HIGH-LEVEL IMAGE UNDERSTANDING VIA BAYESIAN HIERARCHICAL MODELS

by

Colin C. McCulloch

Institute of Statistics and Decision Sciences
Duke University

Date: _____

Approved:

_____
Dr. Valen Johnson, Supervisor

_____
Dr. Robert Wolpert

_____
Dr. Peter Mueller

_____
Dr. Ronald Jaszczak

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Institute of Statistics and Decision Sciences
in the Graduate School of
Duke University

1998

ABSTRACT

(Statistics)

HIGH-LEVEL IMAGE UNDERSTANDING VIA BAYESIAN HIERARCHICAL MODELS

by

Colin C. McCulloch

Institute of Statistics and Decision Sciences
Duke University

Date: _____

Approved:

_____
Dr. Valen Johnson, Supervisor

_____
Dr. Robert Wolpert

_____
Dr. Peter Mueller

_____
Dr. Ronald Jaszczak

An abstract of a dissertation submitted in partial
fulfillment of the requirements for the degree
of Doctor of Philosophy in the
Institute of Statistics and Decision Sciences in the Graduate School of
Duke University

1998

Abstract

The tasks performed by medical image analysis technicians, including registration and segmentation, have become increasingly difficult with the advent of three-dimensional imaging systems. To identify features in these large images, the technician must typically engage in the tedious chore of examining numerous lower dimensional representations of parts of the data set, for instance slices though the volume or volume-rendered views. The pursuit of automatic image understanding, previously sought after in two-dimensional images for objective anatomical measurement and to reduce operator burden, therefore has become proportionally more valuable in these larger image datasets.

A statistical framework is proposed to automate image feature identification and therefore facilitate the image understanding tasks of registration and segmentation. Features are delineated using an atlas image, and a probability distribution is defined on the locations and variations in appearance of these features in new images from the class exemplified by the atlas. The predictive distribution defined on feature locations in a new image from the class essentially balances the two notions that, while each individual feature in the new image should appear similar to its atlas representation, contiguous groups of features should also remain faithful to their spatial relationships in the atlas image. A joint hierarchical model on feature locations facilitates reasonable spatial deformations from the atlas configuration, and several local image measures are explored to quantify feature appearance. The hierarchical structure of the joint distribution on feature locations allows fast and robust density maximization and straightforward Markov Chain Monte Carlo simulation. Model hyperparameters can be estimated using training data in the form of manual feature observations.

Given Maximum *a posteriori* estimates an analysis is performed on *in vitro* mouse brain Magnetic Resonance images to automatically segment the hippocampus. The model is also applied to time-gated Single Photon Emission Computed Tomography cardiac images to reduce motion artifact and increase signal-to-noise.

Acknowledgements

I am deeply grateful to Dr. Valen Johnson for his guidance and support. His constant enthusiasm and firm understanding of the truly important challenges to be tackled in this field gave my work a strong foundation and a clear direction. I also appreciate the generous help given to me by Dr. Ron Jaszczak, Dr. Peter Mueller, and Dr. Robert Wolpert. I would also like to thank Jacob Laading for our lively coffee-break discussions. And thank-you to everyone at ISDS for providing a friendly and stimulating research atmosphere.

Finally I would like to thank Maria for her limitless support. In challenging times she has been an invaluable resource, and in good times she has been an immeasurable source of joy.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Image Analysis Applications

The past decade has seen a surge in research aimed at automatically processing image data. Some of the standard techniques are familiar, for instance data compression for sending video over the internet or simple smoothing techniques to remove noise. Others are more specialized, for instance automatic fingerprint recognition for security monitoring. The Bayesian statistical community has made significant contributions to image analysis research by developing methodologies for problems of particular interest in the medical sciences, among others. A brief overview is given in this section of the classical image analysis problems: image restoration, reconstruction, and understanding.

The goal of image restoration is to reproduce the underlying "true" image contained in an observed realization corrupted with noise. The Bayesian solution to this problem characterizes prior knowledge about the true image x in terms of a marginal distribution $p(x)$ and states that the corrupted realization y is drawn from a conditional distribution $p(y|x)$. Restorations of x are inferred from the posterior on the true image given the realization, $p(x|y) \propto p(x)p(y|x)$. Several examples of image

priors and applications are given in Ripley (1991), Titterington (1997), Godsill and Kokaram (1997), Besag (1986), and Geman and Geman (1984).

Image reconstruction is employed in the "computed tomography" medical imaging modalities (single photon emission computed tomography, SPECT, and X-ray computed tomography, CT), since the scanner merely outputs data in the form of separate projected views of the body being imaged from different perspectives. Data in this form must be thoroughly massaged to produce a recognizable image for a clinician to examine. The true three-dimensional representation of the subject matter is reconstructed by integrating together the set of lower-dimensional projections in a manner dependent on the geometry of the scanning system. The Bayesian paradigm can be applied here by again representing the unobserved true image as x and the observed projection data as y. The system noise and geometrical characteristics are incorporated in the conditional distribution $p(y|x)$ and reconstructions of the true image can be inferred from the posterior distribution $p(x|y)$. Example methodologies and applications are given in Geman and McClure (1987) and Bowsher *et al.* (1995).

