TOPICS IN IMAGE RECONSTRUCTION
FOR EMISSION TOMOGRAPHY

A thesis presented
by
Andrew Eric Gelman
to
The Department of Statistics
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Statistics

Harvard University
Cambridge, Massachusetts
May, 1990

Thesis Advisor: Prof. Donald Rubin

Abstract

This thesis reviews and develops image reconstruction for positron emission tomography (PET) with some exploration of the related issues of statistical design and analysis of medical experiments that use PET. Following overviews in the first two chapters on image analysis in general and PET in particular, the discussion of image reconstruction focuses on estimation of hierarchical Bayesian models applicable to raw data. Chapters 2 and 3 derive reasonably fast algorithms for hierarchical maximum-likelihood and Bayesian image estimates, and Chapters 4 and 5 integrate some regularization and smoothing estimates into the Bayesian framework.

In addition, this thesis covers several more specific statistical issues associated with three key difficulties of PET: the data are indirectly observed, the spatial resolution is limited, and the reconstructed image is constrained to be nonnegative. These difficulties explain some of the advantages and drawbacks of linear, maximum-likelihood, and Bayesian image estimates. Finally, the ideas presented here are focused by the analysis of a specific medical imaging study performed at Massachusetts General Hospital and described in Chapter 6. Chapter 7 adapts the models and techniques developed in earlier chapters to the applied problem by deriving expressions for bias and variance of local averages of various PET estimates. The final chapter discusses the difficulties of handling real PET data and concludes with an example of how knowledge of biological spatial structures affects the design and analysis of medical imaging experiments.

Contents

Introduction

Basics

Chapter 1. Selective Overview of Image Analysis.

Some ubiquitous concepts and jargon.

Chapter 2. Introduction to Emission Tomography Reconstruction.

Enough details of positron emission tomography (PET) to understand the mathematics.

Fourier-based reconstruction from projections.

Derivation and implementation of maximum likelihood reconstruction.

Chapter 3. Overview of Bayesian Image Reconstruction.

An example: summary of a method of Geman and McClure.

Comments on Bayesian image models and methods of Bayesian reconstruction.

Implementation of Bayesian methods for PET.

Theoretical Developments

Chapter 4. Regularization and Bayesian estimation for images.

Similarities and differences between these approaches.

Counting degrees of freedom in indirect data.

Chapter 5. Linear Smoothing and Conditional Autoregressive Models.

Gaussian conditional autoregressive (CAR) processes.

Linear smoothing in Bayesian estimation with a nonstationary CAR prior distribution.

Application

Chapter 6. Exploratory Analysis of a Tomography Study.

Background of the study done at Massachusetts General Hospital (MGH).

Analysis of effects in brain regions.

Chapter 7. Bias and Variance of Estimates.

Filtered back-projection image estimate.

Maximum likelihood estimate of region averages.

Bayesian estimates.

Chapter 8. Image Reconstruction and Experimental Design.

Applying maximum likelihood and Bayesian estimates to real PET data.

Probability models of spatial structure.

Experimental design for medical imaging.

Conclusion.

Acknowledgements

Thanks go to my advisor, Donald Rubin, who suggested the topic of this thesis and, through invaluable discussions, assisted in every aspect of this work. Nathaniel Alpert and his colleagues at Massachusetts General Hospital were very helpful in answering questions and supplying tomography data. The U.S. National Science Foundation provided financial support through a graduate fellowship and grants DMS-8504332 and SES-8804533. I also thank Herman Chernoff for his detailed comments, and many colleagues and friends for suggestions and assistance.

Notation Appearing in More than One Chapter

Symbol	Meaning	Section where Defined
$g = (g_i)$	Vector of gray levels of pixels	1.1
N	Number of pixels in an image	1.1
$Y = (Y_j)$	Vector of data	2.1.1
$X = (X_j)$	$X = E(Y\|g)$	2.1.1
n	Length of data vector (number of tubes)	2.1.1
$R = (R_j)$	Vector of random coincidence rates	2.1.1
$\alpha = \mathrm{diag}(\alpha_j)$	Diagonal matrix of attenuation corrections for tubes	2.1.1
$A = (a_{ij})$	Matrix of angle factors	2.3
P	Forward projection operator	2.3
B	Blurring operator	2.3
F	Filtering operator for filtered back-projection	2.3
$\bar{\alpha} = (\bar{\alpha}_i)$	Vector of average attenuation corrections for pixels	2.3.2
$h = (h_k)$	Vector of regional parameters of an image	2.3.4
W	Matrix transforming h to a regionally-constant image	2.3.4
Ω_k	Region k of an image	2.3.4
$Z = (z_{ij})$	Matrix of unobserved counts in pixels and tubes	2.3.2
$Z = (z_i)$	$z_i = \sum_j z_{ij}$	3.3.1
θ	Hyperparameters of an image model	3.3.1

Introduction

This thesis discusses image reconstruction for positron emission tomography (PET), toward the general goal of improving the design and analysis of medical imaging experiments. To fix ideas and focus attention on realistic issues, this work is presented in the context of one study of the physiological effects within the brain of different eye-tracking tasks. PET is a useful tool in this study because it measures concentrations of selected chemicals within a functioning organ, the brain, without requiring any physical intrusion. The output of this noninvasive procedure is a three-dimensional map (or *image*) of metabolism within a brain and is obtained by a mathematical transfomation of the observed indirect data. Current methods of PET image reconstruction are somewhat unsatisfactory, for at least two reasons. First, storing and handling three-dimensional information about the brain is difficult, and three-dimensional images are generally stored inconveniently as arrays of two-dimensional slices. Second, statistical estimation and the interpretation of estimates is difficult for a complicated image with many parameters estimated from a large array of indirect, noisy data. This thesis attacks the second problem, reviewing and providing careful implementations of some PET reconstructions and estimates of their errors, with an eye toward improving estimates by using spatial knowledge about the brain. Ultimately this work should be unified by probabilistically incorporating knowledge of the brain's three-dimensional structure into PET images, with enough understanding of the uncertainty of the reconstructions to apply the principles of experimental design to future PET studies.

Statistical issues

A typical PET experiment gives each of several patients a series of cognitive tasks and during each task produces a three-dimensional scan of the concentration of a radioactive isotope within the brain. The experiment's goal is to understand the effect of the tasks on some aspect of brain physiology. (For example, the experiment described in Chapter 6 of this thesis studies the effect of eye tracking on blood flow in the brain.) A series of statistical issues, starting with image reconstruction, arise in such a medical imaging experiment. In biological imaging, one must connect image estimation and analysis with prior knowledge of the spatial structure under study; this issue influences many phases of an experiment, starting with the design. For a PET study, one must pick an isotope activated by a process that is differentially activated by the cognitive tasks of interest. The brain comprises several dozen anatomical and physiological regions, and the physical PET scan can be configured to observe a few selected regions of interest with especially high accuracy. At the next step, the spatial resolution of the reconstructed image can be improved by knowledge of the locations of regions in the brain within which isotope concentration should be roughly uniform. Finally, the results of a medical imaging experiment generally are expressed in terms of previously mapped anatomical regions; understanding of the functions of these regions should allow a vast set of image data to be usefully reduced.

A key difficulty arises from the limited spatial resolution of PET scanners: the data do not in general allow us to reconstruct fine detail, and the standard reconstruction method now in use actually estimates a *blurred* map of isotope concentration. In fact, if one refuses to make any assumptions, then the limited resolution allows no choice but to reconstruct a blurred image; however, if a set of assumptions can be expressed probabilistically, one can estimate the unblurred image directly by Bayesian methods. This step of course leads to two concerns: first, how sensitive is a Bayesian image estimate to its assumptions, and second, can the PET data be summarized by a set of plausible images rather than a single image? These concerns correspond to the statistical issues of bias and variance of estimation.

The ultimate goal of a statistical study of PET is to combine the tasks of experimental design, image reconstruction, and data reduction, in the following ways. The same biological understanding that suggests appropriate summaries of an image (such as averages within known anatomical regions) can be formalized into assumptions that lead to more accurate image estimates. Expressions for bias and variance of estimated images and image summaries should lead to answers to design questions such as how many patients and tasks are recommended to estimate an effect of a given size. Finally, for a particular problem, one can choose an estimation method that leads to a cost-effective design to estimate the image summaries of substantive interest.

Outline of this work

The first three chapters of this thesis provide the basics of PET reconstruction—a review of recent literature in statistical image estimation from projection data, and some contributions not in the previous literature. Chapter 1 gives a brief overview of image analysis, to put the rest of the thesis in perspective. Chapter 2 introduces the physical set-up of PET and the relevant mathematics. Filtered back-projection—a Fourier-based method that is currently used in practical PET reconstructions—is derived following descriptions in standard references but with additional care devoted to the realistic problem of finite data. Likelihood-based methods of indirect image reconstruction and their difficulties are then reviewed and explored. The derivation presented here of the iterative EM algorithm for maximum likelihood image estimation follows previous sources, with two original extensions. First, the algorithm in this thesis is the first likelihood-based method to account for "random coincidences" and "attenuation"—two features of real PET data ignored in the literature on EM for PET. Second, the reconstruction algorithm is implemented in terms of high-level "projection" operators; this form is conceptually clear as well as easy to implement on computers that currently use the filtered back-projection algorithm.

Chapter 3 of this thesis discusses Bayesian estimation for PET, starting with a review of the recently popular approach to image estimation based on Gibbs probability models and the stochastic relaxation algorithm. This method is then used for PET by applying a homogeneous Bayesian model to the whole image, or separately to each region of the image. Once again some recent literature is followed, but with innovations: the implementation presented here accounts for the form of real PET data and is expressed in terms of the projection operators. In addition, the suggestion is made to average over rather than estimate hyperparameters in a Bayesian model of a single image; the averaging can be done conveniently—both conceptually and computationally—in both the EM and stochastic relaxation algorithms.

Chapters 4 and 5 complete the synthesis of recent PET reconstruction methods from a Bayesian perspective. In discussing regularization, or penalized likelihood estimation, Chapter 4 describes two methods in the literature: maximum entropy and the aborted EM algorithm. Depending on details of the algorithms, each can be interpreted as fully or partly Bayesian; we discuss the validity of various confidence statements implied by these methods and how they approximate fully Bayesian statements. In particular, a replacement for the aborted EM algorithm is proposed that is conceptually more satisfying and should be easier to implement. Chapter 5 reviews conditional autoregressive (CAR) spatial models (a subclass of the Gibbs models of Chapter 3) and their application to image estimation. An original contribution here is to show that linear smoothing yields maximum posterior probability estimates under Gaussian CAR models, which are again applied to the indirect PET problem through the EM and stochastic relaxation algorithms.

The final three chapters apply image reconstruction methods to a specific medical experiment. Chapter 6 describes a recent study at Massachusetts General Hospital of ten three-dimensional PET images, each summarized by average blood flow in a set of anatomical brain regions. An exploratory analysis yields some substantive and statistical conclusions, and shows the limitations of the design of the study and of the filtered back-projection method used to reconstruct the images. Chapter 7 derives the bias and variance of estimates of regional averages of images for three reconstruction methods: filtered back-projection, maximum likelihood estimation of regional averages, and general Bayes estimation. The bias and variance expressions, based on a weighted least-squares approximation, are original, and for practical use are described using computable algorithms as well as formulas in matrix algebra.

The maximum-likelihood and Bayesian estimates are based on assumptions that cannot always be tested by PET data; because of this problem the bias of these estimators must be studied when their assumptions are wrong. For the medical study being considered, the relevant assumptions not testable by PET involve the exact locations of anatomical brain regions in different patients, and the pattern of isotope concentration within the regions. Chapter 8 explores how assumptions about brain structures may be encoded probabilistically and then subsumed into the Bayesian estimation and error analysis. This final chapter concludes with an outline of a method, as yet untried, to connect image estimation to experimental design, using knowledge of the uncertainty in images derived from PET data; design suggestions should allow the investigator to choose a method of image reconstruction to suit a particular problem.

This thesis refers constantly to the PET problem and its details. Grounding the theoretical work in a practical problem is essential in a problem as complex as image reconstruction; however, many of the techniques reviewed and developed here should be useful in a variety of problems in spatial statistics.

Chapter 1

Selective Overview of Image Analysis

We define an *image* as a quantity defined on a two- or three-dimensional space. Before getting to specifics of estimation of medical images, we will describe the field of digital *image processing*: computer manipulation of image data. Image processing can be said to comprise four tasks: enhancement, restoration, coding, and understanding (Lim, 1988).

Altering an image for human viewers, either for visual clarity or to emphasize some special feature, is called *image enhancement.*

Example 1.1: Adaptive histogram equalization

Consider a monochrome image that ranges in brightness from black to white. Assume, however, that the image is composed of several regions, each of fairly uniform intensity. The features within any given region will be easier to see if the brightness scale in that region is expanded to cover the whole range of intensity. That is, the darkest areas in a region become black, the brightest areas become white, and intermediate intensities are scaled correspondingly. This procedure is *adaptive* because its transformation adapts to the differences in different regions of an image. *Histogram equalization* refers to spreading out the frequencies of various intensity levels (within a region) to cover the entire range from black to white.

The task of *image reconstruction* (or *restoration*) is to estimate an image from indirect (or blurred) data and is an example of an *inverse problem* (Meehan, 1990). This thesis primarily discusses reconstruction, specifically in the medical imaging context. Here the image is the blood flow in a bodily organ, as a function of two or three spatial coordinates. The indirect data are collected by electronic instruments outside the organ being studied.

Image coding means compressing a discrete image signal so it can be more easily stored or transmitted. This is important in applications such as television that involve a steady stream of detailed images; however, we will not discuss this topic here.

Finally, *image understanding* is the task of symbolically summarizing an image. This is needed when image data are required, not to make a cleaner or better image, but rather for some other purpose. For example, a medical scan may be performed with the goal of diagnosing an illness. This may be done using an estimated image, but the image understanding task is here one of discriminating between sick and healthy patients, and can be done statistically.

A companion task of image processing is the design of the data recording mechanism, so as to best estimate (or summarize) the image. This topic is rarely discussed in the literature and will not be discussed here.

Before narrowing our focus, we will briefly mention two statistical topics that are similar to image analysis. *Spatial statistics* in general studies variables that are indexed by two- or three-dimensional location (Ripley, 1981). Such a data structure may differ from an "image" by being measured, and perhaps defined, on a fixed discrete set of points. (An image, in contrast, is defined on continuous space, even if it may be discretized for computation or viewing.) An example of spatial statistics is the analysis of spatial autocorrelations among yields of a square array of plots.

A *shape* is another spatial data structure that is not exactly an "image" (Bookstein, 1986). A shape is typically estimated and summarized by the relative locations of a set of "landmark" vertices. A researcher may wish to summarize a series of shapes and track their changes over time. Typical applications in biology and medicine are the shapes of leaves and skulls; the tree-dimensional pattern of anatomical regions in the brain is a shape structure that is important for tomography.

1.1 Some topics in image reconstruction

In this section, we use a succession of brief examples to introduce some important and ubiquitous concepts, models, and methods.

Example 1.2: Gray levels and pixels

Consider a monochrome image of intensities defined on a square region. For practical purposes the continuous range of brightness may be digitized into 256 *gray levels*, whose extremes of 0 and 255 represent black and white, respectively. The image can be further discretized by dividing the spatial region into a grid of N picture elements, or *pixels*, and summarizing the image by "pixel intensities," each of which is the average gray level (rounded to the nearest integer) within a pixel. The original continuous image is now coded as a vector $g = (g_1, \ldots, g_N)$ of discrete gray levels.

An assumption of continuity or local smoothness of the image can be expressed as follows: neighboring pixels are likely to have similar intensities. Two pixels are *neighbors* if they are close enough for this purpose; the set of all neighbors of a pixel is its *neighborhood.* A maximal set of pixels, all of whose pairs are neighbors, is called a *clique.* Clearly, the exact labeling of neighbors depends on how fine the pixel grid is, and also on substantive understanding of the continuous image. If the two-dimensional image has been coded as a vector of length N, the neighborhood structure may be coded as an $N \times N$ matrix.

Example 1.3: Image of categories

In this type of image, each pixel is assigned one of l labels. The number of labels l is fairly small ($l = 2$ for a *binary image*) and, unlike the previous example, the categories are not an approximation of a continuous scale. The categories could, for example, be different kinds of terrain in a satellite photo. In another example, zeroes and ones in a binary image could stand for black and white, respectively, in a display of a monochrome image, such as would appear in a newspaper. (The spatial grid of pixels here, however, may very well be just a discretization of a continuous space.) An estimate of a categorical image may be judged by the proportion of pixels whose estimated labels differ from the truth; this discrepancy is called the *error rate.* Many writers have commented that this simple measure can be misleading; Ripley (1988, p. 101) gives an example of this.

Example 1.4: Smoothing

Consider an unknown image g of gray levels, to be estimated from a "data" image Y that is a pixel-by-pixel sum of truth plus *white noise.* (That is, the noise is composed of independent Gaussian random variables in each pixel with zero mean and common variance.) Assume that the true image, unlike the noise, is "smooth"—it varies slowly over space. One could consider two approaches to averaging and presumably cancelling some of the noise here: using an algorithm to *smooth* the data image or modeling the true, smooth image. (Of course, smoothing algorithms may be useful in estimating an image model.)

An example of a smoothing algorithm is a moving average applied to the data image. This assigns, to each pixel, a weighted average of the gray levels of all neighboring pixels in the data image. This of course requires a definition of "neighborhood"; it would typically be an $m \times m$ square of pixels centered at the target pixel, for some specified odd integer m. This linear *filter* will necessarily reduce the local differences between pixel gray levels, while barely affecting differences between gray levels of pixels that are far apart. This procedure smoothes any sharp features in the data image, whether the features come from the noise or the true image. A moving average smoother is not necessarily a good procedure for estimating a moving average model of the true image.

In the case of additive independent Gaussian noise, one can consider an image to be parameterized by the intensities at its pixels. Then the maximum likelihood estimate of the true gray level of a pixel is the data gray level at that pixel; thus, the data image is a maximum likelihood estimate of the truth. A less noisy estimate of a smooth true image may be obtained by balancing the goals of smoothness and fitting the data; this is called *regularization* and will be discussed in more detail in Chapter 4. (Mitter et al. (1987) discuss this topic in more detail; here we refer to regularization in general, not the specific form presented in that paper.)

Example 1.5: Maximum penalized likelihood

Continuing Example 1.4, label the unknown true image g; we must create an estimate $\hat{g}$ based on the data image Y. We wish to balance the log-likelihood $LL(\hat{g}|Y)$ with some roughness penalty $f(\hat{g})$. The maximum penalized likelihood estimate is the $\hat{g}$ that maximizes:

$$LL(\hat{g}|Y) - \lambda f(\hat{g}),$$

where λ is a fixed constant.

This estimate can also be considered the mode of a posterior distribution; in this case λ is a parameter in a prior distribution for g. If the prior distribution for true image (and thus the data image) is stationary, then the smoothness parameter λ can itself be estimated from the implicit replication in the data (Ripley, 1988, p. 6).

Example 1.6: Entropy

Entropy is a measure of uniformity of an image. The entropy, or negative of the Kullback-Leibler information, of a vector $g = (g_1, \ldots, g_N)$ of gray levels, is:

$$S(g) = -\sum_i g_i \log(g_i/m_i),$$

where $m = (m_1, \ldots, m_N)$, a predefined measure on $\{1, \ldots, N\}$, is usually assumed constant. The entropy is maximized by the uniform image $g = m$. A regularization that maximizes entropy subject to a likelihood constraint (or, equivalently, an entropy-penalized maximum likelihood estimate) can reliably produce smooth image estimates (Gelman, 1989). Entropy cannot strictly be called a measure of spatial smoothness, however, because it ignores the locations of the pixels, being independent of a permutation of pixel labels.

Example 1.7: Inhomogeneous image

Let the true image be composed of several disjoint spatial regions, each of which is considered a realization of an independent stationary process. (For instance, each region in an agricultural satellite photograph could be a field of a single crop.) This situation presents problems for the smoothing approaches described above, for two reasons. First, the appropriate amount of smoothing will depend on the region of the image; and second, the smoothing process itself may blur the boundaries between regions. When the boundaries are known, as is sometimes the case in medical imaging, the estimation problem may be subdivided, with the advantage that noise estimates may still be pooled across regions. When the locations of the boundaries are not known, their estimation is a problem of *edge detection*, in some applications, a task more important than estimation within the regions.

Let us step back now and consider the statistical problem of parameter estimation, applied to images. Assume that the data Y are n-dimensional (that is, the data can be written as a list of n numbers), and are used to estimate an image g of N dimensions (that is, N pixels), considered an approximation to the infinite-dimensional continuous true image. Standard asymptotic theory discusses estimation when $N < n$, both dimensions are fixed, and the variance of the estimate of g approaches zero. The *method of sieves* discusses the case when $n \to \infty$ and N is allowed to increase too, but at a slower rate (Geman, 1982). Often, however, one desires an image estimate of higher resolution than can be provided by the data; that is, the discretization is so fine that $N > n$. In this case, some assumptions must be made to tie the N parameters together. An assumption of smoothness can lead to a regularization estimate. Alternatively, one could put a probability model on the true image, as illustrated in the following example.

Example 1.8: Ising model

Consider a binary image $g = (g_1, \ldots, g_N)$, in which $g_i = +1$ or -1. The *Ising model* (described in many sources, including Ripley, 1988, p. 80) is a probability distribution on such images:

$$P(g) \propto \prod_{i,j} \exp(\beta g_i g_j),$$

where the product is taken over all pairs i, j of nearest neighbors in the square lattice. The parameter β must be specified; positive values favor equal labels on neighboring pixels, so that images with large patches of 1's and -1's have high probability. Conversely, negative values of β favor opposite labels on neighboring pixels, yielding checkerboard-type images with high probability. Large absolute values of β induce long-range dependence in the pixel labels; this may be good or bad from a modeling perspective (Besag, 1986).

The Ising model is an example of a *Gibbs distribution* on the pixel values. In general, a Gibbs distribution is identified by a joint probability density that factors by cliques. In the Ising model, the cliques are pairs of orthogonally (but not diagonally) adjacent pixels. Independence of pixels is a degenerate example of a Gibbs distribution in which each pixel is a clique. At the other extreme, any arbitrary distribution has the Gibbs form, if we allow the entire set of pixels to be a single clique.

The Ising model, like all Gibbs distributions, is a *Markov field,* as identified by the following conditional independence property: the probability density of a pixel value, given the rest of the image, depends only on its neighbors (Besag, 1974). We will discuss these Markov models in more detail in Chapter 3.

With a (prior) probability model of the true image, and a probabilistic relation between the truth and the data, the Bayesian approach yields a posterior probability distribution for the true image. Typically the unnormalized posterior density can be evaluated for any possible image, but the entire N-dimensional distribution has no simple form, and generally cannot even be directly sampled from. Algorithms have been developed for indirectly sampling from Gibbs distributions; *stochastic relaxation* and the related *Metropolis algorithm* allow a random walk through a distribution of images (Geman and McClure, 1985; Metropolis et al., 1953). One can sample from a desired distribution by running one of these algorithms for a long time and stopping. *Simulated annealing* is a modification of stochastic relaxation that finds the mode of a distribution by randomly walking through distributions that are more and more concentrated about the posterior mode (Geman and Geman, 1984). The practical utility of this algorithm has been questioned, because: it is computationally slow, the distribution may not be well summarized by its mode, and one might not fully believe in an N-dimensional prior distribution chosen partly for its simple form (Haslett, 1985; Besag, 1986).

1.2 Medical imaging

Medical imaging is relevant to the rest of this thesis; it also illustrates the specialized methods for image reconstruction that can arise from a specific problem. Medical imaging refers to the use of external measurements to estimate anatomy or physiology inside a bodily organ (typically, the brain). Three common forms of medical imaging are *magnetic* *resonance imaging* (MRI or NMR), *computerized axial tomography* (CT or CAT, also called *transmission tomography*), and *positron emission tomography* (PET). MRI is a very accurate procedure that uses magnetic oscillation to estimate the locations of different chemical elements in the brain and thus infer the anatomical structure. CT transmits x-rays through the head in order to estimate the mass density inside. PET, which is our focus in this thesis, provokes and then records radioactive emissions inside the brain to estimate blood flow. Most of the methods here are useful, in some form, for all of these problems; PET is our running example.

