Chapter 2

Likelihood, Estimation and Testing

2.1 Likelihood and log-likelihood.

In this and the following section, we review briefly the basic ideas and results of likelihood inference: details may be found in any standard mathematical statistics text for beginning graduate students. A vector of data random variables, $\mathbf{Y}$, whose value $\mathbf{y}$ is observed, has one of a family of probability distributions $\{P_\theta; \theta \in \Theta\}$, indexed by a *parameter* θ in *parameter space* Θ. The goals of estimation are to make inferences about which P_θ gave rise to the observed $\mathbf{y}$, and to assess the uncertainty associated with this inference.

The *likelihood* is $L_{\mathbf{y}}(\theta) = P_\theta(\mathbf{y})$, a function of θ. The likelihood provides the connection between the data $\mathbf{y}$ and the probability model P_θ. A *statistic* is a function of the data random variables $\mathbf{Y}$, an *estimator* $T = T(\mathbf{Y})$ is a statistic taking values in Θ, while the *estimate* is $T(\mathbf{y})$, the value taken by the estimator that is used to estimate θ.

For example, suppose Y_i, $i = 1, ..., n$ are independent identically distributed Bernoulli random variables, $B(1, \theta)$, the indicators of success in n independent trials, each with success probability θ. Then $P_\theta(y) = \theta^y(1-\theta)^{1-y}$ $(y = 0, 1)$ for each trial, and $L(\theta) = \prod_1^n(\theta^{y_i}(1-\theta)^{1-y_i})$. The log-likelihood is

$$\ell(\theta) = \log L(\theta) = (\sum_1^n y_i)\log(\theta) + (n - \sum_1^n y_i)\log(1-\theta). \tag{2.1}$$

Note that the (log)-likelihood depends only on the value of $T = \sum_1^n Y_i$, the total number of successes, which has a binomial $B(n, \theta)$ distribution. The likelihood based on the binomial probability of the observed value t of T is

$$\begin{aligned} L(\theta) &= P_\theta(T = t) = \frac{n!}{k!(n-k)!}\theta^t(1-\theta)^{n-t} \\ \ell(\theta) &= \log L(\theta) = \text{const} + t\log(\theta) + (n-t)\log(1-\theta). \end{aligned} \tag{2.2}$$

Up to an additive constant which does not depend on θ, the log-likelihood (2.2) is the same as that of equation (2.1). A statistic T for which this is the case is said to be *sufficient*. It is immaterial whether one considers the likelihood based on the full data $\mathbf{Y} = (Y_1, \ldots, Y_n)$ or that based on a sufficient statistic such as T. Log-likelihoods are defined only up to an additive constant; only ratios of likelihoods are relevant for inference.

The *maximum likelihood estimate* (MLE) maximizes the likelihood as a function of θ, to give the value of θ that "best explains" the data $\mathbf{y}$. To obtain the MLE, we maximize $P_\theta(\mathbf{y})$, or $\log P_\theta(\mathbf{y})$ with respect to θ. For example, differentiating the log-likelihood (2.2) with respect to θ

$$\ell'(\theta) \quad = \quad \frac{t}{\theta} \; - \; \frac{n-t}{1-\theta}$$

Maximizing (2.2) by setting the derivative $\ell'(\theta)$ equal to 0 gives the MLE $\hat{\theta} = t/n$. In general, the equation $\ell'(\theta) = 0$ is known as the *likelihood equation*.

An estimator, $T(\mathbf{Y})$, is *unbiased* if, for any $\theta \in \Theta$, $\mathbf{Y} \sim P_\theta \implies \mathrm{E}(T(\mathbf{Y})) = \theta$, where $\mathrm{E}(\cdot)$ denotes expectation. We rewrite this definition as $\mathrm{E}_\theta(T(\mathbf{Y})) = \theta$ for all $\theta \in \Theta$, the subscript indicating the "true" θ-value—the value indexing the probability distribution with respect to which the expectations are evaluated. The *bias* of estimator $T(\mathbf{Y})$ is $b_T(\theta) = \mathrm{E}_\theta(T(\mathbf{Y})) - \theta$. An unbiased estimator is "correct, on average", over repetitions of the experiment. For example, if T is binomial $B(n, \theta)$, then $\mathrm{E}_\theta(T) = n\theta$, so the MLE is unbiased. However, unbiasedness alone is a very weak criterion. Some unbiased estimators may have poor properties, while many "good" estimators are biased. In particular many MLEs are biased, but under very broad conditions the bias decreases as the sample size increases.

A more important criterion is that an estimator should have small *mean square error* (mse). The mse of estimator $T(\mathbf{Y})$ is $\mathrm{mse}_\theta(T) = \mathrm{E}_\theta((T(\mathbf{Y}) - \theta)^2)$. If T is unbiased, $\mathrm{mse}_\theta(T) = \mathrm{var}_\theta(T)$, while, in general,

$$\mathrm{mse}_\theta(T) \quad = \quad \mathrm{var}_\theta(T) \; + \; (b_T(\theta))^2.$$

For example, for the unbiased maximum likelihood estimator T/n of the binomial parameter θ,

$$\begin{aligned} \mathrm{mse}(T/n) \quad &= \quad \mathrm{var}(T/n) \quad = \quad \mathrm{var}(T)/n^2 \\ &= \quad n\theta(1-\theta)/n^2 \quad = \quad \theta(1-\theta)/n \end{aligned} \tag{2.3}$$

Consider an n-sample $\mathbf{Y}^{(n)} = (Y_1, ..., Y_n)$, where the components Y_i are independent and identically distributed, and a sequence of estimators (T_n) where $T_n = T(\mathbf{Y}^{(n)})$. Then the sequence of estimators (T_n) is *consistent for* θ if, for every $\theta \in \Theta$, and every $\epsilon > 0$, $P_\theta(|T_n - \theta| > \epsilon) \to 0$ as $n \to \infty$. In the binomial example, equation (2.3), the mse of the MLE, T/n, tends to 0 as $n \to \infty$, and hence the sequence of estimators, (T/n), is consistent.

Clearly, maximization of $L(\theta)$ is equivalent to maximization of $\ell(\theta) = \log(L(\theta))$. Moreover, if $\alpha(\theta)$ is a one-one function of θ then $\widehat{\alpha} = \alpha(\widehat{\theta})$. Likelihood is a pointwise function of θ; transformation of the parameter space Θ does not alter the likelihood.