The methods of image understanding aim to represent the image in a more meaningful way than the original array of pixel intensities. Image understanding methods are invariably based on feature identification, where a pre-defined feature possessing known properties, for example an intensity profile, is located in the image. Target tracking is an excellent example of this type of process, where for example the radar profile of an airplane is located in the first image of a series and its position is tracked through the rest of the series by imposing constraints on feature properties like the speed of the feature and its possible feature profiles, etc. (Grenander and Miller, 1994).

Another image understanding task, more central to this paper, is image registration. Here, the challenge is to map a set of features in one image to their homologous

locations in another for the purpose of averaging image information or for generating an anatomical map of one of the images conditional on a pre-defined map of the other. Applications for image registration are abundant in the medical imaging community. For example, there has been recent interest in trying to produce a map of the human brain which delineates function rather than anatomy. To this end, experiments are run with the intention of mapping the regions of the brain responsible for one type of thought process. For instance, a subject is shown a "test" figure followed by a "control" figure while undergoing a Positron Emission Tomography (PET) scan. The PET scan measures regions of high blood flow in the brain, which are interpreted as regions of thought activity. The image taken while the test figure is shown is subtracted from the control image, and the resulting intensity map is interpreted as the difference in thought processing between the test and control states. A major limitation of these experiments is that it is impossible to control all the extraneous activity in a subject's brain, which may cloud the activation regions truly associated with the test state. As a result, experimenters turn to multi-subject studies to reduce this effect by averaging each test state over several images taken of different subjects. However, care must be taken when comparing images across subjects because of their widely differing brain shapes. A registration between subjects must be performed to line up anatomical features of matching homology. Another limitation is the effect of subject motion inside the scanner. This problem is less difficult because the transformation needed to register scans of a single subject is just a six parameter family in three-dimensions: three rotations and three translations.

Intra-subject image registration can become more difficult when the two images of the subject are taken with different modalities. This kind of analysis is done, for instance, to make available an anatomical MR map of a structure of interest that a clinician is currently observing in a PET scan for its functional behavior (Studholme

et al., 1997).

Another image understanding problem is automated image segmentation, where the aim is to construct contours in the image to partition it into structures of interest. (In three-dimensional images the contours are surfaces and the regions are volumes.) Image segmentation is commonly performed on medical images and is almost exclusively done manually. The most automated currently accepted methods for segmentation employ human-directed intensity thresholding-type operations. Although adequate results are achieved, the method is highly dependent on the skill of the operator and therefore not reproducible. It is also extremely time consuming in the case of three-dimensional images since a technician must typically examine numerous lower dimensional representations of parts of the data set, for instance slices though the volume or volume-rendered views.

It should be pointed out that image segmentation can be a free by-product of a registration analysis. This is because if one of the registered images has been correctly segmented then, since the registration maps all the features in that image into their homologous locations in the other, the segmentation of the first image is also transfered to the unsegmented image. This is the approach taken in atlas-based methods, discussed in section 1.2.4.

In the next section, a brief review is given of some of the important contributions made to this field of image analysis, with emphasis given to the challenges of image understanding and also to the notable methodologies developed from a statistical point of view.

1.2 Previous Work

Many of the algorithms and models in the current image understanding literature can be classified into a few categories: pixel-based methods, boundary- and surface-

based methods, and landmark-based methods. A short discussion of each of these categories is given in the following sections. In each section an overview of some of the methodologies will be given along with a more in depth description of one or two particularly promising directions of research.

1.2.1 Pixel-Based Methods

The pixel-based statistical methods for image analysis grew out of Besag's pioneering work in 1974. In it the author layed down the basic principles that must be adopted when defining any joint probability distribution in terms of the conditionals of each random variable given all the others. If these rules are not followed when constructing a model through the conditional probability approach, then a valid joint probability distribution on all the random variables cannot result.

A network of random variables $x = (x_1, \ldots, x_n)$ can have an associated "neighborhood system" N where the neighbor connections in N are defined by the full conditional distributions on each element of x. If the full conditional on x_i only depends on a subset $\{x_k, k \in S_i\}$ of the whole vector x, then the elements of this subset are neighbors of x_i. In this case the full conditional distribution on x_i can be written

$$p(x_i|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = p(x_i|\{x_k, k \in S_i\}).$$

In the paper, Besag re-expressed the Hammersley-Clifford theorem (due to an unpublished 1971 paper by these authors), which defines rules on the form of the full conditionals to ensure a valid joint distribution $p(x)$ on the whole network. Essentially, the joint distribution on x was re-written

$$p(x) = \frac{1}{Z} \exp\left(-\sum_{c \in C} V_c(x)\right) \tag{1.1}$$

where Z is a partition function independent of x. The index of the sum is taken over all the "cliques" in the network, where a clique is defined as a sub-set of x in

which every member is a neighbor of every other member. The potential function $V_c(x)$ only depends on x through the members of the clique c. This form of the joint immediately specifies constraints on the neighborhood structure and the form of the full conditionals on elements in the network. Namely, the full conditional on x_i must include all the potential functions $V_c(x)$ on cliques of which x_i is a member.

