve in Figure 6 graphs $\hat{E}_i$ for $t(\Theta) = \Theta$, that is for the posterior expectation of Θ. We see that $\hat{E}_i$ is nearly 0 for $|\Theta_i| \le 2$, reflecting the preponderence of null genes. We don't obtain a healthily nonzero $\hat{E}_i$ until $|\Theta|$ exceeds at least 3.

The dashed curve in Figure 3 is "Tweedie's estimate" (Efron, 2011), an f-modeling construction,

$$\hat{E}\{\Theta_i|X_i\} = X_i + \hat{f}'(X_i)/\hat{f}(X_i), \tag{4.8}$$

where $\hat{f}$ is a smooth estimate of density for the X values, and $\hat{f}'$ its derivative. Estimating $E\{\Theta|X\}$ is a favorable venue for f-modeling, as discussed in Section 6 of Efron (2014).

Our second biomedical example concerns an intestinal surgery study on $N = 800$ cancer patients. In addition to the primary site, surgeons also removed "satellite" nodes for later testing. The data set comprised pairs

$$(n_i, X_i), \qquad i = 1, 2, \ldots, 800, \tag{4.9}$$

where n_i is the number of satellites removed and X_i is the number of these found to be malignant. The n_i's varied from 1 to 40. Nearly 40% of the cases had $X_i = 0$, but for the remainder of them, $p_i = X_i/n_i$ had a roughly uniform distribution over $[0, 1]$, with a small mode at $p_i = 1$.

We assume the binomial model (1.3),

$$X_i \sim \text{Binom}(n_i, \Theta_i), \tag{4.10}$$

where Θ_i is the ith patient's individual probability for any one satellite being malignant. The Θ_i's are unobservable, but we can estimate their density $g(\theta)$ using a g-model (2.1)–(2.2). Here we took $\mathcal{T} = \{0.01, 0.02, \ldots, 0.99\}$, $m = 99$, Q the 99×5 natural spline R matrix `ns(`$\mathcal{T}$`,df=5)` (with columns standardized to have mean 0 and sum of squares 1), and penalty term $c_0\|\alpha\|$ (3.2), $c_0 = 1$.

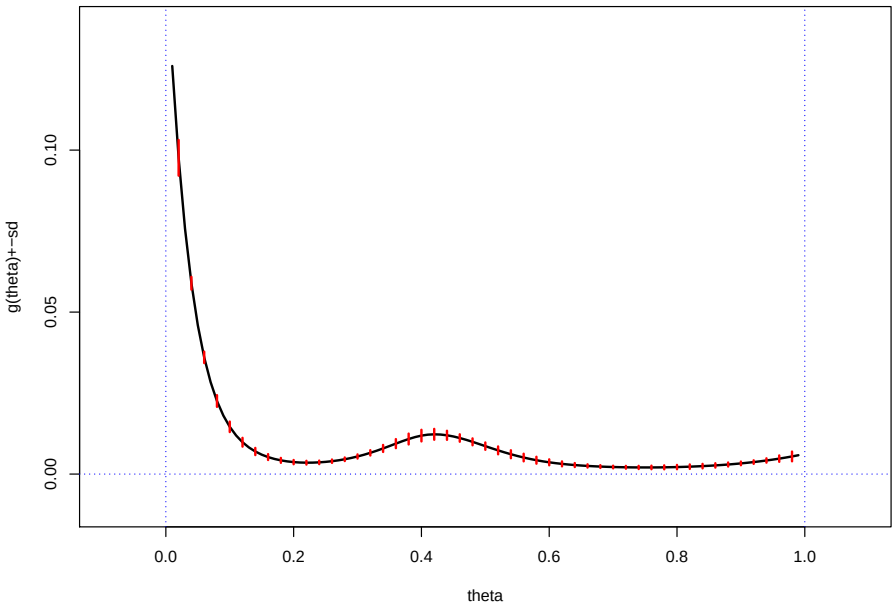

Figure 7: Estimated prior density $\hat{g}(\theta)$ for the sugery data. Vertical bars indicate $\pm$ one standard deviation, computed from formula (3.9) with α_0 equal the penalized MLE $\hat{\alpha}$.

Figure 7 shows the penalized maximum likelihood estimate $\hat{g}(\theta)$ for the distribution of Θ. There is a large mode near $\Theta = 0$, with 50% chance of $\Theta \leq 0.1$ and the remaining 50% spread almost evenly over $[0.1, 1.0]$. The curve was estimated with reasonable accuracy, median coefficient of variation 0.16, the standard deviation being computed using formula (3.9), $\alpha_0 = \hat{\alpha}$.

A parametric bootstrap simulation was run as a check on formula (3.9): for each of 1000 runs, 800 simulated realizations $\hat{\Theta}_i^*$ were sampled from density $\hat{g}$; each Θ_i^* gave an $X_i^* \sim \text{Binom}(n_i, \Theta_i^*)$, with n_i the ith sample size in the original data set; finally, $\hat{\alpha}^*$ was calculated as the penalized MLE based on $X_1^*, X_2^*, \ldots, X_{800}^*$, given $\hat{g}^* = g(\hat{\alpha}^*)$. Table 2 compares the standard deviations and biases of the 1000 $\hat{g}^*$'s with those from formula (3.9). The agreement is excellent.