All of these methods create indirect, noisy data. The indirect aspect—that data are not recorded in the same space as the image—means that special techniques are needed for reconstruction, such as maximum likelihood image estimates using missing data (Shepp and Vardi, 1982). A different estimation approach for CT and PET estimation uses the fact that, in each of these cases, the data are noisy versions of linear *projections* of the true image. *Filtered back-projection*—a linear method based on Fourier analysis—creates exact band-limited reconstruction from noise-free data and provides a fast approximate method from real data. The next chapter describes these methods in detail.

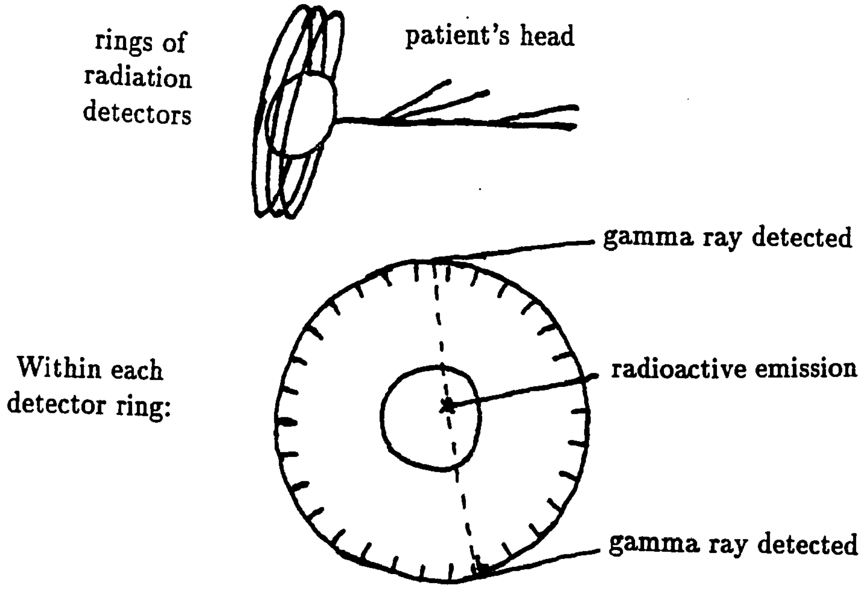
Chapter 2

Introduction to Emission Tomography Reconstruction

This chapter outlines the mathematics and some of the physics of positron emission tomography (PET). A discussion of some simple methods of reconstruction will point up some of the statistical difficulties of this problem. Section 2.2 briefly discusses Fourier-based reconstruction methods, including *filtered back-projection*, which is used in current practice. Section 2.3 introduces likelihood-based methods; maximum likelihood and Bayesian PET reconstructions based on the EM algorithm have appeared in many journals. Shepp and Vardi (1982) and others suggested the algorithm presented as steps 1–4 in Section 2.3.2 of this paper, but ignored the correction for random coincidences. Carson (1986) suggested applying EM to regional parameters (Section 2.3.3 below), and Levitan and Herman (1987) and others applied EM to maximum posterior estimation for PET (Section 3.3.2 below); all of these papers discussed the simplified problem with no random coincidences. These sources apparently did not apply the EM steps using the implementation of steps *i*–*v* in Section 2.3.2 and may have ignored some of the scattering and attenuation effects.

In general, image reconstruction comprises three tasks. First, one must model the noise process; that is, the relation between the true image and the data. Second, one must decide which features of the true image are to be reproduced in a picture. Third, one must use some algorithm to produce a reconstruction. This medical application falls into the category of *image reconstruction from projections*, because the incomplete observed

Figure 2.1: Positron emission tomography



data are random functions of linear projections, from various angles in space, of the true image. Furthermore, these random functions are known Poisson random variables, and so the first task of estimating the noise process is complete.

The physical procedure of PET is as follows: a person inhales or is injected with a radioactive tracer, which the blood carries to the organ of interest, typically the brain. Here the radioactive isotope disintegrates, emitting positrons. These in turn hit nearby electrons, and the resulting annihilation releases two gamma ray photons traveling in nearly opposite directions. Finally, these photons hit a ring of detectors surrounding the organ, and the hits along this ring are recorded. Figure 2.1 illustrates the PET procedure.

Each emission within the brain yields a pair of photons, either of which may be absorbed on its way to the detector ring. If this does not happen, and the rays of emission intersect the detector ring, then a pair of (effectively) simultaneous counts will be recorded. This *coincidence* localizes the emission to a cylindrical area between the detectors. A series of emissions are detected, and are assumed to come from independent locations in the organ. The coincidence detector pairs are used to estimate the density function of the emission locations.

2.1 Physical and mathematical summary

Here we introduce a particular PET experiment at Massachusetts General Hospital (MGH), as described by Elliott (1989); further details appear in Chapter 6. The experiment measures metabolic activity in the brain, through the injection of radioactive carbon dioxide with a half-life of 2 minutes. (The measurements on a single person take over two hours; the radioactive isotope is continuously injected.) Typically, measurements are taken before and after a brain stimulus.

Three stacked rings, 2.8 cm apart, each with 96 detectors, encircle the head; each detector is about 1 cm × 1 cm in cross-section. To improve spatial sampling of data, the whole ring wobbles: the detector ring's orientation remains unchanged while its position moves in a circle of diameter about 1 cm every second. The resulting continuous-time data are summarized by grouping data into 13 wobble locations for each projection. We will label the cylindrical region defined by a detector pair and a wobble position as a *tube*; tomography data are observed at tubes. The head takes up only part of the area within the dector ring, and as a result, coincidences are only recorded in 18 detector pairs (yielding 18 × 13 tubes) for each projection angle. In addition, the whole experiment is done three times with different vertical positions of the detector rings.

Sometimes photons from two different emissions will simultaneously hit two different detectors; this event is called a *random coincidence* (Hoffman et al., 1981). The rate of such events is estimated from the total absorption rates at the separate detectors. Not all emissions in the direction of the detectors are recorded; this effect is called *attenuation* (Huang et al., 1979). Typically, about 5,000,000 coincidence counts are recorded, of which about a third are "random coincidences."

The continuous density function of blood flow is a monotonic non-linear function of the concentration of radioactive tracer and is approximated over a 128 × 128 grid, on each of nine vertical levels.

A computer program creates two-dimensional pictures using the coincidence rates, adjusted for random coincidences and attenuation factors. The filtered back-projection algorithm (described in Section 2.2 below) is typically used in practice.

2.1.1 Mathematical description

Assume an emission density function $g(x,y,z)$ (in units of emissions/volume), that is a function of the Cartesian coordinates x, y, z, and tubes indexed by $i = 1, 2, \ldots, n$. (For convenience we use the same notation g to describe the following three quantities: the continuous three-dimensional image $g(x,y,z)$, the continuous two-dimensional image $g(x,y)$, and the discretized, usually two-dimensional, image coded as a vector $g = (g_i)$.) Physical theory and experiment allow us to express the observed emissions as independent random variables:

$$Y_j \sim \text{Poisson}(X_j),$$

where the Poisson means X_j are integrals over the brain volume:

$$\begin{aligned}
X_j &= R_j + \iiint g(x,y,z)\, a_j(x,y,z)\, \alpha_j(x,y,z)\, dx\, dy\, dz, && (2.1)\\
R_j &= \text{rate of random coincidences appearing in tube } j,\\
a_j(x,y,z) &= \text{probability that an emission at } (x,y,z) \text{ will release}\\
&\ \text{gamma rays in paths leading to tube } j \text{, and}\\
\alpha_j(x,y,z) &= \text{conditional probability that such a pair of emitted gamma}\\
&\ \text{rays will reach the detectors and be recorded.}
\end{aligned}$$

Physical theory tells us that, approximately:

- R_j can be estimated accurately from the rates of emissions observed separately at the two detectors that define tube j.

- $a_j(x,y,z)$ is an angle factor—the conditional probability that an emission from point (x,y,z) will yield gamma rays in paths leading to tube j—that depends only on the geometry of the experiment, and can be regarded as known.

- $\alpha_j(x,y,z)$ does not depend on (x,y,z) and so can be written simply as α_j, which depends largely on the integrated mass density within the path of tube j.

The function $g(x,y,z)$ varies with the blood flow and metabolism of the organ. We wish to estimate the continuous function $g(x,y,z)$ from the observed Y_j's and the known R_j's, a_j's, and α_j's. We will follow the practice of the lab at MGH and approximate $g(x,y,z)$ by a set of two-dimensional images $g(x,y)$, at nine levels z. Each image plane is estimated from its own set of detector data.

Furthermore, we will approximate the two-dimensional angle factors. As pictured in Figure 2.3 (in Section 2.2.2), a tube j can be located at angle θ and distance t from the origin, and the projection of a point (x,y) on to angle θ is $t_\theta(x,y) = x\cos\theta + y\sin\theta$. Due to scattering and the finite widths of detectors, emissions at angle θ reach a range of tubes. The angle factors can be approximated:

$$a_{t,\theta}(x,y) = \frac{1}{\#\text{proj}} K\left(\frac{t - t_\theta(x,y)}{d}\right), \text{ where} \tag{2.2}$$

#proj is the number of projection angles θ,

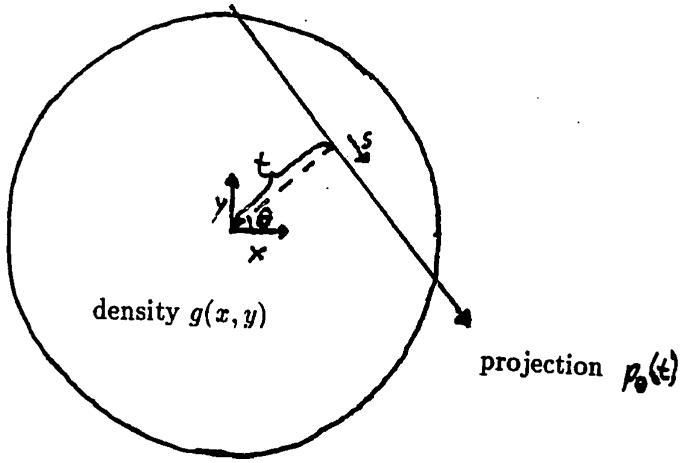
K is a kernel function, and

d is comparable to the width of a detector and the "resolution limit."

The blurring kernel in the conditional probability (2.2) arises because the gamma rays are not emitted in exactly opposite directions, and their positions are measured coarsely as tube locations; the function K and its scale d can be measured experimentally from the counts observed from a point source. In the MGH experiment, K is approximately Gaussian and $d \approx 4\,\text{mm}$.

Note that in this approximation, $a_{t,\theta}(x,y)$ is independent of θ, reflecting the assumption that emissions reach all detector angles in equal expected numbers.

Figure 2.2: Continuous PET projections



2.2 Fourier image reconstruction from projections

The standard method for reconstruction of tomography images is *filtered back-projection*, an algorithm based on the mathematics of the two dimensional Fourier transform. We will derive this method, which is an exact solution for the idealized case of continuous projection data, and then discuss the approximate method used for real, discrete data.

The Fourier method yields a two-dimensional density estimate that is linear in the adjusted coincidence rates $(Y_j - R_j)/\alpha_j$; the method is not completely satisfactory with finite data, because it ignores the probabilistic data recording mechanism (the Poisson distributions of the counts).

2.2.1 Solution for the continuous problem

The continuous Fourier method exactly reconstructs a two-dimensional emission density function $g(x,y)$, given an infinite number of detectors and error-free projection data, as described in Kak (1984).

In the continuous idealization, projections $p_\theta(t)$ are observed for every polar coordinate (θ,t), as pictured in Figure 2.2. A *projection* is an integral of the emission density along a straight line at an angle θ to the x-axis, a minimum distance t from the origin. We can express this integral in terms of rotated coordinates $(s,t) = (x\sin\theta - y\cos\theta,\ x\cos\theta +$

$y \sin\theta$):

$$p_\theta(t) = \int g(x(s,t), y(s,t))\, ds \tag{2.3}$$

This map from image $g(x,y)$ to projections $p_\theta(t)$ is called the *Radon transform* or *continuous projection operator.* In the notation of equations 2.1 in Section 2.1.1, the third dimension is ignored, the projections $p_\theta(t)$ are adjusted expected counts $(X_j - R_j)/\alpha_j$, and the kernel K in equation 2.2 is a Dirac delta function; that is, exact projections are observed.

We define the continuous Fourier transforms:

$$\begin{aligned} P_\theta(\omega) &= \int p_\theta(t) e^{-i\omega t}\, dt \\ G(\omega_x, \omega_y) &= \iint g(x,y) e^{-i\omega_x x - i\omega_y y}\, dx\, dy. \end{aligned}$$

To relate the projections $p_\theta(t)$ to the image $g(x,y)$, we use the "projection-slice theorem," derived here:

$$\begin{aligned} P_\theta(\omega) &= \iint g(x(s,t), y(s,t)) e^{-i\omega t} dt\, ds \text{ (from equation 2.3)} \\ &= \iint g(x,y) e^{-i\omega(x\cos\theta + y\sin\theta)}\, dx\, dy \text{ (changing variables)} \\ &= G(\omega\cos\theta,\ \omega\sin\theta). \end{aligned}$$

One can now linearly reconstruct a two-dimensional image from exact knowledge of all its projections, as follows:

1. For each θ, calculate the Fourier transform of the projections at the angle θ, for $t \in (-\infty, \infty)$. This yields $P_\theta(\omega)$ for each θ.

2. Combine these to get $G(\omega_x, \omega_y)$.

3. Apply the inverse Fourier transform to get $g(x,y)$.

These steps can equivalently be done by the filtered back-projection algorithm, which is based on the following formula:

$$g(x,y) = \frac{1}{(2\pi)^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} G(\omega_x, \omega_y) e^{i\omega_x x + i\omega_y y} d\omega_x d\omega_y \tag{2.4}$$

$$
\begin{aligned}
&= \frac{1}{(2\pi)^2}\int_0^{2\pi}\int_0^{\infty} G(\omega\cos\theta,\omega\sin\theta)e^{i\omega(x\cos\theta+y\sin\theta)}\omega\, d\omega\, d\theta \\
&= \frac{1}{(2\pi)^2}\int_0^{\pi}\int_{-\infty}^{\infty} G(\omega\cos\theta,\omega\sin\theta)e^{i\omega(x\cos\theta+y\sin\theta)}|\omega|\, d\omega\, d\theta \\
&= \frac{1}{2\pi}\int_0^{\pi} \mathrm{IFT}[P_\theta(\omega)|\omega|](x\cos\theta+y\sin\theta)\, d\theta,
\end{aligned}
$$

where IFT is the inverse Fourier transformation.

This corresponds to the following algorithm:

1. For each θ, calculate $P_\theta(\omega)$.

2. For each θ, calculate the inverse Fourier transform of $P_\theta(\omega)|\omega|$.

3. *Back-project* this by averaging these functions over θ to get $g(x,y)$ (using equation 2.4).

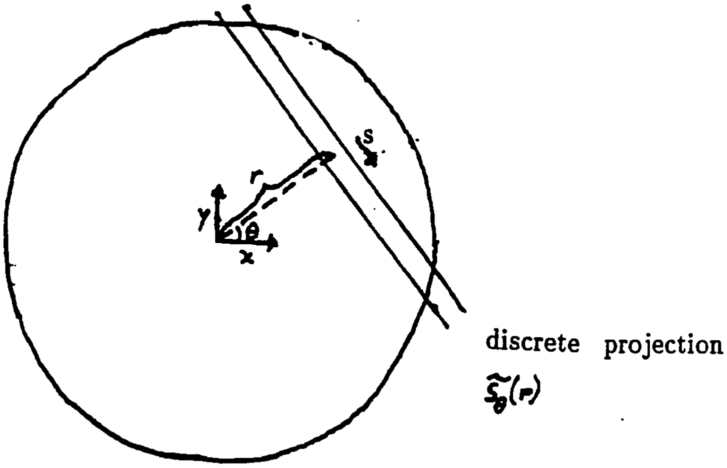
This method can be effectively adapted to the problem with finitely many detectors.

2.2.2 Finite number of projections

The above formulas still hold, with some slight alterations, for the finite case. Now, of course, we cannot hope to reconstruct an arbitrary continuous function $g(x,y)$. Rather, a linear algorithm in this discrete-data PET problem reconstructs the Fourier components of $g(x,y)$ with low spatial frequency, given exact data at a finite number of angles and projections.

To simplify the geometry, we assume that the diameter of the detector ring is much larger than that of the organ under study. Our data are the same as in the continuous case, except that projections are of tubes of finite width $2R$, and we ignore the wobbling of the detector ring. The approximation that tubes have equal width is reasonable if the support of g—that is, the cross-section of the organ under study—is less than half the diameter of the detector ring, as occurs in practice. (The wobble increases the number of projections observed and was irrelevant in the last section, which assumed continuous data.) To simplify notation, we also ignore attenuation and random coincidences, for which

Figure 2.3: PET projections of finite width



this procedure corrects with the obvious linear transformation. The following variables are pictured in Figure 2.3:

- The number #proj of projection angles equals the number of detectors in the ring. Each projection angle corresponds to a series of tubes centered at coordinates $r = \pm R, \pm 3R, \pm 5R, \ldots$, or $r = 0, \pm 2R, \pm 4R, \ldots$. Measurements for each projection angle θ are a series of counts $s_\theta(r)$, where r is the coordinate of the center of the tube. Measurements are taken at #proj projection angles, which equals the number of detectors.

- Expected counts are $\tilde{s}_\theta(r)$.

- The emission density $g(x,y)$ has support in the unit circle.

For an emission near the center of the length of a detector tube specified by $(r, \theta,$ the probability that the gamma ray paths will remain in the tube is proportional to the distance to the edge of the tube and must be normalized by the number of projection angles:

$$a_{r,\theta}(t) = \frac{2}{\#\mathrm{proj}}\left(1 - \frac{|t-r|}{R}\right), \text{ for } t \in [r - R, r + R].$$

This expression is a function of θ because $t = x\cos\theta + y\sin\theta$ and is a specific form, accounting for finite detector widths but not for scattering, of equation 2.2. The counts in the tube can come from the range $t \in [r-R, r+R]$ in (s,t) coordinates, and so their expected number is an integral over this range:

$$\begin{aligned}
\tilde{s}_\theta(r) &= \int_{r-R}^{r+R} dt \int_{-L(t)}^{L(t)} g(x(s,t), y(s,t))a(t,r)\, ds, \\
\text{where } L(t) &= \sqrt{1-t^2} \text{ (The organ lies within the unit circle.)} \\
&\approx \sqrt{1-r^2} \text{ for a large detector ring, which we assume.}
\end{aligned}$$

Using the projection $p_\theta(t)$ defined in the continuous case above,

$$\begin{aligned}
\tilde{s}_\theta(r) &= \int_{r-R}^{r+R} p_\theta(t) \frac{2}{N_d}\left(1 - \frac{|t-r|}{R}\right) dt, \qquad (2.5) \\
\text{and } s_\theta(r) &\sim \text{independent Poisson}(\tilde{s}_\theta(r)).
\end{aligned}$$

We will assume that the expected counts $\tilde{s}_\theta(r)$ are actually observed, and define, for each angle, the discrete Fourier transform. For each θ, there will be approximately $1/R$ tubes that intersect the unit circle and thus may receive counts. The Fourier frequencies are then $\omega = 0\frac{\pi}{2}, 1\frac{\pi}{2}, \ldots, \frac{1}{R}\frac{\pi}{2}$, and the Fourier transform of interest is:

$$S_\theta(\omega) = \sum_r \tilde{s}_\theta(r) e^{-i\omega r}, \text{ for } \omega = k\frac{\pi}{2}, \;\; k = 0, 1, \ldots, \frac{1}{R}.$$

Using equation (2.5):

$$S_\theta(\omega) = \int_{-1}^{1} p_\theta(t) e^{-i\omega t} e^{i\omega\delta(t)} \frac{2}{N_d}\left(1 - \frac{|\delta(t)|}{R}\right) dt, \qquad (2.6)$$

where $\delta(t) = t - r$, and r is the t-coordinate of the center of the tube in which t lies. The sawtooth function $\delta(t)$ maps the interval $[-1,1]$ repeatedly on to $[-R, R]$.

Consider the two functions:

$$\begin{aligned}
f_1(t) &= p_\theta(t) e^{-i\omega t} \\
\text{and } f_2(t) &= e^{i\omega\delta(t)}\left(1 - \frac{|\delta(t)|}{R}\right),
\end{aligned}$$

which appear inside the integral (2.6), for the range $t \in [-1,1]$.

The periods of these functions are generally different. We will make the approximation that the functions are uncorrelated. It seems likely, but is not clear, that this assumption introduces no systematic errors in the reconstruction. The approximation simplifies equation 2.6 to:

$$\begin{aligned} S_\theta(\omega) &\approx \int_{-1}^{1} p_\theta(t) e^{-i\omega t} dt \cdot \frac{1}{2}\int_{-1}^{1} e^{i\omega\delta(t)} \frac{2}{N_d}\left(1 - \frac{|\delta(t)|}{R}\right) dt \qquad (2.7) \\ &= P_\theta(\omega)\frac{1}{N_d}\int_{-1}^{1} e^{i\omega\delta(t)}\left(1 - \frac{|\delta(t)|}{R}\right) dt \\ &= P_\theta(\omega)\int_{-R}^{R} e^{i\omega t}\left(1 - \frac{|t|}{R}\right) dt \\ &= RP_\theta(\omega)\left(\frac{\sin(\omega R/2)}{\omega R/2}\right)^2. \end{aligned}$$

Note that the factor $\left(\frac{\sin(\omega R/2)}{\omega R/2}\right)^2$ is near 1 for the entire range of Fourier frequencies ω, from 0 to $\frac{\pi}{2R}$.

If we assume that $p_\theta(t)$ is band-limited (that is, its Fourier transform is zero above the frequency $\frac{\pi}{2R}$), then equation 2.7 allows us to derive $P_\theta(\omega)$ from $S_\theta(\omega)$:

$$P_\theta(\omega) = \frac{1}{R}\left(\frac{\omega R/2}{\sin(\omega R/2)}\right)^2 S_\theta(\omega),$$

and we can use the continuous formulas to reconstruct $g(x, y)$ from $P_\theta(\omega)$. In particular, we can apply the following back-projection algorithm:

1. For each θ, calculate $S_\theta(\omega)$.

2. For each θ, calculate the inverse Fourier transform of $S_\theta(\omega)|\omega|\frac{1}{R}\left(\frac{\omega R/2}{\sin(\omega R/2)}\right)^2$.

3. Back-project this.

Note that the first two steps above can also be accomplished by convoluting

$$\tilde{s}_\theta(r) \text{ with } h(r) = \text{IFT}\left[\frac{1}{R}|\omega|\left(\frac{\omega R/2}{\sin(\omega R/2)}\right)^2\right].$$

This convolution is called "filtering."

2.2.3 Filtered back-projection in practice

The wobble slightly alters the case of finitely many projections, improving the approximations above. Measurements of projections $s_\theta(r)$ are made at more finely-spaced intervals of r, allowing the Fourier transform $S_\theta(\omega)$ to be defined at higher frequencies ω. Expression 2.6 becomes a sum over wobble positions, and the integrand can be factorized so that $f_1(t)$ remains unchanged while $f_2(t)$ becomes a sum that approaches a constant as the number of wobble positions increase. Approximation 2.7 becomes exact in this case, as does the back-projection algorithm. Results are still band-limited, however, due to the finite number of projection angles.

The MGH lab uses the filter $h(r) = IFT[|\omega|]$, cut off at $\frac{\pi}{2R}$, ignoring the correction $\left(\frac{\omega R/2}{\sin(\omega R/2)}\right)^2$ due to the resolution limit expressed in 2.2.2. Cutting off the filter and not correcting for blur enhances the low frequencies in the projections in an attempt to smooth the image estimate.

2.2.4 Limitations of the Fourier reconstruction

Later chapters will discuss how reconstructions can be analyzed and improved by using substantive knowledge of the image context. We restrict our discussion here to limitations in the reconstruction of an arbitrary image.

Resolution limit. Given perfect data at finitely many projection angles, the Fourier reconstruction of a point source is, approximately, a circular kernel with diameter comparable to the width of a detector. Natterer (1986, chapter 3) gives the resolution limit as a fraction ($2\pi/\#$projectons) of the diameter of the image. For the 96 projection angles of the MGH lab, this gives a resolution of 1.3 mm in a 20 cm diameter head slice; the practical implications of this number are not clear. The reconstruction of an image is then, at best, a blurred version of the truth. The resulting image estimate is biased, and fine details of the true image will be missed. In any case, the resolution limit, which makes it impossible to filter out the blur, comes from finite spatial sampling, not from finite detector widths.

