1
Introduction: Why Mixed Models?

Big ideas have many names and applications. Sometimes the mixed model is called the model for repeated measurements, sometimes a hierarchical model. Sometimes the mixed model is used to analyze clustered or panel data, sometimes longitudinal data.

Mixed model methodology brings statistics to the next level. In classical statistics a typical assumption is that observations are drawn from the same general population and are independent and identically distributed. Mixed model data have a more complex, multilevel, hierarchical structure. Observations between levels or clusters are independent, but observations within each cluster are dependent because they belong to the same subpopulation. Consequently, we speak of two sources of variation: between clusters and within clusters.

Mixed model is also well suited for the analysis of longitudinal data, where each time series constitutes an individual curve, a cluster. Mixed model is well suited for biological and medical data, which display notorious heterogeneity of responses to stimuli and treatment. An advantage of the mixed model is the ability to genuinely combine the data by introducing multilevel random effects. Mixed model is a nonlinear statistical model, due mainly to the presence of variance parameters, and thus it requires special theoretical treatment. The goal of this book is to provide systematic coverage and development of all spectra of mixed models: linear, generalized linear, and nonlinear.

The aim of this chapter is to show the variety of applications for which the mixed model methodology can be useful, or even a breakthrough. For example, application of mixed modeling methodology to shape and image analysis seems especially exciting and challenging.

Mixed models can be used for the following purposes:

- To model complex clustered or longitudinal data.

- To model data with multiple sources of variation.
- To model biological variety and heterogeneity.
- As a compromise between the frequentist and Bayesian approaches.
- As a statistical model for the penalized log-likelihood.
- To provide a theoretical basis for the Healthy Akaike Information Criterion (HAIC).
- To cope with parameter multidimensionality.
- As a statistical model to solve ill-posed problems, including image reconstruction problems.
- To model shapes and images.

An important feature of this book is that it provides numerical algorithms as a realization of statistical methods that it develops. We strongly believe that an approach is not valuable without an appropriate efficient algorithm. Each chapter ends with a summary points section that may help the reader to quickly grasp the chapter's major points.

1.1 Mixed effects for clustered data

The mixed effects approach copes with clustered data that can be viewed as a sample of samples. To illustrate, let us consider the relationship between price (x) and sales (y). Let $\{(x_k, y_k), k = 1, ..., K\}$ be the sample of observations collected on price and sales for several commodities. Plotting y versus x reveals that the relationship is close to linear with a negative slope; see the left-hand panel in Figure 1.1. In classical statistics it is assumed that pairs (x_k, y_k) are independent and identically distributed (iid) with the regression line $E(y|x) = \alpha + \beta x$. However, one may argue that we deal with clustered data, where each cluster is a commodity. In the right-hand panel, we connect observation points for each commodity and obtain a reverse picture—increase in price leads to increase in sales. A paradox?

Classical statistics assumes the model

$$y_k = \alpha + \beta x_k + \varepsilon_k, \quad k = 1, ..., K, \tag{1.1}$$

where the $\{\varepsilon_k\}$ are independent and identically distributed random variables with zero mean and constant variance σ^2. In other words, it is assumed that the data are collected from similar, homogeneous commodities. As follows from the right panel, the commodities are not homogeneous and vary substantially in terms of price and sales. An adequate model for the sales problem would be to assume that each commodity has its own commodity-specific sales (in statistical language, intercept); namely,

$$y_{ij} = \alpha_i + \beta x_{ij} + \varepsilon_{ij}, \quad i = 1, ..., N, \; j = 1, ..., n_i. \tag{1.2}$$

Note that we use a double index now because we are dealing with clustered/panel /tabular data: i corresponds to the ith commodity, j corresponds to the jth observation of the ith commodity, n_i is the number of observations for the ith commodity, and α_i is the commodity-specific intercept. The total number of observations is $K = \sum_{i=1}^{N} n_i$. Regarding the error terms $\{\varepsilon_{ij}\}$, we assume that, as previously, they are iid with the variance σ^2.

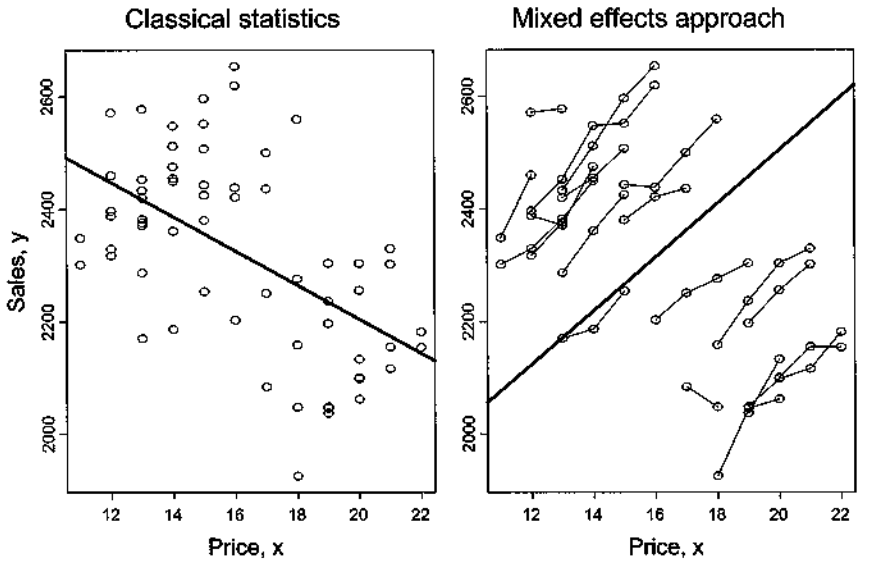

FIGURE 1.1. Classical and mixed effects approaches lead to reverse conclusions. Left: In the classical approach, it is assumed that observations are independent and identically distributed, resulting in a negative relationship. The straight line shows simple regression estimated by ordinary least squares. Right: In the mixed effects approach, it is assumed that each commodity represents a cluster and therefore that an increase in price for a specific commodity leads to an increase in sales. The straight line shows the linear mixed effects model with population-averaged slope and commodity-specific intercept.

Obviously, model (1.2) is more complex than the classical regression model (1.1), and in a special case, $\alpha_i = \alpha$, we come to (1.1). The central assumption of the mixed effects model is that intercepts $\{\alpha_i, i = 1, ..., N\}$ are random and belong to a general population that can be expressed in the second equation as

$$\alpha_i = \alpha + b_i, \tag{1.3}$$

where α is the population-averaged sale (intercept) and b_i is the random effect, or deviation of the commodity-specific sale from the population-averaged sale. Thus, on the one hand, we allow commodity-specific sales, but on the other hand, we assume that commodities represent the country market economy, and therefore one can speak of how an increase in price affects sales across all commodities. Coupled models (1.2) and (1.3) define a linear mixed effects model, parameters α and β are fixed effects (population-averaged parameters), and b_i is the random effect with zero

mean and variance σ_b^2 independent of $\{\varepsilon_{ij}\}$. This is a *hierarchical* model or a model with random coefficients. The model defined by equations (1.2) and (1.3) can be combined into one as

$$y_{ij} = \alpha + \beta x_{ij} + \eta_{ij}, \quad i = 1, ..., N, \; j = 1, ..., n_i, \tag{1.4}$$

where $\eta_{ij} = \varepsilon_{ij} + b_i$ is the composite random error. As follows from (1.4), observations on the same commodity (within a cluster) correlate with the correlation coefficient:

$$\rho = \frac{\text{var}(b_i)}{\text{var}(b_i + \varepsilon_{ij})} = \frac{\sigma_b^2}{\sigma^2 + \sigma_b^2}, \tag{1.5}$$

but observations on different commodities (from different clusters) do not correlate. In a mixed effects model, there are two sources of variation: the within (or intra) -cluster variation, σ^2, and the between (or inter) -cluster variation, σ_b^2. Recall that classical regression assumes one variation. As follows from (1.5), the larger the variation between commodities, the higher the correlation within each cluster. If $\sigma_b^2 = 0$, the correlation is zero and $\alpha_i = \alpha$, ordinary linear regression. For the data in Figure 1.1, $\rho = 0.99$, so the major source of variation is the variation between commodities. That is why the slope has different signs in the two approaches.

Observations $\{y_{i1}, y_{i2}, ..., y_{i,n_i}\}$ can also be interpreted as *repeated measurements.* Therefore, model (1.4) is sometimes called the model for repeated measurements. An important example of clustered data is that of longitudinal data when subjects are observed over time. In fact, the pioneering work by Laird and Ware (1982) on the linear mixed effects model was concerned with this kind of data. Model (1.4) belongs to the family of linear mixed effects (LME) models and is studied extensively in Chapters 2 through 4. Specifically, model (1.4) is called the LME model with random intercepts, and it has many nice properties (see Section 2.4). There is more on ignoring random effects in the LME model in Section 3.9.

Summing up, ignoring clustered structure may lead to false analysis. The linear mixed effects model is an adequate model for clustered (repeated) data that involve two sources of variation, within and between clusters.

1.2 ANOVA, variance components, and the mixed model

The mixed model may be viewed as a combination of analysis of variance (ANOVA), variance component (VARCOMP), and regression models. For example, the simplest, one-way ANOVA model deals with tabular data:

$$y_{ij} = \beta_i + \varepsilon_{ij}, \quad i = 1, ..., N, \; j = 1, ..., n_i, \tag{1.6}$$

where N is the number of units (subjects or clusters), n_i is the number of observations per unit, and $\{\varepsilon_{ij}\}$ are independent and identically distributed (iid) errors with zero mean and variance σ^2. An important, sometimes not well emphasized assumption of the ANOVA model is that $\{\beta_1, ..., \beta_N\}$ are fixed parameters. Consequently, for each unit, observations $\{y_{i1}, y_{i2}, ..., y_{in_i}\}$ can be treated as replicates because they are iid with the mean β_i. A traditional hypothesis in the framework of the ANOVA model is that the units are the same, or $H_0 : \beta_1 = ... = \beta_N$.

The ANOVA model is a special case of the linear regression model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.7}$$

where $\mathbf{y}$ is a $K \times 1$ vector of observations, $\mathbf{X}$ is a $K \times m$ design matrix, and $\boldsymbol{\beta}$ is an $m \times 1$ vector of parameters. For example, the one-way ANOVA model (1.6) can be expressed in the regression form (1.7) if the $\{y_{ij}\}$ are arranged in the vector $\mathbf{y}$ so that $K = \sum_{i=1}^{N} n_i$, the elements of the design matrix $\mathbf{X}$ are 0 or 1, and $m = N$. A classic reference, where various ANOVA models are represented as a linear model, is Searle (1971a). All ANOVA models have two important features: (a) parameters $\{\beta_i, i = 1, ..., m\}$ are estimated by ordinary least squares, and (b) the F-test is the workhorse for linear hypothesis testing. Models (1.6) and (1.7) can also be called fixed effects models.

We come to a *different* statistical model when the $\{\beta_i\}$ are assumed random, say iid normally distributed (independent of ε_{ij}) with the common mean β and variance σ_β^2. Representing $\beta_i = \beta + b_i$, we arrive at the variance components (VARCOMP) model:

$$y_{ij} = \beta + b_i + \varepsilon_{ij}, \tag{1.8}$$

where b_i is called a *random effect*. The ANOVA is a fixed effects model and VARCOMP is a random effects model. Although models (1.6) and (1.8) seem similar, they have different statistical properties. In ANOVA, observations do not correlate; in VARCOMP, observations correlate within each unit and the correlation coefficient is equal to $\sigma_\beta^2/(\sigma^2 + \sigma_\beta^2)$. According to the Gauss–Markov theorem, for model (1.6) the ordinary least squares coincides with the MLE and is efficient, but this does not hold for model (1.8). Moreover, if n_i are different, there is no closed-form solution for the MLE. The null hypothesis $H_0 : \beta_1 = ... = \beta_N$ for the ANOVA model transforms into $H_0 : \sigma_\beta^2 = 0$ and the F-test cannot be applied directly, as it requires substantial modification (see Section 3.5). When the number of units is relatively small (say, $N < \min n_i$), the ANOVA model is preferable. When the number of units is relatively large (say, $N > \max n_i$), the VARCOMP model may be better. The VARCOMP model has a long history (Rao, 1973; Harville, 1977; Searle et al., 1992).

The mixed model may be viewed as a combination of the ANOVA and VARCOMP models. For example, consider the problem of measuring the blood pressure for $i = 1, ..., N$ people at time points $t_{i1}, t_{i2}, ..., t_{i,n_i}$. If y_{ij} denotes the blood pressure of the ith person at time t_{ij}, the VARCOMP model (1.8) may be adequate because it reflects the fact that the blood pressure changes from person to person, but for the same time, one can speak of the population-averaged blood pressure, β. Now we realize that besides blood pressure for each person, we have information about gender, age, and so on. Also, to reflect the fact that measurements are made over a fairly long period of time, we incorporate t_{ij} into the vector of complete covariates $\mathbf{x}_{ij}$. Then the expanded VARCOMP model transforms into the mixed effects model,

$$y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + b_i + \varepsilon_{ij}. \tag{1.9}$$

The similarity with the regression model (1.7) becomes evident.

In general, the linear mixed effects (LME) model is written as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, ..., N, \tag{1.10}$$

where $\mathbf{b}_i$ is a vector of random effects such that $\text{cov}(\mathbf{b}_i) = \sigma^2\mathbf{D}$ and $\mathbf{Z}_i$ is the design matrix. For example, for model (1.9), the random effect is scalar and $\mathbf{Z}_i = \mathbf{1}$. The variance parameters, σ^2 and $\mathbf{D}$, are unknown and are subject to estimation along with the population-averaged parameter $\boldsymbol{\beta}$.

By combining vectors $\{\mathbf{y}_i\}$ and matrices $\{\mathbf{X}_i\}$ into $\sum n_i \times 1$ vector $\mathbf{y}$ and $\sum n_i \times m$ matrix $\mathbf{X}$, and letting $\mathbf{Z} = \text{diag}(\mathbf{Z}_1, ..., \mathbf{Z}_N)$, model (1.10) can be written as one equation, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}$. Although some authors prefer to work with this one-equation LME model, such representation is excessive because observations across i are independent.

Although model (1.10) looks like a linear model, the fact that the variance parameters are unknown makes it a nonlinear statistical model with elaborated estimation methodology. Usually, we assume that the random effects and the error term have a normal distribution, so that model (1.10) can be written more compactly as

$$\mathbf{y}_i \sim \mathcal{N}(\mathbf{X}_i\boldsymbol{\beta}, \sigma^2(\mathbf{I} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i')), \quad i = \mathbf{1}, ..., N, \tag{1.11}$$

meaning that $\mathbf{y}_i$ has a multivariate normal distribution with mean $\mathbf{X}_i\boldsymbol{\beta}$ and covariance matrix $\sigma^2(\mathbf{I} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i')$. If $\mathbf{D}$ were known, as follows from the Gauss–Markov theorem, the generalized least squares estimator,

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} \mathbf{X}'(\mathbf{I} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i')^{-1}\mathbf{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}'(\mathbf{I} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i')^{-1}\mathbf{y}_i\right),$$

would be efficient. But the variance-covariance matrix of the random effects is unknown, and its estimation becomes a central theme in the framework of the mixed effects model. Two families of estimators for the variance parameters are considered: maximum likelihood (Chapter 2) and quadratic noniterative distribution-free estimators, including MINQUE, variance least squares, and method of moments (Chapter 3).

The LME model and its generalizations are studied in the first three chapters of the book. In the first chapter we discuss computational aspects of maximum likelihood, the second chapter is about statistical properties, and in the third chapter we consider several generalizations and important special cases of the LME model. In Chapter 5, meta-analysis, a very special case of the mixed model, is studied; this model is not covered by (1.10) and therefore requires special treatment.