Bias looks small in Table 2, but some caution is necessary: this does not include definitional bias, which may be substantial for the surgery data because of the large proportion of $X_i = 0$

Table 2: Satellite binomial analysis. Standard deviation and bias from formula (3.9) (with α_0 = MLE) compared with simulation estimates from 1000 parametric bootstrap replications. (All columns except first multiplied by 100.)

		StDev		Bias	
θ	$g(\theta)$	formula	simul	formula	simul
.01	12.048	.887	.967	−.518	−.592
.12	1.045	.131	.139	.056	.071
.23	.381	.058	.065	.025	.033
.34	.779	.096	.095	−.011	−.013
.45	1.119	.121	.117	−.040	−.049
.56	.534	.102	.100	.019	.027
.67	.264	.047	.051	.023	.027
.78	.224	.056	.053	.018	.020
.89	.321	.054	.048	.013	.009
.99	.576	.164	.169	−.008	.008

cases. Adding $\theta_0 = 0$ to $\mathcal{T}$ resulted in a more L-shaped estimate $\hat{g}$, though still with about 50% probability mass spread fairly evenly over $[0.1, 1.0]$.

5 Covariates and deconvolution

The previous theory carries over to the situation where each observation X_i is accompanied by an observed covariate vector u_i, say of dimension d. Now we assume a one-parameter exponential family of conditional densities for each X_i, rather than (1.2),

$$p(x_i|\eta_i) = e^{\eta_1 x_i - \psi(\eta_i)} p_0(x_i), \tag{5.1}$$

where x_i ranges over the sample space of X_i. Here η_i is the "natural" or "canonical" parameter, with the derivatives of ψ providing moments of X_i,

$$\dot{\psi}(\eta_i) \equiv \mu_i = E\{X_i|\eta_i\} \quad \text{and} \quad \ddot{\psi}(\eta_i) \equiv V_i = \text{Var}\{X_i|\eta_i\}. \tag{5.2}$$

(The form of the exponential family can depend on i, $p_i(x_i|\eta_i)$, but we will suppress the extra subscript.)

Letting

$$\eta_i = \Theta_i + u_i'\gamma, \tag{5.3}$$

where Θ_i is an unobserved realization from $g(\alpha)$ (2.2), and γ an unknown d-dimensional parameter vector, we observe

$$X_i \sim p(x_i|\eta_i) = p(x_i|\Theta_i + u_i'\gamma) \tag{5.4}$$

independently for $i = 1, 2, \ldots, N$, and wish to estimate the $(p+d)$-parameter vector (α, γ). Without the Θ_i's, (5.3)–(5.4) would be a standard generalized linear model (GLM). With the Θ_i's, it amounts to a generalized linear *mixed* model (GLMM). Most of the GLMM literature assumes g normal, as in Waclawiw and Liang (1993), but here we allow the wider specification (2.2). (On the other hand, the random effect Θ_i, which is univariate here, is usually allowed to be multivariate in normal-theory developments.)

Returning to the discrete setup (2.1), $\mathcal{T} = \{\theta_1, \theta_2, \ldots, \theta_m\}$, let

$$\eta_{ij} = \theta_j + u_i'\gamma, \qquad p_{ij} = p(X_i|\eta_{ij}) \tag{5.5}$$

as in (5.1), and define the m-vector

$$P_i(\gamma) = (\cdots p_{ij}(\gamma) \cdots)'. \tag{5.6}$$

The marginal probability for observation X_i is then

$$f_i(\alpha, \gamma) = P_i'(\gamma)g(\alpha), \tag{5.7}$$

with $g(\alpha) = \exp\{Q\alpha - \psi(\alpha)\}$ as in (2.2).

We wish to calculate the $(p+d)$-vector of first derivations and the $(p+d) \times (p+d)$ matrix of second derivatives of $l_i(\alpha, \gamma) = \log f_i(\alpha, \gamma)$. Because α and γ separate in (5.7), formulas (2.12) for $\partial l_i/\partial\alpha$ and (2.16) for $\partial^2 l_i/\partial\alpha^2$ apply as given, after setting $p_{ij} = p_{ij}(\gamma)$ in (2.10). All of the other required calculations are collected in the next lemma.

