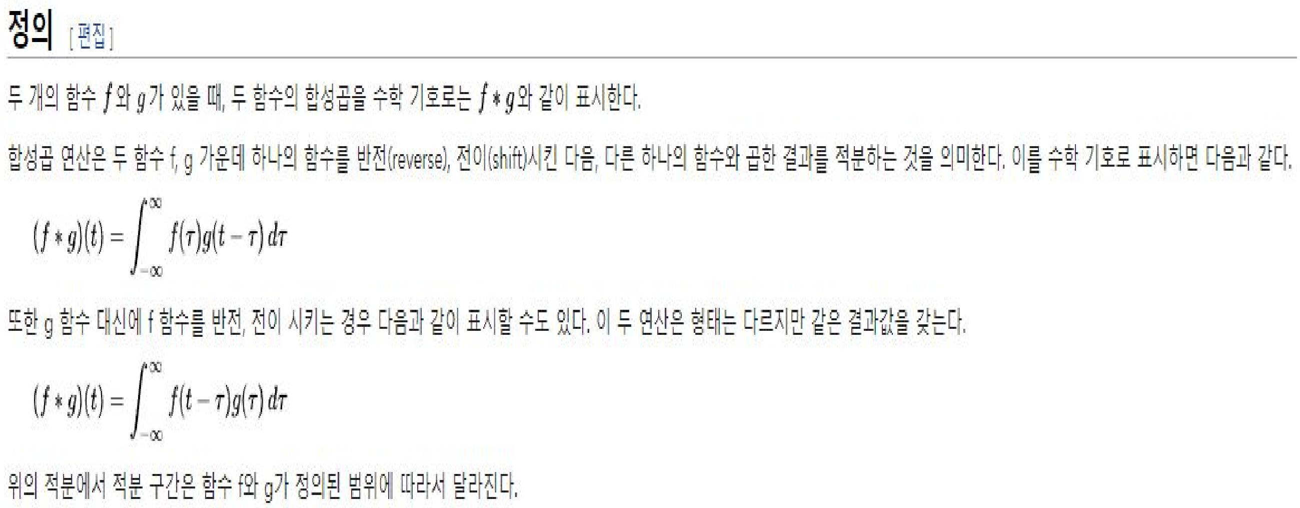
Chapter 21. 경험적 베이즈 추정 전략(Empirical Bayes Estimation Strategies)

2019-07-31 신재혁

21.1 베이즈 디컨볼루션(Bayes Deconvolution)

1) 컨볼루션(Convolution)?



정의 [편집]

두 개의 함수 f와 g가 있을 때, 두 함수의 합성곱을 수학 기호로는 $f*g$와 같이 표시한다.

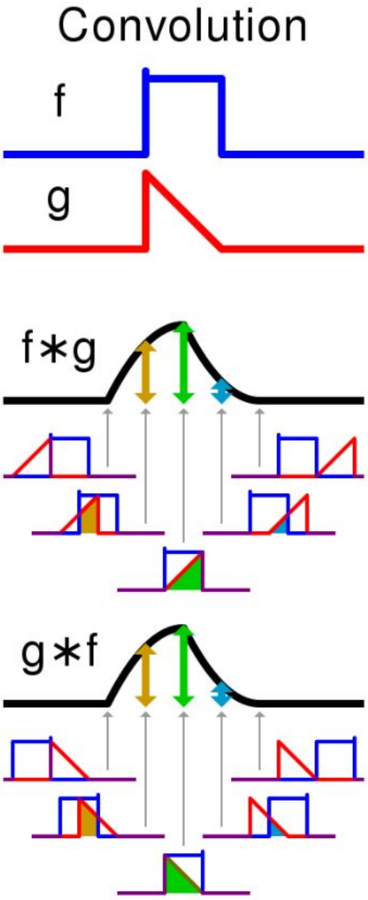
합성곱 연산은 두 함수 f, g 가운데 하나의 함수를 반전(reverse), 전이(shift)시킨 다음, 다른 하나의 함수와 곱한 결과를 적분하는 것을 의미한다. 이를 수학 기호로 표시하면 다음과 같다.

$$(f*g)(t)=\int_{-\infty}^{\infty} f(\tau)g(t-\tau)\,d\tau$$

또한 g 함수 대신에 f 함수를 반전, 전이 시키는 경우 다음과 같이 표시할 수도 있다. 이 두 연산은 형태는 다르지만 같은 결과값을 갖는다.

$$(f*g)(t)=\int_{-\infty}^{\infty} f(t-\tau)g(\tau)\,d\tau$$

위의 적분에서 적분 구간은 함수 f와 g가 정의된 범위에 따라서 달라진다.



2) 베이즈 디컨볼루션(Bayes Deconvolution)?

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

evidence= observable random vriable: $X_i \overset{\text{ind}}{\sim} p_i(X_i|\Theta_i).$

marginal distribution of X_i: $f_i(X_i) = \int_{\mathcal{T}} p_i(X_i|\theta_i) g(\theta_i)\, d\theta_i$

이러한 설정에서 ***관측된 X를 토대로 미지의 사전분포 g에 대해서 추정하고 싶다.***

만약 $p_i(X_i|\Theta_i)$ = $\mathcal{N}(\Theta_i, \sigma^2)$ 와 같이 theta에 대해 간단한 분포로 표현된다면,

$$\begin{aligned} f_i(X_i) &= \int_{\mathcal{T}} p_i(X_i|\theta_i) g(\theta_i)\, d\theta_i \\ &= g * \mathcal{N}(0, \sigma^2) \end{aligned}$$

으로 표현할 수 있다.

따라서, g를 구하는 문제는 위 합성곱 효과를 제거하면, 즉 디컨볼루션(Deconvolution)을 하면 구할 수 있다. 이러한 문제를 **베이즈 디컨볼루션(Bayes Deconvolution)**이라고 한다.

3) 왜 g를 알고 싶은가?
: 베이즈 사후 기댓값을 추정하길 원하거나, 사후 밀도를 그리길 원하는 등 다양한 이유가 있을 수 있다. 추후 예제로 설명.

4) 베이즈 디컨볼루션(Bayes Deconvolution)의 방법?
크게 g-모델링과 f-모델링으로 나뉜다.