Sampling error. The Fourier-based reconstruction methods are based on various idealizations, the most extreme being the assumption of exact projection data $\tilde{s}_\theta(r)$. When

Possion counts are substituted for these projections, the resulting image reconstruction becomes an unbiased linear estimate of the Fourier reconstruction given noise-free data, and can be better understood by calculating the variances and covariances of estimation. Chapter 8 discusses the bias and variance of the image estimate in more detail, and Chapter 10 shows some images from the MGH experiment.

2.3 Likelihood for the PET inverse problem

In the context of maximum likelihood reconstruction, this section introduces ideas that will be used in later chapters. We begin our statistical study of the two-dimensional PET problem by setting up notation. The observed data are a vector $Y = (Y_1, \ldots, Y_n)$; each Y_j is a count at a detector pair and wobble location (of the form $s_\theta(r)$ in Section 2.2.2 above) with the distribution

$$Y_j \sim \text{ Poisson } (X_j).$$

These random variables are assumed independent, and the vector $X = (x_1, \ldots, x_n)$ of their means can be expressed in terms of the emission density function g:

$$X = R + \alpha A g,$$

where R is a vector of random coincidence frequencies, $\alpha = \text{diag}(\alpha_j)$ is a diagonal matrix of attenuation factors, and A is an all-nonnegative linear transformation of conditional probabilities. (R, α, and A are assumed known.) We will discretize the unknown image of interest into N pixels:

$$g = (g_1, \ldots, g_N).$$

The tranformation A can now be written as an $n \times N$ matrix $A = (a_{ij})$ of angle factors that can be approximated by equation 2.2.

For clarity, we split the angle factor matrix A into simpler linear operators: $A = BP$. The *forward-projection* operator P assigns each pixel (x, y) of an image to a location t in each projection angle θ: $t = x\cos\theta + y\sin\theta$. For each angle, the resulting function of t is the sum of contributions from all image pixels. The computation required is on the order of (#pixels × #projection angles). The *blur* operator B convolutes a vector $h_\theta(t)$ in projection space with the kernel $K(t)$ of equation 2.1, for each projection angle θ. In addition, we define the *back-projection* operator P^t; this is forward-projection in reverse, assigning to each pixel a contribution from all of the projection angles, requiring the same computation time as forward-projection. The projection operators, as well as the blur, can be speeded by operating in the frequency domain, using Fourier transforms and interpolation (Stearns, 1990).

As detailed in Section 2.2, P and P^t are not inverses: the inverse of a forward projection is a *filtered* back-projection. Labeling the filtering operation as F, the filtered back-projection reconstruction from data Y, corrected for random coincidences R and attenuation α, is:

$$\tilde{g} = P^t F \alpha^{-1}(Y - R). \tag{2.8}$$

Because of the resolution limit, the filter is inexact: $P^t F \neq P^{-1}$.

2.3.1 Exact solutions

The case $N \geq n$. The likelihood of the observations Y is maximized by any solution g of the equation $R + \alpha A g = Y$. Solutions of this linear system lie on an $(N - n)$-dimensional hyperplane in the N-dimensional space of images g. There will typically be an infinity of all-nonnegative solution vectors g, or no such solutions.

Figures 2.4a and 2.4b show the structure of the PET likelihood. The transformation $g \rightarrow R + \alpha A g$ maps each positive axis in image space to a ray in the positive orthant of data space, emanating from the point R. (A positive axis in image space corresponds to a point source image of positive intensity.) The convex hull of all these projected axes, shaded in the figures, is the region in data space corresponding to all-nonnegative images. Exact all-nonnegative solutions of $R + \alpha A g = Y$ exist if the data vector Y lies in this region, as in Figure 2.4a. If Y falls outside the shaded region, as in Figure 2.4b, no exact

Figure 2.4: Constraints on the PET likelihood: the case $N \geq n$

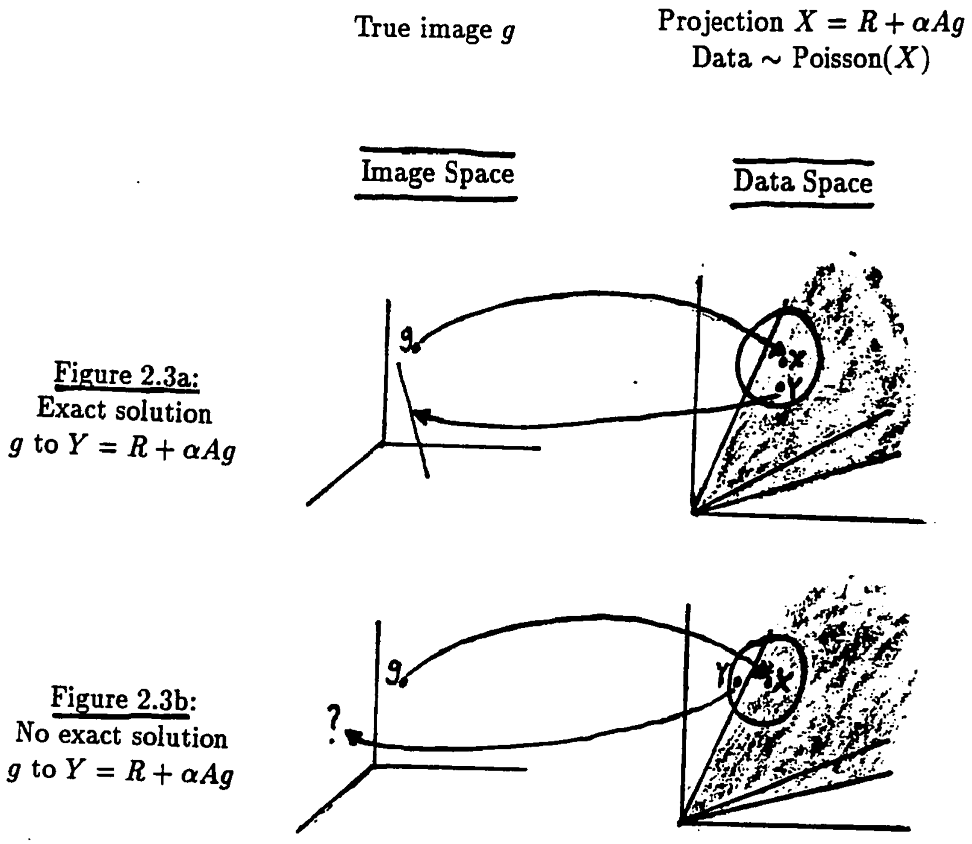


Figure 2.3a:
Exact solution
g to $Y = R + \alpha Ag$

Figure 2.3b:
No exact solution
g to $Y = R + \alpha Ag$

Shaded areas cover data expectations corresponding to all-nonnegative images g. **Circles around** X represent the variability of data Y about their expectations.

solutions exist. Note that the vector $X = R + \alpha Ag$ of expected data necessarily lies within this hull, and so Y is only likely to be outside it if X is sufficiently near the edge of the hull and the Poisson sample size is small enough. ("Small," in this context, may be much larger than a typical PET data set.)

For "small" sample size, some of the principal components of data help the likelihood to identify the true image, while others are nearly useless. (Counts at different wobble positions of the same tube are an example of less useful data.) Geometrically, useful degrees of freedom correspond to directions in which the shaded convex hull is wide, while less useful data correspond to narrow directions in the hull. If the width of the hull in some direction is small compared to the Poisson sampling standard deviation of the data, then the likelihood will be uninformative in that direction. This topic will be elaborated upon in Chapter 4.

The case $N < n$. The shaded convex hull now lies in an N-dimensional hyperplane of data space that maps one-to-one with the positive orthant of image space.

2.3.2 Maximum likelihood reconstruction

This can be done with various parameterizations of the image. We will first assign a nonnegative parameter to each pixel intensity.

Use of the EM algorithm

If they exist, exact all-nonnegative solutions g of $R + \alpha Ag = Y$ may be found by linear programming techniques (Shepp, 1989). If no such solutions exist, we can use the EM algorithm to find the all-nonnegative vector g, generally unique, that maximizes the likelihood (Dempster et al., 1977; Shepp and Vardi, 1982; Vardi et al., 1985). We imagine missing data: a matrix $Z = (z_{ij})$ of the number of emissions from pixels i, recorded in tubes j, and a vector $Z_R = (z_{Rj})$ of the number of random coincidences recorded in tube j. Then $Y_j = z_{Rj} + \sum_i z_{ij}$, and the EM algorithm becomes:

1. Start with a guess $\hat{g}$.

2. E-step: for this Poisson likelihood, we need only calculate the quantities:

$$\begin{aligned} E(z_{ij}|Y,\hat{g}) &= \frac{Y_j \alpha_j a_{ij} \hat{g}_i}{R_j + \alpha_j \sum_{i'} a_{i'j} \hat{g}_{i'}}, \text{ for each } i \text{ and } j, \\ \text{and } E(z_{Rj}|Y,\hat{g}) &= \frac{Y_j \alpha_j R_j}{R_j + \alpha_j \sum_{i'} a_{i'j} \hat{g}_{i'}}, \text{ for each } j. \end{aligned} \tag{2.9}$$

3. M-step: Find a new estimate, g^*, to maximize the complete-data likelihood with the above expectations substituted in. The components of this estimate are: $g_i^* = \sum_j E(z_{ij}|Y,\hat{g})/\bar{\alpha}_i$, where $\bar{\alpha}_i = \sum_j \alpha_j a_{ij}$.

4. Relabel $\hat{g} = g^*$ and repeat steps 2 and 3 above.

Implementation for PET

The steps in this iteration, presented algebraically above, may be computed using no computer storage or new types of computer operations beyond those of the filtered back-projection or Fourier methods now in use. This implementation requires the operations of forward-projection, back-projection, and blurring, defined in Section 2.3 above.

i. Start with a guess $\hat{g}$.

ii. Create the vector $R + \alpha A \hat{g}$ (the denominator of equation 2.9): forward-project g, then blur, multiply by attenuation factors, and add random coincidences in projection (data) space.

iii. Normalize the data y by dividing, component-by-component, by the vector that was just calculated in step *ii*.

iv. Create the image vector $g^* = E(Z|Y,\hat{g})$: take the result of step *iii* above, multiply by attenuation factors, blur, back-project, multiply by $\hat{g}$, and divide by average attenuation factors in image space.

The image of average attenuation factors can be computed before step *i*: take a unit vector in projection space, multiply by attenuation factors, blur, and back-project.

v. Relabel $\hat{g} = g^*$ and repeat steps *ii–iv* above.

Advantages and problems with the maximum likelihood estimate

By virtue of being maximum likelihood, this estimate will always be in the allowed parameter space of all-positive images and, if $N < n$, is efficient as sample size approaches infinity (that is, as the coefficient of variation of the Poisson variables approaches zero).

By defining pixels small enough that $N > n$, one must estimate more parameters than there are data points. It is then hopeless to uses a purely data-based function such as the likelihood to pick a unique image estimate. In fact, if N is less than n but still large, then the likelihood for moderate sample size will not effectively distinguish among all-positive images. As a result, the maximum likelihood estimate will be unacceptably variable and have a noisy appearance (Vardi et al., 1985; Snyder et al., 1987). The difficulty comes from the uninformative components of data described in Section 2.3.1 above.

If no exact all-nonnegative solutions g of $R + \alpha Ag = Y$ exist, then the maximum likelihood estimate will lie on the boundary of the space of all-positive images; that is, some of its pixel intensities will be zero. In such a problem there will typically be smooth images with nearly the same likelihood, and the noisy appearance of the maximum likelihood image will contain many distinctive features not strongly supported by the data. This behavior can get worse as the number of pixels estimated is increased (Gelman, 1989).

Because of its high variance and possible indeterminacy, the maximum likelihood estimate by itself is unsatisfactory for PET; however, the various improvements that have been suggested, discussed in the rest of this chapter and in Chapters 3–5, use iterative procedures similar to the EM algorithm presented above.

2.3.3 Parameters covering larger regions

There is nothing in the above discussion that requires one to estimate the image intensity within small "pixels." If instead the image is piecewise constant within several larger "regions," the nonnegative intensities at these regions can be estimated using the EM algorithm, as above (Carson, 1986). If the regions are few enough, the resulting maximum likelihood image estimate may be acceptably stable for moderate sample size.

If the image is not constant within regions, we can treat the piecewise-constant image estimate as an estimate of *average* regional intensities. This will in general be a biased estimate, with bias depending on the discrepancy between the true image and the piecewise-constant image of true regional averages. Section 7.2 discusses the bias of regional estimates in detail.

2.3.4 General image parameters

A different sort of constraint may be placed on the estimated image by covering it with overlapping pixels of nonnegative intensity. This reduces the "shaded area" of data space in Figure 2.4 in volume but not necessarily in dimension.

Example 2.1: Saturated and unsaturated models

Consider a square array of N "little" pixels, to be modeled by a moving average of "big" pixels, each twice as large on a side. A *saturated* model covers the array with four overlapping overlays of $N/4$ big pixels each; each little pixel intensity is then expressed as an average of four independent big pixel intensities. In contrast, an *unsaturated* model uses two overlays of $N/4$ big pixels each, differing in position by a translation along the diagonal of one little pixel. Here, each little pixel intensity is an average of two independent moving average parameters, and the whole image is covered by $N/2$ independent parameters.

Such models could of course be elaborated with "big pixels" of differing size, allowing an arbitrary reduction in the number of independent parameters of a two-dimensional image. In addition their contribution to the image need not be as piecewise-constant functions; that is, they may be arbitrary kernels (Snyder and Miller, 1985). In general, such a model describes an image by an all-nonnegative vector parameter h, with components h_k defined over regions Ω_k:

$$g_i = \sum_{\{k:i\in\Omega_k\}} w_{ik}h_k. \tag{2.10}$$

The weights w_{ik} sum to one: $\sum_k w_{ik} = 1$ for all i.

Here $g = Wh$, and the likelihood of h is based on the equation $X = R + \alpha AWh$. Estimating h is then the same mathematical problem as estimating g, with the matrix A replaced by AW. The effectiveness of this estimation by moving averages comes from the nonnegativity constraint on the restrictive parameter space of h and from parsimonious parameterization.

Chapter 3

Overview of Bayesian Image Reconstruction

The Bayesian approach has been widely advocated in recent statistical and image reconstruction literature (for example, Geman and Geman, 1984; Besag, 1989; and many of the references in Venetoulias, 1988). Joint probability distributions on pixels are a convenient and flexible framework for modeling images and estimates of images, given data. Of course, this approach requires the effort of specifying a prior model and computing and summarizing the joint distribution. This chapter will discuss, in detail, the steps needed for Bayesian reconstruction of an image of (continuous) gray levels, starting in Section 3.1 with the method presented in Geman and McClure (1985), and continuing in Section 3.2 with a review of other, related models and methods in the recent literature.

Many researchers have suggested Bayesian methods to improve the maximum likelihood image estimate for emission tomography (Geman and McClure, 1985; Lange et al., 1987; and others). Section 3.3 of this paper discusses how to implement Bayesian methods in our PET problem by extending the implementation of the EM algorithm using projection operators as detailed in Section 2.3.2, allowing for prior distributions with unknown hyperparameters. Chapters 4 and 5 discuss related theoretical issues while deriving reconstruction algorithms for specific families of Bayesian posterior distributions. Finally, Chapter 7 derives the bias and variance of regional averages of maximum likelihood and Bayesian PET images.

3.1 An example of Bayesian reconstruction

A Bayesian analysis allows stable estimation of the many parameters of an image from a fixed amount of data. Such a method can be difficult because of computational costs; it also requires a prior distribution that reflects the population of images under study. Geman and McClure (1985) describe a method of creating and computationally summarizing a posterior distribution of possible images g, given a likelihood function based on data Y. This section describes this approach, applied to the simplified two-dimensional PET problem. For exposition, we will also follow the algorithm for the simple example of estimating the four entries in a 2×2 table with margins observed with error. In general, the algorithm proceeds in the following steps, to be described in the corresponding subsections below.

1. Discretize the image space.

2. Determine the likelihood function; that is, the probabilistic rule that derives the observed data from the true image.

3. Set up a Gibbsian distribution for the population of possible images (also see step 6).

4. Use stochastic relaxation to take a random walk through the posterior distribution of images.

5. If desired, use simulated annealing to reach the image of maximum posterior probability.

6. In practice, it will be necessary to estimate parameters in the Gibbsian prior distribution.

3.1.1 Pixels

In this approach, the region over which the image is defined must be made discrete. The two-dimensional tomography image space is usually divided into a grid of square pixels, with the number of pixels being the same order of magnitude as the number of detector "tubes" or data points. Emission density is modeled as constant within each pixel.

3.1.2 Likelihood function

For physical reasons, detector counts are assumed to have independent Poisson distributions. For the PET geometry, $E(Y) = X = R + \alpha Ag$, as detailed in Chapter 2 above. The likelihood is:

$$P(Y|g) \propto \prod_{\text{tubes } j} X_j^{Y_j} \exp(-X_j).$$

Example 3.1: 2×2 **table**

In this illustrative example, the image space consists of 4 discrete cells, and the data are noisy measurements of the table margins. Consider this set of data: $\begin{array}{|c|c|c}\hline & & 9 \\ \hline & & 39 \\ \hline \multicolumn{1}{c}{11} & \multicolumn{1}{c}{41} & \end{array}$. (Due to noise, the row sum $(9 + 39)$ does not equal the column sum (11+41).) The four data points here are Poisson random variables whose expectations are the row and column sums of the table. We wish to estimate the vector $g = (g_1, \ldots, g_4)$ of table entries, which we parameterize as: $\begin{array}{|c|c|}\hline g_1 & g_2 \\ \hline g_3 & g_4 \\ \hline\end{array}$. Two possible images consistent with the noisy marginal data are: $\begin{array}{|c|c|}\hline 2 & 8 \\ \hline 8 & 32 \\ \hline\end{array}$ and $\begin{array}{|c|c|}\hline 0 & 10 \\ \hline 10 & 30 \\ \hline\end{array}$. The likelihood is:

$$\begin{aligned} P(Y|g) &\propto (g_1 + g_2)^9 (g_3 + g_4)^{39} (g_1 + g_3)^{11} (g_2 + g_4)^{41} \\ &\quad \cdot \exp[-(g_1 + g_2) - (g_3 + g_4) - (g_1 + g_3) - (g_2 + g_4)]. \end{aligned} \tag{3.1}$$

3.1.3 Gibbs distribution of images

A distribution defined on an array of variables is called Gibbsian or Markovian if it satisfies a certain conditional independence property, which we define after the following example.

Example 3.2: 2×2 **table**

Consider the following prior distribution on the four parameters of the table:

$$P(g) \propto \exp(-\beta[(g_1 - g_2)^2 + (g_3 - g_4)^2 + (g_1 - g_3)^2 + (g_2 - g_4)^2])$$

The parameter β must be positive; we will assume at first that it is known. This is an example of an *intrinsic conditional autoregressive* model (Kunsch, 1987), which will be discussed in more detail in Chapter 5.

This Normal distribution forces positive correlations between neighboring cells of the table, and so favors a smoother "image." It is an improper distribution, as it is unaffected by a location shift in the parameters g_i. When combined with the likelihood (3.1), however, a proper posterior distribution results.

Note that g_1 and g_4 are conditionally independent in this prior distribution, given g_2 and g_3. Similarly, g_2 and g_3 are conditionally independent, given g_1 and g_4. This is an example of the Markov property, which we will now define.

Definition of Markov field and Gibbs distribution

These expressions characterize certain joint distributions (*random fields*) on a network of variables. In this network, each variable has a (possibly empty) set of neighbor variables. The Markov property states that the distribution of a variable, conditional on all the others in the network, must be the same as its distribution conditional only on its neighbors. (In the 2×2 table above, any two cells in the same row and column are neighbors.) Recall that a clique is a set of variables, all of which are neighbors. (In the above example, the largest cliques are the four pairs of orthogonally but not diagonally adjacent cells.)

Now consider a picture as a single realization of a Markov random field, with each pixel intensity being a variable in the field. Consider this realization g as the combination of its values ((g_s), for all $s \in C$) at all cliques C in the network. (This defines the picture redundantly, because the values of variables shared by overlapping cliques must agree in this specification.) Now, assume the *positivity* condition: if the values at all of the cliques of an image have positive probability, then that image as a whole has positive probability. (This condition states that there are no global restrictions on the space of images.) If this condition holds for all images defined over the network, then the Gibbs-Markov equivalence theorem holds, stating that the joint probability of an image can be factored into probabilities of cliques (Besag, 1974):

$$P(g) \propto \prod_{\text{cliques } C} P_C((g_s), \text{ for all } s \in C).$$

Taking logarithms, this can be also be expressed in terms of a *potential function* U, defined on cliques:

$$\log P(g) = -\sum_{C} U_C((g_s), \text{ for all } s \in C).$$

($\log P(g)$ is here defined up to an additive constant.) This joint distribution is called Gibbsian. In the example of the 2 × 2 table, the potential function can be written:

$$U_{ij}(g_i, g_j) = \beta(g_i - g_j)^2, \text{ for neighboring cells } i \text{ and } j.$$

A prior distribution for tomographic images

In the square grid of pixels, Geman and McClure (1985) define the neighborhood of each pixel to be the four orthogonally adjacent and four diagonally adjacent pixels; cliques are then squares of four adjacent pixels. The potential function must be low for states (images) of high probability. In order to favor smooth images, Geman and McClure specify:

$$U_{ij}(g_i, g_j) = \begin{cases} \beta\phi(g_i - g_j), & \text{if cells } i \text{ and } j \text{ are orthogonal neighbors} \\ \frac{1}{\sqrt{2}}\beta\phi(g_i - g_j), & \text{if } i \text{ and } j \text{ are diagonal neighbors,} \end{cases}$$

$$\text{and } \phi(x) = -\frac{1}{1 + (x/\delta)^2}.$$

The associated prior distribution is $P(g) \propto \exp(-\sum U_{ij}(g_i, g_j))$ and gives low probability to large differences $|g_i - g_j|$ between neighbors; however, unlike with the Gaussian potential function $(g_i - g_j)^2$, this probability density does not approach zero as that difference approaches infinity. This allows the possibility of sharp boundaries in the image.

3.1.4 The conditional (posterior) distribution of images

Given the data Y, the distribution of the true image g is:

$$P(g|Y) \propto P(Y|g)P(g).$$

This is still Markovian, but the cliques are enlarged: in general, two pixels are neighbors here if they are neighbors in the prior distribution *or* in the likelihood. In the example of a contingency table with margins observed with error, each row and column of the table is a clique in the likelihood. In the PET geometry, the likelihood function factors by "tubes"; two pixels are neighbors if they share a tube. Since the tubes are observed at a full range of angles, this means that just about all pairs of pixels are neighbors in the likelihood; if all pairs of pixels are neighbors, then the set of pixels forms one clique.

We would like to characterize the posterior distribution of images, possibly by sampling from it. In general, however, its high dimensionality and complicated structure precludes any direct treatment.

Stochastic relaxation

This algorithm is a modification of the Metropolis algorithm for sampling from a distribution specified by a known, unnormalized density function (Metropolis et al., 1953). Stochastic relaxation performs a random walk through a distribution of images. As detailed below, it creates a sequence of pictures $g^{(1)}, g^{(2)}, g^{(3)}, \ldots$:

1. Start with a picture $g^{(0)}$ that has positive probability.

2. At each step, change the intensity of g at just one pixel. As many steps are performed, cycle through the pixels in any order.

3. Say that, at step t, pixel 1 is being altered. Then, choose $g_1^{(t)}$ from the distribution $P(g_1|g_2 = g_2^{(t-1)}, g_3 = g_3^{(t-1)}, \ldots, g_N = g_N^{(t-1)}, Y)$.

4. The sequence of images $g^{(t)}$ produced by this algorithm is an ergodic Markov process whose stationary distribution is the desired $P(g_1, \ldots, g_N|Y)$.