In accordance with (5.2) and (5.5), define

$$\mu_{ij} = \dot{\psi}(\eta_{ij}), \qquad V_{ij} = \ddot{\psi}(\eta_{ij}), \tag{5.8}$$

and

$$A_{ij} = (X_i - \mu_{ij})p_{ij}, \qquad B_{ij} = \left[(X_i - \mu_{ij})^2 - V_{ij}\right] p_{ij}. \tag{5.9}$$

Also let A_i and B_i be the corresponding m-vectors

$$A_i = (A_{i1}, A_{i2}, \ldots, A_{im})' \quad \text{and} \quad B_i = (B_{i1}, B_{i2}, \ldots, B_{im})'. \tag{5.10}$$

Lemma 4. *In terms of definitions* (5.5)–(5.10)*, we have:*

$$\partial l_i/\partial\gamma = u_i \frac{A_i'}{f_i}\, g, \tag{5.11}$$

$$\partial^2 l_i/\partial\gamma^2 = u_i \left[\frac{g' B_i}{f_i} - \left(\frac{g' A_i}{f_i}\right)^2\right] u_i', \tag{5.12}$$

and

$$\partial^2 l_i/\partial\alpha\partial\gamma = Q' \operatorname{diag}(g) \left(I_m - \frac{P_i}{f_i}\, g'\right) \frac{A_i}{f_i}\, u_i', \tag{5.13}$$

I_m the $m \times m$ identity matrix. Here $g = g(\alpha)$, $f_i = f_i(\alpha, \gamma)$, etc.; $\partial^2 l_i/\partial\gamma^2$ is a $d \times d$ matrix, and $\partial^2 l_i/\partial\alpha\partial\gamma$ $p \times d$.

See Remark E for the derivations. Summing over i in (5.11)–(5.13) gives the $\ddot{l}$ expressions for the total likelihood $l(\alpha, \gamma) = \sum l_i(\alpha, \gamma)$. A Bayesian restatement of these results appears at the end of Remark E.

A GLMM analysis of the surgery data (4.9) was carried out incorporating four covariates, "sex", "age", "smoke", and "prog": sex was coded 0 = female, 1 = male; age in years; smoke coded 0 = nonsmoker, 1 = smoker; prog a pre-operative prognosis score, with large values more favorable. The columns of the 800×4 matrix of covariates $\boldsymbol{u}$, ith row u_i', were standardized to have mean 0 and variance 1.

We assume a binomial model

$$X_i \overset{\text{ind}}{\sim} \text{Binom}(n_i, \pi_i), \tag{5.14}$$

where

$$\pi_i = 1/(1+e^{-\eta_i}) \qquad \text{with } \eta_i = \Theta_i + u_i'\gamma. \tag{5.15}$$

The change in notation from (4.10) is necessary to accommodate definition (5.3), where now Θ_i is a random effect *on the logit scale*; Θ_i can be thought of as a *frailty* parameter, with large values indicating a patient more prone to malignant satellite nodes. The Θ_i were assumed drawn from $g(\alpha) = \exp\{Q\alpha - \psi(\alpha)\}$ (2.2), with $\mathcal{T} = \{0.01, 0.02, \ldots, 0.99\}$ and $Q = ns(\mathcal{T}, df = 5)$, its columns standardized to have mean zero and sum of squares one.

Table 3: GLMM analysis of surgery satellite node data. *Row 1*: Maximum likelihood estimates $(\hat\alpha, \hat\gamma)$. *Rows 2 and 3*: Means and standard deviations from 100 parametric bootstrap simulations. *Row 4*: Standard deviations obtained from Lemmas 2 and 4.

	Alpha					Gamma			
	al1	al2	al3	al4	al5	sex	age	smoke	prog
1. MLE	−2.67	−10.09	−6.86	−11.23	−1.84	.192	−.078	.089	−.698
2. Mean	−3.51	−8.40	−7.11	−11.52	−1.12	.195	−.080	.067	−.694
3. StDev	.79	1.16	1.43	.69	.80	.070	.066	.063	.077
4. Formula	.79	1.35	1.39	.63	.85	.071	.073	.077	.093

The MLE $(\hat\alpha, \hat\gamma)$ in model (2.2), (5.4) was found by numerical maximization of the log likelihood

$$l(\alpha, \gamma) = \sum_{i=1}^{N} \log f_i(\alpha, \gamma) = \sum_{i=1}^{N} \log P_i'(\gamma) g(\alpha), \tag{5.16}$$

using the R function `nlm`. The top row of Table 3 shows $(\hat\alpha, \hat\gamma)$. Rows 2 and 3 give the means and standard deviations from 100 parametric bootstrap replications, drawing the X_i^*'s from (2.2), (5.14), (5.15) with $(\alpha, \gamma) = (\hat\alpha, \hat\gamma)$. Some estimation bias is evident, particularly for the first coordinate of $\hat\alpha$ but this did not translate into large biases for $g(\hat\alpha)$.

The observed Fisher information matrix $-\ddot{l}(\hat\alpha, \hat\gamma)$ was computed using Lemma 2 and Lemma 4, and inverted to provide estimated standard errors for $\hat\alpha$ and $\hat\gamma$, row 4 of Table 3. Comparison with row 3 shows reasonable agreement between formula and simulations, except perhaps for the smoke and prog coefficients of $\hat\gamma$ where the formula overestimates variability.