In pixel-based statistical image models, each element of the random vector x usually represents one pixel in the image. In restoration problems, the random variable x_i associated with pixel i is the "true" image intensity to be estimated from a noisy realization y of the image. In segmentations analyses x is a set of region specifiers to be inferred from the observed image y. In both cases, statistical models (called priors in the Bayesian literature) on x are usually formulated to enforce beliefs about the smoothness of the true image or region specifiers, where smoothness is quantified by correlating nearby pixels. Besag's fundamental result illustrates how this correlation notion must be formulated. If a full conditional specification on every element x_i is proposed (eg. to correlate pixel i with the pixels closest to it), then this specification must be expressible in terms of the cliques which contain pixel i and potentials on those cliques.

Generally, when defining the neighborhood structure in these "Markov Random Field" (MRF) models, only pixels close to one another are defined to be neighbors. For example, a first-order neighborhood system in a square lattice in two dimensions indicates that each pixel has the four nearest pixels as neighbors. Under this neighborhood system the largest cliques have two members, and models are generally defined so that samples from the full conditional distribution on each pixel are readily calculable for direct implementation of a Gibbs sampling algorithm (Geman and Geman, 1984).

Johnson (1994) investigated a MRF model with a neighborhood system equal to

the entire image. In that model the random variables x (on a hexagonal lattice partially because of its symmetry for defining cliques potentials) were region identifiers which, taken together, defined a segmentation of the observed image. Three clique potentials were defined on configurations of region identifiers. These potentials were used to encourage segmentations which had (1) a small number of different regions, (2) regularly shaped regions, and (3) no regions with disconnected sections. The first potential was defined on a clique the size of the entire lattice,

$$V_1(x) = \alpha K^2 \tag{1.2}$$

where K is the number of distinct regions in the graph and α is an arbitrary hyperparameter.

The second potential was defined on a circular ring clique one pixel wide made up of the sites in the hexagonal lattice oriented at a radius d around a central site. The potential on this "regularity" clique was set to an arbitrary value ϕ if any pixel in the ring with a given region identifier was separated within the ring from other pixels having the same identifier. The clique potential was set to zero if this condition did not occur. This regularity potential simply encouraged groups of region identifiers to have regular shapes by discouraging a region from growing fingers thinner than the size of the regularity clique. The threshold finger size could be adjusted by the size of the regularity clique used.

The final clique potential, again on a clique equal to the whole graph, made it impossible for a region to split into two disconnected partitions. This infinite potential was necessary since splitting a region would require changing all the region identifiers in one of the sections, which violates the Markovian property of the underlying Gibbs distribution.

A data model for the observed image intensity scene given a realization of region identifiers from the Gibbs distribution was set to be a Poisson/Gamma conjugate

pair, creating a three level hierarchical model for observed image data. The middle level of the hierarchy was necessary to define the intensity variability within each segmented region. Full conditional distributions were derived for all the Poisson and Gamma parameters, but full conditionals for the the Gibbs parameters (ϕ, α, etc.) were intractable due to the unknown form of the partition function. Higdon *et al.* (1997) gives one solution to this problem. The model was quite successful at segmenting a two dimensional test image and the Hoffman brain phantom (Hoffman *et al.*, 1990).

1.2.2 Boundary and Surface-Based Methods

Many of the boundary-based methods are based on a simple image edge-detectors like the squared gradient magnitude,

$$\left(\frac{\partial Q(x,y)}{\partial x}\right)^2 + \left(\frac{\partial Q(x,y)}{\partial y}\right)^2 \tag{1.3}$$

in the two-dimensional image Q, or the negative absolute value of the image Laplacian,

$$-\left|\frac{\partial^2 Q(x,y)}{\partial x^2} + \frac{\partial^2 Q(x,y)}{\partial y^2}\right|. \tag{1.4}$$

These methods generally involve models for parameterized curves that are attracted to regions in the image having a high value of (1.3) or (1.4). A possible boundary in the image is characterized by a balance of the fit of the curve to its parameterization with its fit to the edge-detector.