Example 3.3: 2×2 **table**

We will describe the application of stochastic relaxation to this example, with the variables and probability distributions given above. We will assume a starting image $g^{(0)} = \begin{array}{|c|c|}\hline 2 & 8 \\ \hline 8 & 32 \\ \hline\end{array}$, and for our first step we will update the variable g_3. Then we can write, $g^{(1)} = \begin{array}{|c|c|}\hline 2 & 8 \\ \hline g_3 & 32 \\ \hline\end{array}$. We must sample from the distribution:

$$\begin{aligned} P(g_3|g_1 = 2, g_2 = 8, g_4 = 32, Y) \;\propto\; & (g_3 + 2)^{11}(g_3 + 32)^{39}\exp(-g_3) \\ & \cdot \exp(-\beta[(g_3 - 2)^2 + (g_3 - 32)^2]). \end{aligned}$$

If g_3 is assumed to be restricted to a finite number of values, their relative conditional probabilities can be easily calculated, and a sample may be drawn. The restriction to a finite space may be done for computational convenience rather than for substantive reasons. To do stochastic relaxation, we can continue by successively updating $g_4, g_1, g_2, g_3, g_4, \ldots$, not necessarily in a regular order.

Stationary distribution of stochastic relaxation

The sequence of pictures $g^{(}t)$ produced by stochastic relaxation form an irreducible, aperiodic Markov process (all its transition probabilities are positive). Then, by a fundamental theorem of Markov chains (Feller, 1968, Section XV.7), $g^{(t)}$ approaches a stationary distribution as $t \to \infty$. Under the stationary distribution, which we label $f(g)$, the probability that a stochastic relaxation update of pixel 1 will change its state from $g_1 = a$ to $g_1 = b$ is:

$$f(g_1 = a, g_2, \ldots, g_N) \cdot P(g_1 = b|g_2, \ldots, g_N, Y). \qquad (3.2)$$

Similarly, the probability of the reverse switch—from image $(b, g_2, \ldots, g_N)$ to $(a, g_2, \ldots, g_N)$—is:

$$f(g_1 = b, g_2, \ldots, g_N) \cdot P(g_1 = a | g_2, \ldots, g_N, Y). \tag{3.3}$$

Note that these two probabilities (3.2) and (3.3) are conditional on the rest of the image, $g_2, \ldots, g_N$, and the data Y. Since f is a stationary distribution, the probabilities must be equal:

$$\begin{aligned} \frac{f(g_1 = a, g_2, \ldots, g_N)}{f(g_1 = b, g_2, \ldots, g_N)} &= \frac{P(g_1 = a | g_2, \ldots, g_N, Y)}{P(g_1 = b | g_2, \ldots, g_N, Y)} \\ &= \frac{P(g_1 = a, g_2, \ldots, g_N | Y)}{P(g_1 = b, g_2, \ldots, g_N | Y)}. \end{aligned}$$

This is satisfied if and only if $f(g) = P(g|Y)$ for every image g. (The above argument can be applied to each of the N pixels.)

The updating sequence of stochastic relaxation requires relative conditional probabilities. Using the Markov property of the posterior distribution, we can simplify:

$$P(g_1 | g_2, \ldots, g_N, Y) = P(g_1 | \text{neighbors of } g_1, Y),$$

and so on. This calculation is computationally suited to distributions with small, local neighborhoods.

Output pictures

The limiting distribution of this algorithm is the desired posterior distribution of images; unfortunately, $g^{(t)}$ and $g^{(t+s)}$ are dependent, for any finite s. We want independent samples from $P(g|Y)$; these can be approximated with pictures $g^{(t)}$, for several well-separated values of t, from a single random walk of stochastic relaxation. Geman and McClure (1985) suggest starting with a simple approximate picture, running the algorithm for 25 super-steps (with a super-step defined as an update of all N pixels in haphazard order), and using the outcomes of five successive super-steps: $g^{(26N)}, g^{(27N)}, \ldots, g^{(30N)}$. These are averaged for an estimate of the mean $E(g|Y)$. Presumably, one would need to wait longer between samples to get a good estimate of the variability of g in this distribution.

3.1.5 Simulated annealing

The iterations of stochastic relaxation can be gradually altered so that the algorithm converges to the mode of a distribution. To find the maximum value of $f(g)$, simulated annealing applies stochastic relaxation to the probability density proportional to $[f(g)]^{\gamma}$. The random walk starts at $\gamma = 1$, and gradually increases γ to infinity as the algorithm proceeds. Geman and Geman (1984) prove that if γ increases gradually enough, then the random walk converges almost surely to an absolute maximum of $f(g)$. In practice, they suggest increasing γ faster than the speed at which convergence is guaranteed, and concluding the random walk in finite time at the value $\gamma = \infty$. (At this point, the algorithm steps only to adjacent points of higher probability.) Ripley (1988, p. 99) gives some practical suggestions for speeding simulated annealing.

3.1.6 Estimating hyperparameters

The prior distribution used above typically depends on some unknown hyperparameters which must be estimated from image data. Geman and McClure (1985) estimate their image parameter β by maximum likelihood, using the EM algorithm, averaging over the unobserved image vector g. Their goal is to maximize:

$$\log P(Y|\beta) = \log \int P(Y|g,\beta)P(g|\beta)\, dg.$$

For standard hierarchical models (including this one), $P(Y|g,\beta) = P(Y|g)$, independent of β, and

$$P(g|\beta) = \frac{1}{Z_\beta} e^{-\beta V(g)}.$$

From Dempster et al. (1977) the maximium likelihood estimate is the β that satisfies:

$$E(V(g)|\beta) = E(V(g)|Y,\beta). \tag{3.4}$$

The E step of the EM algorithm corresponds to calculating the expectations in equation 3.4, conditional on a guessed parameter $\hat{\beta}$; the M step solves the equation to yield an updated guess of β.

Example 3.4: 2×2 **table**

The prior distribution $P(g|\beta)$ here is 4-dimensional Normal, so

$$\begin{aligned} P(g|\beta) &\propto \beta^2 e^{-\beta V(g)} \\ \text{and} \quad E(V(g)|\beta) &= -\frac{d}{d\beta}\log Z_\beta \\ &= \frac{2}{\beta}. \end{aligned}$$

To use the result (3.4), we need $E(V(g)|Y,\beta)$. Through stochastic relaxation we can estimate this quantity, to any degree of accuracy, for any fixed value of β. As an estimate, we can use the average value of $V(g)$ from several states in the iteration; e.g.,

$$\overline{V(g)} = \frac{1}{5}(V(g^{(26N)}) + V(g^{(27N)}) + \ldots + V(g^{(30N)})).$$

We can now apply the EM algorithm to estimate β:

1. Start with a guess $\hat{\beta}$.
2. E-step: approximate $\hat{V} = E(V(g)|Y,\hat{\beta})$, by $\overline{V(g)}$ as above.
3. M-step: find a new estimate β^* to solve the equation: $E(V(g)|\beta^*) = \hat{V}$. (For the 2×2 table, this equation yields: $\beta^* = 2/\hat{V}$.)
4. Relabel $\hat{\beta} = \beta^*$ and repeat steps 2 and 3 above.

PET example

Geman and McClure (1985) use EM to estimate the parameter β for a problem that is similar to PET. Following the algorithm presented above, we find that the M-step is more difficult than in the simple Normal example. For PET, $E(V(g)|\beta)$ is no longer a simple formula and must be evaluated numerically. For any fixed β, we can estimate $E(V(g)|\beta)$ by stochastic relaxation on the prior distribution $P(g|\beta)$. To estimate $E(V(g)|\beta)$ as a function of β, we must do this stochastic relaxation for several possible values of β. (For a fixed family of prior distributions, this can be done "off line" and needs not be repeated for each new data set.) For the M-step we must then invert this function to find β^*. As a procedure which yields only a point estimate, this EM application seems unappealingly complicated and slow.

Estimating other hyperparameters

The above PET model also has a scale parameter δ for local differences between intensities g_i in neighboring cells. Geman and McClure (1985) report that results do not vary much with δ values chosen comparable to the level of typical variation between neighboring pixel intensities.

3.2 Comments on Bayesian image reconstruction

This section covers in more generality some of the aspects of Bayesian reconstruction illustrated above. We elaborate on the choices made in the above example, discussing some topics briefly or only with a reference, and indicating what will be covered more deeply in later chapters.

3.2.1 The discrete image space

The Fourier reconstruction discretizes a continuous image of intensities (gray levels) in the frequency domain; pixels partition the spatial domain. In the latter case one defines a continuous image as having constant intensity within pixels. The shape of the pixel grid can be chosen for computational convenience; most work involves square grids, but a circularly symmetric grid requires fewer angle factor calculations for PET reconstructions (Silverman et al., 1990). A grid of hexagons has been suggested for models of images with internal edges (Besag, 1989).

We allow the pixels to outnumber the number of data degrees of freedom. Because pixel selection is arbitrary, we should hope that our reconstructions are not sensitive to the number of pixels, at least if that number is large enough, for any given set of data. In a Bayesian context, these reconstructions depend on a family of prior models: a distribution for each pixellization. The requirement that posterior results do not depend on the pixel grid restricts the choice of these prior families.

3.2.2 The likelihood function

For the PET problem, it is considered fairly accurate to describe the tube counts, corrected for random coincidences, as independently Poisson distributed (Alpert et al., 1982). For computational convenience, these may often be approximated by Normal distributions on the original or the square root scale. Transforming data and image intensities by the square root also stabilizes the variance in the likelihood, which may be desirable (Johnson et al., 1989). Generally, image data are like tomography data in that their likelihood depends on linear functions (such as projections) of the true image. One difficulty of tomography is that its likelihood involves functions of the image whose neighborhoods are tubes, which are not spatially localized.

3.2.3 Image models

This section briefly introduces some of the statistical literature on image modeling. In general, models can be seen as constraints for estimated images or as families of true images for which we desire efficient estimates. In the Bayesian context, an image model is a prior distribution on image space. One family of image models, suggested by Besag (1974, 1986), are conditional autoregressions; these will be discussed fully in Chapters 5 and 6. We will here briefly discuss some other simple models.

Smooth functions

Rather than model an image pixel-by-pixel, we may fit a smooth function onto continuous image-space. Functions such as two-dimensional polynomials may be fit by a relatively small number of parameters. Such a model may be too restrictive to fit an entire image, but may be useful to "detrend" before fitting a pixel-level model or to fit a small, smooth region within the larger image.

Moving average models

By analogy with time series analysis, we can embed pixel intensities in a moving average model by expressing them as different averages of independent underlying parameters. There is a greater variety of such models in two dimensions than in one, as illustrated by the saturated and unsaturated models presented in Section 2.3.4. The reduction in parameters should make estimation of such a model easier. However, estimation of spatially spread-out parameters may cause computational instability, as pointed out by Silverman et al. (1990) in the context of sieve models.

More general autocorrelation models

A model of spatially-indexed Gaussian variables is determined by its two-dimensional variogram (or the covariance function, if the process is stationary; see Section 5.1 for more details). One may explicitly specify these functions (or fit from an empirical variogram or correlogram), or derive them from a conditional probability model. Cliff and Ord (1981, chapter 6) discuss the autocorrelations of some spatial models. These do not generally have simple mathematical forms; for example, a Gibbs distribution, as described in detail in Section 3.1.3 of this paper, is simply defined by local conditional probability distributions, but has a complicated autocorrelation matrix.

More complicated spatial models can be formed by adding the random variables of two simpler models (Besag, 1974, 1989). Posterior properties of the summed model can be evaluated using stochastic relaxation on the two component variables, even if the full model has a complicated correlation and neighborhood structure.

Finally, we can create a spatial model on pixels from a continuous model, as follows. Start with a two-dimensional Fourier power spectrum (for a two- dimensional image) and alias this to obtain a spectrum on the pixel lattice. Then multiply by the spectrum of a two-dimensional moving average at the pixel level, so as to model pixel intensities that are averages within pixel areas, rather than mere samples of the image function. Finally, apply an inverse Fourier transform to yield the homogeneous covariance structure in two dimensions.

Segmentation models

The simplest image distributions are spatially homogeneous, based only on the concept of spatial smoothness. In practice, though, we may wish to model discontinuities, and subdivide an image into disjoint, locally smooth regions. If the locations of these regions are known, we can easily apply an independent stationary distribution within each region. In the more difficult problem, the locations of the regions, or the boundaries between them, must themselves be estimated. We may attack this problem by defining a binary "edge" random variable between every pair of adjacent pixels in the image. If an edge variable equals 1, this defines a boundary, and the adjacent pixels are considered to be in different regions of the image. A Markov model may then be defined on the space of edges and pixels, including rules to ensure that the boundaries form continuous paths; there will still be a local neighborhood structure. Geman and Geman (1984) and Johnson et al. (1989) discuss such models.

Because of poor spatial resolution, it does not seem possible to estimate boundaries directly from PET data. However, it may be possible to use the sharp boundaries of an MRI scan to estimate boundaries for a PET scan on the same patient; the two images can have similar region boundary patterns, but unrelated images within the regions; for example, MRI and PET brain scans of the same person may have this property[1].

[1] Augustine Kong, private communication.

3.2.4 Estimation of the image

In a Bayesian context, one may summarize one's knowledge of the true image by presenting independent samples from its posterior distribution. Different single-image estimates correspond to minimization of different loss functions. The posterior mean image minimizes the sum of squared errors by pixel, while the posterior mode minimizes the all-or-nothing loss criterion over the whole image. One can define more complicated losses, such as all-or-nothing over each pixel, which is minimized by the pixel-by-pixel mode; however, the human observer's criterion of image restoration quality has yet to be captured by any single mathematical loss function.

Stochastic relaxation

As described in the previous section, stochastic relaxation allows one to randomly walk through a posterior distribution, and the related algorithm of simulated annealing converges to a mode image. These methods have been criticized as slow and computationally expensive (Besag, 1986). This may change, however, with the spread of parallel processing, or even careful implementation on ordinary computers (Ripley, 1988, p.95).

Approximate and other methods

Usually presented as alternatives to simulated annealing reconstructions of the mode image, approximate Bayesian methods save computation time and may be more robust to departures from the prior model (Haslett, 1985; Besag, 1986). This robustness may be important in the case of a Markov model that has been specified for computational convenience. One may accept such a model as fitting "local" behavior without trying to use it to fit a whole image. The various local or approximate methods in the literature typically take an iterative algorithm such as stochastic relaxation, subsitituting an approximation for the exact iterative step.

3.2.5 Parameter estimation for an image model

Probability distributions in the statistical image reconstruction literature typically depend on a few parameters, which we will label as a vector θ. For example, in the density:

$$f(g) \propto \exp(-\frac{1}{2\sigma^2}[\sum_i (g_i - \mu)^2 + 2\rho \sum_{\text{neighbors } i,j} (g_i - \mu)(g_j - \mu)]),$$

the three parameters $(\mu, \sigma, \rho) = \theta$ determine the level, conditional variance, and local correlation of the stationary image process. In general, such parameters will not be known and so must be estimated from the data in order to successfully estimate the image g. This can be difficult; for example, the above density function contains a normalizing constant that is a function of the parameters and is the determinant of a $N \times N$ matrix. It is generally computationally impossible to calculate normalizing constants for spatial models; Besag (1977) and Whittle (1954) give approximations for some Gaussian processes on an infinite lattice.

Estimating parameters, given the true image

Methods have been proposed to estimate image parameters when evaluation of the likelihood function is impossible. We first consider the problem of approximately estimating image parameters directly from an image. (This should be easier than estimation from indirect, noisy data.) If the image model is Markovian, Besag (1974) suggests the *coding* method of estimating its parameters, using the conditional independence property:

1. Choose a subset S of pixels that are jointly independent, conditional on the other pixels in the image. For example, if each pixel has four (orthogonally adjacent) neighbors, then a set of half the pixels, forming the black squares of a checkerboard arrry, will be conditionally independent.

2. Estimate the image parameters from the intensities of the pixels in S, conditional on the intensities of the other pixels. This estimation should be easy, because the conditional likelihood is a product of individual likelihoods for pixels in S.

3. Repeat these steps for other conditionally independent subsets of the image grid, and combine the parameter estimates. (In the example of orthogonally-adjacent neighbors, there will be two subsets S to average estimates over.)

Another approximate method is to maximize the *pseudo-likelihood* of the image:

$$\mathrm{PL}(g|\theta) \propto \prod_i P(g_i|\theta,\, g_j, \text{ for all } j \neq i).$$

The normalizing factor of this density is easily calculated as the product of N one-dimensional normalizing constants, and the pseudo-likelihood is then generally easily maximized (Besag, 1975). The pseudo-likelihood, multiplied by a prior density on θ, is proportional to a pseudo-posterior density $\mathrm{PP}(\theta|g)$ if desired. (The parameter θ is typically assigned a noninformative prior distribution.)

Estimating image parameters from data

We now move to the harder task of estimating parameters θ and image g from data Y. We can estimate θ and g simultaneously by applying stochastic relaxation to the $N+1$ variables $(g_1, \ldots, g_N, \theta)$, conditional on Y. This requires simulation from probability densities like

$$P(g_j|\text{rest of image}, \theta, Y) \tag{3.5}$$

$$\text{and} \quad P(\theta|g, Y). \tag{3.6}$$

Sampling from (3.5), conditional on θ, is just step 3 in the stochastic relaxation algorithm described above in Section 3.1.4. To sample from (3.6), note that $P(\theta|g,Y) = P(\theta|g)$; this sampling may be only approximate (e.g., from the pseudo-posterior distribution defined above).

3.2.6 Summarizing uncertainty in estimates

In general, a Bayesian analysis can be summarized by a posterior confidence region containing the possible estimates of highest posterior probability. Unfortunately, such a summary is difficult for a distribution on images defined on more than two or three pixels. Beyond a point estimate, one can use stochastic relaxation to create several samples from the posterior distribution. It is not clear how long the algorithm must run to yield approximately independent samples, for any particular set of data.

For a more direct approach, we can approximate the sampling variability of any image estimate by resampling the data from its estimated sampling distribution and then studying the variablility of the resulting image estimates, one for each resampled data set. We use this method in Chapter 6 and discuss its theoretical validity in Chapter 7.

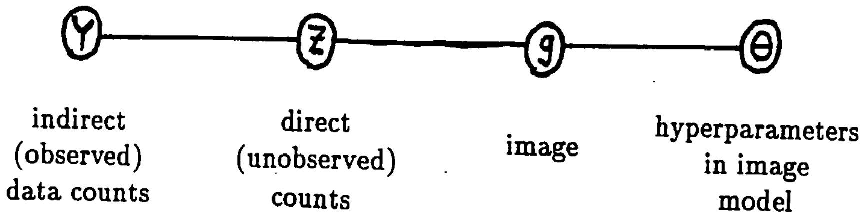
3.3 Implementation of Bayesian methods for PET

The methods suggested in this chapter can be most conveniently implemented for the PET problem by embedding the problem in a hierarchical probability model. This section summarizes this approach, which uses data augmentation to complete the hierarchical model and stochastic relaxation to simulate from the joint posterior distribution. The computer implementation of stochastic relaxation for the PET likelihood is an extension of the EM algorithm introduced in Section 2.3.2 and can be programmed using the methods of conventional (Fourier) reconstruction. We also adapt this algorithm to find maximum posterior density images or maximum likelihood estimates of regional parameters.

3.3.1 The hierarchical probability model

We observe data Y, from an unknown image g, drawn from a prior distribution with parameters θ. To these we add the vector parameter $Z = (z_1, \ldots, z_N)$, where z_j is the number of emissions from pixel j that are recorded in the detector tubes. In the notation of Section 2.3.2, $z_j = \sum_i z_{ij}$. Adding the new parameter Z is called *data augmentation* (Tanner and Wong, Johnson et al., 1989) and places the parameters (Y, Z, g, θ) in a convenient Markovian model with neighborhoods (Y, Z), (Z, g), and (g, θ), as shown in Figure 3.3.1.

Figure 3.1: The hierarchical probability model


Y
Z
g
θ
indirect (observed) data counts
direct (unobserved) counts
image
hyperparameters in image model

Sampling from the posterior distribution

We wish to sample from $P(g|Y)$ and will do so by looking at g in samples from $P(Z,g,\theta|Y)$. This joint distribution can be sampled by stochastic relaxation, in three steps:

1. $P(Z|Y,g,\theta) = P(Z|Y,g)$ is a multinomial distribution, with total sample size parameter $m = \sum_j y_j$, and parameters $m_i = E(z_i|Y,g)$ for the pixels i. These expectations can be easily derived from the result of the E-step of the EM algorithm of Section 2.3.2: $E(z_i) = \sum_j E(z_{ij})$.

2. $P(g|Y,Z,\theta) = P(g|Z,\theta)$ may be simulated by stochastic relaxation for a Markovian model on the components of y, as discussed in Sections 3.1.4 and 3.2.4. In this case, we may think of our larger procedure as stochastic relaxation on $(Z, g_1, \ldots, g_N, \theta|Y)$. The conditional likelihood $P(Z|g,\theta) = P(Z|g)$ is of Poisson form and can be approximated by a Normal density on the square root scale:

$$\sqrt{z_i} \approx N(\sqrt{\bar{\alpha}_i g_i},\ 1/4), \tag{3.7}$$

where $\bar{\alpha}_i = \sum_j \alpha_j a_{ij}$, in the notation of Section 2.3.2. This is conjugate to a Normal prior distribution $P(\sqrt{g}|\theta)$ on the square-root image vector $\sqrt{g} = (\sqrt{g_i})$; in this parameterization, sampling from $P(g|Z,\theta)$ can sometimes be done directly. For example, the posterior mean and variance of a Gaussian conditional autoregression may be calculated by (noniterative) smoothing, as detailed in Section 5.2.2.

3. $P(\theta|Y, Z, g) = P(\theta|g)$ may be approximately sampled from using the pseudo-posterior distribution of Section 3.2.5.

3.3.2 Maximum posterior estimation

The EM algorithm of Section 2.3.2 can be also be adapted to Bayesian maximum posterior estimation. If paramters θ are known, this requires only altering the M-step to maximize the complete-data posterior density function, with the expectations of the E-step substituted in. To obtain the maximum posterior estimate in a problem with unknown hyperparameters θ, we treat both Z and θ as missing data in the EM algorithm, which follows the steps:

1. Start with a guess $\hat{g}$.

2. E-step (likelihood): Calculate $E(Z|Y, \hat{g})$: the E-step in Section 2.3.2.

3. E-step (prior): Determine the function $E(\log P(g)|Y, \hat{g}) = E(\log P(g)|\hat{g})$. The expectation is taken over the unknown θ.

4. M-step (posterior): Find a new estimate, g^*, to maximize the complete-data posterior density with the above expectations substituted in.

5. Relabel $\hat{g} = g^*$ and repeat steps 2–4 above.

We may combine the Normal approximation (equation 3.7) with a conjugate Normal prior distribution $P(\sqrt{g}|\theta)$ by altering the M-step of this algorithm to maximize the (approximate) complete-data posterior density of $\sqrt{g}$ rather than g. The resulting EM algorithm converges to the posterior mode of $\sqrt{g}$.

For examples of easy implementation of the new E-step (prior) and M-step (posterior), see Sections 4.3 and 5.2.2. For another example, the "regional" maximum likelihood estimate, presented just below, can be interpreted as a maximum posterior estimate, given a flat prior distribution on the regional parameters.

3.3.3 Maximum likelihood with regional parameters

We can easily apply the EM iteration to maximum likelihood estimation over the more general regions, in which the image g is parameterized by a vector h of regional weighted averages, as detailed in Sections 2.3.3 and 2.3.4. The new algorithm requires three new steps, which we will present in terms of the PET implementation—that is, the second algorithm of Section 2.3.2. One step comes before the EM iteration, one alters its step *iv*, and one comes at the end of the iteration:

o. Start with a guess of the vector $\hat{h}$ that characterizes the image. Calculate $\hat{g} = W\hat{h}$ (equation 2.10).

iv. Follow step *iv*, but correct for average attenuation using the average attenuation image in regions rather than pixels.

v. Regionally sum g^* to yield $h^* = W^t g^*$. Expand h^* to create the new image estimate $\hat{g} = Wh^*$.

This iteration leads to the maximum likelihood estimate of h, assuming constant intensity within regions.

We can generalize this model by placing a prior distribution on the image intensity within regions. In this case, the steps above can be incorporated into the algorithm of Section 3.3.2 to yield the maximum posterior estimate for regional averages.

Chapter 4

Regularization and Bayesian Estimation for Images

In an *ill-posed problem*, the desired solution to an equation varies greatly as the conditions of the equation are slightly perturbed. The method of regularization yields an approximate solution that is less sensitive to the conditions (Tikhonov and Arsenin, 1977). In the statistical context, a regularized estimator allows some bias (with an imperfect fit to the data) so as to lower its sampling variance. In the PET context, the maximum likelihood estimate may be considered to be the ill-posed solution, to be replaced by a smoother maximum penalized likelihood image. Section 4.1 discusses the application of penalized likelihood to image reconstruction and how this may sometimes, but not always, be considered a Bayesian method. The following section defines the approximate "degrees of freedom" addressed by a set of indirect, noisy data address, and estimate this number for PET. Finally, we discuss implementations of penalized likelihood for PET reconstruction and, in particular, suggest an improvement to the approximate regularization method of running EM and stopping before convergence.