The estimated frailty distribution $g(\hat\alpha)$ is a very close match to $\hat\gamma$ in Figure 7, after transforming Θ_i in (5.3) to Figure 7's probability scale by $[1 + \exp(-\Theta_i)]^{-1}$. (This is the estimated conditional distribution of π_i in (5.14)–(5.15) given covariate vector $\gamma = 0$.) Looking at Table 3, we see that the sex and prog coefficients are significantly different from zero, with prog a particularly strong predictor. Figure 8 graphs the conditional distributions of the binomial probability π_i in (5.14)–(5.15) given the best or the worst levels of prog. Taken together, Figure 7 and Figure 8 indicate large individual differences (frailties) and even larger covariate effects on π, the probability of a positive node.

6 Fourier deconvolution

Stefanski and Carroll (1990) used Fourier analysis to produce an elegant solution to the "additive" deconvolution problem (1.4). Here we will discuss their approach in terms of the normal i.i.d.

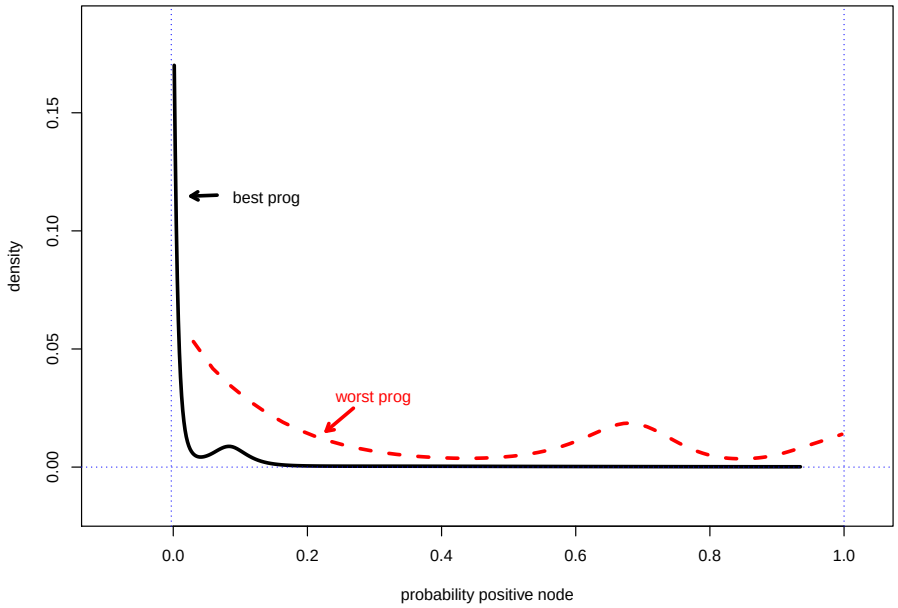

Figure 8: Estimated conditional distributions of the binomial parameter π_i (5.14)–(5.15), given the best and worst categories of the covariate prog.

model (1.6), where $X_i \sim \mathcal{N}(\Theta_i, 1)$. In this case the marginal density $f(x) = \int \phi(x-\theta)g(\theta)d\theta$ (ϕ the standard normal density) relates to the prior $g(\theta)$ via

$$\mathcal{F}(f) = \mathcal{F}(g)e^{-t^2/2}, \tag{6.1}$$

where $\mathcal{F}$ indicates Fourier transform.

The Stefanski–Carroll algorithm begins by smoothing the empirical density of the observed sample $X_1, X_2, \ldots, X_N$ with a "sinc" kernel, giving

$$\hat{f}(x) = \frac{1}{N\lambda}\sum_{i=1}^{N} \operatorname{sinc}\left(\frac{X_i - x}{\lambda}\right), \tag{6.2}$$

$\operatorname{sinc}(x) = \sin(x)/x$. The deconvoluted density estimate for the prior is then

$$\hat{g}(\theta) = \mathcal{F}^{-1}\left\{\mathcal{F}\left(\hat{f}\right)e^{t^2/2}\right\}, \tag{6.3}$$

$\mathcal{F}^{-1}$ being the inverse Fourier transform. Writing (6.1) as $g(\theta) = \mathcal{F}^{-1}\{\mathcal{F}(f)e^{t^2/2}\}$ motivates (6.3).

A pleasant surprise is that $\hat{g}(\theta)$ in (6.3) can be calculated directly as a kernel estimate from the sample $X_1, X_2, \ldots, X_N$,

$$\hat{g}(\theta) = \frac{1}{N}\sum_{i=1}^{N} k_\lambda(X_i - \theta), \tag{6.4}$$

where the kernel k_λ is given by

$$k_\lambda(x) = \frac{1}{\pi}\int_0^{1/\lambda} e^{t^2/2}\cos(tx)\ dt. \tag{6.5}$$

Large values of λ smooth $\hat{f}(x)$ more, reducing variance but possibly increasing bias, and similarly for $\hat{g}(\theta)$ as an estimator of $g(\theta)$.