1.3 Other special cases of the mixed effects model

Another important special case of linear mixed effects model (1.10) is the regression model with random coefficients,

$$\mathbf{y}_i = \mathbf{X}_i\mathbf{a}_i + \boldsymbol{\varepsilon}_i, \quad \mathbf{a}_i = \boldsymbol{\beta} + \mathbf{b}_i, \quad i = \mathbf{1}, ..., N. \tag{1.12}$$

For example, Swamy (1971) studied this model in connection with the analysis of cross-sectional (panel) data where $\mathbf{y}_i$ is a time series of length n and i is an index economic sector. One comes to (1.12) letting $\mathbf{Z}_i = \mathbf{X}_i$ in the LME model (1.10). An interesting special case of model (1.12) is when the data are balanced, $\mathbf{X}_i = \mathbf{Z}$. For

balanced data, the ordinary and generalized least squares lead to the same estimate. This model is studied in Section 2.3. In Chapter 4 we study the growth curve model, where $\mathbf{a}_i = \mathbf{A}_i\boldsymbol{\beta} + \mathbf{b}_i$ and $\mathbf{A}_i$ is the design matrix. Sometimes only a subvector $\mathbf{a}_i$ can be specified, so that other coefficients may be anything. For example, in model (1.9), only the intercept is random; this models is studied in Section 2.4, while a more general family of growth curve models is studied in Section 4.2.

Another special case of the LME model is when $n_i = \mathbf{1}$, which leads to a linear regression with *heteroscedastic* errors, $y_i = \boldsymbol{\beta}'\mathbf{x}_i + \eta_i$, where η_i has zero mean and variance $\text{var}(\eta_i) = \sigma^2(1 + dz_i^2)$ and d is the parameter to estimate. Many examples and treatments of the regression model with heteroscedastic errors may be found in the book by Carroll and Ruppert (1988). A nonlinear regression model with heteroscedastic errors and a nonlinear variance function defined as $\text{var}(\eta_i) = \sigma^2 w_i(\boldsymbol{\beta}, \boldsymbol{\theta})$ can be studied in the framework of the nonlinear marginal mixed model of Chapter 6.

1.4 Compromise between Bayesian and frequentist approaches

The goal of this section is to convince the reader that the mixed model may serve as a compromise between the frequentist (classical) and Bayesian approaches. Both the Bayesian and mixed model approaches are based on a hierarchical statistical model, but in the former the values for all parameters must be specified, whereas in the latter, parameters are estimated from the data.

Specifically, let $\mathbf{y}$ be the data observed. In the *Bayesian* approach, the model is specified in *hierarchical* fashion as

$$\mathbf{y}|\boldsymbol{\theta} \sim L(\mathbf{y}|\boldsymbol{\theta}), \tag{1.13}$$

$$\boldsymbol{\theta} \sim G(\boldsymbol{\theta}). \tag{1.14}$$

Equation (1.13) defines the conditional distribution of $\mathbf{y}$ given $\boldsymbol{\theta}$ through density L. The second equation, (1.14), defines *a priori* the distribution of $\boldsymbol{\theta}$ through density G. Since G is usually a member of a family of distributions, the parameter that specifies G is called the *hyperparameter*. Thus, unlike the frequentist approach, the Bayesian approach assumes that parameter $\boldsymbol{\theta}$ is random and densities L and G must be specified completely. The main computational concern in the Bayesian framework is calculation of the normalization constant

$$A = \int L(\mathbf{y}|\boldsymbol{\theta})G(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{1.15}$$

in the *posterior density*

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{1}{A}L(\mathbf{y}|\boldsymbol{\theta})G(\boldsymbol{\theta}). \tag{1.16}$$

Obviously, computation of A is required to ensure that the area under the surface defined by (1.16) is 1. Much effort has been spent on developing integration techniques for (1.15). In particular, one of the most popular approaches, based on the Markov Chain Monte Carlo (MCMC) technique, is realized in BUGS software (http://www.mrc-bsu.cam.ac.uk/bugs).

The major criticism of Bayesian theory is the requirement for complete specification of the prior distribution G. It is worthwhile to note that G directly affects the posterior density (1.16) because it acts as a factor. Consequently, sensitivity to the choice of the prior distribution in the Bayesian approach is substantial.

In the mixed model approach, the model is also specified as a hierarchical model, (1.13) and (1.14), but it is allowed to have nonrandom parameters, τ, namely,

$$\mathbf{y}|\boldsymbol{\theta} \sim L(\mathbf{y}|\boldsymbol{\theta},\tau), \tag{1.17}$$

$$\boldsymbol{\theta} \sim G(\boldsymbol{\theta},\tau). \tag{1.18}$$

In the Bayesian framework, τ is known and is the hyperparameter. When τ is unknown we come to the frequentist model, where τ is estimated, for example, by maximum likelihood. As in Bayesian theory, integration becomes a technical problem because ML maximizes the marginal likelihood,

$$L(\tau) = \int L(\mathbf{y}|\boldsymbol{\theta},\tau)G(\boldsymbol{\theta},\tau)d\boldsymbol{\theta}. \tag{1.19}$$

In the framework of the mixed model, we call $\boldsymbol{\theta}$ random (or subject-specific) and τ fixed effects (or population-averaged) parameters. Random effects are unobservable and are integrated out in (1.19), but τ is estimated. Thus, the normalizing constant, (1.15), plays the role of the likelihood in the mixed model. After $\widehat{\tau}$ is computed, we apply standard Bayesian formulas, such as posterior density, posterior mean, and so on. In the language of the mixed model, the posterior mean is called the estimate of the random effect. We refer the reader to Sections 3.7 and 8.15, where these quantities are estimated.

In summary, a mixed model combines major features of the frequentist and Bayesian approaches. Symbolically,

$$\text{mixed model} = \text{Bayesian} + \text{frequentist}.$$

On the one hand, as in the Bayesian approach, mixed model assumes a hierarchical (conditional) model where the parameter is treated as random. On the other hand, the hyperparameter, τ, is not specified arbitrarily as in the Bayesian approach, but is *estimated* from the data. As such, a mixed model is more flexible than the Bayesian approach.

We illustrate the difference between the Bayesian and mixed model approaches by a linear model under a normal distribution,

$$\mathbf{y}|\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I}_n), \tag{1.20}$$

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0},\sigma_\beta^2\mathbf{I}_m) \tag{1.21}$$

(Lindley and Smith, 1972; Smith, 1973). These equations are special cases of the general Bayesian model (1.13) and (1.14). In words, if the vector of regression coefficients $\boldsymbol{\beta}$ were known, $\mathbf{y}$ would have a multivariate normal distribution with mean $\mathbf{X}\boldsymbol{\beta}$ and variance σ^2. As follows from (1.21), the prior distribution for $\boldsymbol{\beta}$ is also normal with zero mean and variance σ_β^2. To complete the Bayesian specification, one needs to provide distributions for the variance parameters, σ^2 and σ_β^2. Typically, a gamma distribution with the density $\Gamma^{-1}(\alpha)\lambda^\alpha t^{\alpha-1}e^{-\lambda t}$ is used for this purpose,

where α and λ are the known positive (hyper-) parameters. The idea behind the choice of α and λ is to obtain a noninformative prior. When the hyperparameter belongs to a bounded set, the noninformative prior is constant. For example, if a probability p is the hyperparameter, it is reasonable to assume that the prior density of p is 1 on $(0,1)$. Things are complicated when the hyperparameter is not bounded, such as variance. For example, in BUGS the default values are $\lambda = 1/1000$ and $\alpha = 1/1000$. Since for the gamma distribution, $E = \alpha/\lambda$ and var$= \alpha/\lambda^2$, this choice implies that the *a priori* mean equals 1 and the variance equals 1000. In terms of the variance parameters for our linear model, such a choice would mean that $\sigma^2 = \sigma_\beta^2 = 1000$. Apparently, this choice of the hyperparameters is arbitrary.

Now we turn our attention to the mixed model approach. It uses the same hierarchical models (1.20) and (1.21) but the variance parameters are assumed unknown. We can estimate σ^2 and σ_β^2 either by maximum likelihood (ML) or by noniteratively using unbiased quadratic estimators (Chapter 3). For example, using the ML approach, the pair of models (1.20) and (1.21) imply the model

$$\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \sigma^2(\mathbf{I} + d\mathbf{X}\mathbf{X}')),$$

where $d = \sigma_\beta^2/\sigma^2$ is the scaled variance parameter. In the Bayesian approach, parameters σ^2 and d have to be specified through known distributions. In the mixed model approach, we treat them as unknown parameters to be estimated from maximum likelihood. The log-likelihood, up to a constant $-n \ln \sqrt{2\pi}$, takes the form

$$l(\sigma^2, d) = -0.5n \ln \sigma^2 - 0.5 \ln \left|\mathbf{I} + d\mathbf{X}\mathbf{X}'\right| - 0.5\sigma^{-2}\mathbf{y}'(\mathbf{I} + d\mathbf{X}\mathbf{X}')^{-1}\mathbf{y}.$$

Differentiating with respect to σ^2, we obtain $\sigma^2 = n^{-1}\mathbf{y}'(\mathbf{I} + d\mathbf{X}\mathbf{X}')^{-1}\mathbf{y}$. Plugging it back into l, the variance-profile log-likelihood function simplifies to a function of one argument,

$$l(d) = -0.5n \ln \mathbf{y}'(\mathbf{I}_n + d\mathbf{X}\mathbf{X}')^{-1}\mathbf{y} - 0.5 \ln \left|\mathbf{I} + d\mathbf{X}\mathbf{X}'\right|.$$

A number of algorithms may be used to maximize this function and to obtain the MLE, $\widehat{d}$. So the hyperparameters in the Bayesian approach are estimated in the mixed model. After parameter values are determined, we compute the posterior distribution, which is also normal with mean $\widehat{\boldsymbol{\beta}} = \widehat{d}\mathbf{X}'(\mathbf{I}_n + \widehat{d}\mathbf{X}\mathbf{X}')^{-1}\mathbf{y}$. Using the dimension-reduction formula of Section 2.2.3 or Appendix 13.2, we can express $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \widehat{d}^{-1}\mathbf{I}_m)^{-1}\mathbf{X}'\mathbf{y}$, as in Lindley and Smith (1972) but with the estimate instead of an arbitrary d.

1.5 Penalized likelihood and mixed effects

Penalized likelihood is encountered in many applications as a way to make a problem solvable by replacing an ill-posed problem with a well-posed problem. This methodology has to be proven to make a great deal of improvement in a variety of applications, from applied mathematics and computer science to engineering. However, a substantial drawback of the penalized likelihood approach is the need to know the *penalty coefficient* (sometimes called a *regularization parameter*). Strictly speaking, an ill-posed problem is merely reduced to another problem of choosing the

penalty coefficient. Our aim in this section is to show how the penalized likelihood may be derived from a hierarchical statistical model so that the penalty coefficient term may be estimated along with the parameter of interest. Here we suggest the solution in general terms, and in the following sections we illustrate it by various examples.

Let $\mathbf{y}$ be an n-dimensional vector of observations with the density function L dependent on a k-dimensional parameter $\mathbf{b}$, where k may be large. Denote $l(\mathbf{b};\mathbf{y})$ as the log-likelihood and $L(\mathbf{b};\mathbf{y})$ as the likelihood. If n is close to k, the maximum likelihood solution

$$\max_{\mathbf{b}} l(\mathbf{b};\mathbf{y}) \tag{1.22}$$

turns into an ill-posed problem. To improve (1.22) a penalty term is introduced, so instead one maximizes the penalized log-likelihood,

$$\max_{\mathbf{b}} [l(\mathbf{b};\mathbf{y}) + \rho g(\mathbf{b})], \tag{1.23}$$

where ρ is a nonnegative penalty coefficient and $g(\mathbf{b})$ is a penalty function. Typically, a quadratic term is used, $g(\mathbf{b}) = -\|\mathbf{b}\|^2$, so the penalized log-likelihood reduces to minimization of

$$-l(\mathbf{b};\mathbf{y}) + \rho \|\mathbf{b}\|^2. \tag{1.24}$$

Sometimes, the penalized log-likelihood is used not in a statistical but in an applied mathematics framework as a regularization technique (Tikhonov and Arsenin, 1977). For example, let $\mathbf{y}$ be an $n \times 1$ normally distributed vector and $\mathbf{X}$ an $n \times k$ matrix such that $\mathbf{y} = \mathbf{X}\mathbf{b} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is the error term with independent identically distributed (iid) components $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. For this linear model we have $l(\mathbf{b};\mathbf{y}) = -(2\sigma^2)^{-1} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2$, up to a constant term. If matrix $\mathbf{X}'\mathbf{X}$ is singular (e.g., when $k > n$), (1.22) is an ill-posed problem because $\mathbf{b}$ is not unique. On the other hand, if ρ is a fixed positive number, the penalized negative log-likelihood yields a unique solution, $\mathbf{b} = (\mathbf{X}'\mathbf{X} + \nu\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$, where $\nu = 2\sigma^2\rho$.

What is the value of ρ? The answer is important because if $\rho = 0$, we come to the previous ill-posed problem. If $\rho \to \infty$, we have $\mathbf{b} = \mathbf{0}$. Thus, by varying ρ, one obtains a variety of solutions, from unstable MLE to trivial $\mathbf{0}$.

To estimate ρ we assume that $\mathbf{b}$ is random, so that $L(\mathbf{b};\mathbf{y})$ is the conditional likelihood. Let G be a density, so the density of $\mathbf{b}$ is $\omega^{-k} G(\omega^{-1}\mathbf{b})$, where ω is a positive scale parameter. Symbolically, this scheme may be expressed as a hierarchical statistical model,

$$\mathbf{y}|\mathbf{b} \sim L, \quad \mathbf{b} \sim G. \tag{1.25}$$

Since only observations on $\mathbf{y}$ are available, we need to deal with the marginal distribution

$$\int_{R^k} L(\mathbf{b};\mathbf{y}) \omega^{-k} G(\omega^{-1}\mathbf{b}) d\mathbf{b},$$

where random $\mathbf{b}$ is integrated out. Letting $g = \ln G$, the marginal log-likelihood takes the form

$$l(\omega) = -k \ln \omega + \ln \int_{R^k} e^{l(\mathbf{b};\mathbf{y}) + g(\omega^{-1}\mathbf{b})} d\mathbf{b}. \tag{1.26}$$

The MLE, $\widehat{\omega}$ turns l into a maximum. Now the Laplace approximation comes into play to show the link between maximum and penalized likelihood (see Section 7.7.1

for more details),

$$\int_{R^k} e^{h(\mathbf{b})} d\mathbf{b} \simeq (2\pi)^{k/2} e^{h_{\max}} \left| -\left. \frac{\partial^2 h}{\partial \mathbf{b}^2} \right|_{\mathbf{b}=\mathbf{b}_{\max}} \right|^{-1/2}, \tag{1.27}$$

where $h_{\max} = h(\mathbf{b}_{\max})$ and $|\mathbf{H}|$ is the determinant of the negative Hessian at the maximum, $\mathbf{H} = -\partial^2 h / \partial \mathbf{b}^2$. Applying this approximation to (1.26), one obtains

$$l(\omega) \simeq -k \ln \omega + l(\mathbf{b}; \mathbf{y}) + g(\omega^{-1}\mathbf{b}) - 0.5 \ln |\mathbf{H}| .$$

Finally, assuming that $\ln |\mathbf{H}|$ changes little with $\mathbf{b}$, the marginal log-likelihood (1.26) can be approximated as

$$l(\omega) \simeq -k \ln \omega + l(\mathbf{b}; \mathbf{y}) + g(\omega^{-1}\mathbf{b}). \tag{1.28}$$

In the particular case when ω is known, maximization of the marginal log-likelihood is almost equivalent to maximization of

$$l(\mathbf{b}; \mathbf{y}) + g(\omega^{-1}\mathbf{b}). \tag{1.29}$$

In an important special case, when the marginal distribution of $\mathbf{b}$ is normal ($G = \mathcal{N}$) with zero mean, we have $g(\omega^{-1}\mathbf{b}) = -0.5\omega^{-2} \|\mathbf{b}\|^2$, so the maximum likelihood estimation of the hierarchical statistical model (1.25) is almost equivalent to the minimization of the penalized log-likelihood (1.24) with $\rho = 1/(2\omega^2)$. Finally, to estimate $\mathbf{b}$ and ω simultaneously, we maximize the right-hand side of (1.28), which is a well-posed problem. In the literature on mixed models, (1.28) is called *quasi-likelihood* (Breslow and Clayton, 1993). This likelihood approximation plays an important role in estimation in the generalized linear mixed models of Chapters 7 and 8, respectively. Typically, besides random effects $\mathbf{b}$, we have fixed effects (or population-averaged) parameters $\boldsymbol{\theta}$, but their presence does not alter the reasoning, described above.