2.2 Estimation, information, and testing

In likelihood inference, a key entity is the expected log-likelihood $\mathrm{E}_{\theta_0}(\log(P_\theta(\mathbf{Y})))$. This notation denotes that the true value of the parameter θ is θ_0, and it is the distribution under θ_0 with respect to which expectations are taken. The expected log-likelihood is thus a function both of the true θ_0 and the hypothesized θ. From the convexity of the function $-\log(\cdot)$, it follows by Jensen's inequality that

$$\begin{aligned}
\mathrm{E}_{\theta_0}(\log(P_{\theta_0}(\mathbf{Y}))) - \log(P_\theta(\mathbf{Y})) &= \mathrm{E}_{\theta_0}\left(-\log\left(\frac{P_\theta(\mathbf{Y})}{P_{\theta_0}(\mathbf{Y})}\right)\right) \\
&\geq -\log \mathrm{E}_{\theta_0}\left(\frac{P_\theta(\mathbf{Y})}{P_{\theta_0}(\mathbf{Y})}\right) \\
&= -\log\left(\sum_{\mathbf{y}} \frac{P_\theta(\mathbf{y})}{P_{\theta_0}(\mathbf{y})} P_{\theta_0}(\mathbf{y})\right) \\
&= -\log(\sum_{\mathbf{y}} P_\theta(\mathbf{y})) \\
&= -\log(1) = 0 \qquad (2.4)
\end{aligned}$$

Thus the expected log-likelihood is maximized with respect to θ by $\theta = \theta_0$: the expected log-likelihood is maximized at the true value of the parameter. The non-negative difference

$$K(\theta;\theta_0) = \mathrm{E}_{\theta_0}(\log(P_{\theta_0}(\mathbf{Y})) - \log(P_\theta(\mathbf{Y})))$$

is known as the *Kullback-Leibler information* (Kullback and Leibler, 1951). One of the fairly immediate consequences of equation (2.4) is that under very broad conditions MLEs are consistent.

A related result is the *Cramèr-Rao lower bound* which says that (subject to some regularity conditions) no unbiased estimator can have a variance smaller than

$$\left[\mathrm{E}_{\theta_0}\left(-\frac{\partial^2}{\partial\theta^2}\log(P_{\theta_0}(\mathbf{Y}))\right)\right]^{-1}$$

The quantity within the square brackets is known as the *Fisher information.* The larger the information, the smaller the variance can be. Subject to a few additional conditions, MLEs are asymptotically approximately Normal (Gaussian), with mean θ_0, the true parameter value, and variance the inverse of the Fisher information. This says that, *in large samples,* MLEs are the *best estimators.* The required regularity conditions will be satisfied for most of the examples discussed in this monograph. A condition which may sometimes fail is that the true value θ_0 should lie in the interior of the parameter space Θ.

Of course, the value of θ_0 is unknown, but at least for large samples, the MLE $\hat{\theta}$ is close to the true value θ_0. Thus, $\hat{\theta}(\mathbf{y})$ may be substituted for θ_0 in the Fisher information, to obtain an estimate of the variance of the MLE. In fact, the expectation in the Fisher information can be hard to compute. Then, at least

for large samples, an alternative is the *observed information*

$$-\frac{\partial^2}{\partial\theta^2}\log(P_{\theta_0}(\mathbf{Y}))$$

evaluated by substituting the observed $\mathbf{y}$ for $\mathbf{Y}$ and $\hat{\theta}(\mathbf{y})$ for θ_0. The theoretical details and justification may be found in a mathematical statistics text.

To provide an example which should be familiar to readers, we return to the case of a binomial random variable: T is $B(n,\theta)$. As before (equation (2.2))

$$\ell(\theta) \;=\; \text{const} \;+\; T\;\log(\theta) \;+\; (n-T)\;\log(1-\theta)$$

and the MLE is T/n which has expectation θ and variance $\theta(1-\theta)/n$. Now,

$$\ell''(\theta) = -\frac{T}{\theta^2} - \frac{(n-T)}{(1-\theta)^2}.$$

Since $\mathrm{E}_\theta(T) = n\theta$, and $\mathrm{E}_\theta(n-T) = n(1-\theta)$, the Fisher information is $n/\theta(1-\theta)$. Thus in this example, the MLE has the smallest possible variance.

In practice, we estimate the variance as

$$\hat{\theta}(1-\hat{\theta})/n \;=\; t(n-t)/n^3$$

where t is the observed value of T. In fact, the same result is given by substituting $\hat{\theta} = t/n$ for θ in $-1/\ell''(\theta)$, without going through the expectation step. It is not in general true that the two methods of obtaining an estimated variance of the MLE give identical formulae.

Just as the maximum likelihood estimate is the value of the parameter that best explains the observed data, the maximized value of the likelihood is a measure of how well this parameter value is supported by the data, relative to how well other parameter values are supported by the observation of these data. Accordingly, we define

$$L(\Theta_0) \;=\; \max_{\theta\in\Theta_0}(L(\theta))$$

as a measure of support for the hypothesis $\theta \in \Theta_0 \subset \Theta$, and

$$\Lambda(\Theta_1 : \Theta_0) \;=\; L(\Theta_1)/L(\Theta_0)$$

as a measure of the relative support for the two hypotheses $\theta \in \Theta_1$ and $\theta \in \Theta_0$.

In the case when $\Theta_0 \subseteq \Theta_1$, $\Lambda \geq 1$, and $2\log_e \Lambda \geq 0$. Again subject to regularity conditions, asymptotically, if $\theta \in \Theta_0$ is true, then $2\log_e \Lambda$ is approximately distributed as a chi-squared (χ^2) random variable, with degrees of freedom equal to $\dim(\Theta_1) - \dim(\Theta_0)$. If the true value θ_0 is not in the hypothesis space Θ_0 but is in Θ_1, then $2\log_e \Lambda \to \infty$ at a rate which depends on the minimum Kullback-Leibler information:

$$\inf_{\theta\in\Theta_0} K(\theta;\theta_0) \;=\; \inf_{\theta\in\Theta_0} (\mathrm{E}_{\theta_0}(\log(P_{\theta_0}(\mathbf{Y})) \;-\; \log(P_\theta(\mathbf{Y}))\;))$$

The regularity conditions in order that these results hold are essentially the same as the ones needed for the asymptotic results about MLEs. They will hold in the examples we discuss.