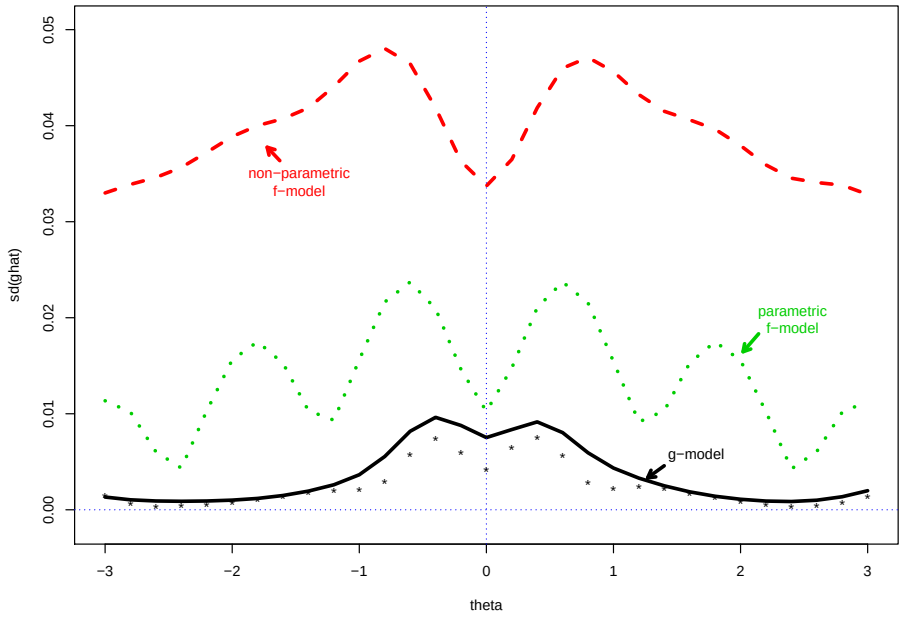

Figure 9: Standard deviations of $\hat{g}(\theta)$ for the artificial example (2.28)–(2.29). *Solid curve*: g-model (2.30), $N = 1000$, $Q = ns(\mathcal{T}, 5)$, $c_0 = 1$ in (3.7). *Dashed curve*: nonparametric f-model Fourier estimate (6.4), $\lambda = 1/3$. *Dotted curve*: parametric f-model using Poisson GLM, structure matrix $ns(\mathcal{X}, 5)$.

Fourier deconvolution was applied to the artificial example of Figure 1. The choice $\lambda = 1/3$ made the average bias of $\hat{g}(\theta)$ in the simulation experiment that follows about the same as that seen in the g-modeling simulation of Figure 2. Accuracy, however, was much worse. Figure 9 compares the standard deviations of $\hat{g}(\theta)$ (6.4) with those obtained using the g-model (2.30). Their median ratio was 20.4.

Part of the disparity is no more than the difference between nonparametric and parametric estimation. We can improve Fourier's performance by using a more efficient smoother at step (6.2).

Returning to the i.i.d. case, where the sample space of observations X_i is $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$, and the realizations Θ_i take values in $\mathcal{T} = \{\theta_1, \theta_2, \ldots, \theta_m\}$, let $\boldsymbol{k}_\lambda$ be the $m \times n$ matrix having jkth value $k_\lambda(x_k - \theta_j)$. We can express the Fourier estimate $\hat{g} = (\hat{g}_1, \hat{g}_2, \ldots, \hat{g}_m)'$ (6.4) as

$$\hat{g} = \boldsymbol{k}_\lambda \bar{f}, \tag{6.6}$$

with $\bar{f} = (\bar{f}_1, \bar{f}_2, \ldots, \bar{f}_i)'$ the *empirical density* that puts weight y_k/N on x_k.

We can reduce variability by replacing the nonparametric estimate $\bar{f}$ in (6.6) with a parametric estimate, say $f(\hat{\beta})$. Section 4 of Efron (2014) does this by taking $f(\beta)$ to be a Poisson GLM. The "parametric f-model" curve in Figure 9 gives the resulting standard deviation for $\hat{g} = \boldsymbol{k}_\lambda f(\hat{\beta})$, when the GLM structure matrix is $ns(\mathcal{X}, 5)$, a natural spline matrix with five degrees of freedom. Now f-modeling is more competitive with g-modeling (though the f-models had an unpleasant tendency to go negative in the tails of $\hat{g}$).

As a rough summary of recommendations concerning practical deconvolution methods:

- Nonparametric f-modeling is disparaged as overly variable.
- Parametric f-modeling can be attractive in i.i.d. additive noise situations (particularly for "classic" empirical Bayes problems of the type discussed in Section 6 of Efron, 2014).

- Parametric g-modeling performed best in the context of this paper, and has the advantage of applying to general non-i.i.d. situations such as the binomial example of Section 4 and Section 5.

7 Proofs and details

The remarks of this section expand on points raised previously in the text.

Remark A. *Continuous formulation* Instead of the discrete setting (2.1), the sample space $\mathcal{T}$ for Θ can be made continuous, say as an interval of the real line. The definition (2.2) becomes

$$g_\theta(\alpha) = e^{Q'_\theta\alpha - \phi(\alpha)}, \tag{7.1}$$

where Q_θ is a smoothly defined $p \times 1$ vector function of $\theta \in \mathcal{T}$, and $\phi(\alpha) = \log(\int \exp\{Q'_\theta\alpha\}d\theta)$.