Generally, any penalized log-likelihood may be derived through a mixed model. For a linear model, the penalized log-likelihood is exact; for a nonlinear model, the penalized log-likelihood is an approximation of the original log-likelihood. The Laplace approximation is the key to proving this link.

In the following sections we show some applications of this general result.

1.6 Healthy Akaike information criterion

The Akaike (1974) information criterion (AIC) became very popular as a criterion for model selection. The rationale behind this criterion is the divergence between the true distribution and a candidate measured in terms of the Kullback–Leibler information criterion, Kullback (1968). It was shown that based on this criterion, the model should be chosen such that

$$\text{AIC} = -2l_{\max} + 2k \tag{1.30}$$

reaches a minimum, where $l_{\max}$ is the log-likelihood maximum and k is the number of unknown parameters. The smaller the AIC, the better the model. The AIC is

especially useful for nonnested models; if the models are nested, standard statistical hypothesis techniques are applied. It is worthwhile to notice that (1.30) has the form of a penalized negative log-likelihood. For example, consider a linear regression model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.31}$$

where $\boldsymbol{\beta}$ is the k-dimensional parameter vector, components ε_i are independent normally distributed random variables with zero mean and variance σ^2, and $i = 1, 2, ..., n$. Assuming that all candidate models use the same number of observations (n =const), it is elementary to check that up to a constant,

$$\text{AIC} = n \ln \widehat{\sigma}^2 + 2k, \tag{1.32}$$

where $\widehat{\sigma}^2 = n^{-1} \left\| \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{LS} \right\|^2$ is the regression variance and $\widehat{\boldsymbol{\beta}}_{LS}$ is the least squares estimate.

Several researchers noted that there can be appreciable bias in the AIC estimate. For example, Hurvich and Tsai (1991) suggested using the term $k+(k+1)(k+2)/(n-k-2)$ instead of $2k$. Sclove (1987) and Dayton (1998) consider a generalization of the Akaike information criterion expressed as $-2l_{\max} + a(n)k$, where $a(n)$ is a function of the sample size.

Although many researchers demonstrated that the AIC is a useful quantity to characterize the information property of a statistical model, Ishiguro et al. (1997) and Mittelhammer et al. (2000), among others, pointed out a weakness of this criterion. In particular, the AIC works poorly in the case of multicollinearity. To illustrate, let us consider the problem of finding the right linear regression model using a set of independent (explanatory) variables or covariates $\{x_j, j = 1, ..., J\}$, where the number of candidate covariates, J, is quite large (perhaps even larger than the number of observations, n). Assume that an analyst has come to a satisfactory set of $k-1$ explanatory variables $x_1, ..., x_{k-1}$ and wants to try to add new variables u or v, one at a time. Consider the situation when both sets, $\{x_1, ..., x_{k-1}, u\}$ and $\{x_1, ..., x_{k-1}, v\}$, yield the same, or a very close, residual sum of squares and consequently, $\widehat{\sigma}^2$. Then, in terms of the AIC, the two models are indistinguishable because as follows from (1.32), they produce the same AIC value. For example, due to the multicollinearity between $x_1, ..., x_{k-1}$ and u, the first model yields large standard errors and low t-statistics for the least squares estimates and assume that the second model still has satisfactory t-statistics. Clearly, the second model would be better, but the AIC fails to identify it, especially when the design matrix is ill-conditioned. The model selection criterion developed below is free of this drawback.

We turn our attention to the penalized log-likelihood (1.28). Assuming that the prior distribution of parameters is normal, we obtain

$$l \simeq -\frac{k}{2} \ln \omega^2 + l(\mathbf{b}; \mathbf{y}) - \frac{1}{2\omega^2} \|\mathbf{b}\|^2 .$$

The maximum of the log-likelihood function over the variance is attained at $\omega^2 = \|\mathbf{b}\|^2 / k$, so the healthy Akaike information criterion takes the form

$$\begin{aligned} HAIC &= H - 2l_{\max} + 2k \\ &= H + AIC, \end{aligned} \tag{1.33}$$

where

$$H = k\left(\ln(\left\|\widehat{\mathbf{b}}_{ML}\right\|^2/k) - 1\right). \tag{1.34}$$

The AIC works well when two models are compared with different numbers of estimated parameters, k, but it fails to discriminate models with the same k and quality of fit when the models are ill-conditioned (ill-posed problems). To illustrate, let us come back to our linear regression example. Consider a linear regression with the number of explanatory variables equals k and the variance $\widehat{\sigma}^2$. Now add a new variable, which is highly correlated with the other variables. The result of such addition in terms of the AIC may not be well reflected because $\widehat{\sigma}^2$ will not change, due to multicollinearity. To the contrary, due to $|\mathbf{X}'\mathbf{X}| \approx 0$, the OLS estimate after addition becomes unstable, which would lead to a large value, $\left\|\widehat{\boldsymbol{\beta}}_{LS}\right\|^2$. The instability will be picked up immediately by HAIC because H becomes large. Now it is clear why the term *healthy* is used to reflect that HAIC works well for ill-posed estimation problems as well.

The healthy AIC works in both directions: when the number of parameters, k, increases and when k is constant. In the latter situation, between two models with the same log-likelihood value, the healthy AIC chooses the model with the shorter estimate length.

1.7 Penalized smoothing

Several authors have pointed out a close relationship between penalized smoothing and the mixed model (Zeger and Diggle, 1994; Wang, 1998; Zhang et al., 1998; Ruppert et al., 2003). To illustrate the connection, we start with the following simplified problem: Let $y_1, y_2, ..., y_n$ be time series data as observations at time $i = 1, 2, ..., n$ (in fact, y may be any equidistant data). We want to find $\mu_1, \mu_2, ..., \mu_n$ such that

$$y_i = \mu_i + \varepsilon_i, \quad i = 1, ..., n, \tag{1.35}$$

where the $\{\varepsilon_i\}$ are iid random variables with zero mean and constant variance σ^2. Clearly, without any restriction on $\{\mu_i\}$, this problem has a trivial solution, $\mu_i = y_i$. To restrict $\{\mu_i\}$, several cost functions have been suggested. The most popular is the *bending energy* cost function (for further discussion see, e.g., Chalmond, 2003). Then total criterion takes the form

$$\sum_{i=1}^{n}(y_i - \mu_i)^2 + \rho\sum_{i=2}^{n-1}(\mu_{i+1} - 2\mu_i + \mu_{i-1})^2, \tag{1.36}$$

where ρ is a positive parameter, the penalty coefficient. The first term in (1.36) is the usual sum of squares, and the second term is the penalty on the curvature of $\{\mu_i\}$. Indeed, if the second term is zero, then $\mu_{i+1} = 2\mu_i - \mu_{i-1}$, and by induction we express $\{\mu_i, i = 3, ..., n\}$ through μ_1 and μ_2 as $\mu_{i+1} = i\mu_2 - (i-1)\mu_1 = i(\mu_2 - \mu_1) + \mu_1$. But this is a linear function of i, so the second term puts a penalty on the nonlinearity of $\{\mu_i\}$. From calculus, $\mu_{i+1} - 2\mu_i + \mu_{i-1}$ can be viewed as a discrete approximation of the second derivative, so the second term may be viewed as a discretization of the commonly used function $\int[\mu''(x)]^2 dx$ to penalize the nonlinearity.

Several different terminologies are used in the literature for the problem specified by equations (1.35) and (1.36), such as scatter plot smoothing and spline regression.

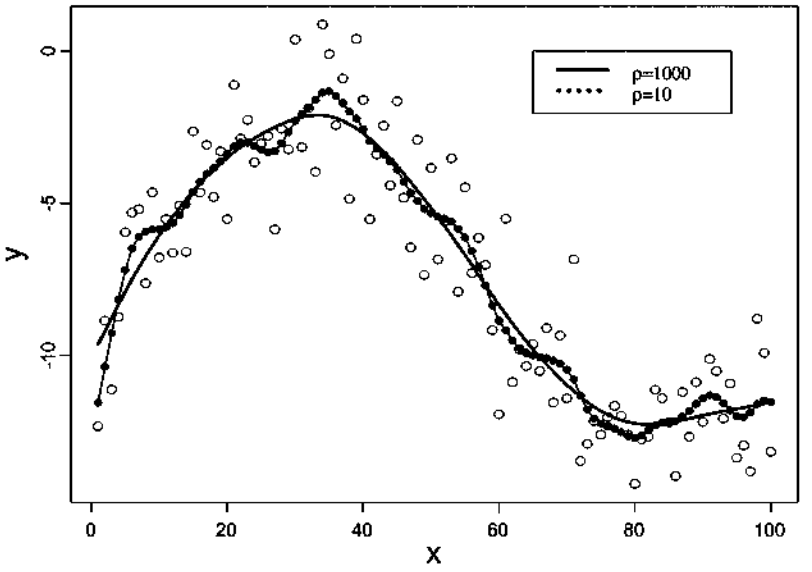

FIGURE 1.2. Penalized smoothing (1.40) with two values of the penalty coefficient, ρ. The larger the penalty coefficient, the smoother the average curve.

Since (1.36) is a quadratic function, its minimization can be expressed through matrix inverse. Indeed, introduce an $n \times (n-2)$ matrix $\mathbf{Q}$ with elements 1 and -2 parallel to the main diagonal; for example,

$$\mathbf{Q} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

for $n = 6$. Then it is elementary to see that the ith element of vector $\mathbf{Q}'\boldsymbol{\mu}$ is $\mu_i - 2\mu_{i+1} + \mu_{i+2}$, and therefore the second term in sum (1.36) can be represented as $\boldsymbol{\mu}'\mathbf{Q}\mathbf{Q}'\boldsymbol{\mu}$, so that the function to minimize takes the form

$$\|\mathbf{y} - \boldsymbol{\mu}\|^2 + \rho\boldsymbol{\mu}'\mathbf{Q}\mathbf{Q}'\boldsymbol{\mu}. \tag{1.37}$$

Let $\mathbf{X}$ be the $n \times 2$ matrix with the first column 1 and the second column $1, 2, ..., n$. It is elementary to see that $\mathbf{Q}'\mathbf{X} = \mathbf{0}$, so in (1.37) we can make a substitution $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} + \mathbf{b}_n$ and come to an equivalent minimization problem,

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{b}_n\|^2 + \rho\mathbf{b}_n'\mathbf{Q}\mathbf{Q}'\mathbf{b}_n, \tag{1.38}$$

over $\boldsymbol{\beta}$ and $\mathbf{b}_n$, a $n \times 1$ vector. Differentiating with respect to $\boldsymbol{\beta}$, we obtain $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and

$$\widehat{\mathbf{b}}_n = (\mathbf{I}+\rho\mathbf{Q}\mathbf{Q}')^{-1}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}). \tag{1.39}$$

Equivalently, in terms of (1.39), one can show that the solution to penalized smoothing (1.36) is given by

$$\widehat{\boldsymbol{\mu}} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \widehat{\mathbf{b}}_n = (\mathbf{I} + \rho\mathbf{Q}\mathbf{Q}')^{-1}\mathbf{y}. \tag{1.40}$$

As follows from (1.40), if $\rho = 0$, we come to the trivial solution $\widehat{\boldsymbol{\mu}} = \mathbf{y}$. When $\rho \to \infty$ we obtain the least squares prediction $\widehat{\boldsymbol{\mu}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$. In Figure 1.2 we show data generated with a penalized smoothing with $\rho = 1000$ and $\rho = 10$. Clearly, the first value is more satisfactory. The choice of the penalty coefficient is crucial. Several *ad hoc* methods are available to choose ρ, such as cross-validation or Akaike information (Hurvich, 1998; Jacqmin-Gadda et al., 2002; Ruppert et al., 2003). Below we illustrate how this parameter may be chosen based on a linear mixed effects model.

Now we construct a linear mixed effects (LME) model that leads to automatic choice of the penalty coefficient. Since $\mathbf{X}$ is the fixed effects matrix, we may treat $\mathbf{b}$ as a random effect with uncorrelated components, yielding the following LME model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon},$$

where

$$\mathbf{Z} = \mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}, \quad \mathbf{b} \sim \mathcal{N}(\mathbf{0}, \sigma^2 d\mathbf{I}_{n-2}), \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_n).$$

This model is a special case of the general LME model (1.10) where $N = 1$ and $\mathbf{D} = d\mathbf{I}_m$ (we use a subindex at the identity matrix to show its size). In brief, this model can be written as $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2(\mathbf{I} + d\mathbf{Z}\mathbf{Z}'))$. Several methods of estimation may be suggested: ordinary or restricted maximum likelihood of Chapter 2 or distribution-free quadratic estimation such as variance least squares, MINQUE, or the method of moments of Chapter 3. As follows from Section 3.7, after d is estimated, there are two equivalent ways to estimate $\boldsymbol{\beta}$ and $\mathbf{b}$ in LME model: using the closed-form formulas

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'(\mathbf{I} + \widehat{d}\mathbf{Z}\mathbf{Z}')^{-1}\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I} + \widehat{d}\mathbf{Z}\mathbf{Z}')^{-1}\mathbf{y}, \tag{1.41}$$

$$\widehat{\mathbf{b}} = \widehat{d}(\mathbf{I}_{n-2} + \widehat{d}\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \tag{1.42}$$

or as the minimizers of the penalized function,

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b}\|^2 + d^{-1}\|\mathbf{b}\|^2.$$

To show the equivalence among (1.41), (1.42), and (1.40), where $\rho = 1/d$, we use the dimension-reduction formula of Section 2.2.3. Then, since $\mathbf{Z}'\mathbf{X} = \mathbf{0}$ (1.41) simplifies to the OLS estimate and prediction from the LME model, $\mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\widehat{\mathbf{b}}$ yields (1.40).

In a nonequidistant case, $x_1 < x_2 < \ldots < x_n$ instead of $\mu_{i+1} - 2\mu_i + \mu_{i-1} = \delta_i$, we have

$$\frac{\mu_{i+1} - \mu_i}{x_{i+1} - x_i} - \frac{\mu_i - \mu_{i-1}}{x_i - x_{i-1}} = \delta_i, \quad i = 2, \ldots, n-1, \tag{1.43}$$

where μ_0 and μ_{n+1} are fixed and unknown and $\delta_i \sim \mathcal{N}(0, \sigma^2 d)$. This model can be applied in a more general setting of spline (or semiparametric) regression with covariates $\mathbf{U}$: for example, $\mathbf{y} = \mathbf{U}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}$, where components of vector μ satisfy (1.43). Again, introducing an appropriate band matrix $\mathbf{Q}$, we reduce the model to LME model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}$, where $\mathbf{X}$ is composed of two vectors, $\mathbf{1}$ and $\mathbf{x}$, augmented by matrix $\mathbf{U}$. This method can be applied to a regression coefficient as

well; for example, $y_i = \boldsymbol{\beta}'\mathbf{u}_i + \mu_i x_i + \varepsilon_i$, where the $\{\mu_i\}$ satisfy (1.43) and $\mathbf{u}_i$ is a vector of adjustment covariates.