In particular, much of the data in genetics is multinomial, consisting of counts of outcomes of various types. It is therefore useful to consider the case of the general multinomial model. Suppose there are r possible outcomes, having probabilities $p_i, i = 1, \ldots, r$, and a vector of parameters θ, so p_i is $p_i(\theta)$. The log-likelihood is

$$\ell = \text{const} + \sum_{i=1}^{r} n_i \log p_i \tag{2.5}$$

For the general model, $\sum_{i=1}^{r} p_i = 1$ with no other constraints:

$$\ell = \sum_{i=1}^{r} n_i \log p_i = \sum_{i=1}^{r-1} n_i \log p_i + n_r \log(1 - \sum_{i=1}^{r-1} p_i)$$

$$\frac{\partial \ell}{\partial p_i} = \frac{n_i}{p_i} - \frac{n_r}{p_r}$$

for $i = 1, \ldots, r-1$, giving the MLE $\widehat{p_i} = n_i/n$. The maximized value of the log-likelihood is

$$\hat{\ell} = \sum_{i=1}^{r} n_i \log \widehat{p_i} = \sum_{i=1}^{r} n_i \log n_i - n \log n \tag{2.6}$$

The dimension of the general hypothesis space is $r - 1$ since the p_i are constrained to sum to 1.

Under a constrained model, the outcome probabilities p_i will be functions of some parameters θ_j, where normally the dimensionality of θ is less than $r - 1$. To estimate θ we must solve the equations

$$\frac{\partial \ell}{\partial \theta_j} = \sum_{i=1}^{r} \frac{n_i}{p_i} \frac{\partial p_i}{\partial \theta_j} \quad \text{for all } j$$

It is also possible to find the Fisher information:

$$\frac{\partial^2 \ell}{\partial \theta_j \partial \theta_k} = -\sum_{i=1}^{r} \frac{n_i}{p_i^2} \frac{\partial p_i}{\partial \theta_j} \frac{\partial p_i}{\partial \theta_k} + \sum_{i=1}^{r} \frac{n_i}{p_i} \frac{\partial^2 p_i}{\partial \theta_j \partial \theta_k}$$

Now $\mathrm{E}(n_i) = np_i$, so

$$\begin{aligned} \mathrm{E}\left(-\frac{\partial^2 \ell}{\partial \theta_j \partial \theta_k}\right) &= \sum_{i=1}^{r} \frac{np_i}{p_i^2} \frac{\partial p_i}{\partial \theta_j} \frac{\partial p_i}{\partial \theta_k} - \sum_{i=1}^{r} \frac{np_i}{p_i} \frac{\partial^2 p_i}{\partial \theta_j \partial \theta_k} \\ &= n \sum_{i=1}^{r} \frac{1}{p_i} \frac{\partial p_i}{\partial \theta_j} \frac{\partial p_i}{\partial \theta_k} - n \sum_{i=1}^{r} \frac{\partial^2 p_i}{\partial \theta_j \partial \theta_k} \end{aligned}$$

$$= n\sum_{i=1}^{r}\frac{1}{p_i}\frac{\partial p_i}{\partial\theta_j}\frac{\partial p_i}{\partial\theta_k} - n\frac{\partial^2\sum_{i=1}^{r}p_i}{\partial\theta_j\partial\theta_k}$$

$$= n\sum_{i=1}^{r}\frac{1}{p_i}\frac{\partial p_i}{\partial\theta_j}\frac{\partial p_i}{\partial\theta_k} \tag{2.7}$$

since $\sum_{i=1}^{r} p_i \equiv 1$. Equation (2.7) is sometimes known as Fisher's formula.

2.3 Population allele frequencies

In this section, we consider three examples of the above formulation, in the context of estimation of population allele frequencies. Consider a single genetic locus, with k alleles A_j and having population frequencies q_j, $j = 1,\ldots,k$. Now in a random-mating population, the allelic types of the maternal and paternal genes in an individual are independent. Thus the probability an individual is homozygous A_jA_j is q_j^2, while the probability the individual is heterozygous A_jA_l $(j < l)$ is $2q_jq_l$. These genotype frequencies are known as Hardy-Weinberg proportions, and a population exhibiting genotypes in these proportions is said to be in Hardy-Weinberg equilibrium.

First suppose the alleles A_j are codominant, and a random sample of n individuals is taken from a population assumed to be in Hardy-Weinberg proportions. Suppose that n_{jl} $(j \le l)$ individuals are observed to be of genotype A_jA_l. As above (equation (2.5)), the log-likelihood is

$$\ell = \text{const} + \sum_{j=1}^{k} n_{jj}log(q_j^2) + \sum_{1\le j<l\le k} n_{jl}\log(2q_jq_l)$$

$$= \text{const} + \sum_{j=1}^{k} m_j log(q_j)$$

where $m_j = 2n_{jj} + \sum_{l<j} n_{lj} + \sum_{j<l} n_{jl}$, is the number of A_j alleles among the $2n$ alleles of th n sampled individuals. Hence the MLE of q_j is $m_j/2n$, the sample proportions of the allelic types. The MLE has variance $q_j(1-q_j)/2n$.

Most natural populations show some degree of subdivision or structure, and so do not exhibit Hardy-Weinberg equilibrium. The deviation from Hardy-Weinberg proportions may be small and detectable only from large samples. Testing Hardy-Weinberg proportions is straightforward in the case of a random sample of individuals typed at a locus with codominant alleles. Under the general model, there are $\frac{1}{2}k(k+1)$ genotypic counts n_{jl} with the maximum log-likelihood given by equation (2.6), while assuming Hardy-Weinberg proportions, there are k allelic counts m_j with the same multinomial form of maximum log-likelihood. The dimension of the larger hypothesis space is $\frac{1}{2}k(k+1)-1$, and of the smaller is $k-1$. If Hardy-Weinberg proportions do hold in the population, then twice the difference in log-likelihoods is distributed as a chi-squared random variable on $\frac{1}{2}k(k-1)$ degrees of freedom ($\chi^2_{\frac{1}{2}k(k-1)}$).

A locus with two alleles is said to be *diallelic*, although the alternative *biallelic* is now used increasingly in the literature. As an example of the use of Fisher information, consider the case of a diallelic trait locus, with a recessive allele with allele frequency q. Assuming Hardy-Weinberg proportions, there are two phenotypic categories ($r = 2$), with population frequencies $p_1(q) = q^2$, $p_2(q) = 1 - q^2$. Suppose n individuals are sampled, and t are found to be of the recessive phenotype. Since $p_1(q)$ is a 1-1 transformation of q, over the parameter space $0 \le q \le 1$, the MLE of q is $\widehat{q} = \sqrt{(\widehat{p_1})} = \sqrt{t/n}$. This may also be verified by direct differentiation of the log-likelihood

$$\ell(q) \;=\; t\,\log(q^2) \;+\; (n-t)\,\log(1-q^2)$$

Now also

$$\frac{\partial p_1}{\partial q} \;=\; 2q \;\text{ and }\; \frac{\partial p_2}{\partial q} \;=\; -\,2q,$$

so using Fisher's equation (2.7)

$$\begin{aligned} \mathrm{E}\left(-\frac{\partial^2 \ell}{\partial q^2}\right) &= n\left(\frac{1}{q^2}(2q)^2 \;+\; \frac{1}{1-q^2}(-2q)^2\right) \\ &= \frac{4n}{(1-q^2)} \end{aligned}$$

Thus the large-sample variance of the MLE is $(1-q^2)/4n$, which is $(1+q)/2q$ times larger than the variance $q(1-q)/2n$ obtained if the genotypes were observable. Of course, when there are only two phenotypes, there are no degrees of freedom to test for Hardy-Weinberg proportions.