For example, Kass *et al.* (1988) defined the "active contour model" for the placement of a line in an image under internal line continuity constraints and external image forces. Image forces were set to be either proportional to the image intensity or one of the image measures (1.3) or (1.4) so that the "snake" was drawn to regions of

quickly changing intensity. The continuity of the line was maintained by an internal force which was a function of its spline energy, defined as a weighted linear combination of its first and second directional derivatives along its length. This constraint essentially forced the snake to behave somewhat like a membrane and somewhat like a thin plate, depending on the weights in the linear combination. Under these internal line forces and external image forces, the snake was drawn to locations in the image which were well-delineated contours of large intensity change. These regions were deemed to be reasonable estimates for object boundaries.

Davatzikos *et al.* (1996) used an active contour to delineate the boundary of the human corpus callosum in a mid-sagittal MR image. The resulting boundaries of eight males and eight females were used to define an elastic warp on the interior of the region which matched all subjects to a common corpus callosum shape. The calculated warps allowed the authors to draw conclusions on gender differences in the size and shape of the region.

Active contour models have been extended for mapping the human brain cortex in two dimensions (Davatzikos and Prince, 1995) and for mapping the deep sulci in three-dimensions using higher-dimensional active ribbons (Le Goualher and Barillot,). Other applications of similar models can be found in Sandor and Leahy (1997), Davatzikos and Prince (1996), Thompson and Toga (1996), Cootes *et al.* (1994), Cootes *et al.* (1994), and McInerney and Terzopoulos (1996).

There have been several models proposed in the statistical literature for outlining shapes with parametrized boundary models, for example Grenander and Miller (1994), Clifford and Nicholls (1994), and Hurn and Rue (1997). Much of the recent work has been on creating efficient Markov chain monte carlo (MCMC) jump diffusion samplers (Green, 1995) for object recognition by boundary birth and death proposals.

One particularly elegant and accessible statistical approach to segmenting images by inferring boundaries was given by Phillips and Smith (1994). To segment several regions from face images (head, face, eyes, etc.), a parameterized curve was defined to outline each region, and prior distributions were set on the parameters of each curve. For instance, the curve outlining the head was set to be roughly elliptical by parameterizing the lengths of the major and minor axes in the horizontal and vertical direction from the center (four parameters), and four additional control points on the curve at directions diagonal to these axes. Normal priors were set on all distances with expectations defined to incorporate prior knowledge about the size and shape of the region with respect to the size of the image. A hierarchical model was implemented by constraining the face boundary to lie inside the head boundary, the eyes inside the face, and so on.

Given these segmented regions, an admittedly crude data model for the image intensities was defined. The intensity at each pixel inside a particular region r was modeled as an independent normal random variable with unknown region mean μ_r and variance σ_r^2. A Metropolis sampler was used for posterior inference on all the parameters in the model including the region means and variances, and the template parameters for each region contour. Parameter full conditional distributions were blocked appropriately by template contour to increase sampling efficiency. The method was quite successful even with the modest parameterization of the face and the unsophisticated data model for the image intensities.

1.2.3 Landmark-Based Methods

These methods were initiated with the definitions of several types of landmarks proposed by Bookstein (1991), and later discussed by Dryden and Mardia (1996). Bookstein was mainly concerned with analyzing the shape of configurations of points in

images provided by a human observer. He defined three main types of anatomical landmarks: type I landmarks found at the joints of tissues and bones, type II defined by local image properties such as maximal curvatures, and type III defined at extremal points. A fourth type, the quasi-landmark, could be located on a curve between other landmarks and allowed to slip a small distance with respect to another curve. Dryden and Mardia defined three very similar but slightly more general landmark types: biological landmarks assigned by an expert in a biologically meaningful way, mathematical landmarks located according to some mathematical or geometrical property of the image, and pseudo-landmarks constructed around an outline between biological or mathematical landmarks.

The methods reviewed in this section outline automated procedures for detecting landmark locations in an image. These methods generally provide the locations of a small set of landmarks in an image which can be further analyzed by shape statistics as discussed in Mardia and Dryden (1989), Lele and Cole (1996), and Bookstein (1991), or used as control points for various continuum-based image warping algorithms, eg. Bookstein (1991), Bajcsy and Kovacic (1989), and Gee *et al.* (1993).

Landmarks located automatically are mainly Bookstein's type II or III or Mardia and Dryden's mathematical type since the automated choice of landmark location is based largely on local image intensities near the landmark. To locate more subtle landmarks of other types, prior distributions are set to define likely configurations of a set of landmarks. In this way, landmarks which are poorly defined in a particular image can be located by drawing strength from other well defined landmark locations.

Amit and Kong (1994) manually defined a template orientation of landmarks by pruning a pre-calculated set of local intensity maxima in a template image to include only meaningful landmarks. The landmarks were then connected into a graph of triangular cliques. Given a set of local intensity maxima sites in a new image, a dis-

tribution was defined to assign relative probabilities to every possible configuration of the template landmarks on these candidate sites. The distribution essentially rewarded triangular cliques having similar shape to the template orientation regardless of size, rotation, or translation. Constraints on the structure of the graph permitted the generation of high probability orientations in polynomial time, meaning that a set of 70 landmarks could be located in a new image in a few minutes on a modern computer.