Other more complicated LME models may be suggested: for example, to account for autocorrelation (see the literature cited at the beginning of this section).

1.8 Penalized polynomial fitting

One can apply a mixed model to any regression model where penalization is required. For example, here we use this approach for a fully parametric model with a polynomial of high degree. More specifically, without loss of generality, let $x_1 < x_2 < ... < x_n$ and

$$y_i = \boldsymbol{\beta}'\mathbf{u}_i + \sum_{k=2}^{K} b_{k-1}x_i^k + \varepsilon_i, \quad i = 1, ..., n, \tag{1.44}$$

where the $\{\mathbf{u}_i\}$ are design (explanatory) variables and the $\{b_k, k = 1, ..., K-1\}$ are unknown coefficients. For a reason to be explained later, we start from the second degree; the linear part (x) can be represented in the fixed effects ($\mathbf{u}_i$). It is assumed that maximum polynomial degree, K, may be sufficiently large but known. To avoid multicollinearity, instead of x_i^k we can use Legendre orthogonal polynomials $P_k(x_i)$ of the kth degree, so model (1.44) can be replaced by

$$y_i = \boldsymbol{\beta}'\mathbf{u}_i + \sum_{k=2}^{K} b_{k-1}P_k(x_i) + \varepsilon_i. \tag{1.45}$$

By construction, $\sum_{i=1}^{n} P_k(x_i)P_j(x_i) = 0$ for $k \neq j$ and $\sum_{i=1}^{n} P_k^2(x_i) = 1$, which simplifies further computation. Introducing a $(K-1)\times 1$ vector $\mathbf{p}_i = (P_2(x_i), P_3(x_i), ..., P_K(x_i))'$, we come to a regression (conditional) model $y_i|\mathbf{b} = \boldsymbol{\beta}'\mathbf{u}_i + \mathbf{b}'\mathbf{p}_i + \varepsilon_i$, and in conjunction with the *a priori* distribution for the polynomial coefficients, treated as random effects, we arrive at the LME model,

$$\mathbf{y}|\mathbf{b} = \mathbf{U}\boldsymbol{\beta} + \mathbf{P}\mathbf{b} + \boldsymbol{\varepsilon}, \quad \mathbf{b} \sim \mathcal{N}(0, \sigma^2\mathbf{D}). \tag{1.46}$$

There may be several strategies to specify matrix $\mathbf{D}$. First, we can assume that $\mathbf{D}$ is proportional to the identity matrix. Second, $\mathbf{D}$ may be unstructured, but this would involve a large number of estimated parameters, $K(K-1)/2$. Third, we can penalize the high degree, in other words, nonlinearity, as we did in the penalized smoothing model (1.36). Let us take the latter approach. We note that the curvature of the elementary polynomial x^k is associated with the second derivative. Since for fixed x the second derivative of x^k is proportional to $k(k-1)$, we can assume that the diagonal elements of matrix $\mathbf{D}$ are reciprocals of the curvature. For instance, assuming that $\{b_k\}$ do not correlate for $K = 4$, we have

$$\mathbf{D} = d\begin{bmatrix} [2(2-1)]^{-2} & 0 & 0 \\ 0 & [3(3-1)]^{-2} & 0 \\ 0 & 0 & [4(4-1)]^{-2} \end{bmatrix},$$

where d is the scaled unknown variance. This choice means that the variance of $\{b_{k-1}, k = 2, ..., K\}$ is decreasing with k and proportional to $1/[k(k-1)]^2$. Combining this into (1.46), we finally arrive at the LME model,

$$\mathbf{y} \sim \mathcal{N}(\mathbf{U}\boldsymbol{\beta}, \sigma^2(\mathbf{I} + d\mathbf{PDP}')). \tag{1.47}$$

If the scaled variance d were known, we would estimate $\boldsymbol{\beta}$ and $\mathbf{b}$ from

$$\|\mathbf{y} - \mathbf{U}\boldsymbol{\beta} - \mathbf{Pb}\|^2 + d^{-2}\mathbf{b}'\mathbf{D}^{-1}\mathbf{b} \Rightarrow \min_{\boldsymbol{\beta}, \mathbf{b}},$$

so $1/d$ acts as the penalty coefficient. If the scaled variance is large, the contribution of the penalty term is negligible and we come to an unconstrained least squares estimation of model (1.45). Vice versa, if $d \to 0$, we suppress the polynomial part and simply estimate regression $\mathbf{y} = \mathbf{U}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Thus, the d estimation becomes the first priority of the penalized polynomial fitting.

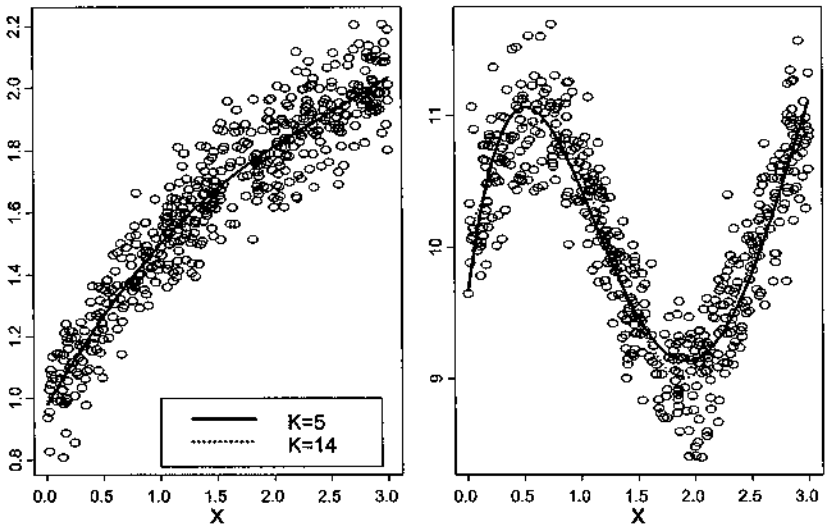

FIGURE 1.3. Two penalized polynomial fittings. The fitting is robust to the choice of the highest degree, K.

Again, several methods are available to estimate d: ordinary or restricted maximum likelihood or noniterative quadratic estimation. In Figure 1.3 we show two penalized polynomial fittings with the penalty coefficient, d, estimated from the linear mixed model (1.47). Points in the left-hand panel were generated as $y_i = x_i^{1.3}(1 + x_i)^{-1} + \varepsilon_i$ and in the right-hand panel as $y_i = \sin(2.5x_i) + 10 + \varepsilon_i$, where the $\{x_i\}$ are randomly distributed on the interval $(0, 1)$ and $\varepsilon_i \sim \mathcal{N}(0, 0.1^2)$, $i = 1, ..., 500$. For this model, the first column of matrix $\mathbf{U}$ is 1 and the second column is $\{x_i\}$. We can draw the following conclusions: (a) the penalized polynomial fitting can adequately approximate nonpolynomial functions such as sin; and (b) since the higher degree is penalized more severely, the choice of K does not make much difference. In particular, polynomials with the highest degree $K = 5$ and $K = 14$ produce almost identical approximation (polynomial curves with $K = 5$ and $K = 14$ overlap).

1.9 Restraining parameters, or what to eat

We have shown above how to restrain (penalize) coefficients in a linear model. In this section we illustrate how a mixed model may be applied to cope with multidimensionality in a nonlinear model: namely, logistic regression with a large number of parameters.

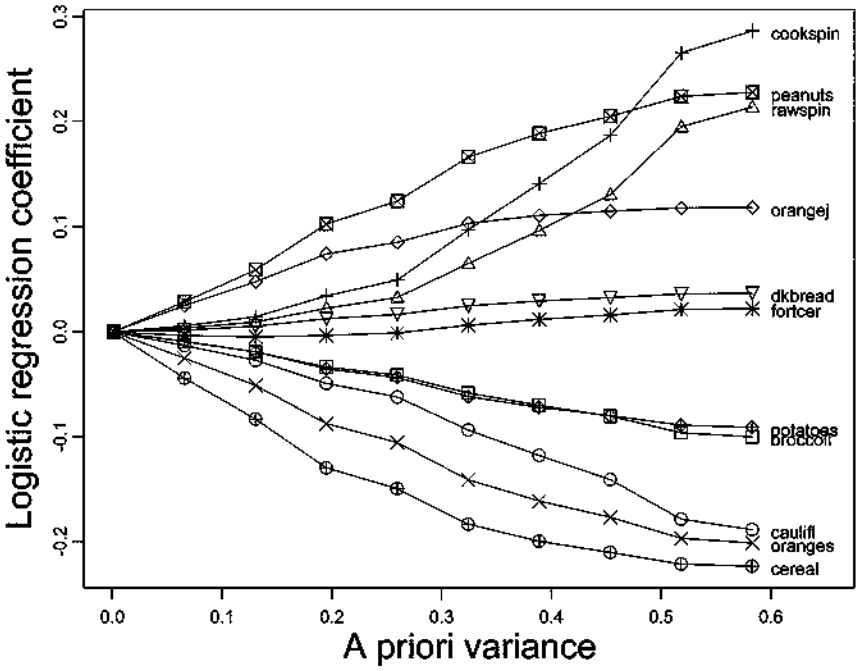

FIGURE 1.4. Coefficients in logistic regression as a function of *a priori* variance σ^2 in the penalized log-likelihood (1.51). Cauliflower, raw oranges, and cereal protect against adenoma, but eating peanuts and spinach increases risk.

A problem with a large number of parameters emerges in nutritional epidemiology (Willett, 1990). To be concrete, let us consider the effect of diet on the health status represented by a binary variable y: If the health status is satisfactory we say that $y = 0$; otherwise, $y = 1$. Let $z_{i1}, z_{i2}, ..., z_{im}$ indicate how much the jth food item was consumed monthly by the ith person, $i = 1, ..., n$. Then, to determine the diet effect, one may relate y_i to $\{z_{ij}, j = 1, ..., m\}$ through logistic regression as

$$\Pr(y_i = 1) = \frac{\exp(\beta_0 + \beta_1 z_{i1} + ... + \beta_m z_{im})}{1 + \exp(\beta_0 + \beta_1 z_{i1} + ... + \beta_m z_{im})}, \quad i = 1, ..., n. \tag{1.48}$$

If a food item increases the probability, $\beta > 0$ ("bad" food); otherwise, $\beta < 0$ ("good" food). Typically, y codes the presence of a disease and quantities z are obtained from a questionnaire. If the number of food items is large (e.g., so large that it exceeds the number of observations), one obtains a wide range of coefficient values with high standard errors. Thus, to obtain meaningful estimates, the food

coefficients should be restrained or penalized. A popular idea in epidemiology to reduce the number of food items is to consider food agglomerates, such as calories, fat, fiber, folate, and so on. This approach is realized in a special DIETSYS program developed under the National Institutes of Health, which replaces a list of original food items with a linear combination representing those agglomerates. Another approach is based on the energy adjustment method (Brown et al., 1994). A big disadvantage of those approaches is that the endpoint recommendation of what to eat is expressed in the agglomerate form, such as "eat less fat food and more vegetables" and therefore is not specific. The approach we discuss here is designed to answer the question: Exactly which food items help to improve health status? To restrain the large number of parameters, a nonlinear mixed model is used.

As an example, we consider a nutritional questionnaire study to reduce the recurrence of colorectal adenoma (Baron et al., 1998). A multicenter study was aimed to investigate the possible beneficial effects of folate intake (mostly from vegetables) based on a questionnaire of patients with at least one recent large-bowel adenoma. It was found that neither cigarette smoking nor folate intake was associated with increased risk of adenoma recurrence. The dependent variable is $y_i = 1$ if for the ith person there was adenoma recurrence and $y_i = 0$ otherwise for $i = 1, 2, ..., n = 751$ people. Thus, according to the logistic regression model (1.48), a large positive coefficient would indicate a risk-increasing food (*bad*) and a negative coefficient would indicate a risk-preventive food (*good*). We do not aim to provide a comprehensive statistical analysis but illustrate how the mixed effects methodology can help to cope with a large number of parameters. Therefore, only $m = 11$ food items were taken into consideration.

Basically, the mixed model is a Bayesian model with unknown food variance, σ^2 as in Section 1.4. More precisely, we treat (1.48) as a conditional model: If $\boldsymbol{\beta} = (\beta_1, ..., \beta_m)'$ were known, then the probability of having an adenoma recurrence is expressed by equation (1.48). *A priori*, we assume that food does not affect recurrence, so we can write

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}). \tag{1.49}$$

This means that the mean of regression coefficients is zero and that, they are independent and have variation σ^2. Equations (1.48) and (1.49) define the generalized linear mixed model (GLMM), to be studied in Chapter 7. To estimate σ^2, we obtain the marginal likelihood with $\boldsymbol{\beta}$ integrated out as

$$l(\beta_0, \sigma^2) = (2\pi\sigma^2)^{-m/2} \int_{R^m} e^{l(\beta_0, \boldsymbol{\beta}) - 0.5\sigma^{-2}\|\boldsymbol{\beta}\|^2} d\boldsymbol{\beta}, \tag{1.50}$$

where $l(\beta_0, \boldsymbol{\beta})$ is the ordinary log-likelihood for model (1.48). Direct integration is prohibitive because dimension m is large (in our case, $m = 11$). Therefore, approximate methods for integral (1.50), such as Laplace approximation or quasi-likelihood, should be used. After estimates $\widehat{\sigma}^2$ and $\widehat{\beta}_0$ are obtained, we derive the posterior means for $\boldsymbol{\beta}$ that maximize the penalized log-likelihood,

$$l(\widehat{\beta}_0, \boldsymbol{\beta}) - 0.5\widehat{\sigma}^{-2} \|\boldsymbol{\beta}\|^2. \tag{1.51}$$

Note that in the Bayesian approach we need to define values for β_0 and σ^2, but in mixed model we obtain them from the data.

In Figure 1.4 we plot posterior regression coefficients as a function of *a priori* variance, σ^2. When the variance is zero, all coefficients are zero. Indeed, the second term in the penalized log-likelihood function (1.51) then prevails, yielding $\boldsymbol{\beta} = \mathbf{0}$. Larger σ^2 implies less penalty and more variation in the regression coefficients. When $\sigma^2 \to \infty$, GLMM converges to ordinary logistic regression. Interestingly, cereal, raw oranges, and cauliflower prevent adenoma, but peanuts and spinach increase the risk.

1.10 Ill-posed problems, Tikhonov regularization, and mixed effects

Mixed models may be considered a tool for solving ill-posed problems. Let $\theta_1, \theta_2, ..., \theta_m$ be system inputs and let $f_1, f_2, ..., f_n$ be a system output. For example, consider an image reconstruction problem based on the Near-Infra Red (NIR) technique. The light goes through a semitransparent body with the absorption density θ_j at location (u_j, v_j) within the body. More details may be found in a recent book by Barrett and Myers (2004). Due to the law of optics, if $\{\theta_j, j = 1, ..., m\}$ were known, the light intensity f_i at detector i on the periphery of the body would be known exactly as a function of $\{\theta_j\}$, or in vector form, $f_i = f_i(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the m-dimensional unknown vector. Vector $\boldsymbol{\theta}$ is called the system vector parameter or, in statistical language, simply the parameter. Having n measurements on the periphery, $\{y_1, ..., y_n\}$, we want to reconstruct the optical properties *within* the body (absorption coefficients), $\{\theta_1, ..., \theta_m\}$, at as many points as possible—this is an inverse problem. An interested reader may read more about statistical aspects of inverse problems in a review paper by Evans and Stark (2002).