As another example, consider the estimation of allele frequencies at a diallelic locus, when, instead of random individuals, we sample parent-offspring pairs. This might arise, for example, if our sample was of mothers with new-born infants. Table 2.1 shows the conditional and joint probabilities of feasible mother-child combinations.

parent genotype	probability	Pr(child|parent) for child genotype			Pr(parent, child) for child genotype		
		A_iA_i	A_iA_j	A_iA_l	A_iA_i	A_iA_j	A_iA_l
A_iA_i	q_i^2	q_i	q_j	q_l	q_i^3	$q_i^2 q_j$	$q_i^2 q_l$
A_iA_j	$2q_iq_j$	$\frac{1}{2}q_i$	$\frac{1}{2}(q_i+q_j)$	$\frac{1}{2}q_l$	$q_i^2 q_j$	$q_iq_j(q_i+q_j)$	$q_iq_jq_l$

TABLE 2.1. *Conditional and joint probabilities of feasible mother-child genotype combinations*

In the case $k = 2$, let n_{ij} be the number of mother-offspring pairs in which the mother has genotype g_i and the offspring has genotype g_j, where $g_0 = A_1A_1$, $g_1 = A_1A_2$ and $g_2 = A_2A_2$. Since $q_1 + q_2 = 1$, every term in Table 2.1 is a product of allele frequencies, and the multinomial log-likelihood reduces to

$$\ell \;=\; \sum_{(i,j)} n_{ij} \log \Pr(g_i, g_j)$$

$$
\begin{aligned}
&= n_{00}\log(q_1^3) + n_{01}\log(q_1^2 q_2) + n_{10}\log(q_1^2 q_2) + n_{11}\log(q_1 q_2) \\
&\quad + n_{12}\log(q_1 q_2^2) + n_{21}\log(q_1 q_2^2) + n_{22}\log(q_2^3) \\
&= (3n_{00} + 2(n_{01} + n_{10}) + n_{11} + n_{12} + n_{21})\log q_1 + \\
&\quad (3n_{22} + 2(n_{21} + n_{12}) + n_{11} + n_{10} + n_{01})\log q_2 \\
&= m_1 \log q_1 + m_2 \log q_2 \qquad (2.8)
\end{aligned}
$$

where m_1 is the number of distinct A_1 alleles, and m_2 is the number of distinct A_2 alleles, in the set of pairs. (By "distinct" we mean that we do not count both copies of an allele which segregates from parent to offspring.) The MLE of q_1 is thus $m_1/(m_1 + m_2)$. Note that

$$
\begin{aligned}
m_1 + m_2 &= 3(n_{00} + n_{01} + n_{10} + n_{21} + n_{12} + n_{22}) + 2n_{11} \\
&= 3n - n_{11}
\end{aligned}
$$

where n is the number of parent-offspring pairs. Although finding the MLE is a matter of "gene-counting", the total number of distinct genes to be counted is not $4n$, since parent and offspring share one gene, nor even $3n$. For each $(g_1, g_1) = (A_1A_2, A_1A_2)$ pair, one gene of allelic type A_1 and one of type A_2 can be counted, but the third distinct gene may be of either type, and does not contribute to the likelihood.

	factor freq.		phenotype frequencies			
	A	B	A	B	AB	0
Data			0.422	0.206	0.078	0.294
H_1 theory	p	q	$p(1-q)$	$p(1-q)$	pq	$(1-p)(1-q)$
H_1 fitted	0.500	0.284	0.358	0.142	0.142	0.358
H_2 theory	p	q	$p^2 + 2pr$	$q^2 + 2qr$	$2pq$	r^2
H_2 fitted	0.295	0.155	0.411	0.194	0.091	0.303

TABLE 2.2. *Data and estimated frequencies for Bernstein's analysis of ABO blood type determination*

As a final example in this section, we consider the classic analysis of Bernstein (1925) who established the mode of determination of the *ABO* blood types using data on population phenotype frequencies. The development in terms of likelihood ratio tests is due to Edwards (1972). Bernstein reported *ABO* blood types on a sample of 502 individuals: 42.2% type A, 20.6% type B, 7.8% type AB and 29.4% type O (Table 2.2). It is a minor mystery of Bernstein's data that these proportions do not give integer counts with a sample of $n = 502$; however we ignore that question here.

Now there were two prevailing hypotheses for the determination of the *ABO* blood types, the first, H_1 being that A and B are independently inherited factors, The frequency of individuals in the sample having the factor A is 0.500 (blood types A or AB), and B is 0.284 (blood types B or AB). As pointed out by Bernstein, independence of the factors would give an AB frequency of $0.500 \times$

$0.284 = 0.142$ much larger than the 0.078 observed. More rigorously, we can perform a likelihood ratio test of H_1 against the general multinomial alternative. For the general alternative, the fitted frequencies are the observed frequencies, and the log-likelihood is

$$\begin{aligned} \hat{\ell} &= 502(.422 \log .422 + .206 \log .206 + .078 \log .078 + .294 \log .294) \\ &= -626.71 \end{aligned}$$

Under the hypothesis H_1 the estimated frequencies are as shown in Table 2.2, and the log-likelihood is

$$\begin{aligned} \ell_1 &= 502(.422 \log .358 + .206 \log .142 + .078 \log .142 + .294 \log .358) \\ &= -647.50 \end{aligned}$$

Twice the log-likelihood difference is 41.58, and would be the value of a χ^2_1 random variable if H_1 were true. Clearly, H_1 is rejected.