4.1 Maximum penalized likelihood

We confine our discussion to the method of maximum penalized likelihood, which gives an image estimate $\hat{g}$ that maximizes $LL(\hat{g}|Y) - \lambda f(\hat{g})$: the log-likelihood LL of the estimate $\hat{g}$ given data Y, minus the penalty function f, weighted by the regularization parameter λ.

This is of course equivalent to the posterior mode of g, based on the prior density function proportional to $\exp(-\lambda f(g))$. However, the regularization method can be applied in a non-Bayesian context, in which case its penalty need not correspond to any empirically plausible distribution of images.

The penalty function is typically chosen to give the regularized estimate a desired smoothness; for example, locally continuous images are favored by the penalty: $f(g) = \sum_{i,j} c_{ij}(g_i - g_j)^2$, where $c_{ij} = 1$ if pixels i and j are neighbors and 0 otherwise. The parameter λ determines the smoothness of the regularization estimate and, conversely, its fit to the data. As λ varies from 0 to ∞, it indexes a family of solutions, ranging from the maximum likelihood image estimate to the globally least-penalized image (typically, a constant).

Titterington (1985) discusses this method for an image estimation problem identical to ours (except that his sampling distribution is Normal, not Poisson) and gives several examples of penalty functions and their implications.

4.1.1 Choice of regularization parameter

By Lagrange multipliers, maximizing the penalized likelihood $LL(g|Y) - \lambda f(g)$ is equivalent to maximizing:

$$LL(g|Y), \text{ subject to the constraint } f(g) \leq f_0,$$

or minimizing:

$$f(g), \text{ subject to the constraint } LL(g|Y) \geq LL_0,$$

where the constraint values f_0 and LL_0 are determined by λ and the likelihood function. The relation among f_0, LL_0, and λ depends in general on the data, so regularizations with LL_0, f_0, or λ fixed are three different procedures. Specifying λ yields the maximum posterior mode corresponding to the prior density proportional to $\exp(-\lambda f(g))$; specifying f_0 or LL_0 does not in general determine a prior distribution on g independent of data and does not correspond to any posterior mode.

To get one estimate $\hat{g}$, we need one value λ (or f_0 or LL_0), which we may estimate from the data in different ways. If we think of λ as a hyperparameter in a hierarchical probability model, we can estimate it from its (Bayesian) posterior distribution, and then estimate g using the posterior mode (say) of λ (following the steps of Section 3.3 for PET). This two-step procedure can be justified in Bayesian terms as an approximation to the marginal posterior mode of g.

4.1.2 Constraining the log-likelihood

A different approach to setting the regularization parameter is to fix the log-likelihood constraint LL_0. Typically this has been set at its unconditional mean, which asymptotically equals $-n/2$, where n is the number of data points. (We assume the likelihood is scaled to equal 1 when the data equal their expectations.) This has seen some practical use (Gull and Daniell, 1978), but suffers from some theoretical and practical difficulties (Titterington, 1984).

The set of images g for which $LL(g|Y) \geq LL_0$ is a confidence region for g whose size is set by the position of LL_0 in the unconditional distribution of the log-likelihood; for example, setting $LL_0 = -n/2$ gives an approximate 50% region. However, setting LL_0 unconditionally on data Y leads to confidence intervals whose size varies widely with Y, rendering them useless for PET data of moderate sample size. The most obvious problem with a fixed log-likelihood constraint for PET comes from the requirement that the image estimate be nonnegative: for any fixed LL_0, there will be a positive probability of observing data Y such that *no* all-nonnegative image g satisfies $LL(g|Y) > LL_0$. Clearly then, the value LL_0 must be chosen based on the data.

Furthermore, many researchers report that setting $LL_0 = -n/2$ oversmoothes the image estimate. This is no surprise, considering that the usual expected value of residual χ^2 statistic is the number of data points minus the number of parameters estimated. To the extent that the image is being estimated at all, the residual χ^2 statistic $-2LL(\hat{g}|Y)$

should average less than n (Gull, 1989).

Finally, the following example illustrates that no fixed constraint L_0 can even approximately correspond to a fully Bayesian model (that is, a model in which the prior distribution exists independent of the data).

Example 4.1: Estimation from direct Normal data

We consider the problem of estimating the vector X from data Y, under the model:

$$Y_i \sim N(X_i, \sigma^2), \quad i = 1, \ldots, n,$$

with independent prior distributions: $X_i \sim N(\mu, \tau^2)$, and the data variance σ^2 known. The regularization parameter is then $\lambda = \tau^{-2}$. Assume n is large enough so that μ is essentially fixed by the data.

A noninformative prior distribution on τ^2 leads to the posterior mode $\hat{\tau}^2 = \max(0, s_Y^2 - \sigma^2)$, where s_Y^2 is the standard deviation of the data. The maximum penalized likelihood estimate of X, with τ^2 known, has components:

$$\hat{X}_i = \frac{\tau^2}{\sigma^2 + \tau^2} Y_i + \frac{\sigma^2}{\sigma^2 + \tau^2} \mu,$$

with fit to data:

$$\begin{aligned} -2LL(\hat{X}|Y) &= \frac{1}{\sigma^2} \sum_i (Y_i - \hat{X}_i)^2 \\ &= \frac{\sigma^2}{(\sigma^2 + \tau^2)^2} \sum_i (Y_i - \mu)^2. \end{aligned}$$

This has unconditional expectation:

$$E(-2LL(\hat{X}|Y)) = \frac{\sigma^2}{\sigma^2 + \tau^2} n,$$

a value that declines from n to 0 as σ^2 decreases from ∞ to 0 (that is, as data sample size increases for each Y_i). Clearly, no fixed constraint value LL_0 can apply to this whole range of scenarios. In fact, even a constraint such as setting LL_0 to the posterior expectation $E(LL(X|Y)|Y)$ will not work over this range, as is apparent when we note that in this case the sampling expectation $E(LL_0) = E(LL(X|Y)) = -n/2$. This fits the Bayesian model only as $\sigma^2 \rightarrow 0$; that is, if there were no data.

Semi-Bayesian interpretation

We have seen that regularization with a likelihood constraint that is fixed exactly or in expectation does not correspond to any fully Bayesian posterior mode. However, we can interpret likelihood-constrained regularization "semi-Bayesianly," in terms of confidence regions determined by a probability model. (To paraphrase Titterington (1984), we give an interpretation, not a justification, of a method currently in use.)

In the PET problem, the data Y tell us about their expectations X, but nothing about the image g, given X; that is, the likelihood of any image g depends only on the smaller vector X. The semi-Bayesian approach follows the following steps:

1. Define a prior distribution on g, or just on X.

2. Apply Bayesian inference to X, yielding a posterior confidence region for the log-likelihood: $LL(X|Y) \geq LL_0(Y)$.

3. Choose the image estimate $\hat{g}$ that minimizes the roughness $f(g)$, while consistent with the confidence region: $LL(g|Y) \geq LL_0(Y)$.

This is semi-Bayesian in that we use a prior distribution, but only to get a posterior region for X. In fact, this prescription does not require that the roughness penalty in step 3 be proportional to the logarithm of the prior distribution in step 1. Choosing a 50% confidence region in step 2 corresponds approximately to the use of the posterior expectation of LL as a constraint. This in turn is not far from the fixed constraint $LL_0 = -n/2$, which equals the posterior mean on average.

4.2 How many parameters can be estimated?

We now return to the question posed earlier: for moderate sample sizes, how much do PET data tell us? With finer and finer discretization, we can estimate any number N of parameters (pixel intensities) of an image. Of course, the n data points y_i allow us to estimate only n of these parameters; in a Bayesian context, the posterior distribution differs from the prior for no more than n degrees of freedom. In fact, as this section shows, we may not be able to effectively estimate even n parameters from small-sample PET data.

Degrees of freedom for smoothing estimates can be defined in many ways (Hastie and Tibshirani, 1990, chapter 3); for our purposes we define the *degrees of freedom* that data Y supply about the parameter vector ξ to be the expected distance between the prior and posterior means, on a metric weighted by the prior precision:

$$DF = E\left(E(\xi|Y) - E(\xi)\right)^t \mathrm{Var}(\xi)^{-1}(E(\xi|Y) - E(\xi)).$$

The outer expectation averages over the prior distribution for ξ and the sampling distribution for Y. (We assume Normal, or approximately Normal, distributions.) As the sampling variance $\mathrm{Var}(Y|\xi)$ ranges from zero to infinity, DF ranges from n to 0, where n is the length of the data vector Y. PET sample sizes such as appear in our experiment place us somewhere between these extremes, as we discuss after this simple example.

Example 4.2: Direct Normal data

We expand the example in Section 4.1.2 by allowing unequal prior variances: we are estimating the vector X of length n from independent observations Y: $Y_i \sim N(X_i, \sigma^2)$, with independent prior distributions, $X_i \sim N(\mu, \tau_i^2)$. We assume μ, σ^2, and τ_i^2 are known for all i. The posterior mean (also posterior mode for this Normal example) vector $\hat{X}$ has components:

$$\hat{X}_i = \frac{\tau_i^2}{\sigma^2 + \tau_i^2} Y_i + \frac{\sigma^2}{\sigma^2 + \tau_i^2} \mu,$$

We are interested in how much the data influence the estimate $\hat{X}$; that is, how far the posterior mean $E(X|Y)$ is from the prior mean $E(X)$. This distance, on a metric weighted by the prior precision, is:

$$\sum_i \frac{1}{\tau_i^2}(\hat{X}_i - \mu)^2 = \frac{\tau_i^2}{(\sigma^2 + \tau_i^2)^2}(Y_i - \mu)^2,$$

with unconditional expectation $DF = \sum_i \frac{\tau_i^2}{\sigma^2+\tau_i^2}$. Conveniently, the following relation also holds with the maximum penalized likelihood estimate $\hat{X}$:

$$\begin{aligned} E(-2LL(X|Y)) &= \sum_i \frac{\sigma^2}{(\sigma^2 + \tau_i^2)^2} E((Y_i - \mu)^2) \\ &= n - DF. \end{aligned}$$

We would like to know the number of data degrees of freedom in the PET problem; that is, the number of independent image parameters that the data allow us to estimate, at any given data sample size. We find that we can apply the rule $DF = n - E(-2LL)$ to indirect PET reconstruction with a Normal prior distribution. In this discussion, we will assume that the pixels are small enough that they outnumber the data ($N > n$) and we approximate the Poisson by a Normal likelihood; this should cause no trouble in any practical example, such as the experiment at the Massachusetts General Hospital, in which tube counts generally exceed 100.

4.2.1 Bayesian estimate

We first consider PET estimation of an unknown image g^{true}, given a Normal prior distribution on the image g. Such a distribution corresponds to a Normal distribution on the linear transformation $X = R + \alpha A g$ (in the notation of Section 2.3). We can normalize to $\tilde{X}$, with components $\tilde{X}_i = X_i/(X_i^{\text{true}})^{1/2}$, where $X_i^{\text{true}} = (R + \alpha A g^{\text{true}})_i$; the approximate Poisson likelihood has a spherical distribution in $\tilde{X}$. Rotating the prior distribution on $\tilde{X}$ to its principal components reduces the problem to the above example of direct Normal data. The linear transformations we have just done do not affect the posterior mode or the value of the likelihood at that point. Thus, we can calculate DF as follows:

1. Sample image g from its prior distribution.
2. Sample data Y from its sampling distribution, given g. Calculate $-2LL(g|Y)$, which has expectation n.
3. Calculate the posterior mode $\hat{g}$ and its χ^2 statistic $-2LL(\hat{g}|Y)$. Save the difference $-2LL(g|Y) - 2LL(\hat{g}|Y)$.
4. Repeat steps 1–3 and average the results to yield a numerical estimate of $DF = n - E(-2LL(\hat{g}|Y))$.

This theoretical calculation uses only the prior distribution and the likelihood—that is, the structure of the experiment—and no specific data.

4.3 Regularization for PET

We can implement penalized likelihoods for image reconstruction in terms of a prior density function $P(g|\lambda) \propto \exp(-\lambda f(g))$, with a noninformative hyperprior distribution on λ. Application of the EM algorithm of Section 3.3.2 yields the regularized maximum posterior image, which can be given a Bayesian interpretation.

Similar to regularization is the *aborted EM algorithm*—running the EM algorithm of Section 2.3.2, starting with the uniform image, and stopping the iteration before convergence, to yield a smooth image that fits the data (Vardi et al., 1985; Veklerov and Llacer, 1987). As discussed in Section 2.3.2, the maximum likelihood image estimate is typically highly variable, and early stopping regularizes the estimate towards the uniform starting image. Deciding at what point to stop the EM iteration in this approach is similar to the choice of a regularization parameter for maximum penalized likelihood. A serious flaw of this method, at least theoretically, is that the final image depends on the choice of the initial image of the iteration; by starting at the uniform image, we regularize toward uniformity.

This method—running EM and stopping early—can be made more rigorous by replacing it with maximum penalized likelihood with a simple quadratic penalty function: $f(g) = \sum_i(\sqrt{g}_i - \mu)^2$. For a Bayesian estimate, we can then run the EM algorithm of Section 3.3.2 toward a posterior mode of $\sqrt{g} = (\sqrt{g_i})$, based on a noninformative hyperprior distribution on μ, λ. (This parameterization offers mathematical convenience and should have only minor effects on a practical image reconstruction.) The E-step (prior) and M-step (posterior) for this model are particularly simple:

3. E-step (prior): Up to a constant,

$$\log P(g) = -\frac{1}{2\tau^2}\sum_i (g_i - \mu)^2,$$

   and so we must substitute $E((1/\tau^2)|\hat{g})$ and $E((\mu/\tau^2)|\hat{g})$ into this expression. Using a noninformative prior distribution on μ, σ^2 yields:

$$E(\log P(g)|\hat{g}) = -\frac{1}{2s^2_{\sqrt{\hat{g}}}}\sum_i (g_i - \overline{\sqrt{\hat{g}}})^2,$$

   where $\overline{\sqrt{\hat{g}}}$ and $s^2_{\sqrt{\hat{g}}}$ are the mean and variance of the $\sqrt{\hat{g}_i}$ values.

4. M-step (posterior): The effect of the prior distribution is to shrink the estimates obtained by the M-step of the algorithm in Section 2.3.2: for each pixel j, the estimate of component $\sqrt{g_i^*}$ is pulled a fraction $1/(1 + 4\bar{\alpha}_i s^2_{\sqrt{\hat{g}}})$ toward the mean $\overline{\sqrt{\hat{g}}}$.

Chapter 5

Linear Smoothing and Conditional Autoregressive Models

Simple linear smoothers have been suggested for use in image reconstruction procedures (Silverman et al., 1990). In some cases, these are optimal Bayesian estimators under conditional autoregressive (CAR) prior models (Gelman, 1990). This chapter begins with a brief review of CAR models, in the context of Normal stationary and non-stationary spatial processes. We then identify a smoothing step in the EM algorithm as a posterior conditional expectation operator based on a nonstationary CAR prior image model; Bayesian estimates based on a CAR prior distribution can be obtained with the EM algorithm using a linear smoother and the algorithms derived in Chapters 2 and 3. We thus unify two methods—smoothing and Bayesian estimation—designed to improve the overly noisy maximum likelihood image estimate.

5.1 Gaussian spatial processes

A multivariate Normal distribution is a simple probability model on a lattice of real-valued variables, given its second-order properties. The two sections of this chapter introduce the Gaussian spatial model and image estimation within this model.

5.1.1 Conditional autoregressive models

We will consider multivariate Normal distributions on gray levels defined on a square grid of pixels: $g = (g_1, \ldots, g_N)$. As described in Cliff and Ord (1981), any such distribution is defined by its conditional means and variances:

$$\begin{aligned} E(g_i | g_j, \text{ all } j \neq i) &= \sum_j c_{ij}(g_j - E(g_j)) + E(g_i), \\ \text{Var}(g_i | g_j, \text{ all } j \neq i) &= \tau_i^2. \end{aligned}$$

This chapter assumes the homogeneous model in which the unconditional means $E(g_i)$ are equal. We define $c_{ii} = 0$ for all i and label the matrix of coefficients $C = (c_{ij})$ and of conditional variances $T = \text{diag}(\tau_j^2)$. Unless g is defined as the logarithm of image intensity, the distribution may have to be truncated to be positive; we will ignore the difficulties this creates.

This distribution can be thought of as Markovian; its neighborhood structure is determined by the nonzero elements in C: g_i and g_j are neighbors if $c_{ij} \neq 0$. Of special interest are the *first order* and *second order* schemes, in which each pixel is a neighbor of the four and eight nearest pixels, respectively. (Some corrections must be made at the edge of the grid, of course.)

The precision of this distribution is $(I - C^t)T^{-1}$. If $\sum_j c_{ij} = 1$ for all i, then the precision matrix is noninvertible and the model is non-stationary (Cliff and Ord, 1981).

5.1.2 Correlations and semivariograms

The CAR model can be understood through its correlation structure as well as its conditional form given above. We will here consider conditional autoregressions on the infinite lattice, with common residual variance $\tau^2 = 1$ and coefficients that can be expressed as $c_{ij} = c_k$, where k is the (two-component vector) difference in the positions of pixels i and j. This model has spectral density

$$f(\omega) = \frac{\tau^2}{1 - \sum_k c_k \cos(\omega \cdot k)}, \tag{5.1}$$

for $\omega \in (-\pi, \pi] \times (-\pi, \pi]$, as shown by Rosanov (1968). Calculating correlations ρ_k (or semivariograms γ_k in the nonstationary case) for this model is tricky, and is discussed in detail in Besag (1981) and Kunsch (1987). We will summarize their results, along with some other calculations, in this section.

Stationary distributions with low correlations

A CAR model with low correlations will have low coefficients c_{ij}; this section derives an approximation of the covariance using:

$$(I - C)^{-1} = I + C + C^2 + \dots. \tag{5.2}$$

The first-order approximation, valid as $C \to 0$, is $\rho_k \approx c_k$, and can be improved by writing the quadratic term as:

$$C^2 = uI + vC + W,$$

giving a matrix $R = (\rho_k)$ of correlations:

$$R \approx I + \frac{1+v}{1+u} C + \frac{1}{1+u} W.$$

For example, Gelman (1990) considers a second-order scheme with coefficients $c_k = c$ for the eight nearest neighbors. In this case, $C^2 \approx 8c^2 I + 3cC$, and so $\rho_k \approx \frac{1+3c}{1+8c^2} c$ for the eight neighbors, and, to first order, correlations are zero at greater distances.

Stationary distributions with high correlations

As $\sum_k c_k \to 1$, the correlations $\rho_k \to 1$, and the power expansion of $(I - C)^{-1}$ is no longer useful. Instead, the correlations can be calculated by solving the recurrence relation

$$\rho_k = \delta_k + \sum_l c_l \rho_{k+l}, \tag{5.3}$$

where $\delta_k = 1$ if $k = 0$ and 0 otherwise (Spitzer, 1964). Besag (1981) studies the solution of this equation for the first-order neighborhood; calculation is tricky even in the simplest case of constant coefficients.

We can determine the asymptotic form of the correlation as $|k| \to \infty$ by approximating the Fourier transform of the spectral density (5.1). For the isotropic first-order neighborhood, Besag (1981) approximates the long-range correlations by a Bessel function proportional to $|k|^{-1/2}e^{-|k|}$.

Nonstationary distributions

A CAR model is nonstationary if its coefficients c_k sum to 1. Such a model is *intrinsic* of order 0 (this is not the same "order" used to describe neighborhoods) and has a well-defined *semivariogram*, whose properties we briefly review (Matheron, 1973; Kunsch, 1987). The semivariogram γ is a function of the lattice vector k and is defined by:

$$2\gamma_k = \mathrm{Var}(g_i - g_{i+k}), \quad \text{for any lattice point } i,$$

and can be calculated from the spectral density function:

$$\gamma_k = \frac{1}{(2\pi)^2} \int (1 - e^{i\omega \cdot k}) f(\omega) d\omega.$$

For a stationary process, $\gamma_k = (1 - \rho_k)\mathrm{Var}\, g_i$.

For fixed conditional variance τ^2, the semivariogram of a stationary CAR model converges to a well-defined limit as $\sum_k c_k \to 1$. For small lags k, one can calculate the semivariogram for the unit-variance model by integration and the use of the recurrence relation (5.3) with the formal substitution (which can be rigorously justified) of $\rho_0 - \gamma_k$ for ρ_k. Spitzer (1964, chapter 15) discusses the first-order neighborhood example in detail and calculates its semivariogram exactly for small lags and approximately for general lags.

For large lags, Kunsch (1987), based on Spitzer (1964), derives the asymptotic form of γ_k for large lags k:

$$\gamma_k \sim \text{constant} \cdot \tau^2 \log|k|, \quad \text{as } k \to \infty.$$

This constant is easily calculated: it is $2/\pi$ and $4/3\pi$ for the symmetric first and second-order models, respectively.

5.2 Bayesian estimation with a CAR prior distribution

Besag (1974) and others have suggested using intrinsic conditional autoregressions as a family of prior distributions on images. In this section we apply this model to PET and derive its simple implementation with EM computations.

5.2.1 Direct Normal likelihood

We first describe the estimation procedure in the simpler context of direct, Normal data. We assume a data vector $y = (y_1, \dots, y_N)$, that depends on the parameter $g = (g_1, \dots, g_N)$:

$$y|g, \Sigma \sim N(g, \Sigma), \qquad \Sigma = \text{diag}(\sigma_i^2).$$

We will assume a nonstationary CAR model (that is, an improper prior distribution) on g with coefficient matrix C and conditional variance $T = \text{diag}(\tau_i^2)$. (This ignores edge effects in the image grid.) This prior distribution is Normal and Markovian, and so is the likelihood, in the trivial sense that each pixel is its own clique in the likelhood. The posterior distribution is then also Normal and Markovian, with the same cliques as in the prior model; that is, the posterior distribution is a CAR model.

The posterior distribution is stationary, with mean and variance:

$$\begin{aligned} E(g|y) &= (\Sigma^{-1} + (I - C^t)T^{-1})^{-1}\Sigma^{-1}Y, \\ \text{Var}(g|y) &= (\Sigma^{-1} + (I - C^t)T^{-1})^{-1}. \end{aligned}$$

The posterior mean can be written as SY, and the posterior variance as $S\Sigma$, where the smoothing operator S is defined by:

$$\begin{aligned} S^{-1} &= I + \Sigma T^{-1}(I - C^t) \\ &= (I + \Sigma T^{-1})(I - \tilde{C}), \end{aligned} \tag{5.4}$$

and $\tilde{C} = (\tilde{c}_{ij})$ is defined by:

$$\tilde{c}_{ij} = \frac{\sigma_i^2}{\sigma_i^2 + \tau_i^2} c_{ji}.$$

The matrix S is then the covariance for a stationary CAR model with coefficient matrix C' and residual variances $(I + \Sigma T^{-1})$. Because the prior distribution is improper, S acts as a weighted averaging operator; that is, $\sum_j s_{ij} = 1$ for all i.

If σ^2 is comparable or smaller than τ^2, then $\tilde{C}$ corresponds to a CAR model with low correlations, and the first-order approximation for the correlations from equation 5.2 holds. The smoothing operator S is then well-approximated by a local average (Gelman, 1990).

A fully Bayesian estimate using this family of prior models requires estimating or averaging over the prior residual variance τ^2, which can be done using the methods of Section 3.2.5.

5.2.2 PET likelihood

We can generalize the above results to the PET problem by using Normal approximations to some conditional distributions. In the notation of Section 3.3, we wish to estimate the image parameter g in the Markovian model of data Y and unobserved direct image data Z, true image g, and image model parameters θ. For a CAR model, $\theta = (C,T)$, which we will assume known. As discussed in Section 3.3, we can then sample from the posterior density $P(Z,g|Y)$, and thus of $P(g|Y)$, by alternately sampling from $P(Z|Y,g)$ (as described in Section 3.3.2) and $P(g|Z,\theta)$.

Taking advantage of the Normal approximation (equation 3.7), we place a nonstationary CAR model on the transformed image $(\sqrt{g_i})$, which reduces to the problem of Section 5.2.1 with direct Normal observations $y_i = \sqrt{z_i/\bar{\alpha}_i}$ with unequal sampling variances $\sigma_i^2 = 1/(4\bar{\alpha}_i)$. (The vector $\bar{\alpha}_i$ is defined in Section 2.3.2.) If θ is unknown, we can use the stochastic relaxation algorithm of Section 3.3.1.