Often, inverse problems are ill-posed. In our example we would like to have as few detectors and as many points as possible, so dimensions n and m are close. Besides, the system is usually noisy, leading to a nonlinear regression problem,

$$y_i = f_i(\boldsymbol{\theta}) + \varepsilon_i, \quad i = 1, 2, ..., n. \tag{1.52}$$

To obtain estimates of $\boldsymbol{\theta}$, the least squares criterion is generally used, $\sum_{i=1}^{n}(y_i - f_i(\boldsymbol{\theta}))^2 \Rightarrow \min$. However, since $m \approx n$ and functions $f_i(\boldsymbol{\theta})$ are nonlinear, estimation (reconstruction) of $\boldsymbol{\theta}$ becomes problematic. Therefore, the problem is called *ill-posed.* A Russian mathematician, Tikhonov (Tikhonov and Arsenin, 1977), suggested augmenting the sum of squares by a quadratic term that leads to the functional

$$T(\boldsymbol{\theta}) = \sum_{i=1}^{n}(y_i - f_i(\boldsymbol{\theta}))^2 + \rho \|\boldsymbol{\theta}\|^2, \tag{1.53}$$

where ρ is called the *regularization parameter* ($\rho > 0$). The original ill-posed problem becomes a well-posed problem. Tikhonov regularization became very popular in applied mathematics and engineering, with a variety of applications: solution of an ill-conditioned linear system, integral equations, density estimation, image reconstruction, and so on. Although several heuristic techniques are available to assess the regularization parameter, such as cross-validation, there is no unified approach

to the selection of ρ (Vogel, 2002). However, selection of the value of the regularization parameter is crucial: If ρ is close to 0, we come again to an ill-posed problem; if ρ is too large, the solution degenerates to $\widehat{\boldsymbol{\theta}} = \mathbf{0}$. Strictly speaking, the problem of ill-posedness is just reduced to another problem: the selection of ρ.

Tikhonov regularization may be treated from a statistical point of view, interpreting the inverse problem as a mixed model written in a hierarchical (two-stage) fashion. Indeed, following the line of the Bayesian approach, we assume that $\boldsymbol{\theta}$ is random and (1.52) is treated as a conditional equation, where the $\{\varepsilon_i\}$ are normally distributed with zero mean and constant (system) variance σ^2. Assume that our prior experience says that the component values of vector $\boldsymbol{\theta}$ are expected to be in the neighborhood of zero with certain variance σ_θ^2; or more precisely, $\theta_j \sim \mathcal{N}(0, \sigma_\theta^2)$. This is called the *prior* distribution for the parameters. Let us assume for awhile that the system variance, σ^2, and parameter variance, σ_θ^2, are known. After observations $\{y_i\}$ are collected, we may ask how our prior distribution changes to become a *posterior* distribution with the density $f(\boldsymbol{\theta}|\mathbf{y}) = C \times f(\mathbf{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta})$, where C is the normalizing constant, $f(\mathbf{y}|\boldsymbol{\theta})$ is the conditional density, and $f(\boldsymbol{\theta})$ is the parameter density (see Section 1.4). Since we assume normal distribution,

$$f(\mathbf{y}|\boldsymbol{\theta}) = (2\pi\sigma)^{-n/2} e^{-\frac{1}{2\sigma^2}\|\mathbf{y}-\mathbf{f}(\boldsymbol{\theta})\|^2}, \quad f(\boldsymbol{\theta}) = (2\pi\sigma_\theta)^{-m/2} e^{-\frac{1}{2\sigma_\theta^2}\|\boldsymbol{\theta}\|^2}.$$

Note that the posterior distribution, $f(\boldsymbol{\theta}|\mathbf{y})$, is not a normal distribution of $\boldsymbol{\theta}$, and the "center" of the distribution would give an idea of where the posterior values are concentrated. Let us take the *mode* of the distribution, where the density takes its maximum. In image processing and reconstruction literature the model is called MAximum a Posteriori (MAP) estimation (Geman and Geman, 1984; Besag, 1986, 1989). Since the variances are known, the MAP estimator for $\boldsymbol{\theta}$ reduces to the minimization problem,

$$\frac{1}{\sigma^2}\|\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})\|^2 + \frac{1}{\sigma_\theta^2}\|\boldsymbol{\theta}\|^2 \Rightarrow \min_{\boldsymbol{\theta}}. \tag{1.54}$$

But this is the Tikhonov functional (1.53) with $\rho = \sigma^2/\sigma_\theta^2$.

Summing up:

1. The Tikhonov regularization procedure can be derived through the Bayesian approach with known variances of the system error and *a priori* parameters assuming a normal distribution.

2. The regularization (penalty) coefficient is the ratio of the system to the parameter variance.

3. The Tikhonov solution is the mode of the posterior distribution, the MAP estimator.

4. The Tikhonov solution assumes that the *a priori* value of the parameter is zero.

We make several comments. First, if the system is not too noisy but there is substantial *a priori* variation in $\boldsymbol{\theta}$, the regularization parameter should be small. Second, the assumption that the *a priori* value of the parameter is zero may be

inadequate. For example, in the NIR problem this would mean that the absorption is zero, which is equivalent to assuming that the body is absolutely transparent (as a vacuum). It would be better to assume that $\theta_j \sim \mathcal{N}(\theta_0, \sigma_\theta^2)$, so that the modified Tikhonov functional takes the form

$$\sum_{i=1}^{n}(y_i - f_i(\boldsymbol{\theta}))^2 + \rho \left\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\right\|^2, \tag{1.55}$$

where θ_0 is the background absorption coefficient. Although the Bayesian interpretation gives Tikhonov regularization a nice statistical interpretation, the problem of selection of the regularization parameter remains. Now we shall show that using a *mixed effects* approach, ρ can be obtained along with $\boldsymbol{\theta}$. The following reasoning follows the line of penalized likelihood of Section 1.5, the only difference being that now we apply it to reducing an ill-posed problem to a well-posed problem.

In the mixed effects approach, we change nothing in the Bayesian approach except for the assumption that σ^2 and σ_θ^2 are unknown along with $\boldsymbol{\theta}$. Thus, the mixed model is written in hierarchical fashion as

$$\mathbf{y}|\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{f}(\boldsymbol{\theta}), \sigma^2\mathbf{I}_n), \quad \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}_0, \sigma_\theta^2\mathbf{I}_m), \tag{1.56}$$

where $\mathbf{I}$ is an identity matrix of the appropriate size and $\boldsymbol{\theta}_0$ is known. Model (1.56) belongs to the family of nonlinear mixed effects model studied in Chapter 8. Since only the observations $\{y_i\}$ are available to estimate the parameters, we need to find the marginal distribution with the likelihood expressed via an integral as

$$L(\sigma^2, \sigma_\theta^2) = (2\pi\sigma^2)^{-n/2}(2\pi\sigma_\theta^2)^{-m/2} \int_{R^m} e^{-\frac{1}{2\sigma^2}\|\mathbf{y}-\mathbf{f}(\boldsymbol{\theta})\|^2 - \frac{1}{\sigma_\theta^2}\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2} d\boldsymbol{\theta}.$$

One could maximize L over the unknown parameters σ^2 and σ_θ^2 to obtain the maximum likelihood (ML) estimation that involves a multidimensional integration. The core of the approximation methods to the ML solution is the Laplace approximation (1.27), implemented in Section 8.8; we also refer the reader to Section 8.15. The easiest way to estimate the variance parameters is to approximate $\mathbf{f}$ by a linear function about $\boldsymbol{\theta}_0$, see Section 8.6. Then model (1.56) simplifies to a LME model $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma^2(\mathbf{I}+d\mathbf{Z}\mathbf{Z}'))$, where $d = \sigma_\theta^2/\sigma^2$ and $\mathbf{Z} = \partial\mathbf{f}/\partial\boldsymbol{\theta}$ is evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. The maximum likelihood algorithm for estimation of σ^2 and d is described in the next section. When the variance parameters are known, estimation of *a posteriori* $\boldsymbol{\theta}$, as follows from the Laplace approximation (1.27), is almost equivalent to minimization of (1.55), where $\rho = 1/d$. Symbolically,

$$\text{MAP estimator} = \text{mixed model ML estimator},$$

but unlike MAP, we do not require values for σ^2 and σ_θ^2 (specifically for their ratio), which are estimated from maximization of L. The appropriate methods are to be studied extensively in Chapter 8. In fact, the choice of ρ based on our mixed model has much in common with what other authors suggested based on the noise level, σ^2 (Kirsch, 1996; Kress, 1999; Colton et al., 2000). Notice that model (1.56) allows a combination of repeated measurements of $\mathbf{y}$, leading to a multilevel mixed model. This nonlinear mixed model technique has been applied to breast image reconstruction by microwave, with promising results (Meaney et al., 2001).

Application of the mixed model methodology to a linear image reconstruction is described in Section 1.11. We provide a constructive procedure to estimate the regularization parameter, ρ in the penalized least squares (1.54) from the data.

1.11 Computerized tomography and linear image reconstruction

Computerized (sometimes called computed) tomography (CT) reconstructs an image from projections. Thus, by measuring signals on the periphery of the body, CT reconstructs what is inside the body. This technique has many applications in radiology, and the interested reader can learn more from Andrews and Hunt (1977), Hall (1979), Herman (1980), Parker (1990), Seeram (1994), and Kak and Slaney (2001) among others. Epstein (2003) provides a comprehensive account of mathematical aspects of image reconstruction with medical applications. An up-to-date and complete discussion of image analysis is given in a book by Barrett and Myers (2004).

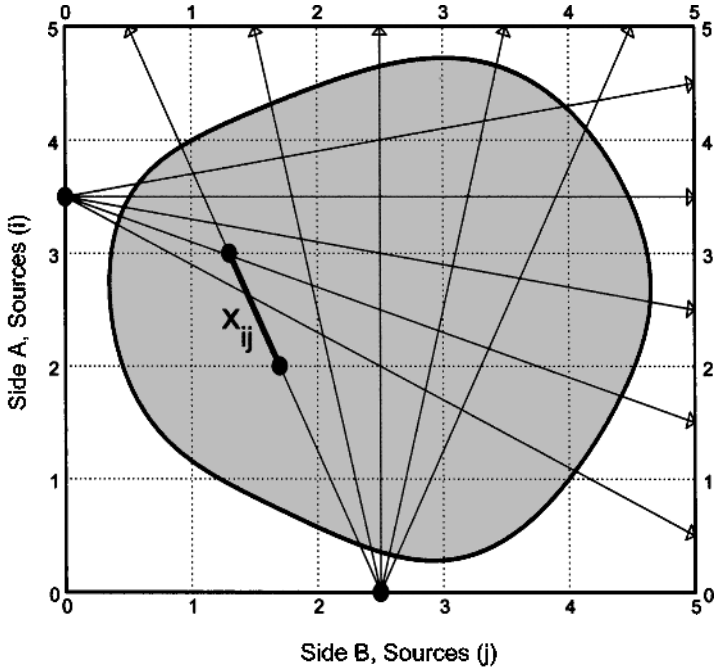

FIGURE 1.5. Principal idea of CT image reconstruction from projections. Beams penetrate the body so that the initial signal intensity is reduced. Measuring the exit intensities at several locations, CT reconstructs the attenuation coefficient in each box. Plotting these attenuation coefficients results in an image.

A CT device consists of several sources and detectors located on the periphery of a square or circle—we refer the reader to Figure 1.5, where the principal idea of a CT

scan is represented schematically. Beams of x-rays or light come out of the source at a given angle, penetrate the body, and are received at detectors on an opposite side. If I_0 is the initial intensity of the beam, which comes in at one end of a homogeneous bar of length x and comes out at the other end with intensity I_1, with a certain degree of approximation we have $I_1 = I_0 e^{-\theta x}$, where θ is called the *attenuation coefficient*. If a nonhomogeneous bar is composed of m homogeneous bars of length x_i and attenuation coefficient θ_i, the intensity at the end is $I_1 = I_0 e^{-\sum_{i=1}^{m} \theta_i x_i}$, or on the log scale, $y = \sum_{i=1}^{m} \theta_i x_i$, where $y = \ln(I_0/I_1)$. This simple formula gives rise to the CT image reconstruction. Imagine that the body is divided into m small boxes (dotted lines in Figure 1.5) and within each box the attenuation coefficient θ_j is constant, $j = 1, ..., m$. If the beam comes out from the source at a given angle, we can compute the length of the ray within each box so that the following representation takes place:

$$y_i = \sum_{j=1}^{m} x_{ij}\theta_j + \varepsilon_i, \quad i = 1, ..., n, \tag{1.57}$$

where i is the number of beams, x_{ij} is the length of the ith beam within the jth box, and ε_i is the iid random term (see Figure 1.5). Since beam angles are predefined, $\{x_{ij}, i = 1, ..., n, j = 1, ..., m\}$ are fixed numbers and can be derived from the CT hardware specification. Having n measurements $\{y_i\}$, we reconstruct (estimate in statistical terminology) m attenuation coefficients $\{\theta_j\}$. Plotting $\{\theta_j\}$ at appropriate locations yields a CT image, so the set of attenuation coefficients $\{\theta_j\}$ is called an *image*. The larger θ_j, the denser the image. This is a linear image reconstruction because it reduces to a linear problem. Special features of this problem are: (a) since we want to see as many pixels as possible, m and n are close; and (b) the number of estimated coefficients, m, is large; for example, to see a 64×64 image, we have $m = 62^2 = 3844$ unknown parameters. This makes the CT problem ill-posed. To improve the least squares solution, several approaches have been put forward, such as Tikhonov regularization and the Bayesian approach. The former requires knowledge of the regularization parameter, and the latter requires complete specification of an *a priori* image.

We apply the mixed effects approach, in which the *a priori* image is not specified completely but is up to some unknown parameters. Then the regularization parameter is estimated from the CT data along with attenuation coefficients. Introducing the $n \times m$ projection matrix $\mathbf{X}$ with elements x_{ij}, we rewrite (1.57) in vector form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \tag{1.58}$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ and $\mathbf{I}_n$ is the $n \times n$ identity matrix. Model (1.58) is an ordinary linear regression model with the efficient least squares (LS) estimator $\widehat{\boldsymbol{\theta}}_{LS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. This estimator is valid if $n > m$. When m approaches n, the LS estimator becomes unstable because matrix $\mathrm{cov}(\widehat{\boldsymbol{\theta}}_{LS}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ becomes unstable as well. Consequently, a small perturbation in data leads to a large perturbation in $\widehat{\boldsymbol{\theta}}_{LS}$. To improve the solution, we use *a priori* information on the image to be reconstructed. For example, we may know how the image may look from previous experiments. Statistically, if $\boldsymbol{\theta}_0$ is the prior image, we write

$$\boldsymbol{\theta} = \boldsymbol{\theta}_0 + \mathbf{b}, \tag{1.59}$$

where $\boldsymbol{\theta}$ is known and $\mathbf{b}$ is the deviation, a random vector. The reader will immediately recognize that $\mathbf{b}$ may be treated as the random effect, so that the couple (1.58) and (1.59) specify a linear mixed effects model, or, more precisely, a linear model with random coefficients. It is simplest to assume that $\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \sigma^2 d\mathbf{I}_m)$, where d is the scaled variance.

A distinctive feature of the LME model from the Bayesian standpoint is that we do not specify variances σ^2 and d, but estimate them along with $\boldsymbol{\theta}$. In the rest of this section we provide a constructive algorithm to estimate $\boldsymbol{\theta}$ and σ^2 and d, which becomes the reciprocal of the penalty coefficient in the Tikhonov regularization (1.55).

The two equations (1.58) and (1.59) can be combined to produce a one-equation statistical model

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\theta}_0, \sigma^2(\mathbf{I}_n + d\mathbf{X}\mathbf{X}')). \tag{1.60}$$

Our plan to estimate σ^2 and d is as follows. First we estimate σ^2 and d by maximum likelihood. Second, we apply the penalized least squares with the regularization parameter d^{-1} to derive an improved *a posteriori* image. For details, we refer the reader to Chapter 2. Another, pedagogical purpose of the following derivation is for the reader to get a flavor of the statistical and matrix algebra techniques to be used throughout the book.