Wilson (1995) took a similar course in defining a probability density on all possible configurations of a set of landmarks in a new image, given a template orientation defined manually. The distribution was defined as a MRF on a graph where the arcs in the graph were manually set in the template to define the neighborhood structure of the field, and the random variable at each node in the graph represented the location of one of the landmarks in the set. Landmark locations were actually defined in the whole scalespace of the image (eg. Lindeberg 1994, ter Haar Romeny *et al.* 1991, and Koenderink 1984) and the clique potential on the exclusively two member cliques were defined in terms of the scalespace distance between the two neighbors. The scalespace metric, defined in Eberly (1994), is too involved to be included here, but it was essentially used to invoke invariance to the size of the landmark configuration. Let the scalespace distance between landmarks i and j be $d_{ss}(x_i, x_j)$ and the analogous distance in the template $(\mu_1, \ldots, \mu_n)$ be $d_{ss}(\mu_i, \mu_j)$. Then the density on any landmark configuration was defined as

$$p(x_1, \ldots, x_n) = \frac{1}{Z} \exp\left(\sum_{c \in C} V_c(x_1, \ldots, x_n)\right) \tag{1.5}$$

where the potential function on each two member clique $c = \{i, j\}$ was defined as

$$\begin{aligned} V_{i,j} = -k[&(d_{ss}(x_i, x_j))^2 + (d_{ss}(\mu_i, \mu_j))^2 \\ &- 2d_{ss}(x_i, x_j) d_{ss}(\mu_i, \mu_j) \cos(x_i - x_j, \mu_i - \mu_j)]. \end{aligned} \tag{1.6}$$

In this equation, $\cos(x_i - x_j, \mu_i - \mu_j)$ is the cosine of the angle between the vectors $x_i - x_j$ and $\mu_i - \mu_j$, and k is an arbitrary hyperparamter. This potential rewarded a clique orientation for having its connecting vector $(x_i - x_j)$ similar to the template vector $(\mu_i - \mu_j)$.

Each landmark was also given a feature type - "boundary", "middle" or "corner". These names referred to the function on the image used to quantify relative probabilities of candidate landmark locations. For instance, a boundary landmark would be attracted to regions in the image having a high value of the gradient magnitude feature function given in equation (1.3). The predictive density on a landmark configuration was defined to be a product of the feature function component and the shape component (equation 1.5), where weighting parameters on the two components were set manually.

McCulloch *et al.* (1996) extended these scalespace landmark models by defining new clique potentials on larger cliques in the configuration. The Procrustes distance (Sibson 1978 and Stoyan 1994) was used as a metric on clique shape similarity between the proposed clique orientation x_c and its corresponding template orientation μ_c. Under this potential, cliques in the MRF could be defined to have any number of members greater than two. In this way, it was possible to represent perceived objects in the image more naturally by cliques in the MRF.

Similar work was undertaken in Mardia *et al.* (1997) in which the shape distribution was made invariant to arbitrary rotation, translation, and scaling by multiplying the $n \times d$ matrix of landmark locations by the Helmert sub-matrix.

1.2.4 Atlas-Based Methods

The methods presented thus far have either incorporated limited information from a template image, or have been entirely "atlas-free." On the other hand, the methods

in this section are concerned with registering the image to a similar but different atlas image. These "atlas-based" methods have been very successful in image understanding applications because the atlas image injects a great deal of structure into the model, and because the higher tasks of feature identification and segmentation are consolidated into the problem of image registration. Given a correct registration of the new image to a pre-analyzed atlas image, the analysis of the atlas can be directly transferred to the new image. The choice of atlas has received some attention (Mazziotta *et al.* 1995, Greitz *et al.* 1991, and Talairach and Tournoux 1988) but in many applications the atlas is taken to be simply one image from the class of images to which all other images are registered.

Medical image registration problems, on which atlas-based methods are founded, can be divided into two categories: intra-subject and inter-subject registrations. For intra-subject registration, generally one image is rotated and translated with respect to the other until some match criterion is maximized. In single modality registration, the image match criterion for a candidate registration is usually based on a correlation measure comparing each pixel in the atlas image with the overlaid pixel in the other image after it has been rotated and translated (Woods *et al.*, 1992). For multimodality registration problems the match criterion is more subtle; several possibilities have been investigated (Studholme *et al.*, 1997).

Inter-patient registration is a much more difficult task because of the necessity of a nonlinear transformation to map the atlas to the new image due to the morphometric variability between subjects. Much work has been done to model this nonlinear transformation by various physical models on a continuum, such as the thin plate spline (Bookstein, 1991), the elastic solid (Bajcsy and Kovacic, 1989), and the viscous fluid (Christensen *et al.*, 1993). Amit *et al.* (1991) modeled the registration deformation as a two-dimensional Gaussian field on the domain of the atlas image.