21.6 디컨볼루션과 f-모델링(Deconvolution and f-Modeling)

1) f-Modeling이란?

marginal distribution of X_i = $$f_i(X_i) = \int_{\mathcal{T}} p_i(X_i|\theta_i)g(\theta_i)\, d\theta_i$$

를, 관측 표본 X1, X2, ... , Xn을 통해 직접 적합화 한 후, 원하는 대답을 f항으로 직접 표현하는 공식을 찾아 적용하는 방법이다.

예를 들어

- 로빈의 공식(6.5)

$$E\{\theta|x\} = (x+1)f(x+1)/f(x).$$

- 지역 거짓 발견율(15.38)

$$\mathrm{fdr}(z) = \pi_0 f_0(z)/f(z)$$

- 트위디의 공식(20.37)

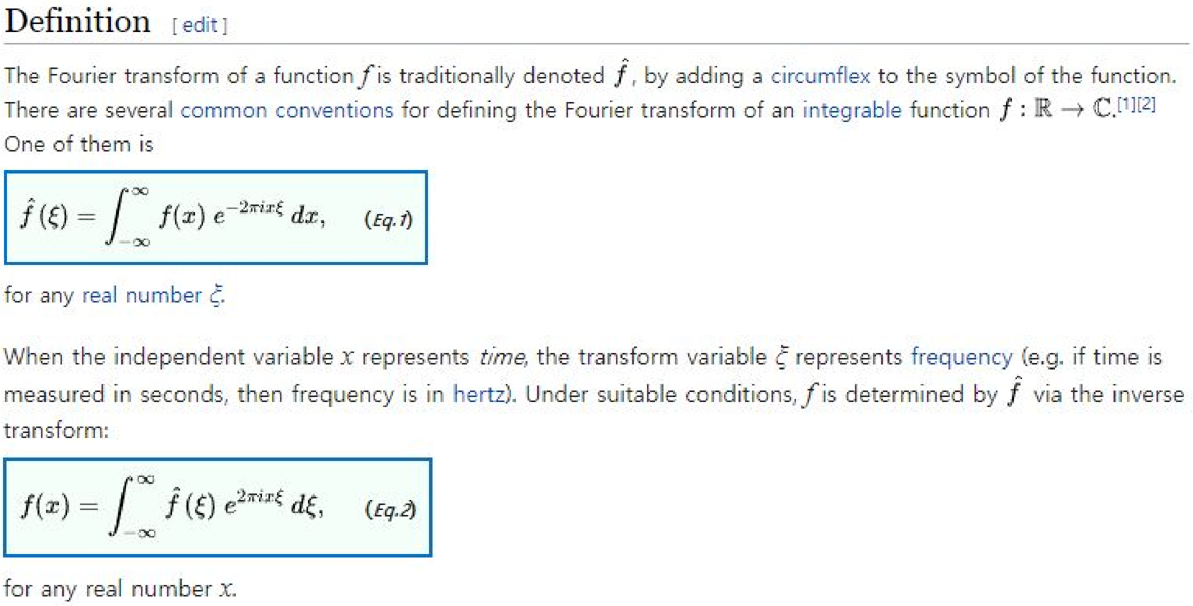
$$E\{\mu|z\} = z + \sigma^2 l'(z) \qquad with\ l'(z) = \frac{d}{dz}\log f(z)$$

등이 있다.

그러나 베이즈 사후 기댓값이나 사후 분포가 아닌, g, 즉 사전분포를 알고자 할 때는 문제가 생긴다.

그럼에도 불구하고, 사전분포 g를 추정하는 f-Modeling 기법이 존재한다.

2) 푸리에 변환(Fourier Transform)?



Definition [edit]

The Fourier transform of a function f is traditionally denoted $\hat{f}$, by adding a circumflex to the symbol of the function. There are several common conventions for defining the Fourier transform of an integrable function $f : \mathbb{R} \to \mathbb{C}$.[1][2] One of them is

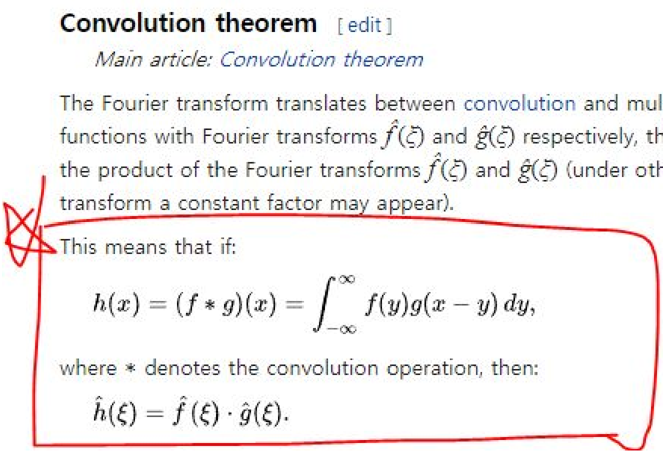
$$\boxed{\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x)\, e^{-2\pi i x \xi}\, dx, \qquad (Eq.1)}$$

for any real number ξ.

When the independent variable x represents *time*, the transform variable ξ represents frequency (e.g. if time is measured in seconds, then frequency is in hertz). Under suitable conditions, f is determined by $\hat{f}$ via the inverse transform:

$$\boxed{f(x) = \int_{-\infty}^{\infty} \hat{f}(\xi)\, e^{2\pi i x \xi}\, d\xi, \qquad (Eq.2)}$$

for any real number x.



Convolution theorem [edit]

Main article: Convolution theorem

The Fourier transform translates between convolution and multiplication of functions. If $f(x)$ and $g(x)$ are integrable functions with Fourier transforms $\hat{f}(\xi)$ and $\hat{g}(\xi)$ respectively, then the Fourier transform of the convolution is given by the product of the Fourier transforms $\hat{f}(\xi)$ and $\hat{g}(\xi)$ (under other conventions for the definition of the Fourier transform a constant factor may appear).