The second hypothesis, H_2 is that A and B are the two non-null alleles of a single system. If the three alleles A, B and O have frequencies p, q and r $(p+q+r=1)$, and if Hardy-Weinberg proportions hold, then the frequencies of the four blood types are $p^2 + 2pr$, $q^2 + 2qr$, $2pq$ and r^2 (Table 2.2). Bernstein pointed out that the sum of the A and O blood type frequencies is $(p+r)^2$, or one minus the square root of this frequency is $(1 - p - r) = q$. Similarly one minus the square root of the sum of the B and O blood type frequencies is p, and the square root of the O blood type frequency is r. The sum of these three numbers should be one. For his data

$$(1 - \sqrt{0.422 + 0.294}) + (1 - \sqrt{0.206 + 0.294}) + \sqrt{0.294} = 0.99$$

which is close to one, suggesting a good fit. Again, more formally, we may perform a likelihood ratio test. However, finding the MLEs of the parameters p, q and r is not simple; in fact, we shall discover in section 2.5 that these MLEs are $\hat{p} = 0.2945$ and $\hat{q} = 0.1547$, with the resulting fitted frequencies given in Table 2.2. The fitted frequencies are all close to the observed ones, and the log-likelihood is

$$\begin{aligned} \ell_2 &= 502(.422 \log .4114 + .206 \log .1942 + .078 \log .0911 + .294 \log .3033) \\ &= -627.52 \end{aligned}$$

Twice the log-likelihood difference between this and the general alternative is now only 1.62. Again, this is the value of a χ^2_1 random variable if H_2 is true, and this hypothesis is not rejected.

Of course, there is also evidence on the ABO blood type determination in the transmission of genes from parents to children. For example, under H_2 an AB parent cannot have an O child, while under H_1 this may happen. Both inheritance patterns and population frequencies can provide information on genetic mechanisms. Bernstein's analysis is perhaps the first example of determination of the genetic model underlying a trait from population frequency data, rather than from the inheritance patterns in pedigrees.

2.4 The EM algorithm; general formulation

Many of the problems in genetic analysis fall within the classical *missing data* or *latent variable* framework. Many data may be missing, in the sense that some pedigree members may be unobserved, or not all marker phenotypes observed even for some available pedigree members. We therefore prefer the term *latent variables* for unobservable features, such as the multilocus haplotypes of individuals (equation (1.5)), or the meiosis indicators that specify the descent of genes in pedigrees (equation (1.2)). An important approach to likelihood analysis, and specifically to maximum likelihood estimation, in such latent variable problems was provided by Dempster et al. (1977). Although their approach had been developed previously in many special cases, they provided the overall framework, giving it the name the *EM algorithm*, or *expectation-maximization* algorithm.

For generality, we denote latent variables by $\mathbf{X}$, bearing in mind that for our examples, these will generally be meiosis indicators, genotypes, indicators of genotypic status or linkage phase, or genotypic or allelic counts. For simplicity, we use summation rather than integration over latent variables, since for the majority of our examples, the latent variables are discrete. The structure of any latent variable problem is that the likelihood $L(\theta)$ from observed data values $\mathbf{y}$ of the data random variables $\mathbf{Y}$ is

$$L(\theta) \;=\; P_\theta(\mathbf{Y}=\mathbf{y}) \;=\; \sum_{\mathbf{x}} P_\theta((\mathbf{X},\mathbf{Y})=(\mathbf{x},\mathbf{y}))$$

Now the joint probability of data and latent variables is

$$P_\theta((\mathbf{X},\mathbf{Y})=(\mathbf{x},\mathbf{y})) \;=\; P_\theta((\mathbf{X}=\mathbf{x}|\mathbf{Y}=\mathbf{y}))P_\theta(\mathbf{Y}=\mathbf{y}).$$

This joint probability, considered as a likelihood of parameter θ, is known as the *complete-data likelihood.* Taking logs and rearranging,

$$\log L(\theta) \;=\; \log P_\theta((\mathbf{X},\mathbf{Y})=(\mathbf{x},\mathbf{y})) \;-\; \log P_\theta((\mathbf{X}=\mathbf{x}|\mathbf{Y}=\mathbf{y})). \tag{2.9}$$

Now define

$$\begin{aligned} H_{\mathbf{y}}(\theta;\theta^*) &= \mathrm{E}_{\theta^*}(\log P_\theta(\mathbf{X},\mathbf{Y}) \mid \mathbf{Y}=\mathbf{y}) \\ G_{\mathbf{y}}(\theta;\theta^*) &= \mathrm{E}_{\theta^*}(\log P_\theta(\mathbf{X} \mid \mathbf{Y}=\mathbf{y}) \mid \mathbf{Y}=\mathbf{y}) \end{aligned}$$

The function $H_{\mathbf{y}}(\theta;\theta^*)$ is the *expected complete-data log-likelihood*, while the Kullback-Leibler information (section 2.2) in the conditional distribution of $\mathbf{X}$ given $\mathbf{Y}=\mathbf{y}$ is

$$K_{\mathbf{y}}(\theta;\theta^*) \;=\; G_{\mathbf{y}}(\theta^*;\theta^*) - G_{\mathbf{y}}(\theta;\theta^*).$$

Taking expectations over $\mathbf{X}$, under model θ^*, conditional upon $\mathbf{Y}=\mathbf{y}$, in equation (2.9) we obtain

$$\log L(\theta) \;=\; H_{\mathbf{y}}(\theta;\theta^*) \;-\; G_{\mathbf{y}}(\theta;\theta^*)$$

since $L(\theta)$ does not depend on the random variable $\mathbf{X}$. Now suppose that $\tilde{\theta}$ maximizes $H_{\mathbf{y}}(\theta;\theta^*)$ over θ, and consider

$$\begin{aligned} \log L(\tilde{\theta}) - \log L(\theta^*) &= (H_{\mathbf{y}}(\tilde{\theta};\theta^*) - H_{\mathbf{y}}(\theta^*;\theta^*)) + \\ &\qquad (G_{\mathbf{y}}(\theta^*;\theta^*) - G_{\mathbf{y}}(\tilde{\theta};\theta^*)) \end{aligned} \tag{2.10}$$

Now $\tilde{\theta}$ maximizes $H_{\mathbf{y}}(\theta;\theta^*)$. Also, for any probability distributions $P_\theta(\cdot)$ indexed by parameter θ, $\mathrm{E}_{\theta^*}(P_\theta(\mathbf{X}))$ is maximized by $\theta = \theta^*$ (equation (2.4)). Thus

$$H_{\mathbf{y}}(\tilde{\theta};\theta^*) \geq H_{\mathbf{y}}(\theta^*;\theta^*) \quad \text{and} \quad G_{\mathbf{y}}(\theta^*;\theta^*) \geq G_{\mathbf{y}}(\tilde{\theta};\theta^*) \tag{2.11}$$

$$\text{so} \quad \log L(\tilde{\theta}) \geq \log L(\theta^*) \tag{2.12}$$

with equality only if $\tilde{\theta}$ and θ^* provide the same conditional distribution for $\mathbf{X}$ given $\mathbf{Y} = \mathbf{y}$.