The development in Section 2 proceeds as before, with subscript j replaced by the continuous variable θ, and sums replaced by integrals over $\mathcal{T}$, e.g.,

$$p_{i\theta} = p_i(X_i|\Theta_i = \theta) \tag{7.2}$$

in place of (2.5), and

$$w_{i\theta}(\alpha) = g_\theta(\alpha)\,(p_{i\theta}/f_i(\alpha) - 1) \tag{7.3}$$

in place of (2.10), where $f_i(\alpha) = \int p_{i\theta}g_\theta(\alpha)d\theta$.

For any function s_θ of θ, define $E_\alpha\{s\} = \int s_\theta g_\theta(\alpha)d\theta$. Then (2.12) becomes

$$\dot l_i(\alpha) = \int Q_\theta w_{i\theta}(\alpha)\; d\theta = E_\alpha\{Qv_i(\alpha)\} \tag{7.4}$$

where

$$v_{i\theta}(\alpha) = p_{i\theta}/f_i(\alpha) - 1. \tag{7.5}$$

Using this notation we can re-express (2.13) as

$$-\ddot l_i(\alpha) = \dot l_i(\alpha)\dot l_i(\alpha)' + \dot l_i(\alpha)E_\alpha\{Q\}' + E_\alpha\{Q\}\dot l_i(\alpha)' - E_\alpha\{Qv_i(\alpha)Q'\}. \tag{7.6}$$

Summing over the observations i gives $w_{+\theta}(\alpha) = \sum w_i(\alpha)$ and

$$\dot l(\alpha) = \int_{\mathcal{T}} Q_\theta w_{+\theta}(\alpha)\; d\theta \tag{7.7}$$

as in (2.14), with similar extensions of Lemma 2.

All of this has a more familiar look than the discrete versions of Section 2. However, the numerical calculation of $\dot l(\alpha)$ and $\ddot l(\alpha)$ will usually get us back to the discrete sums of Lemma 1 and Lemma 2, which are necessary for the numerical implementation of the theory.

Remark B. *Lemma 1* Abbreviating g_i for $g_i(\alpha)$, etc., let $\dot g$ be the $p \times m$ matrix $(\partial g_j/\partial\alpha_k)$, and likewise $\dot f_i$ the $p \times 1$ vector $(\partial f_i/\partial\alpha_k)$. Then $f_i = g'P_i$ (2.7) gives $\dot f_i = \dot g P_i$. But

$$\dot g = Q'D \qquad \big[D = \mathrm{diag}(g) - gg'\big] \tag{7.8}$$

as in (5.7) of Efron (2014) (where the order of indices is reversed), yielding

$$\dot l_i = \dot f_i/f_i = Q'DP_i/f_i = Q'W_i, \tag{7.9}$$

the equality $DP_i/f_i = W_i$ coming, after a little algebra, by direct comparison with (2.10). This verifies (2.12).

Differentiating (7.9) shows that the $p \times p$ second derivative matrix $\ddot{l}_i$ is

$$\ddot{l}_i = Q'\dot{W}_i', \tag{7.10}$$

where $\dot{W}_i$ is the $p \times m$ matrix $(\partial w_{ij}/\partial\alpha_k)$, $k = 1, 2, \ldots, p$ and $j = 1, 2, \ldots, m$. It remains to evaluate $\dot{W}_i$. We can write W_i as

$$W_i = u \cdot v \qquad (u = g \text{ and } v = P_i/f_i - 1), \tag{7.11}$$

the notation indicating coordinatewise multiplication, $W_{ij} = u_j \cdot v_j$. The $p \times m$ derivative matrix $d(u \cdot v)/d\alpha$ is obtained from the identity

$$\frac{d(u \cdot v)}{d\alpha} = \dot{U}\,\mathrm{diag}(v) + \dot{V}\,\mathrm{diag}(u) \tag{7.12}$$

where $\dot{U}$ and $\dot{V}$ are the $p \times m$ matrices $(\partial U_j/\partial\alpha_k)$ and $(\partial V_j/\partial\alpha_k)$.

Letting $\tilde{P}_i = P_i/f_i$, u and v in (7.11) give, using (7.8)–(7.9),

$$\dot{U} = Q'D \quad \text{and} \quad \dot{V} = -Q'D\tilde{P}_i\tilde{P}_i', \tag{7.13}$$

and then

$$\dot{W}_i = Q'D\left\{\mathrm{diag}\left(\tilde{P}_i - 1\right) - \tilde{P}_i\tilde{P}_i'\,\mathrm{diag}(g)\right\} \tag{7.14}$$

from (7.12). This yields

$$-\ddot{l}_i = Q'\left\{\mathrm{diag}(g)\tilde{P}_i\tilde{P}_i' - \mathrm{diag}\left(\tilde{P}_i - 1\right)D\right\}Q \tag{7.15}$$

from (7.10). Finally the identities

$$\begin{aligned} &D\tilde{P}_i = W_i, \qquad \mathrm{diag}(g)\tilde{P}_i = W_i + g_i, \\ &\text{and} \quad D\,\mathrm{diag}\left(\tilde{P}_i - 1\right) = \mathrm{diag}\left(W_i - gW_i'\right) \end{aligned} \tag{7.16}$$

transform (7.15) into expression (2.13) for $-\ddot{l}_i(\alpha)$.