This means that if:

$$h(x) = (f * g)(x) = \int_{-\infty}^{\infty} f(y)g(x-y)\,dy,$$

where $*$ denotes the convolution operation, then:

$$\hat{h}(\xi) = \hat{f}(\xi)\cdot\hat{g}(\xi).$$

3) 푸리에 변환을 이용한 f-modeling과 g의 추정
위의 Convolution Thm에 의해,

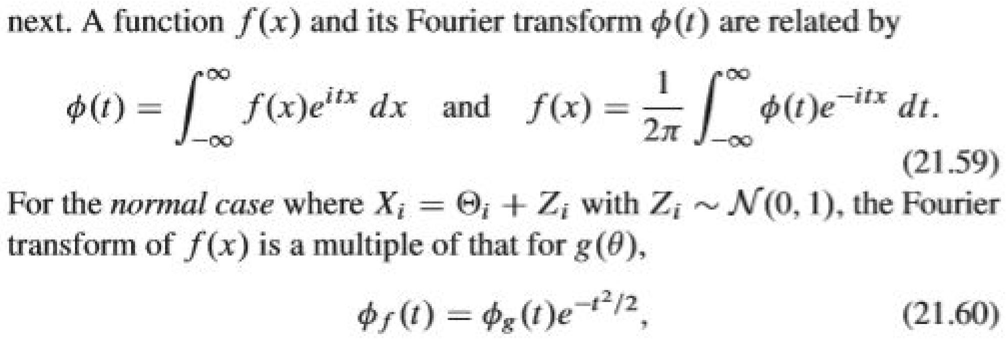
$$f = g * N(0,\sigma^2) \longrightarrow \hat{f} = \hat{g}\cdot\hat{N}(0,\sigma^2)$$

$$\xrightarrow{\sigma=1} \hat{f} = \hat{g}\,e^{-\frac{1}{2}t^2} \quad \text{if } \sigma=1$$

$$\longrightarrow \hat{g} = \hat{f}\,e^{\frac{1}{2}t^2}$$

$$\xrightarrow{\text{inverse transform}} g$$

따라서, (21.60)이 성립한다.



next. A function $f(x)$ and its Fourier transform $\phi(t)$ are related by

$$\phi(t) = \int_{-\infty}^{\infty} f(x)e^{itx}\,dx \quad \text{and} \quad f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \phi(t)e^{-itx}\,dt. \tag{21.59}$$

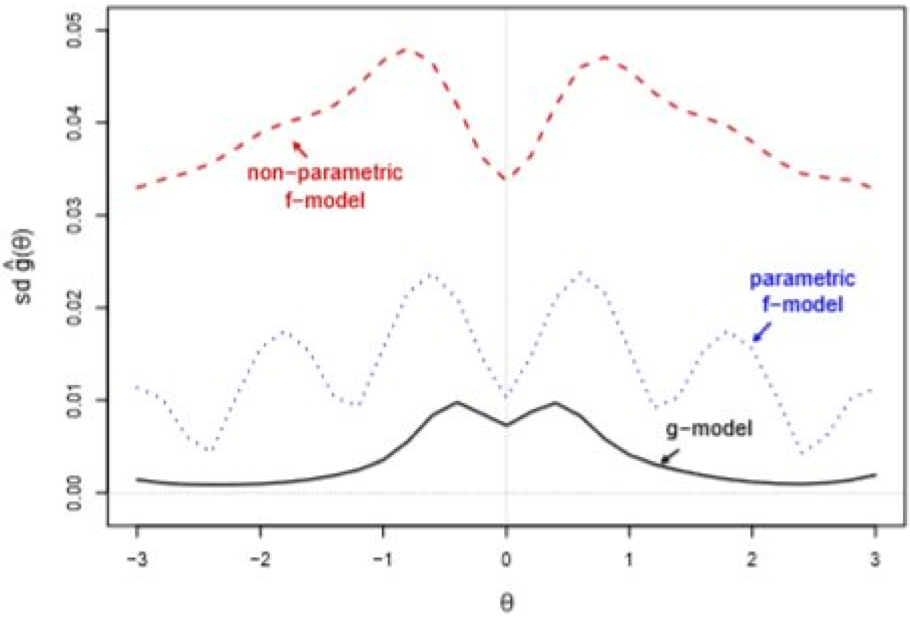
For the *normal case* where $X_i = \Theta_i + Z_i$ with $Z_i \sim \mathcal{N}(0,1)$, the Fourier transform of $f(x)$ is a multiple of that for $g(\theta)$,

$$\phi_f(t) = \phi_g(t)e^{-t^2/2}, \tag{21.60}$$

따라서 특별히 위의 경우에는 f-modeling을 통해 g를 추정할 수 있게 된다.
아래의 흐름을 따른다.

1. 관측 표본 X1, X2, ... , Xn을 통해 marginal distribution f를 직접 적합화하여 추정한다.
2. f의 푸리에 변환 f hat을 계산한다.
3. f hat*exp(t^2 / 2) = g hat을 역 푸리에 변환하여 g를 구한다.

여기에서,
(1) 비모수적 모델: 각 관측 값 Xi에 1/N 확률을 할 당하는 경험적 밀도 f를 사용
(2) 모수적 모델: 관측 표본 X1, X2, ... , Xn 에 포아송 회귀를 사용해 밀도 f를 추정
하여 f를 설정한 후, 위의 흐름대로 g를 구하고 두 가지 경우를 비교해보자.



21.2 g-모델링과 추정(g-Modeling and Estimation)

1) g-Modeling이란?
theta에 대한 직접적인 Modeling을 통해 g를 추정하는 방법. 일반적으로 g를 p-모수적 지수 패밀리(p-parameter exponential family)라고 가정한다. 그러면 g를 다음과 같이 쓸 수 있다.

$$g = g(\alpha) = e^{Q\alpha - \psi(\alpha)}$$

2) 데이터 생성 Process
책에서는 간단한 설명을 위해, 몇 가지 가정을 한다.