Multi-resolution techniques (Bajcsy and Kovacic 1989, Studholme *et al.* 1997, and Lifshitz and Pizer 1990) have been useful in the inter-subject registration problem by directing the optimization over the numerous parameters in the nonlinear transformation to locate large scale image features before turning to the more variable local features. Scalespace is the usual choice for implementing the multi-resolution optimization: high scale, blurred versions of the two images are registered first, after which lower scale representations are incorporated which recapture the finer details of the original images.

Collins *et al.* (1995 and 1996) implemented a multi-scale nonlinear registration procedure on three-dimensional MR brain images. Several grids of varying densities were laid down in the atlas image and the new image to be analyzed, and each grid was associated with one scale in the pre-calculated scalespace of the images. The atlas grids were fixed, and the registration was accomplished by spatially transforming each grid in the new image to maximize a correlation measure on the intensities in a neighborhood around each grid point.

Starting with the coarsest grid and most strongly blurred image, the correlation objective function on the location of each grid point was maximized in turn. Let μ be the location of the grid point in the atlas image Q_0, let the set of its neighboring pixels be N_μ, let x be the grid location in the new image Q, and let the mapping of N_μ to the neighboring pixels around x be $N(x)$. Then the following correlation function of the neighboring pixel intensities around each grid point was used as the objective function,

$$R(x) = \frac{\sum_{i \in N_\mu} Q_0(i) Q(\{N(x)\}_i)}{(\sum_{i \in N_\mu} Q_0(i))^{1/2} (\sum_{i \in N_\mu} Q(\{N(x)\}_i))^{1/2}}. \tag{1.7}$$

The maximization was constrained to retain the continuity of the grids by relaxing the estimated mode back toward the average location of the neighboring grid points by a manually chosen factor.

Once the deformation of the coarse grid was estimated on the high scale image, then this deformed grid was used to initialize the same optimization procedure for the next finer grid on a lower scale image. This procedure continued until the deformation on the final lowest scale grid was optimized.

Impressive segmentation results were achieved by using the recovered image registration to map a manual segmentation of one image onto the other.

1.3 Motivation

In this paper, a statistical framework is developed for automated image registration and segmentation. The models draw on the strengths of several of the methods presented in the previous section by defining a probability distribution on the location and appearance of multiple image features in images from a common class. These image features, called "facets", can be thought of as generalized landmarks. They could be defined to have special anatomical or topological meaning as landmarks invariably have, but moreover they could also be defined implicitly by the use of an atlas image, where the pre-determined location of a facet in that image defines its meaning. The latter method for facet labelling transforms the statistical method from a landmark based approach towards an atlas-based approach since the facet distribution becomes a model of how an intelligent observer locates atlas features in other images from the class. If a sufficient number of facets are used in the model to represent all the salient features in the atlas, then a prediction of these facet locations in a new image from the class is essentially a registration of the new image to the atlas.

The registration of the new image to the atlas is frequently the end goal. However, the registration also facilitates automatic segmentation of the new image given a manual segmentation of the atlas. Each facet is given a segmentation label using the

manual atlas segmentation, and the statistical model is used to predict the locations of each facet in the new image. In this way, all the facets from a particular atlas segmentation region can be used to automatically segment that region in the new image. Both of these applications are explored in Chapters 4 and 5.

By taking a statistical approach rather than merely deriving a numerical algorithm, the methods allow self-calibration through the use of training data. Any tuning parameters, which would normally have to be set manually in a numerical algorithm, are statistical model parameters which can be estimated via posterior distributions given the training data. Furthermore, the often ignored reality of human observer variability can be incorporated into the model to give a more reasonable representation of human performance.

Chapter 2

Model Definition

2.1 Notation/Nomenclature

In this paper, a model is proposed of how an intelligent observer visually locates and recognizes features in an image. The probability model proposed represents these "features" through the new concept of "facets", and the observer's prior knowledge about the image scene is represented through the concept of an "image class". In this section, the new nomenclature used hereafter in this paper is defined.

First, a d-dimensional image is considered a function Q on the finite domain $D \in R^d$ to the real line, $Q : R^d \rightarrow R$. The domain D is the finite extent of the image dataset, and d is usually equal to 2 or 3. For defining the probability model below, it will be important that the image function is defined on all of R^d, so the function is extended from $D \in R^d$ to R^d by defining that its value at $Q(z)$, $z \in R^d \backslash D$ is equal to $Q(z_0)$, $z_0 \in D$ for which the Euclidean distance $||z - z_0||$ is minimized.

The image class concept is used to express the fact that, even on viewing an image for the first time, a human observer immediately draws on a wealth of experience to classify the image by a number of criteria. For example, s/he would immediately realize if the image were a color photograph of a human face, and could also classify

it by a number of other criteria such as gender, age, hair color, etc. For the purposes of this paper, it is assumed that a pre-classification has been performed and the focus is restricted to identifying features in images taken from a common "class". An image class is defined as a set of images of the same structure taken in different cases, where cases might be, for instance, different individuals or different images of the same individual.