Letting $\mathbf{e} = \mathbf{y} - \mathbf{X}\boldsymbol{\theta}_0$, the log-likelihood, up to a constant term, can be written as

$$l(\sigma^2, d) = -0.5\left\{n \ln \sigma^2 + \ln\left|\mathbf{I}_n + d\mathbf{X}\mathbf{X}'\right| + \sigma^{-2}\mathbf{e}'(\mathbf{I}_n + d\mathbf{X}\mathbf{X}')^{-1}\mathbf{e}\right\}, \tag{1.61}$$

where $\mathbf{I}_n$ is the identity matrix of the order indicated. Using the dimension-reduction formulas of Section 2.2.3, we obtain

$$\left|\mathbf{I}_n + d\mathbf{X}\mathbf{X}'\right| = \left|\mathbf{I}_m + d\mathbf{X}'\mathbf{X}\right|, \quad (\mathbf{I}_n + d\mathbf{X}\mathbf{X}')^{-1} = \mathbf{I}_n - d\mathbf{X}(\mathbf{I}_m + d\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

Let $\lambda_1, ..., \lambda_m$ be the eigenvalues and $\mathbf{p}_1, ..., \mathbf{p}_m$ the corresponding eigenvectors of matrix $\mathbf{X}'\mathbf{X}$. Then we can represent

$$\mathbf{e}'\mathbf{X}(\mathbf{I}_m + d\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e} = \sum_{j=1}^{m} \frac{w_j^2}{1 + d\lambda_j}, \quad \ln\left|\mathbf{I}_m + d\mathbf{X}'\mathbf{X}\right| = \sum_{j=1}^{m} \ln(1 + d\lambda_j),$$

where $w_j = \mathbf{e}'\mathbf{X}\mathbf{p}_j$. Then (1.61) simplifies to

$$-0.5\left\{n \ln \sigma^2 + \sigma^{-2}\left[S - d\sum_{j=1}^{m} w_j^2(1 + d\lambda_j)^{-1}\right] + \sum_{j=1}^{m} \ln(1 + d\lambda_j)\right\},$$

where $S = \mathbf{e}'\mathbf{e}$. When d is held fixed, the maximum over σ^2 is computed exactly:

$$\sigma_s^2 = n^{-1}\left[S - d_s\sum_{j=1}^{m} \frac{w_j^2}{1 + d_s\lambda_j}\right].$$

When σ^2 is held fixed, we use the fixed-point iterations

$$d_{s+1} = d_s \frac{\sum_{j=1}^{m} w_j^2(1 + d_s\lambda_j)^{-2}}{\sigma_s^2 \sum_{j=1}^{m} \lambda_j(1 + d_s\lambda_j)^{-1}}, \quad s = 0, 1, 2, ...$$

See Appendix 13.3.4 for a general discussion of optimization algorithms, including the FP algorithm. We can start from $\sigma_0^2 = (S - \sum_{j=1}^m w_j^2/\lambda_j)/n$ and $d_0 = \sum_{j=1}^m w_j^2/\lambda_j^2/(m\sigma_0^2)$. At convergence, we obtain $\widehat{\sigma}_{ML}^2$ and $\widehat{d}_{ML}$.

To obtain the posterior image, $\widehat{\boldsymbol{\theta}}$, after σ^2 and d are determined, we can use a closed-form formula or derive $\widehat{\boldsymbol{\theta}}$ from the penalized least squares (PLS); the equivalence is proved in Section 3.7. The PLS takes the form

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}_0 - \mathbf{X}\boldsymbol{\theta}\|^2 + \widehat{d}_{ML}^{-1} \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 \Rightarrow \min_{\boldsymbol{\theta}}. \tag{1.62}$$

Denoting $\boldsymbol{\theta}_0 - \boldsymbol{\theta} = \mathbf{b}$, we come to the Tikhonov optimization criterion function $\|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 + \rho \|\mathbf{b}\|^2$, where $\rho = 1/\widehat{d}_{ML}$. This is a quadratic function of $\boldsymbol{\theta}$ and the closed-form solution exists. Thus, the final mixed effects (ME) CT image is given by

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 + \left(\mathbf{X}'\mathbf{X} + \widehat{d}_{ML}^{-1}\mathbf{I}\right)^{-1} \mathbf{X}'\mathbf{y}. \tag{1.63}$$

When the variance of the random effect is zero, we obtain $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$; when $d \to \infty$, the ME estimate converges to the LS estimate, $\widehat{\boldsymbol{\theta}}_{LS}$. The covariance matrix, $\mathrm{cov}(\widehat{\boldsymbol{\theta}}) = \widehat{\sigma}_{ML}^2 \left(\mathbf{X}'\mathbf{X} + \widehat{d}_{ML}^{-1}\mathbf{I}\right)^{-1}$, is well-conditioned, and therefore the ME image is stable. As the reader may notice, (1.62) is the Tikhonov regularization (1.55) with the penalty coefficient equal to the reciprocal of the scaled variance estimate.

We may put other restrictions on the reconstructed image. For example, one may assume that the image is fairly smooth. Then, introducing the $(m-1) \times m$ difference matrix,

$$\mathbf{L} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}, \tag{1.64}$$

we come to the model $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\theta}_0, \sigma^2(\mathbf{I}_n + d\mathbf{W}))$, where $\mathbf{W} = \mathbf{X}\mathbf{L}'\mathbf{L}\mathbf{X}'$ is a fixed matrix. Then, nonsmoothed solutions will be penalized with PLS $\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta} - \mathbf{X}\mathbf{L}'\boldsymbol{\theta}\|^2 + \widehat{d}_{ML}^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)'$.

It is straightforward to generalize a mixed model (1.60) to a multilevel clustered model where, for example, repeated imaging data may be combined into one pool to detect differences between visits to the doctor, or to determine a trend, differences in gender, differences in age, and so on.

1.12 GLMM for PET

In this section we consider an image reconstruction method popular in medical applications. The statistical solution involves two components: a statistical model and an estimation algorithm. We emphasize that computational features become integral to successful implementation.

Positron emission tomography (PET) is important in nuclear medicine and has features common with x-ray computerized tomography. The difference with the linear image reconstruction considered above is that the observations are not continuous but are photon counts that imply a nonlinear statistical model. Shepp and

Vardi (1982) described PET as a probabilistic model based on the Poisson distribution. A current review of reconstruction methods for PET may be found in Lewitt and Matej (2003). The idea of PET is as follows: A subject is administered a dose of the molecules labeled with radioactive atoms. These atoms are unstable isotopes leading to the emission of gamma-ray photons, which are detected outside the body by a ring of surrounding detectors. To simplify, we consider a two-dimensional PET system. By counting the number of photons in different directions, PET attempts to reconstruct the decay rate, λ, at each point within the body. The PET image is the distribution of these rates. To make the problem solvable, instead of a continuum of points, imagine that the body is divided into m disjoint boxes. The number of decay events, n_j occurring over a fixed time in box j is random and follows the Poisson law with the rate λ_j, so that $E(n_j) = \lambda_j$, $j = 1, ..., m$. Numbers n_j are unobservable and λ_j are unknown. However, there is a ring of detectors around the body which count the total number of decay events in n cross-section tubes. Let the total number of decay events occurring in the ith tube be k_i, $i = 1, ..., n$. There exists a fixed $n \times m$ matrix $\mathbf{A}$ such that $k_i = \sum_{j=1}^{m} a_{ij} n_j$. This matrix, called the projection matrix, is derived from the geometry of the body: tube angle, size, etc. Assuming that counts n_j are independent, k_i also has a Poisson distribution with the rate $E(k_i) = \sum_{j=1}^{m} a_{ij} E(n_j) = \sum_{j=1}^{m} a_{ij} \lambda_j$. Further, assuming that $\{k_i, i = 1, ..., n\}$ are independent, we come to the likelihood function

$$L(\lambda_1, ..., \lambda_m) = \prod_{i=1}^{n} \frac{\left(\sum_{j=1}^{m} a_{ij} \lambda_j\right)^{k_i}}{k_i!} e^{-\sum_{j=1}^{m} a_{ij} \lambda_j}.$$

The log-likelihood, up to a constant term, is

$$l(\lambda_1, ..., \lambda_m) = \sum_{i=1}^{n} \left(k_i \ln \sum_{j=1}^{m} a_{ij} \lambda_j - \sum_{j=1}^{m} a_{ij} \lambda_j \right). \tag{1.65}$$

To find the maximum likelihood estimate, we need to solve m score equations,

$$\frac{\partial l}{\partial \lambda_p} = \sum_{i=1}^{n} \frac{h_{ip}}{\sum_{j=1}^{m} a_{ij} \lambda_j} - r_p = 0, \quad p = 1, ..., m, \tag{1.66}$$

where $h_{ip} = k_i a_{ip}$ and $r_p = \sum_{i=1}^{n} a_{ip}$. Usually, the EM algorithm is used to maximize (1.65):

$$\lambda_{p,s+1} = \frac{\lambda_{ps}}{r_p} \sum_{i=1}^{n} \frac{h_{ip}}{\sum_{j=1}^{m} a_{ij} \lambda_{js}}, \quad p = 1, ..., m, \tag{1.67}$$

with iterations $s = 0, 1, ...$ The iterations can be started from $\lambda_{p0} = r_p^{-1} \sum_{i=1}^{n} a_{ip}^{-1} h_{ip}$. At convergence, $\lambda_{p,s+1} = \lambda_{ps}$, satisfying the score equations (1.66) and meaning that the EM algorithms converges to the maximum likelihood estimate (MLE). Moreover, as follows from the general properties of the EM algorithm, iterations (1.67) increase the log-likelihood value, l from iteration to iteration. For a general discussion of the optimization algorithms used in statistics, including EM, see Appendix 13.3.4.

An alternative maximization algorithm for l is the Newton–Raphson (NR). We rewrite the estimating equations (1.66) in vector form, but first we need the first and second derivatives,

$$\frac{\partial l}{\partial \boldsymbol{\lambda}} = \sum_{i=1}^{n} \frac{k_i}{w_i}\mathbf{a}_i - \mathbf{r}, \quad \frac{\partial^2 l}{\partial \boldsymbol{\lambda}^2} = -\sum_{i=1}^{n} \frac{k_i}{w_i^2}\mathbf{a}_i'\mathbf{a}_i,$$

where $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_m)'$, $\mathbf{r} = (r_1, ..., r_m)'$, $w_i = \sum_{j=1}^{m} a_{ij}\lambda_s$, and $\mathbf{a}_i$ is the ith row vector of matrix $\mathbf{A}$. Then the NR iterations are

$$\boldsymbol{\lambda}_{s+1} = \boldsymbol{\lambda}_s + \left(\sum_{i=1}^{n} \frac{k_i}{w_i^2}\mathbf{a}_i'\mathbf{a}_i\right)^{-1} \left(\sum_{i=1}^{n} \frac{k_i}{w_i}\mathbf{a}_i - \mathbf{r}\right). \tag{1.68}$$

Noticing that $E(k_i) = w_i$, we obtain the expected NR or Fisher scoring algorithm:

$$\boldsymbol{\lambda}_{s+1} = \boldsymbol{\lambda}_s + \left(\sum_{i=1}^{n} \frac{1}{w_i}\mathbf{a}_i'\mathbf{a}_i\right)^{-1} \left(\sum_{i=1}^{n} \frac{k_i}{w_i}\mathbf{a}_i - \mathbf{r}\right). \tag{1.69}$$

At the final iteration, the inverse matrix is the covariance matrix for the MLE. This matrix will be needed later for various statistical hypothesis testing. Note that the EM algorithm does not produce this matrix, which may partially explain the fact that little statistical testing has been reported in the PET literature.

Our practice shows that whereas the EM algorithm may be very slow (sometimes it requires 1000 iterations) algorithms (1.68) and (1.69) are very fast and require only four or five iterations to obtain the MLE with the same precision. However, an advantage of the EM algorithm is that it does not require a matrix inverse. Since the number of reconstructed nodes/pixels is typically large, a matrix inverse at each iteration may become a limitation. We can modify the NR or FS to avoid the matrix inverse by employing the idea of the Unit Step (US) algorithm (see also Section 7.1.5). The idea of this algorithm is to obtain an approximation of the matrix inverse from above. For example, for the FS algorithm, we have

$$\sum_{i=1}^{n} w_i^{-1}\mathbf{a}_i'\mathbf{a}_i \le \nu^{-1}\sum_{i=1}^{n} w_i^{-1}\mathbf{a}_i'\mathbf{a}_i = \nu^{-1}\mathbf{A}'\mathbf{A},$$

where $\nu = \min w_i$. Then the US algorithm, as an economical version of the FS algorithm, takes the form

$$\boldsymbol{\lambda}_{s+1} = \boldsymbol{\lambda}_s + \nu_s\left(\mathbf{A}'\mathbf{A}\right)^{-1}\left(\sum_{i=1}^{n} k_i w_i^{-1}\mathbf{a}_i - \mathbf{r}\right), \tag{1.70}$$

where $\left(\mathbf{A}'\mathbf{A}\right)^{-1}$ is computed once beforehand. Although the US algorithm is usually slower than NR or FS, it is faster than EM and requires a dozen iterations rather than hundreds or even thousands.

PET is, as are many image reconstruction problems, an ill-posed problem because we want to have the number of pixels as large as possible and the number of measurements as small as possible, so that m is close to n. If no *a priori* information is available, the ML estimate is unstable. The Bayesian approach gained much

popularity for PET image reconstruction (Hebert and Leahy, 1989; Kaufman, 1993; Fessler, 1994; Qi and Huesman, 2001; De Pierro and Yamagishi, 2001; Nuyts and Fessler, 2003). As mentioned above, under the Bayesian approach *a priori* image does not allow unknown parameters, whereas the mixed model does.

In the following parameterization, we assume that the rate, λ, is expressed through an exponential function as $\lambda = e^{\gamma}$, which is convenient because (a) one does not have to care about the positiveness of the rate, and (b) it is easy to penalize γ using a normal distribution. Following the line of the generalized linear mixed model (GLMM) technique of Chapter 7, we write the conditional log-likelihood in the form

$$l(\gamma_1, ..., \gamma_m) = \sum_{i=1}^{n} \left(k_i \ln \sum_{j=1}^{m} a_{ij} e^{\gamma_j} - \sum_{j=1}^{m} a_{ij} e^{\gamma_j} \right),$$

where $\gamma_1, ..., \gamma_m$ are iid random rates specified in the second equation as

$$\gamma_j \sim \mathcal{N}(\gamma_0, \sigma^2), \quad j = 1, ..., m,$$

where γ_0 is known and σ^2 is unknown. To obtain a marginal likelihood, we need to use integration,

$$L(\gamma_1, ..., \gamma_m) = \frac{1}{\sigma\sqrt{2\pi}} \int_{R^m} e^{l(\gamma_1, ..., \gamma_m) - \frac{1}{2\sigma^2} \sum_{j=1}^{m} (\gamma_j - \gamma_0)^2} d\gamma_1 ... d\gamma_m.$$

This mixed model belongs to the family of Poisson models with random intercepts (see Section 7.5). Since m is large, exact integration is prohibitive. Several methods were developed to avoid integration using approximate estimation. Importantly, to obtain *a posteriori* rates, as follows from Laplace approximation (1.27), we maximize the penalized log-likelihood

$$P = \sum_{i=1}^{n} \left(k_i \ln \sum_{j=1}^{m} a_{ij} e^{\gamma_j} - \sum_{j=1}^{m} a_{ij} e^{\gamma_j} \right) - \sigma^{-2} \sum_{i=1}^{n} (\gamma_i - \gamma_0)^2$$

after σ^2 and γ_0 are estimated. If the image is close to the prior image, σ^2 is small and the second term in P overshadows the first. If the image is far from the prior image, the penalizing term is small. After σ^2 is estimated, one can maximize P by the NR or FS algorithm. The inverse matrix at the final iteration gives the covariance matrix of the mixed model MLE. Many *a priori* assumptions may be realized in the mixed model. For example, if one wants to penalize nonsmoothness $\gamma \sim \mathcal{N}(\gamma_0 \mathbf{1}, \sigma^2 \mathbf{L}'\mathbf{L})$, where L is the difference matrix defined in (1.64).