1. 가능한 theta의 값의 공간이 유한하다

$$\mathcal{T} = \left\{\theta_{(1)}, \theta_{(2)}, \dots, \theta_{(m)}\right\}$$

2. 커널 $p_i(\cdot|\cdot)$ 이 i에 종속되지 않는다

ex1) $X_i \sim \text{Poi}(\Theta_i)$

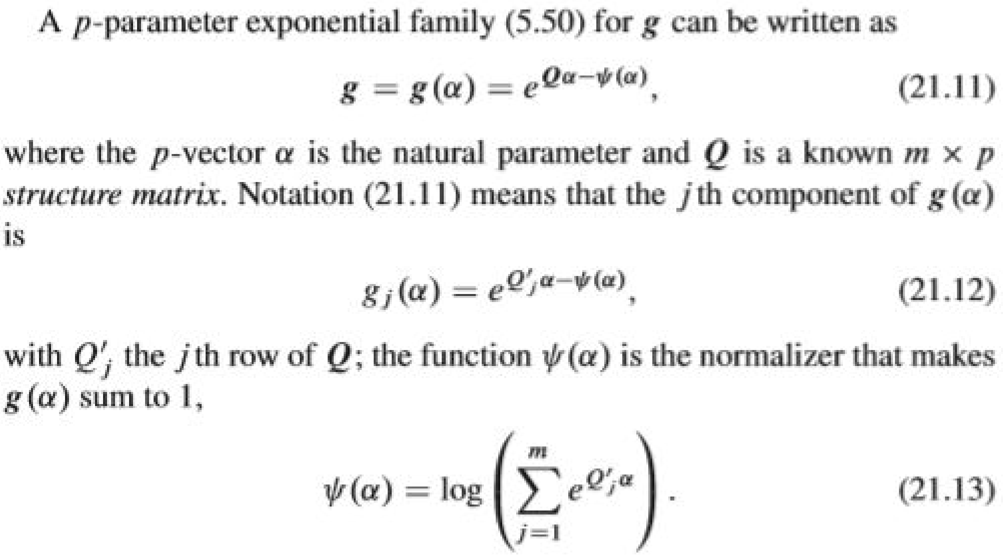
ex2) $X_i \sim \mathcal{N}(\Theta_i, \sigma^2)$

counter ex) $X_i \sim \text{Bi}(n_i, \Theta_i)$

3. X_i 관측치의 표본공간이 유한하고 이산이다

$$\mathcal{X} = \left\{x_{(1)}, x_{(2)}, \dots, x_{(n)}\right\}$$

이러한 가정에서 다음과 같이 수식을 전개할 수 있다.



A p-parameter exponential family (5.50) for g can be written as

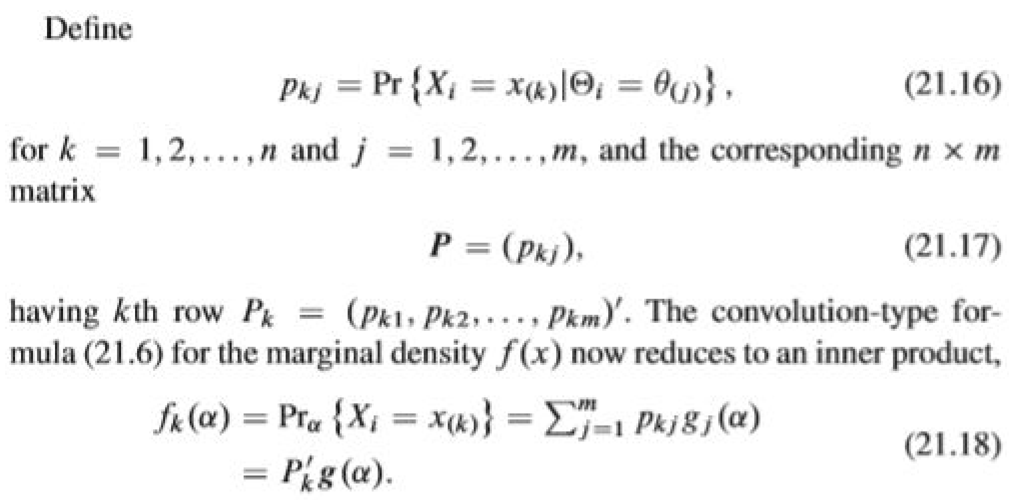
$$g = g(\alpha) = e^{Q\alpha - \psi(\alpha)}, \tag{21.11}$$

where the p-vector α is the natural parameter and Q is a known $m \times p$ *structure matrix*. Notation (21.11) means that the jth component of $g(\alpha)$ is

$$g_j(\alpha) = e^{Q'_j\alpha - \psi(\alpha)}, \tag{21.12}$$

with Q'_j the jth row of Q; the function $\psi(\alpha)$ is the normalizer that makes $g(\alpha)$ sum to 1,

$$\psi(\alpha) = \log\left(\sum_{j=1}^{m} e^{Q'_j\alpha}\right). \tag{21.13}$$



Define

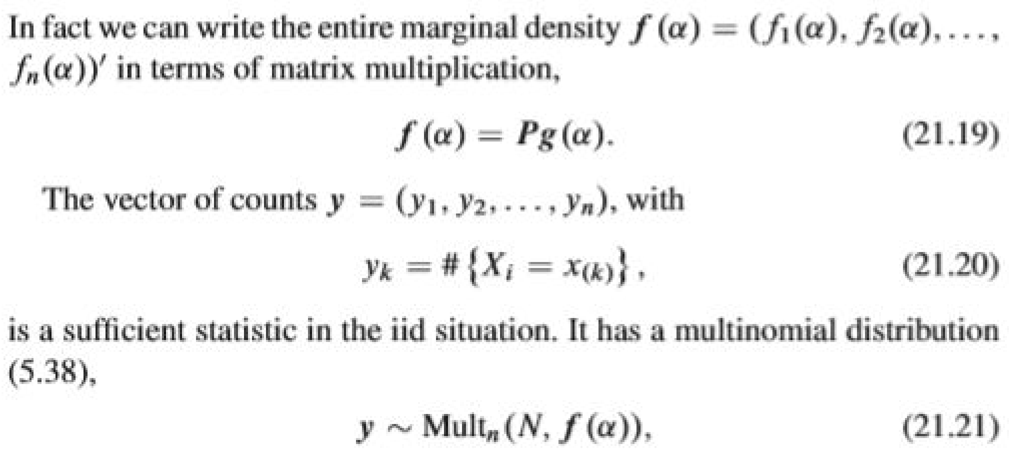
$$p_{kj} = \Pr\left\{X_i = x_{(k)} | \Theta_i = \theta_{(j)}\right\}, \tag{21.16}$$

for $k = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$, and the corresponding $n \times m$ matrix

$$\boldsymbol{P} = (p_{kj}), \tag{21.17}$$

having kth row $P_k = (p_{k1}, p_{k2}, \ldots, p_{km})'$. The convolution-type formula (21.6) for the marginal density $f(x)$ now reduces to an inner product,

$$\begin{aligned} f_k(\alpha) &= \Pr_\alpha\left\{X_i = x_{(k)}\right\} = \textstyle\sum_{j=1}^m p_{kj} g_j(\alpha) \\ &= P_k' g(\alpha). \end{aligned} \tag{21.18}$$