To illustrate a typical classification, one image class of interest in this paper is a set of three-dimensional magnetic resonance (MR) images of *in vitro* new-born mice brains. These are all taken with the "T2*" setting of the MR apparatus, and every mouse brain is physically registered in the apparatus to a common frame of reference before imaging.

In this example, since the brains have all been physically registered, any probability distribution on feature locations in these images need not be invariant to orthogonal transformations (ie. rotations, translations, and scalings). On the other hand, any invariances found in the image class must be incorporated in the probability model. For example, typical class invariances include image brightness and contrast.

A facet can be thought of as a generalized landmark, discussed in section 1.2.3. Whereas landmarks must be salient features in the image, facets can represent any image feature no matter how pronounced it is. For example, in an image class containing photographs of human faces, one facet might be defined as the left corner of the mouth. In this case, the facet could surely be called a landmark (Bookstein type I or Dryden & Mardia Biological) since it has an anatomically important feature associated with it. However, another facet might be defined at a point on the cheek that has no biological name. It is possible to define this type of facet using an atlas image by, for instance, laying down several facets in some regular array on this image

and allowing the image to implicitly label all the facets. As in the case where the facet is the corner of the mouth, a human observer could presumably locate these new implicitly defined facets in a new image drawn from the class using the atlas image for facet definitions. This implicit method of labeling facets is the substance behind the methods in this paper because it allows the model to be applicable in registration and segmentation problems.

Therefore, each facet i has a label (defined either explicitly or implicitly) and a d-dimensional location $x_{j,i}$ in each image Q_j from the class. Along with location, each facet is defined to have a feature value $f_{j,i}$ in that image. The feature value can be simply the image intensity, or some other possibly non-scalar measure of the image, for instance the first spatial derivatives in orthogonal directions or a set of intensities in a small region around the facet. A set of facet locations and feature values in image j are denoted x_j and f_j, respectively.

Facets also have "template" properties which represent the average locations and features of the facets. The template value for the ith facet location is written μ_i, and similarly the feature template value is written ϕ_i. The vector of all template locations and features are denoted μ and ϕ. The interpretation of the random variables μ and ϕ can vary widely in different models, for example in a multivariate normal model on x_j, μ would denote the mean of the distribution. However, in a feature model on f_j that is invariant to image brightness and contrast, ϕ might represent the mean of the set of standardized features f_j.

2.2 General Framework

The goal is to use training data gathered from a human who has observed J images from the class and located p facets in those images to predict facet locations in a new image from the same class. The template (μ, ϕ) can be assumed to be either fully or

partially unknown, or entirely fixed. For the present discussion we assume that both μ and ϕ are unknown. See section 2.3 for other cases. Note however that each facet does have a label, defined either explicitly or implicitly using an atlas image.

The following paragraphs discuss the modeling assumptions used hereafter on the process by which an intelligent observer quotes facet locations and feature values using a training set of J images. The human observer has a list of facets in mind (ie. facet labels) and identifies their positions $x_1, \ldots, x_J$ in a set of J images from the class. The corresponding facet feature values are read from the image data, namely $f_{j,i} = f_{Q_j}(x_{j,i})$ where $f_Q(x)$ denotes a feature function on the image Q at the location x (e.g. the intensity at x, so that $f_Q(x) = Q(x)$, see section 2.4.2 for others).

In this section the image Q_j is assumed unknown. It is, of course, known to the observer since this is the method by which one quotes feature values $f_{j,i}$. However, for modeling purposes, the feature value $f_{j,i}$ is a random variable separate from (but not necessarily independent of) the facet location $x_{j,i}$. For example, in the simple case where the feature value is equal to the image intensity at $x_{j,i}$, one could view this kind of modeling as putting a distribution on the value of the image at $x_{j,i}$. Note that the image Q_j itself is never modeled by a probability distribution, it is simply compounded into the method by which the observer quotes facet locations and feature values.

Therefore the data for p facets identified in J images has the form

$$\{\{(x_{1,1}, f_{1,1}), \ldots, (x_{1,p}, f_{1,p})\} \ldots, \{(x_{J,1}, f_{J,1}), \ldots, (x_{J,p}, f_{J,p})\}\},$$

where $x_{j,p}$ is a d-dimensional vector. This long notation is shortened to

$$\{(x_1, f_1), \ldots, (x_J, f_J)\}.$$

Independence is assumed across images so that the data (x_j, f_j) are drawn from some joint density $p(x_j, f_j|\mu, \phi, \theta)$, where θ is a scale parameter representing the variability

of (x_j, f_j) around the template location and feature (μ, ϕ). This joint distribution can be written in two factors

$$p(x_j, f_j|\mu, \phi, \theta) = p(x_j|\mu, \phi, \theta)p(f_j|x_j, \mu, \phi, \theta).$$

Conditioning on x_j in the second term is natural since it can be used to model, for instance, spatial dependencies in the feature vector f_j. Note that $p(f_j|x_j, \cdot)$ is not a point mass on $(f_{ji} = f_{Q_j}(x_{ji}), ..., f_{jp} = f_{Q_j}(x_{jp}))$ since the conditional distribution does not include Q_j in its conditional arguments right of the bar. Conditional on x_j but not Q_j, f_j is a random variable.