1.13 Maple leaf shape analysis

The mixed model is an adequate statistical model to describe individual variety within a biological category. Indeed, the milestone concepts of the mixed model, the within- and between-subject variation, exactly match the principles of biological variety. Look at Figure 1.6: Nine maple leaves from the same tree have significant

individual variation, but at the same time look similar. In the language of the mixed model, population-averaged parameters specify the common biological type (such as average maple leaf), and subject-specific parameters specify subject individuality. In classical statistics, observations are assumed to be independent and identically distributed; in the mixed effects approach, observations from the same individual constitute a cluster and therefore are correlated.

Shape is perhaps the simplest characteristic of a biological subject. We apply mixed model techniques to shape analysis in Chapter 11. Importantly, ordinary shape analysis deals with one shape, whereas a mixed model processes a sample (ensemble) of shapes simultaneously.

An important step in shape analysis is shape quantification, or in other terms, representation of a two-dimensional geometrical object numerically as a sequence of numbers. Typically, different quantification methods lead to different statistical models.

For example, for this maple leaf analysis, we use the Random Fourier Descriptor (RFD) model (see Section 11.7.2 for details). This model deals with pair coordinates, $\{(x_{ij}, y_{ij}), j = 1, 2, ..., n_i\}$ for each shape $i = 1, 2, ..., N$: for example, the outlines of maple leaf images in Figure 1.6. To obtain these coordinates, a characteristic (original) point on each shape should be identified manually, this point is shown by the circle at the top of each maple leaf. Then a traverse technique is implemented. Moving counterclockwise along the image outline, we record (x, y) coordinates, so that eventually we come to the same point/circle (Gonzalez and Woods, 2002). In Figure 1.7 we plot x and y versus the point for each leaf—these are the data with which the mixed model works.

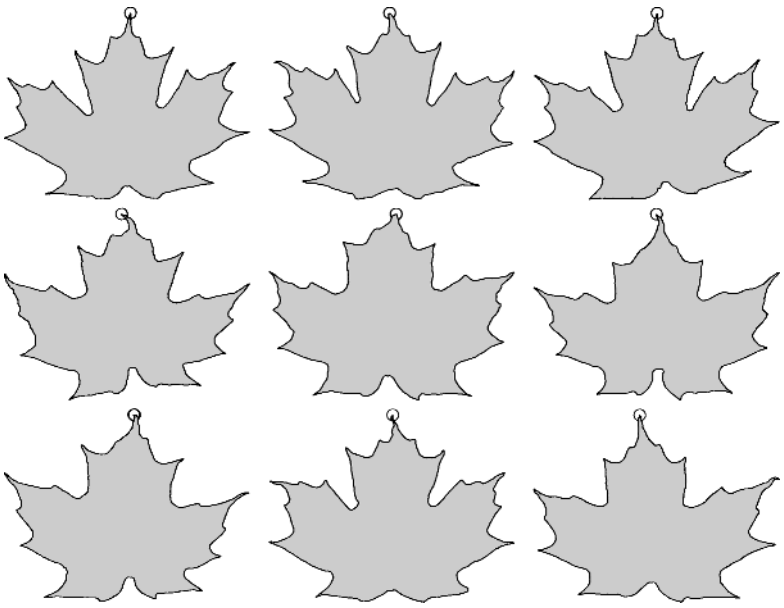

FIGURE 1.6. Nine maple leaf shapes. The circle on the top of each leaf is the starting point where the traverse starts.

An important feature of this shape quantification is that x and y are periodic functions because moving along the shape, one comes to the original point. Therefore, Fourier analysis is an adequate mathematical tool to describe x and y through a linear combination of a finite number of harmonics (see Section 11.7). Ordinarily, such analysis assumes that the Fourier coefficients $\{a_k, k = 1, 2, ..., K\}$ are fixed, where K is the number of harmonics. According to the mixed model methodology, the coefficients vary from shape to shape but stay constant within the population: namely, $a_{ik} = \alpha_k + b_{ik}$, where a_{ik} is the kth Fourier coefficient for the ith shape and α_k is the population-averaged coefficient. The RFD model for shape reduces to a LME model with appropriate formulas and algorithms.

Shape analysis is complicated by the fact that shapes may have different sizes and may be rotated arbitrarily. Fortunately, the traverse method is not affected by rotation, but the size and the specific location of the original point should be taken into account. Thus, before analyzing data in Figure 1.7, normalization and rescaling are required.

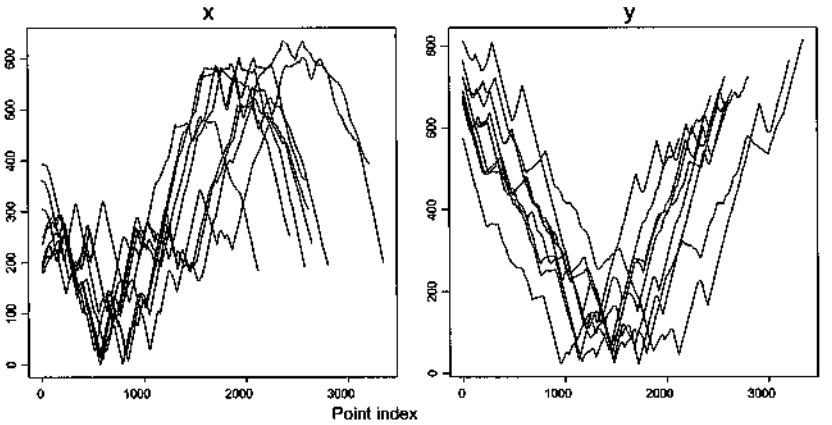

FIGURE 1.7. Quantified maple leaf shapes; x and y coordinates as a function of the traverse point for each shape. Before doing the analysis, these curves must be properly normalized and rescaled.

1.14 DNA Western blot analysis

Western blot analysis (or immunoblotting) is a popular DNA imaging analysis used for the detection of specific proteins. In this technique DNA is electrophoresed through a gel matrix to separate the individual fragments by size. The result of this procedure is a bandlike image. Two typical Western blot images, for a normal patient and a cancer patient, are shown in Figure 1.8.

Special interpretation skills are required to identify blocks and to detect the difference between two sample tissues. Besides general difficulties of interpretation and identification, the variation between samples, laboratories, and patients becomes overwhelming. Needless to say, often the DNA analysis becomes imprecise

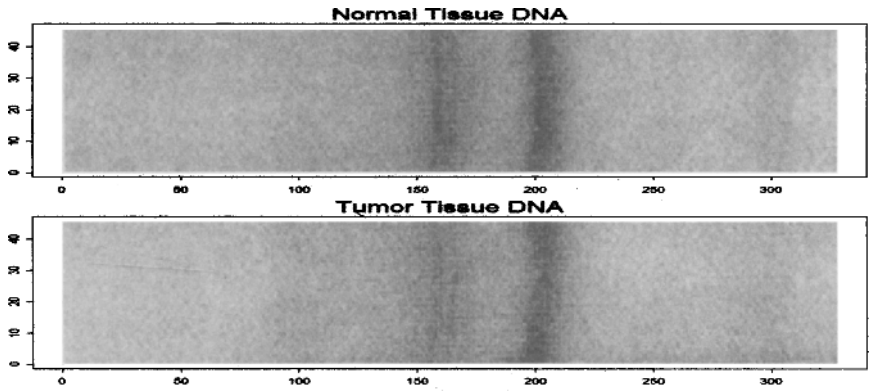

FIGURE 1.8. Typical Western blot 45×327 image for normal- and cancer-patient DNA.

and subjective. Moreover, since the human eye cannot compare hundreds of images, the analysis is reduced to just a few comparisons—biased and false results are unavoidable.

To address sample, laboratory, and patient heterogeneity, a multilevel mixed model should be applied. We quantify the two images and show the result in Figure 1.9. The result of quantification is two matrices with integer values in the range from 0 (absolute black) to 255 (absolute white). The reader may learn more about image quantification in Chapter 12. The columns are interpreted as repeated measurements, and therefore averaging is allowed. Assuming that values are normally distributed (not integers), perhaps the simplest statistical model takes the form

$$\begin{aligned} &\text{Control:} && y_{ij1} = \mu_{j1} + \varepsilon_{ij1}, \\ &\text{Patient:} && y_{ij2} = \mu_{j2} + \varepsilon_{ij2}, \end{aligned} \tag{1.71}$$

where μ_{j1} and μ_{j2} are the mean values at the jth vertical readings and $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$ are iid random variables ($k = 1, 2$) and $i = 1, ..., m = 45$ and $j = 1, ..., n = 327$. The null hypothesis is $H_0 : \mu_{1,1} = \mu_{1,2}, \mu_{2,1} = \mu_{2,2}, ..., \mu_{327,1} = \mu_{327,2}$. In this setting, this hypothesis may be tested by the paired t-test applied to average data, $y_{j1} = \sum_{i=1}^{m} y_{ij1}/m$ and $y_{j2} = \sum_{i=1}^{m} y_{ij2}/m$. Several improvements may be made to model (1.71). First, one may assume that observations along the x-axis are dependent. A parsimonious correlation structure can be described by a Toeplitz (band) matrix assuming that observations follow a stationary random process, see Section 4.3.4. Second, one can address the curvature along the y-axis using the model $y_{ijk} = \mu_{j1} + \nu_k(i - m/2) + \alpha_k(i - m/2)^2 + \varepsilon_{ijk}$. Then ν_k and α_k are nuisance curvature coefficients. Again, we are concerned with the same null hypothesis, H_0.

More important, model (1.71) can be used as a building block to test H_0 when repeated measurements are available, such as from different laboratories, tissue samples, etc. For example, if DNA analysis is available for M_1 controls and M_2 cancer patients, we introduce an additional index p so that $y_{ijpk} = \mu_{jk} + b_{pk} + \varepsilon_{ijk}$, where b_{pk} is the subject-specific random effect. Moreover, one may be interested in the dependence of DNA analysis on age, gender, or other covariates $\mathbf{x}_{pk}$, leading to a linear mixed effects model

$$y_{ijpk} = \mu_{jk} + \boldsymbol{\beta}' \mathbf{x}_{pk} + b_{pk} + \varepsilon_{ijpk}.$$

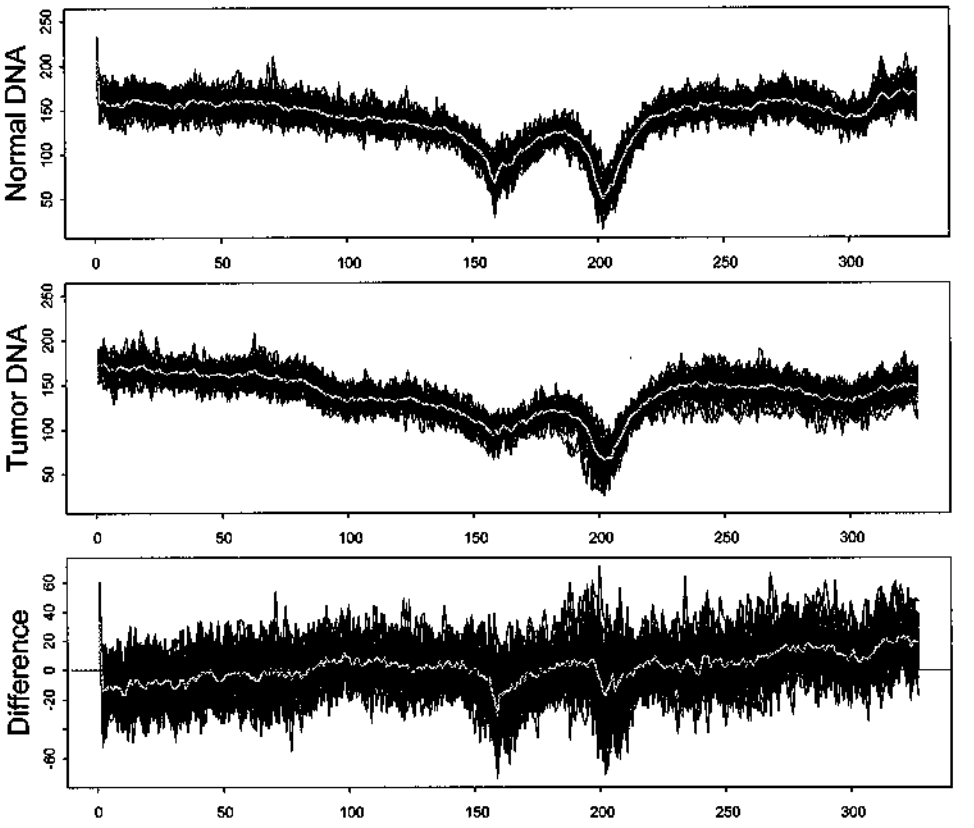

FIGURE 1.9. Quantified Western blot images with the difference. The average across vertical readings is shown by the bold line. The paired t-test produced a p-value of 0.0001.

Obviously, it takes the form (1.10) after combining observations in vectors and matrices. If the covariance matrix ε_{ijpk} is modeled via a Toeplitz matrix, we come to the LME model with linear covariance structure (see Section 4.3).

1.15 Where does the wind blow?

In this section we illustrate how a mixed model may be applied to analyze moving objects. In Figure 1.10 four images of the same sky are taken at 15-second intervals (the camera position was held fixed). From an analysis of these images, we want to determine where the wind blows, or in other words, in what direction/angle the clouds move and with what speed. First, we solve this problem assuming that the shape of the clouds does not change with time. Second, we show how to describe this problem via a nonlinear mixed model under the more realistic assumption that the moving clouds change.

A grayscale image is a $P \times Q$ matrix with integer entries from 0 (absolute black) to 255 (absolute white). Let $M_t(p, q)$ be the intensity of the image at time t at pixel (p, q), where $p = 1, 2, ..., P$ and $q = 1, 2, ..., Q$. In our example, $P = 576$ and $Q = 432$, $t = 1, 2, 3, 4 = T$. See Chapter 12 for more discussion.

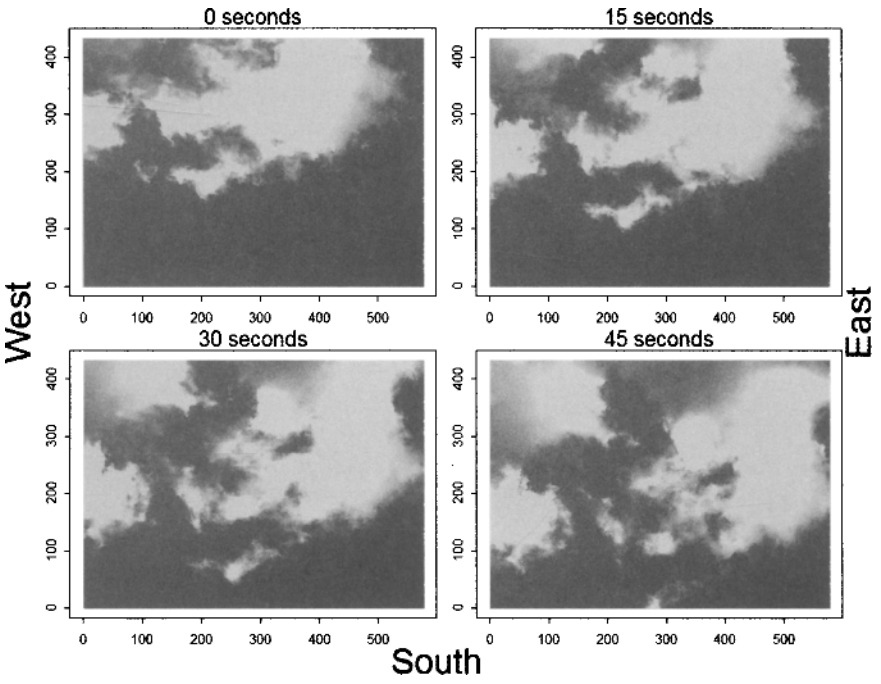

FIGURE 1.10. Pictures of the sky taken at 15-second intervals. Where does the wind blow? We apply the mixed modeling technique to answer this question.