In fact we can write the entire marginal density $\boldsymbol{f}(\alpha) = (f_1(\alpha), f_2(\alpha), \ldots, f_n(\alpha))'$ in terms of matrix multiplication,

$$\boldsymbol{f}(\alpha) = \boldsymbol{P}\boldsymbol{g}(\alpha). \tag{21.19}$$

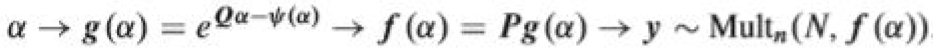
The vector of counts $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$, with

$$y_k = \#\left\{X_i = x_{(k)}\right\}, \tag{21.20}$$

is a sufficient statistic in the iid situation. It has a multinomial distribution (5.38),

$$\boldsymbol{y} \sim \text{Mult}_n(N, \boldsymbol{f}(\alpha)), \tag{21.21}$$

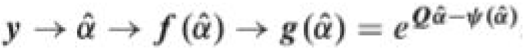
이러한 수식 전개는 g-modeling의 데이터 생성 Process를 다음과 같이 기술할 수 있도록 해준다. (정확하게는 X가 아니라 y)



$$\alpha \rightarrow \boldsymbol{g}(\alpha) = e^{\boldsymbol{Q}\alpha - \psi(\alpha)} \rightarrow \boldsymbol{f}(\alpha) = \boldsymbol{P}\boldsymbol{g}(\alpha) \rightarrow \boldsymbol{y} \sim \text{Mult}_n(N, \boldsymbol{f}(\alpha))$$

3) g-Modeling Estimation

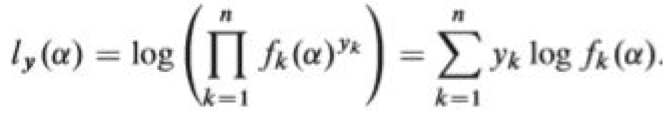
추론은 데이터 생성 Process의 역방향으로 진행된다.



$$\boldsymbol{y} \rightarrow \hat{\alpha} \rightarrow \boldsymbol{f}(\hat{\alpha}) \rightarrow \boldsymbol{g}(\hat{\alpha}) = e^{\boldsymbol{Q}\hat{\alpha} - \psi(\hat{\alpha})}$$

사실, 가운데의 f(alpha hat)을 구하는 것은 불필요하다. alpha의 값만 추정하면 바로 g(alpha hat)으로 계산할 수 있다.

일반적으로 alpha는



$$l_{\boldsymbol{y}}(\alpha) = \log\left(\prod_{k=1}^n f_k(\alpha)^{y_k}\right) = \sum_{k=1}^n y_k \log f_k(\alpha).$$

에 MLE를 적용해서 추정한다.

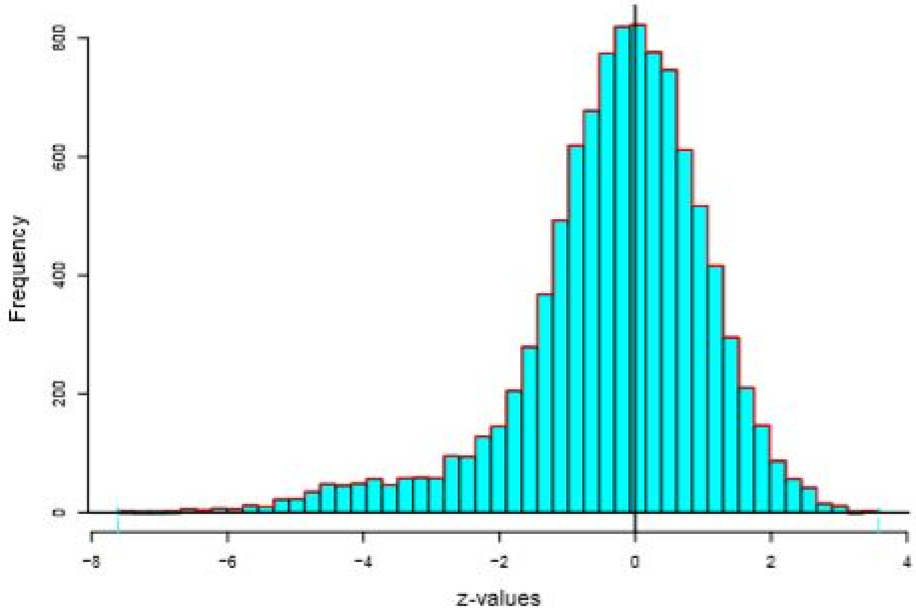
21.4. 예제 1: 인공적 미세배열(9000개의 mu는 0, 1000개의 mu는 N(-3, 1))

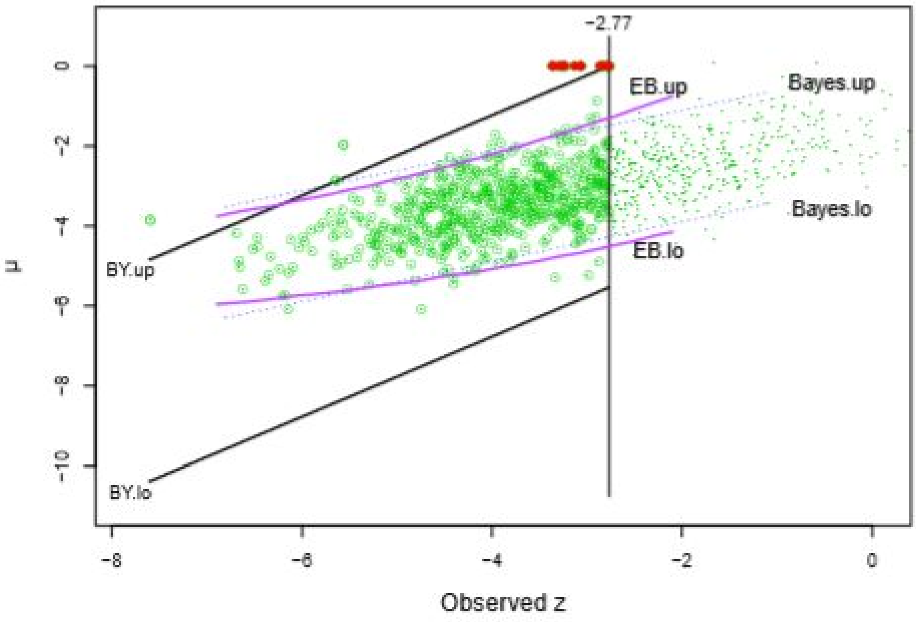
N =10,000 independent observations

$$z_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, 1), \qquad i = 1, 2, \ldots, N = 10{,}000,$$