Thus we have the EM algorithm for finding MLEs (Dempster et al., 1977).
E-step (expectation):

At the current estimate θ^* compute $H_{\mathbf{y}}(\theta;\theta^*) = \mathrm{E}_{\theta^*}(\log P_\theta(\mathbf{X},\mathbf{Y}) \mid \mathbf{Y} = \mathbf{y})$
M-step (maximization):

Maximize $H_{\mathbf{y}}(\theta;\theta^*)$ with respect to θ to obtain a new estimate $\tilde{\theta}$.
E-steps and M-steps are alternated, and, in accordance with equation (2.12) the likelihood is non-decreasing over the process. Where the likelihood surface is unimodal, convergence to the MLE is assured, although it may be slow.

In the case when the complete-data joint probability $P_\theta((\mathbf{X},\mathbf{Y}) = (\mathbf{x},\mathbf{y}))$ is an exponential family of full rank, the EM equations take a particularly simple form. If $T_j(\mathbf{X},\mathbf{Y})$, $j = 1,...,k$ are the natural sufficient statistics, with corresponding natural parameters $\alpha_j(\theta)$, $j = 1,...,k$,

$$\begin{aligned} P_\theta((\mathbf{X},\mathbf{Y}) = (\mathbf{x},\mathbf{y})) &= c(\theta)\ \exp(\sum_{j=1}^{k} T_j(\mathbf{x},\mathbf{y})\alpha_j(\theta)) \\ H_{\mathbf{y}}(\theta;\theta^*) &= \log c(\theta) - \sum_{j=1}^{k} \mathrm{E}_{\theta^*}(T_j(\mathbf{X},\mathbf{y}) \mid \mathbf{Y} = \mathbf{y})\alpha_j(\theta) \\ \frac{\partial H_{\mathbf{y}}}{\partial \alpha_j} &= \frac{\partial \log c(\theta)}{\partial \alpha_j} - \mathrm{E}_{\theta^*}(T_j(\mathbf{X},\mathbf{y}) \mid \mathbf{Y} = \mathbf{y}) \\ &= \mathrm{E}_\theta(T_j(\mathbf{X},\mathbf{Y})) - \mathrm{E}_{\theta^*}(T_j(\mathbf{X},\mathbf{y}) \mid \mathbf{Y} = \mathbf{y}). \end{aligned}$$

Thus to implement EM in this case we compute the conditional expectations of the natural sufficient statistics T_j, give the data $\mathbf{Y}$, under the current estimate θ^* and set them equal to their unconditioned expectations to obtain the new estimates $\tilde{\theta}$. Thus the EM algorithm is often discussed in terms of the E-step "imputing" the latent variables conditional upon the data $\mathbf{Y}$ under the currrent estimates θ^*, and the M-step being the maximization of the complete-data log-likelihood, using these imputed variables. Although for many practical cases this is so, some care is needed. Only in the case of an exponential family of full rank is the expected

complete-data log-likelihood a linear function of the natural sufficient statistics T_j. Even in this case, T_j may not be linear in the latent variables $\mathbf{X}$, so that

$$\mathrm{E}_{\theta^*}(T_j(\mathbf{X},\mathbf{y}) \mid \mathbf{Y}=\mathbf{y}) \quad \neq \quad T_j(\mathrm{E}_{\theta^*}(\mathbf{X} \mid \mathbf{Y}=\mathbf{y}),\ \mathbf{y})$$

An example is given in section 2.6.

This monograph will take a likelihood approach to inference, but some of the methods are closely related to those of Bayesian inference. In Bayesian inference, parameters θ are given a *prior* probability distribution $\pi(\theta)$ which expresses information or belief about parameter values before data $\mathbf{Y}$ are observed. After data are observed, beliefs about θ are expressed via the *posterior* distribution

$$\pi(\theta|\mathbf{Y}) \quad = \quad \frac{\pi(\theta)\mathrm{Pr}(\mathbf{Y};\theta)}{\int_\theta \pi(\theta)\mathrm{Pr}(\mathbf{Y};\theta)d\theta}$$

Bayesian inferences are based on the posterior probability distribution for parameters of interest. Clearly the likelihood $L(\theta) = \mathrm{Pr}(\mathbf{Y};\theta)$ is closely related to the Bayesian posterior.

Bayesian inference is often useful where there are many parameters, only a few of which are of interest. The nuisance parameters are integrated over to provide a marginal posterior distribution for a parameter of interest. This is thus often a convenient way to view a multi-parameter likelihood surface, integrating over nuisance parameters with respect to some prior distribution, rather than maximizing over them to obtain a profile likelihood. From the Bayesian viewpoint, there is no difference between latent variables $\mathbf{X}$ and parameters θ, and the conditional probability distribution of $\mathbf{X}$ given observed data $\mathbf{Y}$ would be referred to as a posterior distribution for $\mathbf{X}$, whereas the probability unconditioned on data would be the prior distribution for $\mathbf{X}$ at a given value of θ. To avoid confusion, we shall refer to the distribution of $\mathbf{X}$ given $\mathbf{Y}$, indexed by parameter θ as the *conditional* distribution, and reserve the word *posterior* for a Bayesian posterior for model parameters θ. We shall, however, refer to the model-based distribution for latent variables $\mathbf{X}$ as a *prior* distribution for $\mathbf{X}$. This should not be confused with a Bayesian prior distribution for model parameters θ.

2.5 Gene counting and the ABO blood types

We have seen in the examples of section 2.3 that, where genotypes are observable, estimating allele frequencies is just a matter of *counting the genes.* In a slightly more general sense, the same is true when genotypes cannot be fully observed. "Counting methods" have been used to estimate allele frequencies since the approach was first introduced by Ceppelini et al. (1955). In fact, these methods are particular instances of the EM-algorithm (section 2.4).

Given a sample of n individuals, the phenotypic counts n_j, $i = 1,\ldots,r$, are multinomial, with probabilities $p_i(\mathbf{q})$, where $\mathbf{q} = (q_1,\ldots,q_k)$ is the vector of underlying allele frequency parameters to be estimated:

$$\ell \quad = \quad \log \mathrm{Pr}(n_1,\ldots,n_r) \quad = \quad \sum_{i=1}^{r} n_i \log p_i(\mathbf{q}) \tag{2.13}$$

The complete-data, consisting of the counts m_j of allelic types of all distinct genes in the sample, are also multinomial:

$$\log \Pr(m_1, \ldots, m_k) = \sum_{j=1}^{k} m_j \log q_j.$$

Determining the conditional expected complete-data log-likelihood (E-step), is simply a matter of determining the expectations e_j of allele counts m_j given the phenotypic counts n_i and current estimates of the allele frequencies q_j. The M-step is even simpler: the new estimate of q_j is the proportion e_j/m^*. Here, m^* is the number of distinct genes in the n individuals: for the case of samples of unrelated individuals, $m^* = 2n$.

current q	current $2q/(1+q)$	recessive phenotype $t_1 = 36$ AA	dominant phenotype $t_2 + t_3 = 64$ AB	BB	new $q = (2t_1 + t_2)/2n$
0.5	0.667	36	42.67	21.33	0.573
0.573	0.729	36	46.64	17.36	0.593
0.593	0.745	36	47.66	16.34	0.598
0.598	0.749	36	47.91	16.09	0.600
0.600	0.750	36	48.00	16.00	0.600

TABLE 2.3. *Sequence of EM iterates for the example of estimation of the frequency of a recessive allele*

We consider two examples of the above, the first being the case of a recessive allele, with allele frequency q. Suppose in a sample size $n = 100$ there are $n_1 = 36$ of the recessive type AA. As seen in section 2.3, the MLE of q is $\sqrt{0.36} = 0.6$. Although the EM algorithm is unnecessary here, it provides a useful example.