Remark C. *Lemma 3 and Theorem 1* Expression (2.13) can be rewritten in the same form as (7.6),

$$-\ddot{l}_i(\alpha) = \dot{l}_i(\alpha)\dot{l}_i(\alpha)' + \dot{l}_i(\alpha)\left(g'(\alpha)Q\right) + \left(Q'g(\alpha)\right)\dot{l}_i(\alpha)' + Q'\,\mathrm{diag}(W_{i\alpha})Q. \tag{7.17}$$

Let $\mathbb{E}$ indicate expectation with respect to the "i.i.d. case" model

$$X_i \overset{\text{iid}}{\sim} f(\alpha), \qquad i = 1, 2, \ldots, N, \tag{7.18}$$

expression (2.21), with α fixed. Familiar MLE theory says that

$$\mathbb{E}\left\{\dot{l}_i(\alpha)\right\} = 0 \quad \text{and} \quad \mathbb{E}\left\{-\ddot{l}_i(\alpha)\right\} = \mathbb{E}\left\{\dot{l}_i(\alpha)\dot{l}_i(\alpha)'\right\}. \tag{7.19}$$

Taking expectations in (7.17) then shows that

$$\mathbb{E}\left\{Q'\,\mathrm{diag}(W_{i\alpha})Q\right\} = 0, \tag{7.20}$$

and that the total expected Fisher information is

$$\begin{aligned}\mathcal{I}(\alpha) &= \mathbb{E}\left\{\sum_{i=1}^{N} \dot{l}_i(\alpha)\dot{l}_i(\alpha)'\right\} = \mathbb{E}\left\{\sum_{k=1}^{n} \dot{l}_k(\alpha) y_k \dot{l}_k(\alpha)'\right\} \\ &= \sum_{k=1}^{n} \dot{l}_k(\alpha)\left(N f_k(\alpha)\right)\dot{l}_k(\alpha)' = Q'\left\{\sum W_k(\alpha)\left(N f_k(\alpha)\right) W_k(\alpha)'\right\} Q,\end{aligned} \tag{7.21}$$

verifying Lemma 3. (Notice that $\dot{l}_k(\alpha)$ is a nonrandom quantity in these calculations.)

Our i.i.d. probability model

$$\alpha \longrightarrow g(\alpha) \longrightarrow f(\alpha) = \boldsymbol{P} g(\alpha) \longrightarrow \boldsymbol{y} \sim \text{Mult}_n\left(N, f(\alpha)\right) \tag{7.22}$$

is a *curved exponential family* (Efron, 1975). In such families, the plug-in estimates of expected and observed Fisher information are equal, this being the first equality in (2.26).

Remark D. *Computational details* All of the numerical calculations began by discretizing the sample space $\mathcal{X}$, for instance to $\mathcal{X} = \{-4.4, -4.2, \ldots, 5.4\}$ for the prostate data and the artificial example of Figure 1, and setting p_{kj} (2.20) equal to the probability that X falls nearest point x_k of $\mathcal{X}$. The columns of structure matrix Q were standardized to have mean zero and sum of squares 1.

The maximum likelihood estimate $\hat{\alpha}$ was obtained using `nlm`, the R language nonlinear maximizer:

$$\hat{\alpha} = \texttt{nlm(qmle,}\alpha_0, \cdots\texttt{)\$est}. \tag{7.23}$$

Here α_0 is a starting value while $\texttt{qmle}(\alpha, \cdots)$, available from the author, computes minus the likelihood of the data for any trial value α. Starting at $\alpha_0 = 0$ worked in our examples, but some exploration of starting values was done to avoid getting trapped in local minima. Each bootstrap replication in Table 3 began with α_0 equal to the original MLE $\hat{\alpha}$, and similarly for the simulations used in Figure 3. The "closest g" in Figure 1 was obtained by setting $y = Nf$ ($f = Pg$ (2.28)), using `nlm` to find the maximizing value $\hat{\alpha}$, and finally taking $\hat{g} = \exp\{Q\hat{\alpha} - \phi(\hat{\alpha})\}$.

Estimate of bias and standard deviation for $\hat{g}(\theta)$ were obtained substituting the MLE $\hat{\alpha}$ for α_0 in formula (3.9) (using R function `qformula`, available from the author).

Remark E. *Lemma 4* From $\log(p_{ij}) = \eta_{ij} x_i - \psi(\eta_{ij})$ we compute the gradient vector

$$\frac{d\log(p_{ij})}{d\gamma} = \frac{\partial \eta_{ij}}{\partial \gamma}(x_i - \mu_{ij}) = u_i(x_i - \mu_{ij}), \tag{7.24}$$

so

$$\frac{dp_{ij}}{d\gamma} = u_i(x_i - \mu_{ij})p_{ij} = u_i A_{ij}. \tag{7.25}$$

Then (5.7), $f_i = g'P_i$, gives $\partial f_i/\partial\gamma = u_i g' A_i$ and $\partial l_i/\partial\gamma = u g' A_i/f_i$, verifying (5.11).