with

$$\mu_i \sim \begin{cases} 0 & \text{for } i = 1, 2, \ldots, 9000 \\ \mathcal{N}(-3, 1) & \text{for } i = 9001, \ldots, 10{,}000. \end{cases}$$





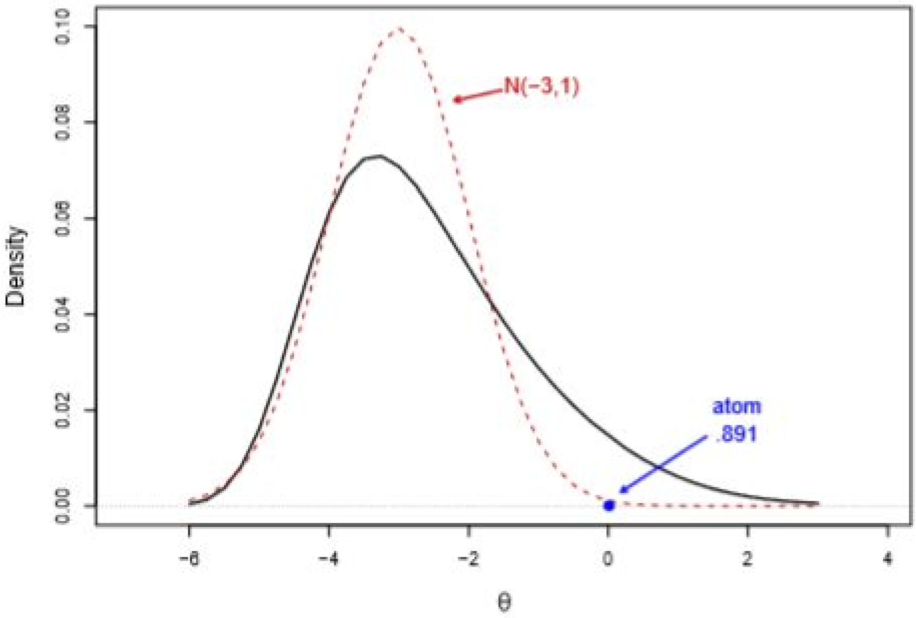


Figure 21.6 The heavy black curve is the g-modeling estimate of $g(\mu)$ for $\mu \neq 0$ in the artificial microarray example, suppressing the atom at zero, $\hat{g}(0) = 0.891$. It is only a rough estimate of the actual nonzero density $\mathcal{N}(-3, 1)$.

21.4. 예제 2: 전립선 연구 데이터(암환자 52명, 정상 50명, N=6033 유전자 활동성)

$n = 102$ men, 52 prostate cancer patients and 50 normal controls. Each man's gene expression levels were measured on a panel of $N = 6033$ genes, yielding a 6033×102 matrix of measurements x_{ij},

$$x_{ij} = \text{activity of } i\text{th gene for } j\text{th man}. \qquad (15.1)$$

For each gene, a two-sample t statistic (2.17) t_i was computed comparing gene i's expression levels for the 52 patients with those for the 50 controls. Under the null hypothesis H_{0i} that the patients' and the controls' responses come from the same normal distribution of gene i expression levels, t_i will follow a standard Student t distribution with 100 degrees of freedom, t_{100}. The transformation

$$z_i = \Phi^{-1}\left(F_{100}(t_i)\right), \qquad (15.2)$$

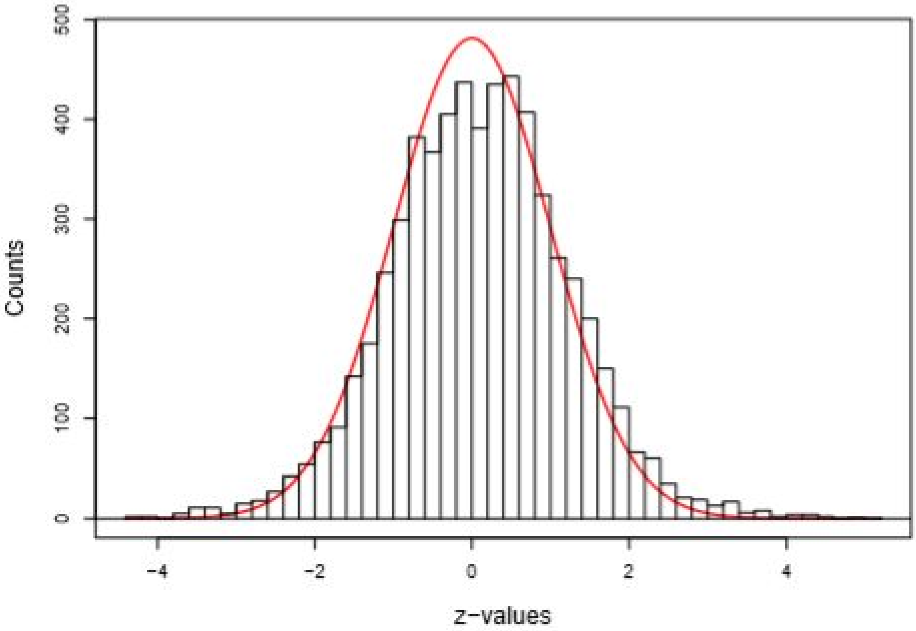
where F_{100} is the cdf of a t_{100} distribution and Φ^{-1} the inverse function of a standard normal cdf, makes z_i standard normal under the null hypothesis:

$$H_{0i} : z_i \sim \mathcal{N}(0, 1). \qquad (15.3)$$

Of course the investigators were hoping to spot some *non-null* genes, ones for which the patients and controls respond differently. It can be shown that a reasonable model for both null and non-null genes is[2†]