The three genotypes are AA, AB and BB, with counts say t_i, $(i = 1, 2, 3)$. Now, $n_1 = t_1$, but the counts of AB and BB are unobservable since B is dominant to A. If these counts, t_2 and t_3, were known, then the number of A alleles is $m_1 = 2t_1 + t_2$, and the MLE of q would be $(2t_1 + t_2)/2n$. Further,

$$\Pr(AB \mid AB \text{ or } BB) = \frac{2q(1-q)}{1-q^2} = \frac{2q}{1+q}$$

so

$$\mathrm{E}_q(t_2 \mid n_2 = t_2 + t_3 = 64) = 64\frac{2q}{1+q}.$$

So now the EM-algorithm implements the sequence of iterates shown in Table 2.3. Starting from an arbitrary initial value $q = 0.5$, the proportion $2q/(1+q)$ is computed, and the 64 individuals of dominant phenotype divided into the expected numbers t_2 and t_3 that are that are AB and BB, respectively (E-step). Then a

current values				phenotype A		phenotype B		...
p	q	$\frac{2r}{p+2r}$	$\frac{2r}{q+2r}$	$\Pr(A) = 0.422$		$\Pr(B) = 0.206$		...
				AA	AO	BB	BO	...
0.3	0.3	0.73	0.73	0.115	0.307	0.056	0.150	...
0.308	0.170	0.77	0.86	0.096	0.326	0.029	0.177	...
0.298	0.156	0.79	0.87	0.091	0.331	0.026	0.180	...
0.295	0.155	0.79	0.88	0.089	0.333	0.025	0.181	...

...	phen AB	phen O	new values	
...	$\Pr(AB) = 0.078$	$\Pr(OO) = 0.294$	p	q
...	AB	OO		
...	0.078	0.294	0.308	0.170
...	0.078	0.294	0.298	0.156
...	0.078	0.294	0.295	0.155
...	0.078	0.294	0.295	0.155

TABLE 2.4. *EM iterates for the estimation of ABO allele frequencies. The iterates of allele frequencies, and the resulting conditional probabilities of genotype AO and BO, given phenotypes A and B, respectively, are shown in the upper left panel. Then are shown the resulting expected genotype frequencies, given the observed phenotype frequencies and current allele frequency estimates (E-step). Finally, in the lower right are shown the new iterates of the allele frequencies (M-step)*

new value of q is estimated as $(2t_1 + t_2)/2n$ (M-step). The process is repeated, and convergence to the MLE $\widehat{q} = 0.6$ is obtained within five steps.

The second example provides the MLEs of the ABO blood group allele frequencies discussed in section 2.3. Here the EM-algorithm is in fact one of the easiest ways to find the MLEs, since there is no explicit solution of the likelihood equation. Now, we must partition both the count of A phenotypes into expected counts of AA and AO genotypes, and the B phenotype into BB and BO genotypes:

$$\begin{aligned} \Pr(AO \mid \text{type } A) &= \frac{2pr}{p^2 + 2pr} = \frac{2r}{p + 2r} \\ \Pr(BO \mid \text{type } B) &= \frac{2qr}{q^2 + 2qr} = \frac{2r}{q + 2r}. \end{aligned}$$

Once the counts are partitioned, according to current estimates of allele frequencies, the new estimate of the A allele frequency p is $\Pr(AA) + (\Pr(AO) + \Pr(AB))/2$, and the new estimate of the B allele frequency q is $\Pr(BB) + (\Pr(BO) + \Pr(AB))/2$. Recall, Bernstein (1925) reported a sample of 502 individuals, with frequencies of the four types, A, B, AB and O, 0.422, 0.206, 0.078, and 0.294, respectively. Table 2.4 shown the sequence of EM-iterates, with convergence being obtained, from starting values $p = q = 0.3$ in four iterations. Again, the details of this example are due to Edwards (1972).

One interesting feature of the sequence of iterates in this example is that the value of p does not change monotonely; there is no reason why it should. What is

guaranteed to change monotonely is the value of the log-likelihood, which, for given allele frequencies may be easily evaluated (section 2.3 and equation (2.13)). For this example, over the iterations, the values of the log-likelihood are −687.1242, −628.9991, −627.5693, −627.5262, −627.5246. Note that, typically of the EM algorithm, the log-likelihood increases rapidly in the first steps, and the parameter values move rapidly to the neighborhood of the MLE, whereas the final convergence is much slower. In examples such as this, where evaluation of the log-likelihood is possible, this provides a better check on convergence than a criterion based on the changes in parameter estimates.

2.6 EM estimation for quantitative trait data

For simple qualitative or quantitative traits, were genotypes observable, estimation of penetrance parameters would also be primarily a matter of "counting". However, even in the simplest cases, explicit EM equations are not readily obtained. There may be no single statistic; the complete-data sufficient statistics may be functions of the genotypes G_i of every individual i. Consider, for example, the simplest possible model for a quantitative trait determined by alleles at a single diallelic locus. (For example, the trait value might be an enzyme level.) The phenotypic value is assumed to have a Gaussian distribution, with mean depending on the genotype at the locus, and variance σ_e^2. The penetrance parameters are the three genotypic means, and the residual variance σ_e^2. The only additional parameter is the allele frequency at determining the trait-locus genotype frequencies. The model for the phenotype Y_i of individual i having genotype G_i may be specified as

$$Y_i = \mu(G_i) + \epsilon_i. \tag{2.14}$$

If sampling unrelated individuals, then the Y_i are independent and identically distributed and this is a simple mixture estimation problem, which can be addressed by EM (see for example, Redner and Walker (1984)). Of greater interest, in the context of genetic analysis, are data observed for members of a pedigree structure. To implement an EM algorithm, we would need to estimate the conditional probabilities that each member of the pedigree is of each of the three genotypes, given current parameter values and the data $\mathbf{Y}$. For related individuals, estimation of the conditional probabilities of genotypes, $\mathbf{G}$, given the observed phenotypic data $\mathbf{Y}$ on the pedigree, is a complex computation equivalent to computation of the total likelihood $\Pr(\mathbf{Y})$. We return to this problem in section 7.4.