Differentiating A_{ij} yields

$$\frac{dA_{ij}}{d\gamma} = u_i B_{ij}, \tag{7.26}$$

where we have used (7.24) and $d\mu_{ij}/d\eta_{ij} = V_{ij}$ (5.8). Differentiating (7.25) gives $d^2p_{ij}/d\gamma^2 = u_i B_{ij} u_i'$ and, from $f_i = g'P_i$,

$$\frac{\partial^2 f_i}{\partial \gamma^2} = u_i(g' B_i) u_i'. \tag{7.27}$$

The identity

$$\frac{\partial^2 l_i}{\partial\gamma^2} = \frac{1}{f_i}\frac{\partial^2 f_i}{\partial\gamma^2} - \frac{\partial l_i}{\partial\gamma}\frac{\partial l_i'}{\partial\gamma}, \tag{7.28}$$

applied to (5.11) and (7.27) then verifies statement (5.12) of Lemma 4.

Differentiating $(\partial f_i/\partial\gamma)' = g' A_i u_i'$ with respect to α, the $p \times d$ matrix $\partial^2 f_i/\partial\alpha\partial\gamma$ equals

$$\frac{\partial^2 f_i}{\partial\alpha\partial\gamma} = \frac{\partial g'}{\partial\alpha} A_i u_i' = Q' D A_i u_i', \tag{7.29}$$

where we have used $\dot{g} = Q'D$ (7.8) (remembering that $\partial g'/\partial\alpha = \dot{g}$ in our notational conventions). Finally, the identity

$$\frac{\partial^2 l_i}{\partial\alpha\partial\gamma} = \frac{1}{f_i}\frac{\partial^2 f_i}{\partial\alpha\partial\gamma} - \frac{\partial l_i}{\partial\alpha}\frac{\partial l_i'}{\partial\gamma}, \tag{7.30}$$

along with $\partial l_i/\partial\alpha$ (3.12) and $\partial l_i/\partial\gamma$ (5.11) result, after some simplification, in statement (5.13) of Lemma 4.

For the surgery example of Section 5, the cross-term $-\partial^2 l(\hat{\alpha},\hat{\gamma})$ in $-\ddot{l}(\hat{\alpha},\hat{\gamma})$ was quite small. Taking it to be 0, i.e., taking $\hat{\alpha}$ and $\hat{\gamma}$ to be independent, had little effect on row 4 of Table 3.

According to Bayes rule, the posterior distribution of Θ_i given X_i and u_i in (5.3)–(5.5) is

$$\Pr_{\alpha,\gamma}\{\Theta_i = \theta_j | X_i, u_i\} = g_j p_{ij}/f_i, \tag{7.31}$$

Therefore the factor $A_i' g/f_i$ in (5.11) equals the posterior expectation

$$A_i' g/f_i = E_{\alpha,\gamma}\{X_i - \mu_i | X_i, u_i\}, \tag{7.32}$$

where μ_i is the random quantity $\dot{\psi}(\Theta_i)$. Likewise

$$B_i' g/f_i = E_{\alpha,\gamma}\{(X_i - \mu_i)^2 - V_i\} \qquad \left[V_i = \ddot{\psi}(\Theta_i)\right]. \tag{7.33}$$

Substituting (7.32)–(7.33) into (5.12) then gives

$$-\frac{\partial^2 l_i}{\partial\gamma^2} = u_i E_{\alpha,\gamma}\{V_i | X_i, u_i\} u_i'. \tag{7.34}$$

References

Butucea, C. and Comte, F. (2009). Adaptive estimation of linear functionals in the convolution model and applications. *Bernoulli* 15: 69–98, MR2546799.

Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.* 83: 1184–1186, MR997599.

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

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

Efron, B. (2011). Tweedie's formula and selection bias. *J. Amer. Statist. Assoc.* 106: 1602–1614, doi: 10.1198/jasa.2011.tm11181.

Efron, B. (2014). Two modeling strategies for empirical bayes estimation. *Statist. Sci.* 29: 285–301, doi: 10.1214/13-STS455.

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.* 19: 1257–1272, MR1126324.

Hall, P. and Meister, A. (2007). A ridge-parameter approach to deconvolution. *Ann. Statist.* 35: 1535–1558, MR2351096.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixed distribution. *J. Amer. Statist. Assoc.* 73: 805–811, MR521328.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R. and Sellers, W. R. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 1: 203–209, doi: 10.1016/S1535-6108(02)00030-2.

Stefanski, L. and Carroll, R. J. (1990). Deconvoluting kernel density estimators. *Statistics* 21: 169–184, MR1054861.

Waclawiw, M. A. and Liang, K.-Y. (1993). Prediction of random effects in the generalized linear model. *J. Amer. Statist. Assoc.* 88: 171–178, `jstor.org/stable/2290711`.