To get a posterior mode instead, we would run the EM algorithm, calculating for the M-step the conditional posterior mode SZ, with S defined by equation 5.4 as the smoother for a direct Normal likelihood. This M-step calculates g_i^* as in step 3 of the iteration in

Section 2.3.2, and then locally smoothes $\sqrt{g_i^*}$ by S. When treated as unknown, θ be estimated by averaging over its posterior distribution using the EM algorithm of Section 3.3.2.

Silverman et al. (1990) suggest this algorithm—incorporating a smoother in the M step but assuming θ is known—and call it "EMS." Their smoother uses a second-order neighborhood for each pixel and, as described in Section 5.2.1 and the discussion following equation 5.2, approximately corresponds to an intrinsic second-order CAR prior distribution with variance at least comparable to the sampling variance. The new derivation described above places the EMS algorithm in a Bayesian context, with the bonus of showing us how to average over an unknown θ by a simple modification of the iterative algorithm.

Chapter 6

Exploratory Analysis of a Tomography Study

The final chapters of this thesis focus on a study of brain activity using PET, performed at McLean Hospital and Massachusetts General Hospital. The study gave inconclusive results, and we examine it with the goal of designing a more effective future experiment. This chapter gives the medical and technical background for the study and displays and analyzes some of its data.

6.1 Medical background and experimental details

Schizophrenia has been empirically linked to problems with eye movement. To understand how eye movement is controlled by the brain, a PET study was done with five normal patients; as we shall see, this was too small a sample to measure effects of interest. Each patient performed two tasks—*smooth pursuit* and *saccadic eye movement*—with a PET scan of carbon dioxide concentration taken for each task. Elliott (1989) describes in detail the eye-tracking tasks, as well as the rest of the PET study discussed in this chapter.

Each patient and experimental condition yielded a three-dimensional PET scan (each comprising 9 two-dimensional slices) of isotope concentrations; the raw images were transformed nonlinearly, pixel by pixel, into blood flow values. (We ignore the details of this transformation and any questions about its accuracy.) The investigators are interested in the locations in the brain that are activated by smooth pursuit but not by saccadic tracking; for this purpose, each three- dimensional blood flow image is summarized by its average within each of 117 anatomical regions of the brain. The outcome of the PET scans, for each patient, is then the contrast in the blood flow intensity maps between saccadic and smooth tasks. In the language of experimental design, the study is a factorial experiment of regions × patients × tasks, and region × task effects are of primary interest. Substantive prior hypotheses include the belief that the *superior colliculus* is activated by saccadic eye movement more than smooth pursuit. (The superior colliculus is a small region, with a cross-section about 10 mm in diameter in each hemisphere, which contains a sensory map of the visual field.) The experiment was designed to confirm this hypothesis by showing an effect in the superior colliculus. Substantive hypotheses were advanced about other brain regions, which we do not discuss here.

6.1.1 Details of the reconstruction

This study required two steps to transform the PET data into medically useful results: (1) combining coincidence counts with other data to reconstruct individual images of isotope concentration and thus of blood flow; and (2) averaging and comparing results across patients. We will describe these steps in reverse order.

Comparing results across patients

Images from different patients are compared on the space of a *standard brain*. Each person's brain is assumed to be an affine (linear) transformation of the standard brain, with this transformation preserving the relative positions and sizes of the different anatomical brain regions. A "brain atlas" partitions the standard brain into about sixty regions, each of which is assumed to have a common biological function for all normal patients.

As detailed in Section 2.1, each person's brain is reconstructed in nine parallel layers, oriented and positioned based on the location of the head in the scanner. The PET layers are interpolated to the coordinates of the standard brain. Transformation parameters for each person are determined based on measured anatomical landmarks measured from a

CT scan, as described in Senda (1988) and Elliott (1989).

This method is imprecise, for at least two reasons. First, people's brains differ beyond the range of an affine transformation and cannot be so easily transformed into invariant "standard" coordinates. Second, the nine reconstruction layers are far enough apart that some brain regions may be largely missed by the PET measurements, so a three-dimensional image interpolated on the standard brain can be misleading, with scans on different patients missing different parts of the brain. This last effect is somewhat under the experimenter's control. In the study discussed here, the superior colliculus was aligned to be in the center of a PET layer for each patient, ensuring accurate measurement for this region.

Reconstructing a single image

Each two-dimensional layer was reconstructed by the Fourier-based filtered back-projection algorithm, described in Section 2.2, using data from detector pairs and wobble locations (tubes), corrected for estimated attenuation and random coincidences. Section 7.1.1 below discusses the blurring and variance inherent in the filtered back-projection reconstruction. Table 6.1 shows a bit of PET data from a single brain layer of one patient: raw counts, random coincidences, and relative attentuation and exposure time factors for the 13 wobble positions of a typical detector pair. One can look at images of individual slices and at the difference between reconstruction images for two tasks applied to the same patient. As discussed in Section 6.2.2, the visible features of these difference images are largely due to sampling variablity; true region × task effects only show up visually when averaged over several patients.

6.2 Analysis of regional summaries

For the MGH study, each three-dimensional PET scan was summarized by the average blood flow in 117 regions of the standard brain. Blood flow in each region was summarized

Table 6.1: Data from a single PET detector pair

Wobble position	Relative exposure times	Coincidence counts	Expected random coincidences
1	.132	838	220
2	.095	591	159
3	.073	462	122
4	.062	391	105
5	.058	382	96
6	.052	312	86
7	.055	315	91
8	.052	269	86
9	.058	314	96
10	.062	346	105
11	.073	395	122
12	.095	508	159
13	.132	736	220

by the average in a 20.4 mm diameter disk of the transformed images. (The cerebral cortex is approximately 20 mm wide.) Five patients and two tasks yielded a $2 \times 5 \times 117$ data array, with about a tenth of the data missing because the scans did not include all the regions for every patient. Following a logarithmic transformation of the all-positive regional blood flows, we fit a linear model of patient, region, task, and interaction effects.

6.2.1 Analysis of variance

Least-squares estimates of the one-way and two-way interactions were calculated by EM iteration of analysis of variance calculations on observed and missing data (Little and Rubin, 1987)[1]. The results are presented in Table 6.2.

6.2.2 Sampling variances

As described in Section 7.1.2, we simulated ten sets of fake data from the Poisson data of each of the $9 \times 2 \times 5$ two-dimensional PET scans in this study. Reconstructing each image and transforming to the standard brain gave ten independent simulated replications of the

[1] The imputed missing values were perturbed by adding independent errors with the variance of the three-way interaction; the resulting estimated ANOVA was not sensitive to the multiply imputed values.

Table 6.2: ANOVA of log (regional blood flow)

Source	df	SS	$\sqrt{MS}$
Total	1069	83.3	.28
Task	1	1.5	1.2
Patient	4	21.6	2.3
Patient × task	4	.4	.32
Region	116	50.7	.66
averaging hemispheres	74	48.9	.81
diffs between hemispheres	42	1.8	.21
Region × task	116	.16	.038
averaging hemispheres	74	.12	.041
diffs between hemispheres	42	.04	.030
Region × patient	414	13.2	.18
averaging hemispheres	261	10.2	.20
diffs between hemispheres	153	2.9	.14
Region × patient × task	414	.49	.035
averaging hemispheres	261	.36	.037
diffs between hemispheres	153	.13	.030

Notes: (1) Because of missing observations, sums of squares do not sum to "Total." (2) The analysis is on the logarithmic scale, so .030 represents a 3.0% change, etc.

study, which we viewed as approximating the sampling variability. A comparison of the simulated difference images with the real difference images between two tasks for a single patient suggests that the signals of task effects on the image are hard to distinguish from the noise from the Poisson data counts; Section 7.1.2 explains the local correlations in the reconstruction of experimental noise that causes the appearance of spatial clustering in real and simulated difference images.

We summarized each simulated standard brain image by its regional averages, thus creating ten hypothetical replications for each cell in the above analysis of variance. The resulting region × patient × task × simulated replication error has a standard deviation of .026. Improved data collection or image reconstruction should reduce this error, with a lower bound of $\sqrt{.035^2 - .026^2}$ for true patient × region × task variation.

6.2.3 Typical region × task effects

We are interested in the region × task effects—the average task effects across patients, localized by region. These are barely larger than the patient × region × task effects, against which they must be compared in order to detect real differences. Thus, a huge number of patients would be needed to estimate typical effects of interest as summarized in this experiment. The simulated replications above suggest that eliminating experimental noise would approximately halve the sample size needed to discover region × task effects amidst the variability among patients. The region × task effects could also be more accurately estimated in the ANOVA framework if their 116 degrees of freedom were structured into meaningful contrasts.

Region × task effects, averaged over patients, varied from $-.04$ to $.04$ for the 74 regions, averaging over hemispheres. To estimate the magnitude of the true effects, we estimated the region × task effects as Normally distributed random effects, treating between-hemisphere and three-way interactions as independent Normal errors, with the remaining effects and interactions estimated as fixed. The resulting random effects standard deviation estimate was about .01.

6.2.4 Results for one region of interest

The study at hand clearly does not allow accurate estimation of the spatial pattern of average task effects across patients, whether expressed as an image or summarized by brain region averages. However, it may be possible to accurately estimate task effects in a few selected regions of the brain with particularly large task effects or small variance among patients. The superior colliculus is one such region: the experimenters expected a substantial negative task effect here, and they carefully aligned this region in the same place for each patient. This experimental setup is designed to minimize the between-patient variance of effects in the aligned area; to take advantage of this, Elliott (1989) dispensed with the analysis of variance and just estimated the task effect and its standard error for the superior colliculus using data from that region only. The difference, (smooth – saccadic), of the logarithms of average blood flow in the superior colliculus, averaged over the five patients, is $-.027$, with a standard error of only .012, based on four degrees of freedom. By contrast, the standard error from the analysis of variance, pooled among regions, is $.020 = \sqrt{2/5} \cdot .030$. The sample variance of task effects across patients was lower for the superior colliculus then all but one of the other 116 regions. The lower variance in the superior colliculus is consistent with the belief that lining up the region reduced variability in the regional average among patients. the region identically across patients helped. The superior colliculus is more accurately estimated for two reasons: first, alignment within a two-dimensional reconstruction slice allowed PET to better observe this region; and second, consistent alignment for all patients reduced variability of the region × task effect across patients.

6.3 Conclusions from the exploratory study

The analysis presented in this chapter suggest ways in which a future PET study may be improved. First, the random effects analysis shows that region × task effects are mostly around 1.0%, which makes them hard to distinguish in the simple ANOVA framework. Second, the residual region × patient × task variance of $(.030)^2$ appears divided roughly evenly between experimental noise and true variability among patients. The analysis could thus be substantially improved by reducing either of these two variances. The variance of regional averages due to experimental noise may quite possibly be lowered by a better reconstruction algorithm, as discussed in Chapter 7. As illustrated above for the case of the superior colliculus, variance of regional averages among patients can be decreased by careful alignment of the brains of different patients. Extending this improvement to all brain regions requires a more accurate three-dimensional model of the variation in the shapes and structures of patients' brains.

Chapter 7

Bias and Variance of Estimates

Chapters 2 and 3 above presented some methods for estimating tomographic images, and Chapter 6 described how image estimates were used to draw substantive conclusions for a specific experiment. To better link image reconstruction to reality, we need to know how an estimate can be wrong, whether because of sampling variation or because the assumptions underlying the estimate are violated.

An image reconstruction can be thought of as a statistical estimator and be described by its sampling properties. Here we discuss the second-order properties of image reconstructions and estimates of regional averages, theoretically and for the eye-tracking experiment described in Chapter 6. We focus on two estimates described in Chapter 2—the standard method of filtered back-projection and the maximum likelihood estimate of image intensities by region—and then generalize to the Bayes estimates of Chapter 3, which in general estimates regional averages more accurately than the regional maximum likelihood estimate. (Either estimate is of course preferable to the indeterminate or wildly variable maximum likelihood estimate over independent pixels.) From the familiar statistical definition, the bias of an image estimate is the expectation of the pixel-by-pixel difference between the estimate and the true image, averaging over the data, conditional on the true image. For Bayesian models, we may further average over a distribution of true images. This chapter primarily considers the bias and sampling variance of the average image intensities within spatially-localized anatomical regions. We provide new algorithms for computing bias and variance of maximum-likelihood and Bayesian estimates using the projection operators developed in Chapter 2.

The resolution limit from projection data sampled at finitely many projection angles, mentioned in Section 2.2.4, means that no assumption-free consistent estimator exists for an image from PET data: the best that can be done is a blurred estimate such as filtered back-projection. Instead of trying to estimate the continuous image, we follow the methods of Chapter 6 and try to estimate regional averages. Here too, PET data do not allow consistent or even asymptotically unbiased estimates of regional averages without strong assumptions about the image function within regions.

The major causes of bias and variance of estimates of regional averages for PET are: (1) error (or uncertainty) in the locations of regions, (2) limited spatial resolution, (3) incomplete three-dimensional sampling, and (4) sampling variation. First, PET, with its limited resolution, does not identify anatomical regions, and so region locations must be interpolated from a brain atlas, as described in Section 6.1.1. Different patients have differently-shaped brains and cannot all be described by a common brain atlas. Mislocation of brain regions leads to bias in regional averages; uncertainty in region locations yields uncertainty in regional summaries, especially for comparison among patients. In addition a method that uses region locations to obtain more accurate reconstructions, but mislocates the regions, yields a biased estimate of the image itself. Second, the bias in an image estimate caused by limited spatial resolution will propagate to regional averages. Third, the coarseness of the three-dimensional sampling means that some regions may be captured completely by a layer of the PET scan, while other features smaller than 1 cm in diameter may fall between layers. Furthermore, different patients with different brain shapes will be aligned differently in the scanner and so will have different features missed in their brain scans; we will ignore three-dimensional difficulties here. Fourth, the Poisson variation of the data counts will of course lead to sampling variation in any image estimate and summary statistics.

7.1 Filtered back-projection

The filtered back-projection method, described above in Sections 2.2 and 6.1.1, reconstructs the convolution of the true image with a blurring filter and is an approximately unbiased estimate of the blurred image. Mathematically, a nonparametric estimate of a continuous image from indirect PET data must be blurred because of the ill-posedness on the Radon transform inversion problem (Natterer, 1986); the filtered back-projection image is blurred by the frequency cutoff of the reconstruction filter and the lack of correction for detector resolution (Section 2.2.3).

7.1.1 Bias

This section decomposes the bias of regional averages of the filtered back-projection image caused by reconstruction blur and errors or ignorance in region locations.

Blur

Blurring biases estimates of regional averages, especially for small regions (the "partial volume effect" of Hoffman et al., 1979), with consequences for experimental design. We approximate the reconstruction blur by a circular Gaussian kernel with standard deviation 4.5 mm (estimated from the reconstruction of a point source of radiation) and examine the bias of the average image intensity of the superior colliculus, approximated as a disk with diameter 15 mm. The average of the reconstructed image in this disk is a mixture of 55% superior colliculus and 45% neighboring regions.

Averaging within a disk. In general, we can estimate the average of the true image within a region by the average of the image estimate within a disk, centered on the region of interest but not necessarily the same size. Averaging in a larger disk reduces sampling variance, but at the risk of incurring bias (see Section 7.1.2 below), while a smaller disk reduces bias from blurring. The bias due to blurring depends of course on the differences between intensity levels of adjoining regions; this simple analysis is appropriate for the bias of the estimate of task effect for a region surrounded by regions with zero task effects. For example, the estimate in Chapter 6 of the superior colliculus by a disk of diameter 20.4 mm is biased by up to −65%; the true task effect for that region is thus probably about three times as high as estimated by the circular region.

Mislocation of regions

Error in the map of region boundaries does not affect the filtered back-projection image, of course, but does bias estimates of average intensity in regions (Mazziota et al., 1981). Typical mislocations are believed to be on the order of 5 mm; the magnitudes of individual mislocations depend on the region and how a patient is aligned in the scanner. Mislocation of a region by a shift in a random direction affects the bias of the estimate of average region intensity as if it were added to the reconstruction blur.

A simple model of ignorance of the position of a region is that the true region is shifted from its expected position by a rigid displacement, generated by a circularly-symmetric Normal distribution. For the simple model, we can read off the combined bias from reconstruction blurring and region mislocation from Figure 7.1 above. For example, a blur of standard deviation 4.5 mm combined with a mislocation error of standard deviation 5 mm yields a blur of 6.7 mm and thus a bias of −62% when estimating a constant disk of unit intensity with diameter 15 mm surrounded by an image of zero intensity.

If mislocation varies among patients (that is, they are aligned differently in the scanner or their brains are configured differently), then the bias due to mislocation will vary, contributing to between-patient variance of region estimates. In the study described in Chapter 6, careful alignment of patients reduced this source of variance to yield a more precise estimate of task effect for the superior colliculus than for other regions.

7.1.2 Variance

The sampling variance of a PET image estimate comes from the Poisson distribution in the likelihood (see Sections 2.1.1 and 2.3). The filtered back-projection reconstruction ($\tilde{g}$ in equation 2.8) is linear and its first and second moments are thus easily calculated (for example, see Alpert et al., 1982). Furthermore, hundreds of counts are observed in most data tubes, so we can approximate the reconstructed image pixels $\tilde{g}_i$, given the data Y, as jointly Normally distributed.

Pixel variances and covariances can be calculated analytically for an all-constant image, say, using the known reconstruction filter. It is easier, however, to simulate the sampling variation of the data and propagate to the reconstructed image, approximating the unknown Poisson expectations by observed data. The simulated correlations between reconstructed pixel intensities are virtually isotropic, and we display correlation as a function of squared distance (in pixel units) in Figure 7.1. (Pixel dimensions are 2.55 mm × 2.55 mm.) The filtered back-projection is a linear combination of approximately Normally distributed counts, and so the pixel estimates have a covariance structure based on moving averages over tubes. The tubes run across the whole image; however, the contributions of the moving average from any pixel cancel for faraway pixels, and the correlation structure of Figure 7.1 can, to a very close approximation, be generated by the isotropic local moving average process with coefficients displayed in Figure 7.2.

We are interested in regional averages rather than individual pixels. Figure 7.3 displays the variance of the logarithms of averages of clusters of nearby pixels for different cluster sizes, normalized by multiplying by cluster size. The normalized curve levels off for areas larger than 20 pixels (an area of 130 mm^2, corresponding to a disk of diameter 13 mm), which fits the essentially local pattern of noise correlation (Whittle, 1956). The average estimated image intensity within a region of area $|\Omega_k| > 130$ mm^2 has sampling variance approximately inversely proportional to $|\Omega_k|$, for the filtered back-projection PET images of Chapter 6. (The numerical level of variance is not directly comparable to the simu-

Figure 7.1: Simulated noise correlations of a filtered back-projection reconstructed image

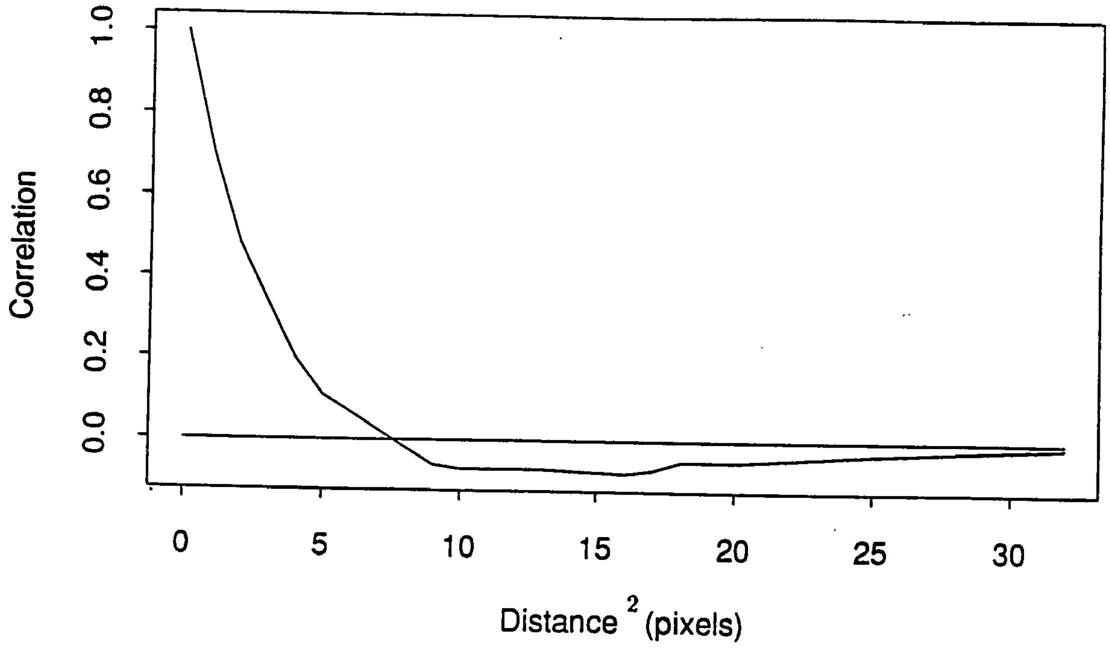


Figure 7.2: Moving average coefficients for reconstruction noise

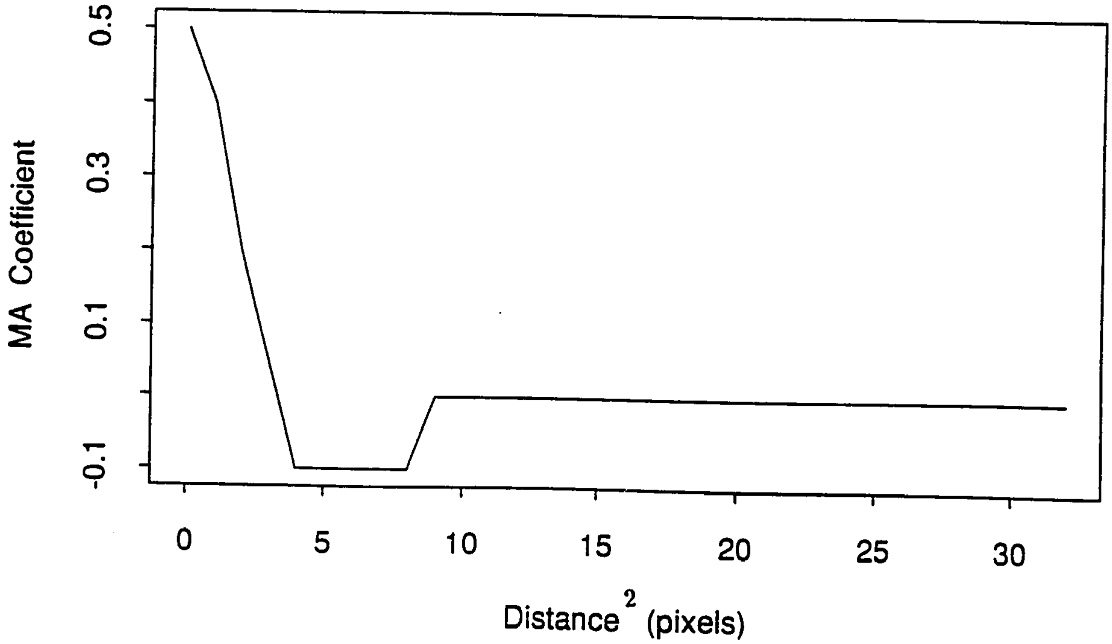
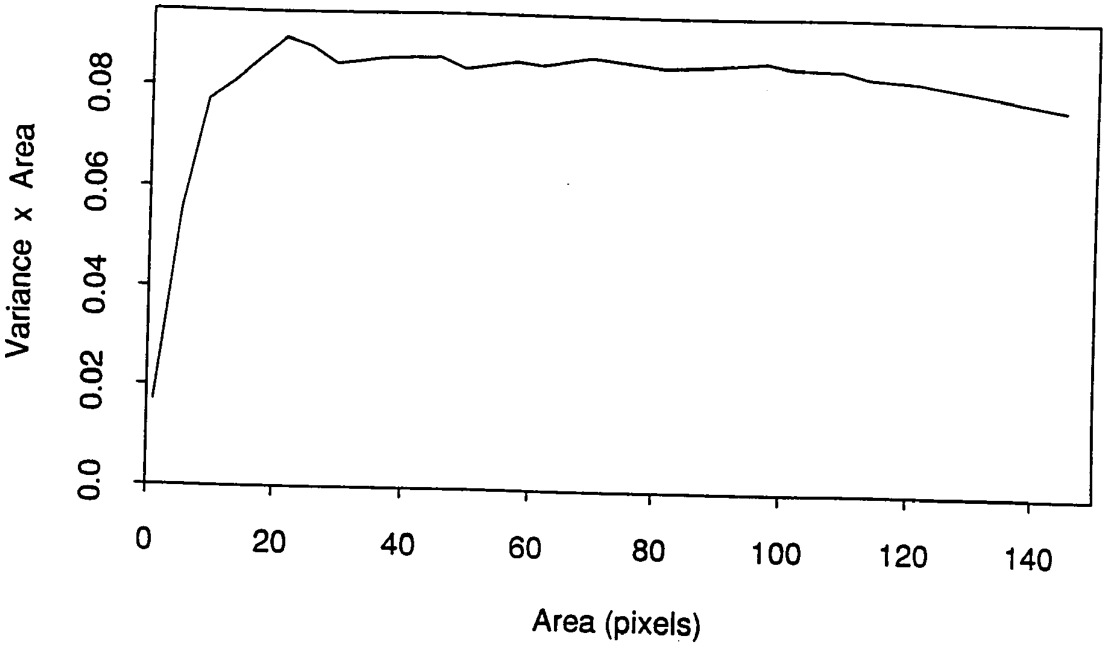


Figure 7.3: Normalized noise variance of reconstructed regions



lated noise variance of Section 6.2.2 because of the nonlinear transformation from isotope concentration to blood flow.)