Estimation for a genetically more complex model turns out to be simpler, statistically. We consider briefly the classical *polygenic model*, where discrete genotypes G_i are replaced by Gaussian random effects Z_i, known as *polygenic values*. Rather than a single-locus trait, we are now considering a phenotype such as height, probably influenced by a very large number of genes throughout the genome. The genotype configuration $\mathbf{G}$ becomes a vector of polygenic values $\mathbf{z}$, and sums become integrals. The founder probabilities $\Pr(G_i)$ of equation (1.4) are replaced by $N(0, \sigma_a^2)$ population densities for Z_i, where the parameter σ_a^2 is known as the *additive genetic variance*. The transmission probabilities $\Pr(G_i|G_{M_i}, G_{F_i})$

(equation (1.4)) become a transmission density for Z_i given $Z_{M_i} = z_{M_i}$ and $Z_{F_i} = z_{F_i}$:

$$Z_i = \frac{1}{2}(z_{M_i} + z_{F_i}) + \eta_i \tag{2.15}$$

where the η_i are independent, identically distributed segregation residuals, $\eta_i \sim N(0, v_\eta)$, independent of Z_{M_i} and Z_{F_i}. If Z_{M_i} and Z_{F_i} are uncorrelated, then

$$\text{var}(Z_i) = (1/4)(\text{var}(Z_{M_i}) + \text{var}(Z_{F_i})) + v_\eta$$

so to maintain constant population variance σ_a^2 of the Z_i over the generations $v_\eta = \sigma_a^2/2$. The transmission equation (2.15) for the offspring value Z_i, given the parental values, may then be rewritten as $Z_i \sim N((z_{M_i} + z_{F_i})/2, \sigma_a^2/2)$. The joint probability of $\mathbf{Z}$ is Gaussian, with mean $\mathbf{0}$ and variance-covariance matrix $\sigma_a^2\mathbf{A}$, where $\mathbf{A}$ is a matrix determined by the pedigree structure, and known as the numerator-relationship-matrix (Henderson, 1976). In fact, $\mathbf{A}$ is the matrix $2\mathbf{\Psi}$, where the (i,k) component $\Psi_{i,k}$ is the coefficient of kinship $\psi(i,k)$ between individuals i and k (section 3.2).

The simplest penetrance model for a quantitative phenotypic value Y_i of individual i is that it is a direct reflection of the polygenic value Z_i. Ignoring all other possible fixed/random effects, the penetrances $\Pr(Y_i \mid G_i)$ become the density $Y_i \sim N(z_i, \sigma_e^2)$, given $Z_i = z_i$, or

$$Y_i = Z_i + \epsilon_i. \tag{2.16}$$

The variance σ_e^2 of the independent, identically distributed residuals ϵ_i is known as the residual or (individual) environmental variance. In this simplest version of the model, there are just two parameters, σ_e^2 and σ_a^2. In a pedigree (or a collection of pedigrees), suppose there are a total of n_{tot} individuals, and that for n_{obs} of them a value of the quantitative phenotype is observed. The complete-data log-likelihood is

$$\begin{aligned}
\log\Pr(\mathbf{Y}=\mathbf{y},\mathbf{Z}=\mathbf{z}) &= \log\Pr(\mathbf{Y}=\mathbf{y} \mid \mathbf{Z}=\mathbf{z}) + \log\Pr(\mathbf{Z}=\mathbf{z}) \\
&= -\frac{1}{2}\left(n_{obs}\log(2\pi\sigma_e^2) + (\mathbf{y}-\mathbf{z})'(\mathbf{y}-\mathbf{z})/\sigma_e^2\right. \\
&\quad \left. + n_{tot}\log(2\pi\sigma_a^2) + \log(|\mathbf{A}|) + \mathbf{z}'\mathbf{A}^{-1}\mathbf{z}/\sigma_a^2\right).
\end{aligned}$$

This is again of exponential family form, with two complete-data sufficient statistics $(\mathbf{y}-\mathbf{z})'(\mathbf{y}-\mathbf{z})$ and $\mathbf{z}'\mathbf{A}^{-1}\mathbf{z}$, which leads to EM equations

$$\begin{aligned}
\sigma_e^{2*} &= \mathrm{E}_{\sigma_e^2,\sigma_a^2}((\mathbf{Y}-\mathbf{Z})'(\mathbf{Y}-\mathbf{Z}) \mid \mathbf{Y}=\mathbf{y})/n_{obs} \\
\sigma_a^{2*} &= \mathrm{E}_{\sigma_e^2,\sigma_a^2}(\mathbf{Z}'\mathbf{A}^{-1}\mathbf{Z} \mid \mathbf{Y}=\mathbf{y})/n_{tot}.
\end{aligned} \tag{2.17}$$

If $\mathrm{E}_{\sigma_e^2,\sigma_a^2}(\mathbf{Z} \mid \mathbf{Y}=\mathbf{y}) = \mathbf{a}$ and $\mathrm{Var}_{\sigma_e^2,\sigma_a^2}(\mathbf{Z} \mid \mathbf{Y}=\mathbf{y}) = \mathbf{V}$, the equations reduce to

$$\begin{aligned}
\sigma_e^{2*} &= (n_{obs})^{-1}((\mathbf{y}-\mathbf{a})'(\mathbf{y}-\mathbf{a}) + tr(\mathbf{V})) \\
\sigma_a^{2*} &= (n_{tot})^{-1}(\mathbf{a}'\mathbf{A}^{-1}\mathbf{a} + tr(\mathbf{V}\mathbf{A}^{-1})).
\end{aligned}$$

We do not pursue this further here. There is a large literature on the use of EM in polygenic models, particularly in plant and animal breeding. For additional details in the context of simple models on complex pedigrees, see Thompson and Shaw (1990; 1992). For more general work in this area, see the references therein. The point of this example is to show that, even in an exponential family of full rank, the natural sufficient statistics may not be linear functions of latent genotypic counts or values. Estimation of $\mathbf{a} = \mathrm{E}_{\sigma_e^2,\sigma_a^2}(\mathbf{Z} \mid \mathbf{Y} = \mathbf{y})$ is straightforward but insufficient. Since the sufficient statistics are quadratic functions of $\mathbf{Z}$, the conditional variances $\mathbf{V}$ are also needed to implement the EM equations.