$$z_i \sim \mathcal{N}(\mu_i, 1), \qquad (15.4)$$

μ_i being the *effect size* for gene i. Null genes have $\mu_i = 0$, while the



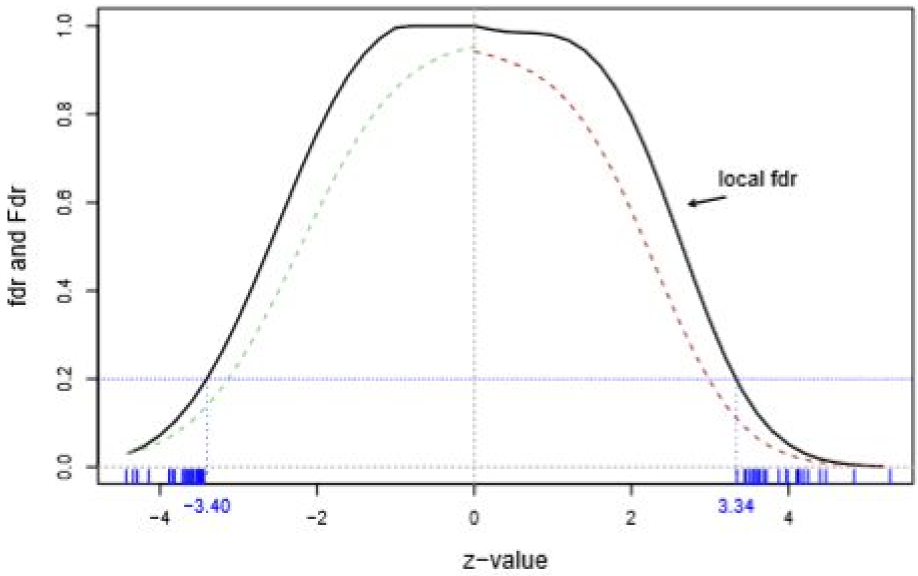


Figure 15.5 Local false-discovery rate estimate $\widehat{\text{fdr}}(z)$ (15.39) for prostate study of Figure 15.1; 27 genes on the right and 25 on the left, indicated by dashes, have $\widehat{\text{fdr}}(z_i) \leq 0.2$; light dashed curves are the left and right tail-area estimates $\widehat{\text{Fdr}}(z)$ (15.26).

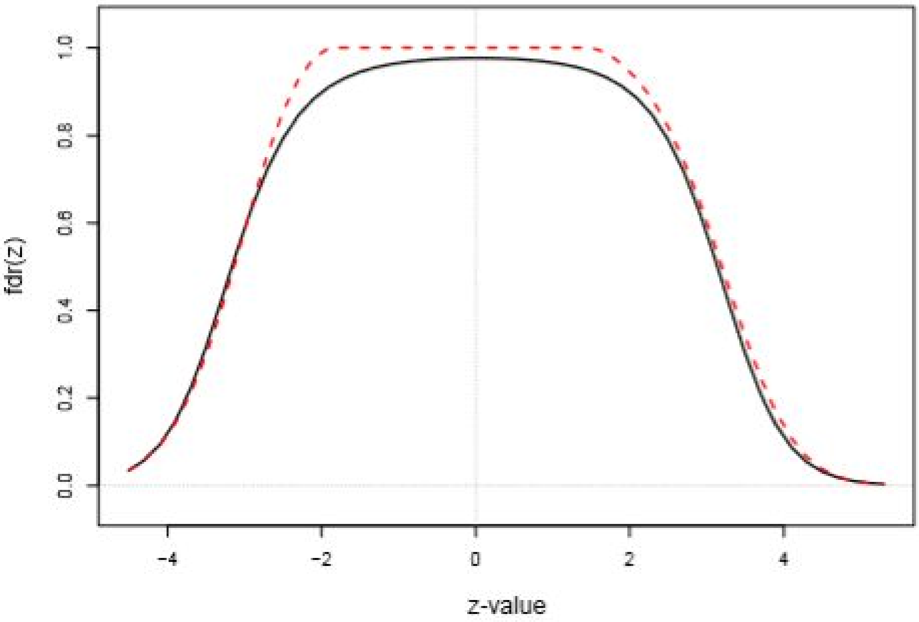
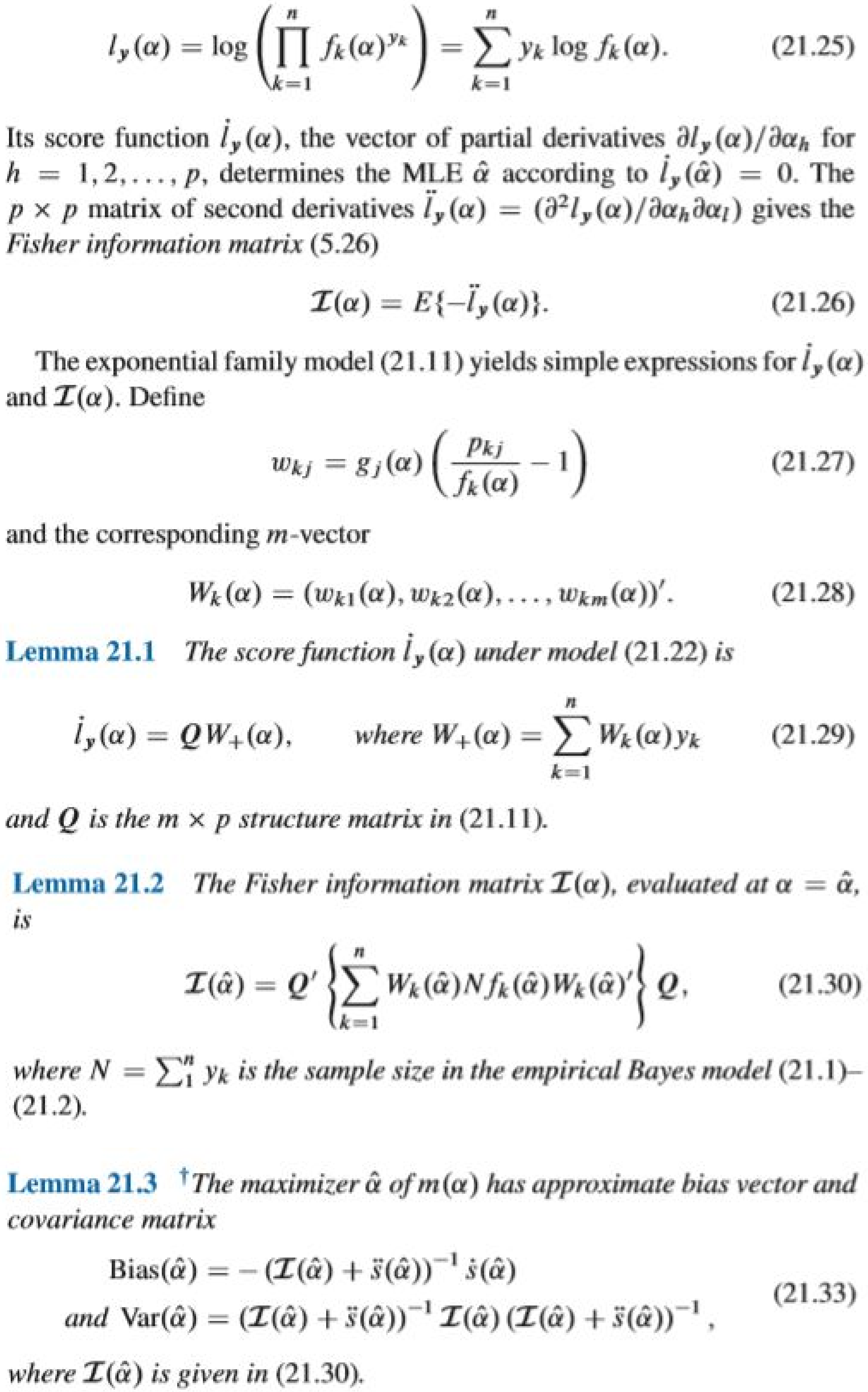