7.2 Maximum likelihood for regional averages

Estimating one parameter per region should be easier and less variable than estimating the intensities of the potentially infinite number of pixels of a continuous image. This section considers a method described in Sections 2.3.3 and 3.3.3: maximum likelihood estimation of the average image intensities within regions. If the true image were constant within known regions, this is of course an asymptotically unbiased estimate of the 50 or so regional parameters, with no bias from lack of resolution or blurring. Of course, the true image is not constant within regions, and the region locations are not exactly known anyway. The maximum likelihood estimate can still be run *as if* the image were constant within regions, and then the resulting bias can be examined.

7.2.1 Weighted least-squares estimate

In this section we approximate the bias by approximating the maximum likelihood estimate of region intensities by a weighted least-squares estimate; that is, we approximate the Poisson by a Normal likelihood, and ignoring the constraint that the image be all-nonnegative. The Normal approximation is reasonable, considering the hundreds of counts typically observed in each data tube, and the nonnegativity constraint becomes irrelevant in a large sample problem. When estimating parameters for 50 regions instead of 10,000 pixels, a PET data set of 5,000,000 counts in 96 projection angles is large enough. If anything, ignoring the nonnegativity constraint makes our approximation conservative, because the constraint decreases the magnitude of bias.

First we recall and define some notation based on Section 2.3:

g is the vector of pixel intensities g_i,

h is the vector of average intensities h_k of regions Ω_k,

$|\Omega|$ is the vector of region areas (in pixels) $|\Omega_k|$.

If g is constant within regions, then we can write $g = Wh$, where W is a matrix of zeroes and ones: $g_j = h_k$ for $j \in \Omega_k$. In general,

$$h = |\Omega|^{-1} W^t g,$$

and we define the discrepancy δ from a regionally-constant image:

$$\delta = g - Wh.$$

The corrected data vector is $\tilde{Y} = \alpha^{-1}(Y - R)$, where the attenuation factors and expected random coincidence rates α and R are assumed known, and we define Σ to be the diagonal matrix of sampling variances of $\tilde{Y}$, given the true image g. From the Poisson likelihood:

$$\Sigma = \alpha^{-2} \mathrm{diag}(E(Y|g)).$$

We approximate $E(Y|g)$ by Y and assume it is known exactly.

Assuming a regionally-constant image—that is, $\delta = 0$—the corrected data $\tilde{Y}$ satisfy $E(\tilde{Y}) = AWh$, and the weighted least-squares estimate of h from the data Y is:

$$\hat{h} = (W^t A^t \Sigma^{-1} AW)^{-1} W^t A^t \Sigma^{-1} \tilde{Y}.$$

7.2.2 Bias

The weighted least-squares estimate has bias

$$\begin{aligned} \text{Bias}(\hat{h}) &= E(\hat{h} - h|g) \\ &= (W^t A^t \Sigma^{-1} AW)^{-1} W^t A^t \Sigma^{-1} Ag - h. \end{aligned}$$

Substituting $g = Wh + \delta$ yields:

$$\text{Bias}(\hat{h}) = (W^t A^t \Sigma^{-1} AW)^{-1} W^t A^t \Sigma^{-1} A\delta, \tag{7.1}$$

which we will express as $U^{-1}V\delta$. This bias is independent of the regional averages h and is proportional to δ. In contrast, the bias of the filtered back-projection estimate of a region depends on the intensities of neighboring regions and does not vanish when $\delta = 0$.

Computation

We can compute the bias of $\hat{h}$ for any image discrepancy vector δ, using the forward-projection, back-projection, and blurring operations of Section 2.3.2. We must first calculate the square matrix $U = W^t A^t \Sigma^{-1} AW$, of rank equal to the number of regions, and its inverse. To calculate one row of U (that is, u_{kl}, for one region k and all regions l):

1. (Multiply by W) Start with an image that is 1 in each pixel of region k and 0 elsewhere.

2. (Multiply by A) Forward-project and blur, yielding a vector in data space.

3. (Multiply by Σ^{-1}) Divide, component-by-component, by the data vector Y and multiply, component-by-component, by the square of the vector α of attenuation factors.

4. (Multiply by A^t) Blur and back-project, yielding an image vector.

5. (Multiply by W^t) Sum this image over regions and record the vector of regional sums.

U is of moderate rank and can be easily computed and inverted once. We can then compute the bias of $\hat{h}$ by multiplying δ by $V = W^t A^t \Sigma^{-1} A$ (following steps 2–5 above) and then multiplying by U^{-1}.

Averaging over a distribution of images

Equation 7.1 expresses the bias as a function of the unwieldy (and inestimable) parameter vector δ. It is of more interest to examine this bias averaged over various distributions of δ. The resulting mean and variance of bias tells us about the robustness of $\hat{h}$ for different models with $\delta \neq 0$.

For convenience, we place a Normal prior distribution on δ:

$$\begin{aligned} E(\delta|h) &= 0 \\ \mathrm{Var}(g|h) &= T. \end{aligned}$$

(The zero conditional expectation merely asserts that no individual pixel intensity is expected beforehand to stand out as specifically higher or lower than others in its region.) We label T_δ as the variance of δ:

$$\begin{aligned} T_\delta &= \mathrm{Var}(\delta|h) \\ &= (I - W|\Omega|^{-1}W^t)\mathrm{Var}(g|h)(I - W|\Omega|^{-1}W^t) \\ &= (I - W|\Omega|^{-1}W^t)T. \end{aligned}$$

Under this model, the bias of weighted least-squares estimate can summarized by:

$$\begin{aligned} E(\mathrm{Bias}(\hat{h})) &= 0 \\ \mathrm{Var}(\mathrm{Bias}(\hat{h})) &= U^{-1}V(I - W|\Omega|^{-1}W^t)TV^tU^{-1}. \end{aligned} \tag{7.2}$$

With knowledge of T and the number and locations of the regions Ω_k, this variance matrix can be calculated, one row at a time, using the operators above.

Nonuniformity within regions

The two forms of discrepancy from an image constant within known regions yield two kinds of covariance matrix T; we first consider variation within known regions. Nonuniformity can be modeled by the methods of Section 3.2.3 (except for segmentation models); each method can be formulated probabilistically, leading to a family of matrices T which can be substituted into equation 7.2.

Mislocation of regions

The second way we can have $\delta \neq 0$ is by incorrectly locating region boundaries. If the regions are mislocated but the image is constant within true regions, the discrepancy $\delta = g - Wh$ will be piecewise constant within areas of overlap of the true and assumed region maps. We can then use equation 7.1 to calculate the bias of $\hat{h}$ due to a typical mislocation of a given image. Samples of mislocations from a probability distribution, combined with the true image g (or a probability distribution on g) yield samples from a distribution of δ, which equation 7.1 transforms to samples of Bias($\hat{h}$).

Alternately, it may be possible to calculate the variance of the bias directly from equation 7.2. In any case, this bias depends on the true region intensities as well as the true region locations. In general, both nonuniformity and mislocation lead to bias; these effects are additive, so the resulting biases in their weighted least squares estimates add too. We can approximate the covariance matrix T for any simple probablilistic model of uncertain region locations, as follows.

1. Start with a Normal model: given the vector h of region averages, assign pixels i and j a covariance of T^0_{ij} if i and j fall in the same brain region, and zero covariance otherwise. For a homogeneous model, T^0_{ij} depends only on the distance between the

pixels.

2. Consider two pixels i and j. To summarize our uncertainty, we place a probability distribution on the regions in which these pixels fall. The largest effect of this distribution on $\text{Var}(g|h)$ will occur for pairs of pixels located near an uncertain region boundary. Here we will approximate T_{ij} by T_{ij}^0 for all pixels i and j except those near a common region boundary.

3. For pixels i and j in regions Ω_1 or Ω_2, define:

$$p_{11} = P(i,j \in \Omega_1), \quad p_{12} = P(i \in \Omega_1, j \in \Omega_2),$$
$$p_{21} = P(i \in \Omega_2, j \in \Omega_1), \quad p_{22} = P(i,j \in \Omega_2).$$

   From these probabilities and the average image intensities h_1 and h_2 in regions Ω_1 and Ω_2, we can determine $T_{ij} = \text{Cov}(g_i, g_j|h)$. After some algebra, this yields:

$$T_{ij} = (p_{11} + p_{22})T_{ij}^0 + (p_{11}p_{22} - p_{12}p_{21})(h_1 - h_2)^2.$$

4. In practice, this expression simplifies for most pixels i and j:

   (a) If i and j are clearly in different regions, then $T_{ij} = 0$ from the specification of T^0 in step 1 above.

   (b) Similarly, if i and j are clearly in the same region, then $T_{ij} = T_{ij}^0$.

   (c) If i is not near a boundary and is clearly in region Ω_1, say, then $p_{21} = p_{22} = 0$, and $T_{ij} = P(j \in \Omega_1) \cdot T_{ij}^0$.

5. We are simplifying expressions by ignoring pixels near the intersection of three or more regions.

7.2.3 Variance

The sampling variance of the regional maximum likelihood estimate is easily calculated from the weighted least-squares form:

$$\text{Var}(\hat{h}|g) = U^{-1}$$

$$= (W^tA^t\Sigma^{-1}AW)^{-1}. \tag{7.3}$$

7.2.4 Comparison with filtered back-projection

Like the weighted least-squares estimate, the filtered back-projection image is a linear multiple of the corrected data vector, allowing simple algebraic expression of its first two moments. The results may be compared with equations 7.2 and 7.3 for the regional maximum likelihood estimate. In the notation of equation 2.8, the filtered back-projection reconstruction is $\tilde{g} = P^tF\tilde{Y}$; the vector of regional averages is then $\tilde{h} = |\Omega|^{-1}W^tP^tF\tilde{Y}$, with bias and variance:

$$\begin{aligned}\text{Bias}(\tilde{h}) &= |\Omega|^{-1}W^tP^tF(AWh+\delta) - h \\ \text{Var}(\tilde{h}) &= |\Omega|^{-1}W^tP^tF\Sigma F^tPW|\Omega|.\end{aligned}$$

The bias vector and covariance matrix are be easily calculated (given h, δ, and Σ), using the forward-projection operator P, the back-projection operator P^t, and the filter F.

7.3 Bayesian estimates

The maximum likelihood estimate of the vector h of regional averages can be improved by generalizing the assumption of constant intensity within regions to allow a probability distribution on the pattern of the intensities within regions. Our Bayesian estimate is then the maximum likelihood estimate of h, averaging over the distribution of the nuisance parameter δ. (We implicitly assign a noninformative prior distribution to h.) This section shows that the Bayesian estimate, even with the wrong prior distribution, is better than the regional maximum likelihood estimate under the stronger wrong assumption that the image is constant within regions; this result leads to bounds on the accuracy of the Bayesian estimate of h.

7.3.1 Weighted least-squares estimate

Following the procedure of Section 7.2.1 above, we approximate the Bayeseian estimate by a weighted least-squares estimate h^* based on a Normal likelihood and prior distribution.

As before, the approximating Normal likelihood is defined by:

$$\begin{aligned} E(\tilde{Y}|h,\delta) &= Ag \\ &= AWh + A\delta \\ \text{Var}(\tilde{Y}|h,\delta) &= \Sigma. \end{aligned}$$

We assign the prior distribution:

$$\begin{aligned} E(\delta|h) &= 0 \\ \text{Var}(\delta|h) &= T_\delta. \end{aligned}$$

Combining, and thus averaging over δ:

$$\begin{aligned} E(\tilde{Y}|h) &= AWh \\ \text{Var}(\tilde{Y}|h) &= \Sigma + AT_\delta A^t. \end{aligned}$$

The weighted least-squares estimate of h, averaging over δ, is:

$$h^* = (W^tA^t[\Sigma + AT_\delta A^t]^{-1}AW)^{-1}W^tA^t[\Sigma + AT_\delta A^t]^{-1}\tilde{Y}.$$

As in Section 7.2.1, it is fruitful to evaluate the sampling properties of this estimate, averaged over some model for δ. We distinguish the modeled prior distribution (7.4), which is used to obtain the Bayesian estimate, from the "true" prior distribution of δ, over which we will average. In general, these two distributions will be different.

We can compare the Bayesian bias and variance to the results (7.2) and (7.3) for the estimate assuming the image is constant within regions ($\delta = 0$). Of course, the Bayes posterior expectation has the lowest model-averaged mean squared error, assuming the model is true. The Bayesian estimate shrinks pixel intensities partly toward local and regional averages, whereas the estimate assuming $\delta = 0$ shrinks pixel intensities all the way to regional averages.

7.3.2 Correct prior distribution

We first consider the simpler case in which the modeled prior distribution is the truth. In this case, the mean squared error of the weighted least-squares estimate, averaged over δ, is simply:

$$\mathrm{mse}(h^*) = (W^tA^t[\Sigma + AT_\delta A^t]^{-1}AW)^{-1}. \tag{7.4}$$

This can be decomposed into bias and variance:

$$\begin{aligned}
\mathrm{Bias}(h^*) &= (W^tA^t[\Sigma + AT_\delta A^t]^{-1}AW)^{-1}W^tA^t[\Sigma + AT_\delta A^t]^{-1}A\delta \\
&= U_*^{-1}V_*\delta && (7.5) \\
E(\mathrm{Bias}(h^*)) &= 0 \\
\mathrm{Var}(\mathrm{Bias}(h^*)) &= U_*^{-1}V_*T_\delta V_*^tU_*^{-1} && (7.6) \\
\mathrm{Var}(h^*|g) &= U_*^{-1}W^tA^t[\Sigma + AT_\delta A^t]^{-1}A\Sigma A^t[\Sigma + AT_\delta A^t]^{-1}AWU_*.
\end{aligned}$$

Unfortunately, these are hard to compute directly, because they require the inversion of the large matrix $[\Sigma + AT_\delta A^t]$. For simple patterns of T_δ, the inverse may be calculated by Fourier methods or approximated.

7.3.3 Incorrect prior distribution

If the modeled prior distribution is wrong, we are interested in the bias and variance of the Bayesian estimate based on the model $P(\delta)$, averaged over the true distribution $P^{\mathrm{true}}(\delta)$.

The sampling variance is unchanged from equation 7.6, no matter what the true prior distribution is, since we assume Σ is known. (In any case, perturbations of the prior distribution should not perturb Σ enough to seriously affect the sampling variance formula.) For simplicity we assign a Normal distribution to $P^{\mathrm{true}}(\delta)$:

$$\begin{aligned}
E(\delta|h) &= 0 \\
\mathrm{Var}(\delta|h) &= T_\delta^{\mathrm{true}}.
\end{aligned}$$

From equation 7.5, the true bias has mean and variance:

$$\begin{aligned} E(\text{Bias}(h^*)) &= 0 \\ \text{Var}(\text{Bias}(h^*)) &= U_*^{-1}V_*T_\delta^{\text{true}}V_*^tU_*^{-1}. \end{aligned}$$

In summary: with the right model, the Bayesian method shrinks δ the optimal amount towards zero; with the wrong model, it is still better than the result obtained using extreme assumption $\delta \equiv 0$. The mean squared errors of the correct Bayesian estimate h^* and the estimate $\hat{h}$ assuming a constant image within regions thus give a lower and upper bound for the mean squared error of the Bayesian estimate with a misspecified prior distribution $P(\delta)$.

7.4 Limit of infinite sample size

It is interesting to follow the bias and variance as the number of data counts approaches infinity; that is, as the elements of the diagonal matrix Σ approach zero while the rest of the parameters remain unchanged. The sampling variance $\text{Var}(h^*|g)$ is proportional to Σ and thus approaches zero in this limit for any estimate; this makes sense, as we are estimating the finite-dimensional parameter h.

In contrast, the bias of region estimates declines but does not approach zero as data counts increase, because of the resolution limit from the finite PET sampling. Even with infinite counts, the discrete data Y do not allow precise estimation of the continuous variable δ. In filtered back-projection, bias comes from the reconstruction blur; the maximum-likelihood region estimates are biased because $\delta \neq 0$ in general; and the Bayesian reconstruction is biased because δ cannot be perfectly estimated, even given the correct prior distribution.

In the limit of infinite data counts, the mean squared error of the maximum likelihood estimate $\hat{h}$ approaches the expected squared bias. From equation 7.2, this is:

$$\inf \text{mse}(h^*) = u^{-1}VT_\delta V^tU^{-1}.$$

Now consider the Bayesian estimate with correct prior distribution. To let data counts approach infinity, we set $\Sigma = 0$ in equation 7.4 and the mean squared error approaches the nonzero limit:

$$\inf \mathrm{mse}(h^*) = (W^t A^t [A T_\delta A^t]^{-1} A W)^{-1},$$

a bound on how accurately h can be estimated, even with infinte data counts and with δ sampled from a known prior distribution.

In the limit of continuous PET data at a continuous range of angles, we find that δ can be estimated, the matrices A and A^t become invertible, and the mean squared error of the Bayesian estimate (with known prior distribution) approaches zero for infinite sample size. The dependence of the accuracy on sample size varies with assumptions about the image—in our model, the prior variance T_δ. Johnstone and Silverman (1990) discuss the accuracy of estimation under a different parameterization of the local variation of the image.

Chapter 8

Image Reconstruction and Experimental Design

This concluding chapter discusses how to apply the statistical theory developed earlier in the thesis to the design and analysis of PET experiments. Section 8.1 discusses practical implementation of statistical methods to real PET data; the difficulties begin with, but are not limited to, the random coincidence and attenuation effects handled in Chapter 2. Understanding of anatomical regions within the brain is key to the analysis and sometimes to the estimation of PET images. Section 8.2 introduces probability models to describe uncertainty and variation in the spatial patterns of brain regions of different patients. Section 8.3 connects image estimation to experimental design, focusing on sources of variation in PET reconstruction and analyis, with an example of the analysis required for the statistical design of a medical imaging study. We conclude with a review of hierarchical Bayesian estimation for medical imaging as an example of estimation from indirect, noisy data.

8.1 Applying likelihood-based methods to real PET data

As discussed in Chapters 2 and 3, maximum likelihood and Bayesian estimates have the advantage of yielding all-nonnegative images using the known Poisson distribution of the data, and their bias and variance can be calculated as described in Chapter 7. Furthermore, maximum likelihood estimates with parameter restrictions (Sections 2.3.3

and 2.3.4) and Bayesian or regularized smoothing estimates avoid the noisy appearance of the simple maximum likelihood estimate of pixel intensities. Finally, the Bayesian approach allows easy incorporation of prior knowledge (for example, about the spatial structure of a brain) and automatically summarizes posterior uncertainty.

Implementing the methods of Section 2.3 and Chapters 3–5 for PET requires computation, data handling, image modeling, and image analysis; the complications of each of these four tasks are described below.

Computation

The core of any maximum likelihood or Bayesian estimate for PET is the iterative EM or stochastic relaxation algorithm, either of which is based on the iterative procedure (steps *i–v*) of Section 2.3.2. The forward and back-projection operators required for this algorithm can easily be adapted from the computer code for the filtered back-projection algorithm and requires no additional storage. One iteration of the EM algorithm (or of stochastic relaxation) takes about twice as long as a back-projection. Stearns (1990) suggests a fast method based on Fourier transforms and interpolation to replace filtered back-projection; it should be similarly possible to speed the steps of the EM algorithm using Fourier transforms. Finally, the blurring operator required for an EM reconstruction of finite-resolution PET data is a convolution that is easily implemented directly or by Fourier methods.

PET data

Chapter 2 discusses the corrections for random coincidences and attenuations, which are the major nuisances of PET data and require alterations in published statistical reconstruction algorithms. In addition, the corrections themselves are noisy, which affects the likelihood in tubes with near-zero counts, requiring special care for these data points. Another common practical problem, missing data, is luckily easy to handle exactly in the EM or stochastic relaxation framework. In addition, the model may be expanded to anticipate problems such as defective or mislocated radiation counter. Finally, one must look at the fit of a model to real data; for PET, one can compare this fit—perhaps expressed as a log-likelihood ratio or a χ^2 statistic—to the fit of the model to a similar set of data simulated from an image estimate. If the fit to real data is much worse, then the model may be ignoring a source of error or uncertainty. Implementation of likelihood-based methods has revealed a lack of fit of the Poisson model to the raw data of the study described in Chapter 6; further study is needed to explain this departure. Unfortunately, real PET data are unwieldy—even a two-dimensional data set from the experiment of Chapter 6 comprises 20,000 tubes and 5,000,000 counts—and so it is difficult to search for recording errors or patterns of outliers.

Image modeling

Image estimation based on homogeneity within brain regions requires knowledge of region locations and boundaries. The simplest example of applying such knowledge is to set the estimated image intensity to zero outside the body. Unfortunately, boundary information is hard to use even when it is known, particularly because of the difficulty of storing—in books or computers—three-dimensional information about spatial structure. For example, labeling the cross-sections of regions in a two-dimensional slice of arbitrary orientation in the brain is difficult practice if given only an atlas of brain regions stored two-dimensionally. The task of identifying brain regions is further complicated by the variation in brain shapes among patients and the inaccuracy of measuring brain anatomy on individuals. Finally, the connection between anatomical regions and brain physiology is not well understood and it may be risky to assume such knowledge in estimating a PET image—after all, one of the purposes of PET is to learn about this connection.

Image analysis

Brain region locations are needed for analysis and summary as well as estimation of medical images, and so variation and uncertainty in spatial structures complicates the analysis stage of a PET study. In addition, the isotope concentrations estimated by PET must be nonlinearly transformed into the physiological variables of interest; inaccuracy in the transformation further degrades estimates of task effects.

8.2 Spatial models of brain structure

Brain structure is the key connection between image estimation and the design and analysis of experiments such as that described in Chapter 6. As detailed in Elliott (1989), the brain is conventionally divided into about 50 anatomical regions in each hemisphere; the "standard." locations and boundaries of these regions are recorded in a brain atlas, but there is of course variation in the size and shape of different brains. (Indeed, an individual brain changes shape as a person ages.) We might hope to estimate region boundaries from PET, perhaps through image segmentation modeling (Section 3.2.3); however, the resolution of PET is too poor to allow accurate estimation of sharp boundaries in a practical setting. We would like to know the location of the regions for each patient under study, or at least know our uncertainty about the region locations.

Information about anatomical brain regions comes from physical skull measurement and CT and MRI scans (see Section 1.2). As described in Section 6.1.1, each individual brain in the study was affinely (linearly) transformed to the standard brain by matching physical *landmarks*—objectively identifiable positions in the bone structure—from CT scans to the brain atlas. To the extent that the variety of brain shapes cannot be summarized by a linear family, these transformations lead to inefficient and biased estimates of regional effects in a study with many patients.

8.2.1 Deformable templates

We sketch a general family of models of spatial structure called *deformable template* models because they fit deformations of a pattern, or template, of brain region boundaries (Chow et al., 1989). In the PET context these models are transformations from the atlas in the standard brain to region locations in real brains. For simplicity, this section describes two-dimensional models; three-dimensional extensions are straightforward.