To fix the idea, we consider the case when only two sky images, M_1 and M_2, are available. If pixel (p, q) moved to a new position $(p + \alpha, q + \beta)$, we could identify α and β from nonlinear least squares by minimizing the mean squared error

$$S(\alpha, \beta) = \frac{1}{PQ} \sum_{p=1}^{P} \sum_{q=1}^{Q} [M_1(p, q) - M_2(p + \alpha, q + \beta)]^2.$$

Although $M_1(p, q)$ and $M_2(p, q)$ are discrete functions, actually matrices, we can find the minimum of S; we refer the reader to Section 12.7.7, where a derivative-free algorithm is discussed. In image analysis, we treat elements as functions of p and q, and therefore we use the notation $M(p, q)$ rather than M_{pq}.

Next we assume that there are $t = 1, 2, ..., T$ images moving with a constant speed. Let $M(p, q)$ be the image of the moving object, which is unknown. Since after time t, pixel (p, q) on image M moved to pixel $(p + \alpha t, q + \beta t)$ on image M_t, we find α and β which minimize the MSE,

$$S(\alpha, \beta) = \sum_{t=1}^{T} \left[\frac{1}{|\mathcal{M}|} \sum_{(p,q) \in \mathcal{M}} [M_t(p + \alpha t, q + \beta t) - M(p, q)]^2 \right], \tag{1.72}$$

where $\mathcal{M}$ is the index set (p, q), so that $1 \leq p + \alpha t \leq P$ and $1 \leq q + \alpha t \leq Q$; $|\mathcal{M}|$ is the number of pair elements in $\mathcal{M}$, or in other words, the number of summation terms. We use MSE rather than a simple sum of squares to account for the number of summation terms; this technique is called affine image registration, see Section 12.7.1 for more details. From (1.72), we immediately obtain the fact that the optimal M is the average,

$$M(p, q) = \overline{M}(p, q) = \frac{1}{T\,|\mathcal{M}|} \sum_{(p,q)\in\mathcal{M}} M_t(p + \alpha t, q + \beta t), \tag{1.73}$$

so M is replaced in (1.72) with (1.73) after each iteration for α and β.

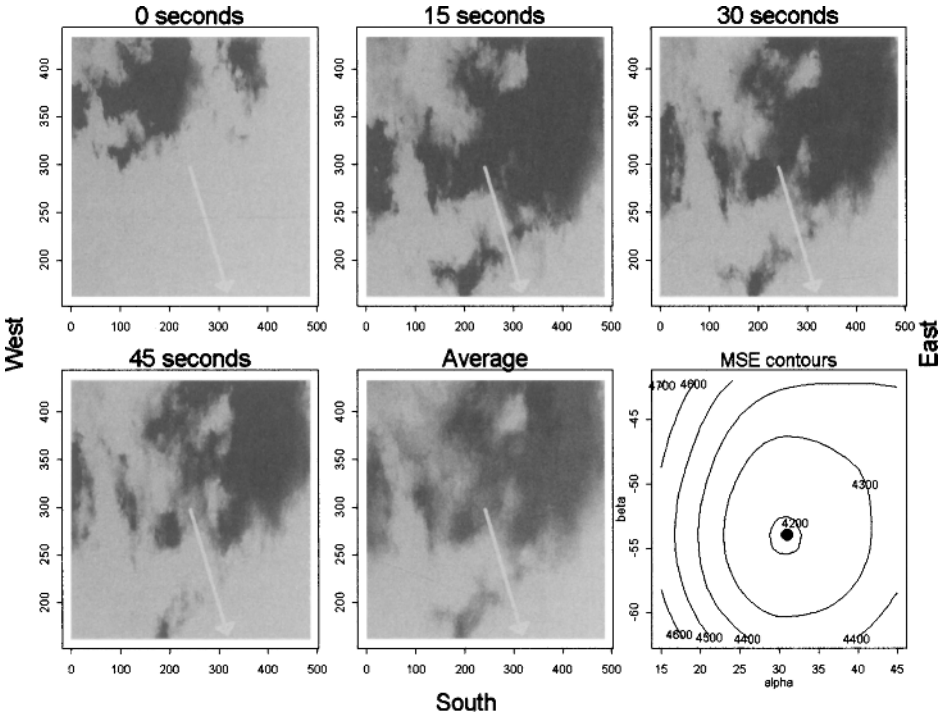

FIGURE 1.11. Reconstructed sky and wind direction indicated by an arrow. The wind blows at -60^0 with speed of 62 pixels per 15 seconds.

After a few iterations we get $\widehat{\alpha} = 31$ and $\widehat{\beta} = -54$, so the angle at which the wind blows is -60^0, indicated by an arrow on the images in Figure 1.11. After α and β are estimated, we estimate the speed as $(\widehat{\alpha}^2 + \widehat{\beta}^2)^{1/2} = 62$ pixels per 15 seconds. In Figure 1.11 we show four images $M_t(p + \widehat{\alpha}t, q + \widehat{\beta}t)$ at $t = 0, 15, 30, 45$, the average image (1.73), and the contours for the mean squared error (1.72) in coordinate system (α, β).

Now we set up a nonlinear mixed effects model (studied in Chapter 8 in a general form). In the least squares criterion (1.72), it was assumed that the moving clouds do not change, which clearly is not true. To account for change, we allow coefficients

α and β to be random, leading us to a statistical model,

$$M_t(p + \alpha(p,q)t, q + \beta(p,q)t) = M(p,q) + \varepsilon(p,q), \tag{1.74}$$

where $\alpha(p,q)$ and $\beta(p,q)$ are random variables with means α and β, or more specifically,

$$\alpha(p,q) = \alpha + b_\alpha(p,q), \quad \beta(p,q) = \beta + b_\beta(p,q). \tag{1.75}$$

In this model, $\varepsilon(p,q)$ is the iid error term with zero mean and variance σ^2, and $b_\alpha(p,q)$ and $b_\beta(p,q)$ are treated as random effects with zero mean and the 2×2 covariance matrix $\sigma^2\mathbf{D}$. The stochastic equations (1.74) and (1.75) define a nonlinear mixed effects model. When $\mathbf{D} = \mathbf{0}$, we come to an ordinary nonlinear regression model and criterion (1.72): otherwise, the population-averaged parameters α and β should be estimated using approximate methods from Chapter 8, such as those based on the Laplace approximation.

1.16 Software and books

There are several statistical packages for linear and nonlinear mixed effects model estimation. The most advanced are `proc mixed` for SAS (SAS Institute, Inc.) and library `nlme` (or `lme4`) for R (R Development Core Team, 2011). Other relevant `R` packages/libraries are `gee` and `MASS` (function `glmmPQL`); all these can be downloaded from `http://www.R-project.org`. The documentation for `R` functions is usually too succinct for immediate programming. For example, there is no explanation of how to extract the variance-covariance matrix of random effects, $\mathbf{D}_* = \sigma^2\mathbf{D}$, from `lme` or `lme4`, or how to keep these functions running in the case of a failure during simulations—we illustrate these features. However, providing details on the use of this software is beyond the scope of this book. The relevant coverage of linear and nonlinear mixed models within S-Plus is given in the book by Pinheiro and Bates (2000). For SAS users we recommend books by Verbeke and Molenberghs (2009) and Vonesh (2012), which have numerous examples.

A number of books on mixed models have been published. Below is a list arranged in order of similarity to this book:

- McCulloch, C.E. and Searle, S.R. (2001). *Generalized, Linear and Mixed Models.* New York: Wiley.
- Vonesh, E.F. and Chinchilli, V.M. (1997). *Linear and Nonlinear Models for the Analysis of Repeated Measurements.* New York: Marcel Dekker.
- Vonesh, E.F. (2012). *Generalized Linear and Nonlinear Models for Correlated Data. Theory and Applications Using SAS.* Cary, NC: SAS Institute.
- Davidian, M. and Giltinan, D.M. (1995). *Nonlinear Models for Repeated Measurement Data.* London: Chapman & Hall.
- Pan, J.X. and Fang, K.T. (2002). *Growth Curve Models and Statistical Diagnostics.* New York: Springer-Verlag.

- Davis, C.S. (2002). *Statistical Methods for the Analysis of Repeated Measurements.* New York: Springer-Verlag.
- Diggle, P., Heagerty, P., Liang, K.-Y., and Zeger, S. (2002). *Analysis of Longitudinal Data.* Oxford, UK: Oxford University Press.
- Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models.* New York: Springer-Verlag.
- Agresti, A. (2002). *Categorical Data Analysis.* New York: Wiley.
- Pinheiro, J.C. and Bates, D.M. (2000). *Mixed Effects Models in S-Plus.* New York: Springer-Verlag.
- Fitzmaurice, G.M., Laird, N.M., and Ware, J.H. (2011). *Applied Longitudinal Analysis.* Hoboken, NJ: Wiley.
- Hedeker, D. and Gibbons, R.D. (2006). *Longitudinal Data Analysis.* Hoboken: Wiley.
- Verbeke, G. and Molenberghs, G. (2009). *Linear Mixed Models for Longitudinal Data.* New York: Springer-Verlag.
- Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G. (Eds.) (2009). *Longitudinal Data Analysis.* Boca Raton, FL: CRC Press.
- Hsiao, C. (2003). *Analysis of Panel Data.* Cambridge, UK: Cambridge University Press.
- Searle, S.R., Casella G., and McCulloch, C.M. (1992). *Variance Components.* New York: Wiley.

1.17 Summary points

- Often, data have a clustered (panel or tabular) structure. Classical statistics assumes that observations are independent and identically distributed (iid). Applied to clustered data, this assumption may lead to false results. In contrast, the mixed effects model treats clustered data adequately and assumes two sources of variation, within cluster and between clusters. Two types of coefficients are distinguished in the mixed model: population-averaged and cluster (or subject)-specific. The former have the same meaning as in classical statistics, but the latter are random and are estimated as posteriori means.
- The linear mixed effects (LME) model may be viewed as a generalization of the variance component (VARCOMP) and regression analysis models. When the number of clusters is small and the number of observations per cluster is large, we treat the cluster-specific coefficients as fixed, and ordinary regression analysis with dummy variables applies, as in the ANOVA model. Such a model is called a fixed effects model. Vice versa, when the number of clusters is large but the number of observations per cluster is relatively small, a random

effects model would be more adequate—then the cluster-specific coefficients are random.

- The mixed model technique is a child of the marriage of the frequentist and Bayesian approaches. Similar to the Bayesian approach, a mixed model specifies the model in a hierarchical fashion, assuming that parameters are random. However, unlike the Bayesian approach, hyperparameters are estimated from the data as in the frequentist approach. As in the Bayesian approach, one has to make a decision as to the prior distribution, but that distribution may contain unknown parameters that are estimated from the data, as in the frequentist approach.

- Penalized likelihood is frequently used to cope with parameter multidimensionality. We show that the penalized likelihood may be derived from a mixed model as an approximation of the marginal likelihood after applying the Laplace approximation. Moreover, the penalty coefficient, often derived from a heuristic procedure, is estimated by maximum likelihood as an ordinary parameter.

- The Akaike information criterion (AIC) is used to compare statistical models and to choose the most informative. The AIC has the form of a penalized log-likelihood with the penalty equal to the dimension of the parameter vector. A drawback of the AIC is that it does not penalize ill-posed statistical problems, as in the case of multicollinearity among explanatory variables in linear regression. We develop a healthy AIC that copes with ill-posedness as well because the penalty term involves the average length of the parameter vector. Consequently, among models with the same log-likelihood value and number of parameters, HAIC will choose the model with the shortest parameter vector length.

- Since the mixed model naturally leads to penalized likelihood, it can be applied to penalized smoothing and polynomial fitting. Importantly, the difficult problem of penalty coefficient selection is solved using the mixed model technique by estimating this coefficient from the data. In penalized smoothing, we restrain the parameters through the bending energy, in polynomial fitting through the second derivative.

- The mixed model copes with parameter multidimensionality. For example, if a statistical model contains a large number of parameters, one may assume that *a priori* parameters have zero mean and unknown variance. Estimating this variance from the data, after Laplace approximation we come to the penalized log-likelihood. We illustrate this approach with a dietary problem in conjunction with logistic regression where the number of food items consumed may be large.

- Tikhonov regularization aims to replace an ill-posed problem with a well-posed problem by adding a quadratic penalty term. However, selection of the penalty coefficient is a problem. Although Tikhonov regularization receives a nice statistical interpretation in the Bayesian framework, the problem of the

penalty coefficient remains. A nonlinear mixed model estimates the penalty coefficient from the data along with the parameter of interest, $\boldsymbol{\theta}$.

- Computerized tomography (CT) reconstructs an image from projections and belongs to the family of linear image reconstruction. Since the number of image pixels is close to the number of observations, CT leads to an ill-posed problem. To obtain a well-posed problem, *a priori* assumptions on the reconstructed image should be taken into account. We show that a mixed model may accommodate various prior assumptions without complete specification of the prior distribution.

- Positron emission tomography (PET) uses the Poisson regression model for image reconstruction and the EM algorithm for likelihood maximization. Little statistical hypothesis testing has been reported, perhaps due to the fact that the EM algorithm does not produce the covariance image matrix. The Fisher scoring or Unit step algorithms are much faster and allow computation of the covariance matrix needed for various hypothesis testing as if two images in the area of interest are the same. To cope with ill-posedness, Bayesian methods and methods of penalized likelihood have been widely applied. The generalized linear mixed model (GLMM), studied extensively in Chapter 7, also follows the line of the Bayesian approach, but enables estimation of the regularization parameter from PET data. A multilevel GLMM model can combine repeated PET measurements and process them simultaneously, increasing statistical power substantially.

- The mixed model is well suited for the analysis of biological data when, on the one hand, observations are of the same biological category (maple leaf), but on the other hand, individuals differ. Consequently, there are two sources of variation: variation between individuals (intersubject variance) and variation within an individual (intrasubject variance). The common biological type corresponds to population-averaged parameters and individuality corresponds to subject-specific parameters. Shape is the simplest biological characteristic. Its analysis is complicated by the fact that shapes may be rotated and translated arbitrarily. Several mixed models for shape analysis are discussed in Chapter 11.

- Image science enables us to derive a large data set of repeated structure; thus, application of the repeated-measurements model, such as a mixed model, seems natural. Until now, image comparison in medicine has been subjective and based on "eyeball" evaluation of a few images (often, just a couple). Statistical thinking in image analysis is generally poor. For example, a proper DNA Western blot image evaluation should be based on several tissue samples analyzed by a multilevel mixed model.

- Mixed models can be applied for statistical image analysis, particularly to analyze an ensemble of images (see Chapter 12). As with shape analysis, two sources of variation are considered, the within-image and between-images variation. Since an image may be described as a large matrix, we may treat the element as a nonlinear function of the index and apply the nonlinear mixed

effects model of Chapter 6. The mixed model can also be applied to study the motion of fuzzy objects such as clouds.