Figure 21.9 The black curve is the empirical Bayes estimated false-discovery rate $\widehat{\Pr}\{\mu = 0|z\}$ from g-modeling. For large values of $|z|$ it nearly matches the `locfdr` f-modeling estimate fdr(z), red curve.

Appendix. Mathematics in g-modeling



$$l_{\boldsymbol y}(\alpha) = \log\left(\prod_{k=1}^{n} f_k(\alpha)^{y_k}\right) = \sum_{k=1}^{n} y_k \log f_k(\alpha). \tag{21.25}$$

Its score function $\dot l_{\boldsymbol y}(\alpha)$, the vector of partial derivatives $\partial l_{\boldsymbol y}(\alpha)/\partial\alpha_h$ for $h = 1, 2, \ldots, p$, determines the MLE $\hat\alpha$ according to $\dot l_{\boldsymbol y}(\hat\alpha) = 0$. The $p \times p$ matrix of second derivatives $\ddot l_{\boldsymbol y}(\alpha) = (\partial^2 l_{\boldsymbol y}(\alpha)/\partial\alpha_h\partial\alpha_l)$ gives the *Fisher information matrix* (5.26)

$$\mathcal{I}(\alpha) = E\{-\ddot l_{\boldsymbol y}(\alpha)\}. \tag{21.26}$$

The exponential family model (21.11) yields simple expressions for $\dot l_{\boldsymbol y}(\alpha)$ and $\mathcal{I}(\alpha)$. Define

$$w_{kj} = g_j(\alpha)\left(\frac{p_{kj}}{f_k(\alpha)} - 1\right) \tag{21.27}$$

and the corresponding m-vector

$$W_k(\alpha) = (w_{k1}(\alpha), w_{k2}(\alpha), \ldots, w_{km}(\alpha))'. \tag{21.28}$$

Lemma 21.1 *The score function $\dot l_{\boldsymbol y}(\alpha)$ under model (21.22) is*

$$\dot l_{\boldsymbol y}(\alpha) = QW_+(\alpha), \qquad \textit{where } W_+(\alpha) = \sum_{k=1}^{n} W_k(\alpha) y_k \tag{21.29}$$

and Q is the $m \times p$ structure matrix in (21.11).

Lemma 21.2 *The Fisher information matrix $\mathcal{I}(\alpha)$, evaluated at $\alpha = \hat\alpha$, is*

$$\mathcal{I}(\hat\alpha) = Q'\left\{\sum_{k=1}^{n} W_k(\hat\alpha) N f_k(\hat\alpha) W_k(\hat\alpha)'\right\} Q, \tag{21.30}$$

where $N = \sum_1^n y_k$ is the sample size in the empirical Bayes model (21.1)–(21.2).

Lemma 21.3 †*The maximizer $\hat\alpha$ of $m(\alpha)$ has approximate bias vector and covariance matrix*

$$\begin{aligned} &\mathrm{Bias}(\hat\alpha) = -\left(\mathcal{I}(\hat\alpha) + \ddot s(\hat\alpha)\right)^{-1} \dot s(\hat\alpha) \\ \textit{and } &\mathrm{Var}(\hat\alpha) = \left(\mathcal{I}(\hat\alpha) + \ddot s(\hat\alpha)\right)^{-1} \mathcal{I}(\hat\alpha) \left(\mathcal{I}(\hat\alpha) + \ddot s(\hat\alpha)\right)^{-1}, \end{aligned} \tag{21.33}$$

where $\mathcal{I}(\hat\alpha)$ is given in (21.30).

Theorem 21.4 *The penalized maximum likelihood estimate* $\hat{\boldsymbol{g}} = \boldsymbol{g}(\hat{\alpha})$ *has estimated bias vector and covariance matrix*

$$\begin{aligned} \text{Bias}(\hat{\boldsymbol{g}}) &= \boldsymbol{D}(\hat{\alpha})\boldsymbol{Q}\,\mathit{Bias}(\hat{\alpha}) \\ \textit{and}\ \ \text{Var}(\hat{\boldsymbol{g}}) &= \boldsymbol{D}(\hat{\alpha})\boldsymbol{Q}\,\mathit{Var}(\hat{\alpha})\boldsymbol{Q}'\boldsymbol{D}(\hat{\alpha}) \end{aligned} \tag{21.38}$$

with Bias($\hat{\alpha}$) *and* Var($\hat{\alpha}$) *as in* (21.33).[3]