Global and local models

We first distinguish between two classes of models of spatial variation: a *global* transformation is deterministically described by a few parameters and so follows the global pattern of the template (brain atlas). Like a global specification, a *local* shape model may be determined by only a few parameters; however, the shape under a local model is only probabilistically constrained and thus has the freedom to fit local data. (A local model may be thought of as a tensor field; Goodall, 1983.) One can fit a global model to the overall size and shape of the brain and then apply a local model to allow for variation among region boundaries in true brains relative to the template pattern.

The affine transformation used for Chapter 6 is an example of a global transformation, mathematically equivalent to a general mapping of the labeled vertices of a triangle. The six scalar parameters of the affine transformation in two dimensions are then identified as two for shape, one for scale, two for translation, and one for rotation. More complicated global transformations are possible; for example, the *bilinear* mapping transforms the labeled vertices of a quadrilateral; four of its eight parameters alter the shape of the transformed object. The affine transformation transforms all interior regions similarly and preserves straight lines; the bilinear does neither (Bookstein, 1986).

Local Markovian shape models

A family of global models that successfully fit the anatomy of real brains, relative to the brain atlas, would be of great use to medical imaging as well as of substantive interest in its own right. Until this happens, however, it seems necessary to fit flexible local models to the variation of spatial structure among brains. There seem to be two ways to construct such models: by transforming the space on which the regions are defined (possibly using a discrete lattice for computational purposes) or by directly transforming elements of the boundaries between regions. Local Markovian models can be applied naturally in either parameterization; to be specific we outline here an example of a local model of region boundaries.

Start with a template—a map of region locations and boundaries in a brain slice. Approximate the region boundaries by polygons and define the log-length l_i^0 and orientation ϕ_i^0 of each directed edge i in the template. We will model the possible interlocking patterns of region boundaries by deforming the edges of these polygons in the template. The local model allows the boundaries to vary under the following constraints: the outside boundary of the template is fixed (it can be fit with a global model); the two-dimensional topology of the network of edges and labeled regions is fixed; and boundaries cannot intersect. For a real map of region boundaries—a deformed template—we define random variables l_i and ϕ_i for the log-lengths and orientations of edges:

$$\begin{aligned} l_i &= l_i^0 + \delta l_i \\ \text{and} \quad \phi_i &= \phi_i^0 + \delta\phi_i, \end{aligned}$$

and place a Gibbs distribution (equivalent to a Markov model) on the deformations (δl_i) and $(\delta\phi_i)$ from the template:

$$P(\delta l, \delta\phi) \propto \prod_{\text{neighbors } i,j} \exp(-U_l(\delta l_i, \delta l_j) - U_\phi(\delta\phi_i, \delta\phi_j)). \tag{8.1}$$

For simplicity, define the edges i and j to be neighbors if they share a vertex and set the potential functions U_l and U_ϕ to be Gaussian with two parameters each to describe variance and local correlation. (The mean deviation from the template in the local model is naturally zero.)

The probabilities (8.1) must be set to zero for those cases forbidden by the previous paragraph, such as intersecting boundaries. Due to these restrictions, this local model is not truly Markovian (or Gibbsian); the approximation may be close for reasonable choices of parameter values in U_l and U_ϕ, however.

Computer simulation

The restrictions of equation 8.1 and the paragraph preceding it define a probability model on maps of region boundaries and can be visualized through the simulation. Unfortunately, the density (8.1) is unnormalized and highly multivariate and cannot be directly sampled. However, an algorithm exists for simulating a *random walk* through a probability distribution, requiring only calculation of a function proportional to the probability density function (Metropolis et al., 1953). This application of the *Metropolis algorithm* loops through the interior vertices of the network of region boundaries, altering one vertex at a time. Like the stochastic relaxation algorithm described in Section 3.1.4, the Metropolis algorithm is computationally well-suited to a Gibbs distribution, requiring local calculation of ratios of conditional probabilities at each step. By altering only interior vertices in simulating the probability distribution of edges, the iterates of the algorithm automatically satisfy the constraints of a fixed outside boundary and unchanged network topology. The further constraint of non-intersecting boundaries can be satisfied by running the Metropolis algorithm and picking out of its stream of iterates those region maps that satisfy the constraint.

8.3 Experimental design and comparison of reconstruction methods

Now that we have some idea of the difficulties of statistical modeling and estimation of PET images, we explore the applied problem of interest: the design and analysis of medical imaging experiments. The design of a study such as that described in Chapter 6 involves parameters such as the number of tasks, patients, and replications, true region × task effects, and various components of variance. The main products of the analysis are summaries of region × task contrasts, averaged over patients and structured by regions (by hemisphere, function, anatomy, etc.). The following sources of error influence the experiment:

- **True region × patient × task variation** is the limiting factor on the precision of such study of region × task effects with a fixed number of patients.
- **Variation in spatial brain structure among patients** (discussed in Section 8.2 above) affects the bias of region estimates (Chapter 7) and, perhaps more importantly, the effectiveness of averaging region estimates among patients.
- **Limited resolution of the PET scan,** a concern even with infinite data counts, blurs image estimates or makes them subject to untestable assumptions. In either case the resulting region estimates are biased (Chapter 7), and there is reduced statistical precision for estimates of region × task effects, especially for small regions.
- **Sampling error** due to finite data (discussed in Section 2.3) is of course the final level of variation in any experiment.

8.3.1 Example of an analysis for experimental design

This section considers a hypothetical future study based on the actual study presented in Chapter 6, but with replications on each patient and task. Assume an anlysis of logarithms of regional averages mimicking that presented in Section 6.2; then experimental design requires the expression of region × task and residual region × task × patient variances in terms of modeled variance components. In addition one would want region × task estimates and variances among patients for selected regions of interest such as the superior colliculus. The following simple model illustrates how the experimental design could begin.

Model and definitions

Assume n_{pat} patients (indexed by p), two tasks, a fixed number of regions (indexed by r), and n_{rep} replications (indexed by j). Label the following true effects, all on the log scale:

$$\begin{aligned}\text{region effects (averaging the two tasks)} &\quad A_r,\\ \text{region} \times \text{task effects} &\quad B_r,\\ \text{region} \times \text{task} \times \text{patient effects} &\quad B_{pr},\\ \text{region} \times \text{task} \times \text{patient} \times \text{replication effects} &\quad B_{prj}.\end{aligned}$$

It is a plausible approximation that the regional average value A_r is constant across patients and replications. It is reasonable to model region × task effects as fixed for some regions of interest and random for the others, and to model higher-level interactions as random effects:

$$\begin{aligned} B_{pr} &\sim N(B_r, \nu^2),\\ B_{prj} &\sim N(B_{pr}, \tau^2).\end{aligned}$$

The variance τ^2 should increase with the time interval between replicates.

Summary of data and image estimates

Two-dimensional images will be estimated using filtered back-projection; assume that the true images of isotope concentration are constant within regions. Ignore three-dimensional effects and the nonlinear transformation from isotope concentration to blood flow. Regional average blood flow is then summarized, as in Chapter 6, by a circular average of the reconstructed image; the diffference y_{prj} in the logarithms of these circular averages between the two tasks is one "data point" for subsequent analysis. Make the reasonable assumption that the data are Normally distributed:

$$y_{prj} \sim N(c_{prj}, \sigma_r^2),$$

with variances σ_r^2 that depend on the region r and may be lower for regions, such as the superior colliculus in Chapter 6, that are carefully positioned to lie wholly within a PET slice.

The expected circular averages c_{prj} are biased estimates of the true task effects B_{prj}, because of ignorance of exact region locations and blurring of the PET image estimate. Express this bias as f_{pr}: the fractional contribution of region r to the circular average of the reconstructed image estimating region r. Unless a region is precisely localized in the same position in the PET scan for all patients, f_{pr} will vary with p. (Section 7.1.1 estimates f_{pr} to be 35% for the superior colliculus in the study described in Chapter 6.) Transforming to the log scale yields the approximate result:

$$c_{prj} = f'_{pr} B_{prj} + (1 - f'_{pr}) \tilde{B}_{prj},$$

where f'_{pr} is defined by:

$$\mathrm{logit}\,(f'_{pr}) = \mathrm{logit}\,(f_{pr}) + A_r - \tilde{A}_r,$$

$\tilde{B}_{prj}$ is a weighted average of the effects B_{psj} in regions s neighboring r, and $\tilde{A}_r$ is defined analogously.

Estimates of task effects

Again following the straightforward methods of Chapter 6, the region × task effects of interest are obtained by averaging summaries of image differences over patients and replications. We will decompose the variance of this summary step by step, starting with the region × patient × task summary $y_{pr\cdot}$, averaging over replications:

$$y_{pr\cdot} \sim N(c_{pr\cdot}, \sigma_r^2/n_{\text{rep}}), \tag{8.2}$$

with the subscripted dot meaning "average over replications." Also,

$$c_{pr\cdot} = f'_{pr} B_{pr\cdot} + (1 - f'_{pr}) \tilde{B}_{pr\cdot}. \tag{8.3}$$

The average $B_{pr\cdot}$ has mean B_{pr} and variance τ^2/n_{rep}; the mean and variance of $\tilde{B}_{pr\cdot}$ should be closer to zero than its components, as it is an average over several neighboring regions. Approximate the mean to be zero and the number of neighboring regions by five; combining equations 8.2 and 8.3 yields:

$$y_{pr\cdot} \sim N(f'_{pr} B_{pr}, [f'^2_{pr} + \tfrac{1}{5}(1 - f'_{pr})^2]\tau^2/n_{\text{rep}} + \sigma^2/n_{\text{rep}}). \tag{8.4}$$

Assume that all the coefficients f'_{pr} are known; then a linear and approximately unbiased estimate of the region × task effect B_r is

$$\hat{B}_r = \frac{\sum_p y_{pr\cdot}}{\sum_p f'_{pr}}.$$

Further calculation, proceeding from expression (8.4), leads to an approximate form for the sampling variance of the region × task estimate:

$$\begin{aligned} \operatorname{Var}(\hat{B}_r) &= \frac{1}{n_{\text{pat}}} \left((vf)_r^2 B_r^2 + \left[1 + (vf)_r^2\right] \nu^2 + \right. & (8.5) \\ & \left. + \left[1 + \frac{1}{5}\left(\frac{1 - f'_r}{f'_r}\right)^2 + \frac{6}{5}(vf)_r^2\right] \frac{\tau^2}{n_{\text{rep}}} + \frac{1}{f'^2_r} \frac{\sigma^2}{n_{\text{rep}}} \right). & (8.6) \end{aligned}$$

where f'_r and $(vf)_r$ are the average and coefficient of variation of the fractions f_{pr} over the n_{pat} patients.

The four terms on the right side of equation 8.5 correspond to four sources of variance in the medical imaging experiment. The first term, driven by $(vf)_r^2$, represents variance among patients in the location of region r. The second term is proportional to ν^2—the variance among patients of true task effects. The third and fourth terms are proportional to the replication variance τ^2 and the sampling variance σ^2, respectively, and also depend on the variables $(vf)_r^2$ and f'_r.

Experimental design

The immediate design problem for the above proposed experiment is to choose n_{pat} and n_{rep}, given rough estimates of the magnitudes of task effects B_r and the variance components detailed above. Further problems within the same image estimation and analysis framework are to align the patients in the PET scanner and to pick the shapes over which to estimate regional averages from PET images, again with the goal of minimizing the variance (8.5). Good design choices require biological knowledge, as coded in the variables f_r' and $(vf)_r^2$, as well as a previous study such as that of Chapter 6 to estimate variance components of true effects.

8.3.2 Comparison of reconstruction methods

The virtues of an image reconstruction method depend of course on the intended use of the output image; a reconstruction may be useful for the visual exploration of a PET scan but not for estimating averages within fixed anatomical regions, or vice-versa. The lack of a clear choice among estimates is especially apparent in light of the limited resolution of PET, which forces all reconstructions to be biased or depend on unverifiable information. In the specific context of linear estimation of region × task effects, however, competing reconstruction methods may be compared by performing competing experimental design calculations, as in Section 8.3.1 above, and then choosing a reconstruction method on which a cost-effective study can be based.

8.4 Conclusion

This thesis reviews and develops image reconstruction for positron emission tomography (PET) with some exploration of the related issues of statistical design and analysis of a factorial experiment using PET. Following overviews in the first two chapters on image analysis in general and PET in particular, the discussion of image reconstruction focuses on estimation of hierarchical Bayesian models. Such models turn out to fit neatly into the indirect form of PET data, even after accounting for realistic data corrections, and can be estimated using iterative EM and stochastic relaxation algorithms. As detailed in Chapters 2 and 3, the Bayesian estimation algorithms can be expressed in terms of operators used in standard linear image reconstruction algorithms; in addition, estimation using the hierarchical model conveniently averages over nuisance parameters. Chapters 4 and 5 integrate some regularization and smoothing estimates into the Bayesian framework, in particular identifying local spatial smoothing as a case of estimation from a nonstationary Gaussian conditional autoregressive prior model.

In addition, this thesis covers several more specific statistical issues associated with three key difficulties of PET. First, PET data are indirectly observed, requiring an inverse of an integral transform to reconstruct an image. Fourier analysis is useful for this task and motivates filtered back-projection, a linear reconstruction method that is exact for continuously observed, noise-free data and is currently used in real PET scanners. Second, the mathematical constraint that an image be a nonnegative function explains some of the unsatisfactory properties of maximum likelihood image reconstruction from indirect, noisy data sources such as PET. The failure of maximum likelihood motivates the aforementioned regularized and smoothed image estimates, which in turn suggest some specific Bayesian methods. A third difficulty is the resolution limit: every PET image estimate must be based on assumptions that cannot be verified from the data, leading to bias in estimation that is serious for filtered back-projection and may be reduced with Bayesian estimation. Interestingly, estimation bias often manifests itself as a blur in the reconstructed image.

Finally, the ideas presented here are focused by the analysis of a specific medical imaging study performed at Massachusetts General Hospital. Chapter 6 presents an exploratory analysis of the effects of substantive interest, which are based on local averages of PET images. Chapter 7 adapts the models and techniques developed in earlier chapters to the applied problem by deriving expressions for bias and variance of local averages of various PET estimates. The final chapter begins with the difficulties of applying the likelihood-based estimates of Chapters 2–5 to real PET data; a direct implementation, not presented here, of the maximum likelihood estimate suggests that the statistical model of PET data, reviewed here in Chapter 2, is incomplete in practice. The thesis concludes with a discussion of statistical design for the medical study and the dependence of design and analysis on knowledge of biological spatial structures, not estimable from PET, which can themselves be modeled probabilistically. In its eight chapters, this thesis derives, unifies, and motivates a family of image estimates and reconstruction algorithms applicable to noisy projection data and also explores the limits of image reconstruction for experiments based on PET.

Bibliography

Alpert, N. M., Chesler, D. A., Correia, J. A., Ackerman, R. H., Chang, J. Y., Finklestein, S., Davis, S. M., Brownell, G. L., and Taveras, J. M. (1982). Estimation of the local statistical noise in emission computed tomography. *IEEE Transactions on Medical Imaging* **MI-1**, 142–146.

Bartlett, M. S. (1975). *The Statistical Analysis of Spatial Pattern.* London: Chapman and Hall.

Besag, J. (1974). Spatial interactions and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society B* **36**, 192–235.

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician* **24**, 179–195.

Besag, J. (1977). Errors-in-variables estimation for Gaussian lattice schemes. *Journal of the Royal Statistical Society B* **39**, 73–78.

Besag, J. (1981). On a system of two-dimensional recurrence equations. *Journal of the Royal Statistical Society B* **43**, 302–309.

Besag, J. (1986). On the statistical analysis of dirty pictures (with discussion). *Journal of the Royal Statistical Society B* **48**, 259–302.

Besag, J. (1989). Towards Bayesian image analysis. Technical Report, Department of Mathematical Sciences, University of Durham, England.

Bookstein, F. L. (1986). Size and shape spaces for landmark data in two dimensions (with discussion). *Statistical Science* **1**, 181–242.

Carson, R. E. (1986). A maximum likelihood method for region-of-interest evaluation in emission tomography. *Journal of Computer Assisted Tomography* **10**, 654–663.

Chow, Y., Grenander, U., and Keenan, D. M. (1989). Hands: a pattern theoretic study of biological shapes. Technical Report, Brown University.

Cliff, A. D., and Ord, J. K. (1981). *Spatial Processes: Models and Applications.* London: Pion.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B* **39**, 1–38.

Dempster, A. P. (1974). The direct use of likelihood for significance testing. In *Proceedings of the Conference on Foundational Questions in Statistical Inference*, eds. O. Barndorff-Nielsen, P. Blaesild, G. Sihon. University of Aarhus.

Elliott, D. (1989). Brain activity patterns during smooth pursuit and saccadic eye movements: a study using positron emission tomography. Ph.D. thesis, Department of Psychology, Harvard University.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications.* New York: Wiley.

Gelman, A. (1989). Constrained maximum entropy methods in an image reconstruction problem. In *Maximum Entropy and Bayesian Methods*, ed. J. Skilling, 429–435. Dordrecht: Kluwer Academic Publishers.

Gelman, A. (1990). Comment on "A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography," by Silverman et al. *Journal of the Royal Statistics Society B*, to appear.

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Geman, S., and Huang, C-R. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *Annals of Statistics* **10**, 401–414.

Geman, S., and McClure, D. E. (1987). Statistical methods for tomographic image reconstruction. *Proceedings of the ISI Meetings.* Tokyo.

Goodall, C. R. (1983). The statistical analysis of growth in two dimensions. Ph.D. thesis, Department of Statistics, Harvard University.

Gull, S. F. (1989). Developments in maximum entropy data analysis. In *Maximum Entropy and Bayesian Methods*, ed. J. Skilling, 53–71. Dordrecht: Kluwer Academic Publishers.

Gull, S. F., and Daniell, G. J. (1978). Image reconstruction from incomplete and noisy data. *Nature* **272**, 686–690.

Haslett, J. (1985). Maximum likelihood discriminant analysis on the plane using a Markovian model of spatial context. *Pattern Recognition* **18**, 287–296.

Hastie, T., and Tibshirani, R. (1990). *Generalized Additive Models.* London: Chapman and Hall, in press.

Hoffman, E. J., Huang, S-C., and Phelps, M. E. (1979). Quantitation in positron emission computed tomography: 1. Effect of object size. *Journal of Computer Assisted Tomography* **3**, 299–308.

Hoffman, E. J., Huang, S-C., Phelps. M. E., and Kuhl, D. E. (1981). Quantitation in positron emission computed tomography: 4. Effect of accidental coincidences. *Journal of Computer Assisted Tomography* **5**, 391–400.

Huang, S-C., Hoffman, E. J., Phelps, M. E., and Kuhl, D. E. (1979). Quantitation in positron emission computed tomography: 2. Effects of inaccurate attenuation correction. *Journal of Computer Assisted Tomography* **3**, 804–814.

Huang, S-C., Hoffman, E. J., Phelps, M. E., and Kuhl, D. E. (1980). Quantitation in positron emission computed tomography: 3. Effect of sampling. *Journal of Computer Assisted Tomography* **4**, 819–826.

Johnson, V. E., Wong, W. H., Hu, X., and Chen, C-T. (1989). Statistical aspects of image restoration. Technical Report #250, Department of Statistics, University of Chicago.

Johnstone, I. M., and Silverman, B. W. (1990). Speed of estimation in positron emission tomography. *Annals of Statistics* **18**, to appear.

Kak, A. C. (1984). Image reconstruction from projections. Chapter 4 in *Digital Image Processing Techniques*, ed. M. P. Ekstrom. New York: Academic Press.

Kunsch, H. R. (1987). Intrinsic autoregressions and related models on the two-dimensional lattice. *Biometrika* **74**, 517–524.

Lange, K., Bahn, M., and Little, R. (1987). A theoretical study of some maximum likelihood algorithms for emission and transmission tomography. *IEEE Transactions on Medical Imaging* **MI-6**, 106–114.

Lange, K., and Carson, R. (1984). EM reconstruction algorithms for emission and transmission tomography. *Journal of Computer Assisted Tomography* **8**, 306–316.

Levitan, E., and Herman, G. T. (1987). A maximum *a posteriori* probability expectation maximization algorithm for image reconstruction in emission tomography. *IEEE Transactions on Medical Imaging* **MI-6**, 185–192.

Lim, J. (1988). *Two-Dimensional Signal and Image Processing.* Lecture Notes for Course 6.344, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

Little, R. J. A., and Rubin, D. B. (1987). *Statistical Analysis with Missing Data.* New York: Wiley.

Matheron, G. (1973). The intrinsic random functions and their applications. *Advances in Applied Probability* **5**, 439–468.

Mazziota, J. C., Phelps, M. E., Plummer, D., and Kuhl, D. E. (1981). Quantitation in positron emission computed tomography: 5. Physical-anatomical effects. *Journal of Computer Assisted Tomography* **5**, 734–743.

Meehan, P. M. (1990). Modeling groundwater flow. Qualifying paper, Department of Statistics, Harvard University.

Metropolis N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical*

Physics **21**, 1087–1092.

Mitter, J., Marroquin, S., and Poggio, T. (1987). Probabilistic solution of ill-posed problems in computational vision. *Journal of the American Statistical Association* **82**, 76–89.

Moran, P. A. P. (1973). A Gaussian Markovian process on a square lattice. *Journal of Applied Probability* **10**, 54–62.

Natterer, F. (1986). *The Mathematics of Computerized Tomography.* New York: Wiley.

Ripley, B. D. (1981). *Spatial Statistics.* New York: Wiley.

Ripley, B. D. (1986). Statistics, images, and pattern recognition (with discussion). *Canadian Journal of Statistics* **14**, 83–111.

Ripley, B. D. (1988). *Statistical Inference for Spatial Processes*, chapters 1, 2, 4, 5. Cambridge University Press.

Rosanov, Y. A. (1967). On Gaussian fields with given conditional distributions. *Theory of Probability and its Applications* **12**, 381–391.

Senda, M. (1988). Stereotactic coordinates transformation program: theory and method. Unpublished, 10 pp.

Shepp, L. A., and Vanderbei, R. J. (1988). New insights into emission tomography via linear programming. Technical Report, AT&T Bell Laboratories.

Shepp, L. A., and Vardi, Y. (1982). Maximum likelihood reconstruction for emission tomography. *IEEE Transactions on Medical Imaging* **MI-1**, 113–122.

Silverman, B. W., Jones, M. C., Wilson, J. D., and Nychka, D. (1990). A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography. *Journal of the Royal Statistics Society B*, to appear.

Snyder, D. L., and Miller, M. I. (1985). The use of sieves to stabilize images produced with the EM algorithm for emission tomography. *IEEE Transactions on Nuclear Science* **NS-32**, 3864–3872.

Snyder, D. L., Miller, M. I., Thomas, L. J., and Politte, D. G. (1987). Noise and edge artifacts in maximum-likelihood reconstructions for emission tomography. *IEEE Transactions on Medical Imaging* **MI-6**, 228–238.

Spitzer, F. (1964). *Principles of Random Walk.* New York: Springer-Verlag.

Stearns, C. (1990). Accelerated image reconstruction for a cylindrical positron tomograph using Fourier domain methods. Ph.D. thesis, Division of Health Sciences and Technology, Massachusetts Institute of Technology.

Switzer, P. (1987). Statistical image processing. Technical Report #15, Department of Statistics, Stanford University.

Tanner, M. A., and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **82**, 528–550.

Tikhonov, A. N., and Arsenin, V. Y. (1977). *Solution of Ill-posed Problems.* New York: Wiley.

Titterington, D. M. (1985). General structure of regularization procedures in image reconstruction. *Astronomy and Astrophysics* **144**, 381–387.

Vanmarcke, E. (1983). *Random Fields: Analysis and Synthesis.* Cambridge, Mass.: MIT Press.

Vardi, Y., Shepp, L. A., and Kaufman, L. (1985). A statistical model for positron emission tomography (with discussion). *Journal of the American Statistical Association* **80**, 8–37.

Veklerov, E., and Llacer, J. (1987). Stopping rule for the MLE algorithm based on statistical hypothesis testing. *IEEE Transactions on Medical Imaging* **MI-6**, 313–319.

Venetoulias, A. (1988). Statistical image processing: an annotated bibliography. Technical Report #TR-64, Sloan School of Management, Massachusetts Institute of Technology.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika* **41**, 434–449.

Whittle, P. (1956). On the variation of yield variance with plot size. *Biometrika* **43**, 337–343.