Next, θ is written in two components, $\theta = (\theta_x, \theta_f)$. The former represents the variability of x_j around μ and the latter represents the variability of f_j around ϕ. Under this assumption, the joint distribution is written

$$p(x_j, f_j|\mu, \phi, \theta) = p(x_j|\mu, \phi, \theta_x)p(f_j|x_j, \mu, \phi, \theta_f).$$

Facet locations are assumed to be independent of template feature values, $p(x_j|\mu, \phi, \theta_x) = p(x_j|\mu, \theta_x)$. This is reasonable since, for example, knowing that two facets i and i' have similar template features ϕ_i and $\phi_{i'}$ actually tells nothing about the locations of those facets $x_{j,i}$ and $x_{j,i'}$ since there can be several locations in an image with similar feature values. And similarly, feature values are assumed independent of the template location, $p(f_j|x_j, \mu, \phi, \theta_f) = p(f_j|x_j, \phi, \theta_f)$ since the information contained in the locations x_j of the facets in this particular image most likely negates the information contained in the template locations μ.

The joint posterior on the template and scale parameters given the whole set of data $(x, f) = \{(x_1, f_1), \ldots, (x_J, f_J)\}$ is written

$$p(\mu, \phi, \theta_x, \theta_f|x, f) = \frac{1}{Z}p(\mu, \phi, \theta_x, \theta_f)\prod_{j=1}^{J} p(x_j|\mu, \theta_x)p(f_j|x_j, \phi, \theta_f)$$

where Z is a normalizing constant independent of μ, ϕ, θ_f, and θ_x. The posterior

predictive distribution for a new set of data $(\tilde{x}, \tilde{f})$ is

$$p(\tilde{x}, \tilde{f}|x, f) = \int p(\tilde{x}, \tilde{f}|\mu, \phi, \theta_x, \theta_f) p(\mu, \phi, \theta_x, \theta_f|x, f) d\mu d\phi d\theta_x d\theta_f. \qquad (2.1)$$

However, remember that the goal of this analysis is to use the training data (x, f) to predict facet locations $\tilde{x}$ in a new image $\tilde{Q}$ from the class. The distribution (2.1) does not condition on that new image, and is therefore not quite the desired distribution. Section 2.5 will show that a reasonable predictive distribution on locations in the new image can be defined based on (2.1).

2.3 Template Definition

The model class allows for general template specification, namely unknown μ and/or unknown ϕ. If both μ and ϕ are unknown, and facet labels are defined explicitly (not using an atlas image), then every facet then can be called a landmark and a morphometric analysis ensues similar to those outlined in Bookstein (1991), Dryden and Mardia (1996), Dryden (1997), and Lele and Cole (1996). In this case, along with seeking posterior distributions on the model scale parameters θ_x and θ_f, posterior distributions on μ and ϕ can also be determined to describe average facet locations and feature values in the class.

However, in the image understanding applications such as image registration and segmentation with which this paper is concerned, a large number of facets is required to adequately probe the structures of the images in the class. In these cases, the length of the facet vector might approach the number of pixels in the image, and it becomes impossible to explicitly label all the facets with anatomically meaningful labels. Analysis can go forward if an atlas image from the class is used to implicitly apply facet labels. Facets are simply arrayed in some regular fashion (for example a d-dimensional grid) in the atlas image and their implicit labels are extracted from

their locations in that image. The data (x, f) gathered from a human observer are still locations and feature values in other images from the class, but in this case the observer looks for structures similar to those found in the atlas image. In these applications it is usually preferred to define μ to be the facet locations on the array used in the atlas, and always condition on μ. This is the approach taken in this paper. The atlas image can be used to define the template feature values, so that ϕ_i is the atlas image evaluated at μ_i. Then, once data has been collected to estimate θ_x and θ_f, the resulting predictive density on facet locations $\tilde{x}$ in a new image $\tilde{Q}$ (given in section 2.5) predicts a mapping of the atlas image features into the new image. This is the framework used in Chapters 4 and 5. Otherwise, the atlas image could be used only as a "free" first datum where ϕ is left to be estimated.

When using an atlas image, one must of course be prepared for missing data, since there is usually an overwhelming number of facets for an observer to identify. This is discussed in section 2.6.

Note that the layout of facets in the atlas image is completely arbitrary. Applications in this paper indicate that a grid should be used, however other schemes have been investigated. See, for instance, McCulloch *et al.* (1996) and Laading *et al.* (1997).

2.4 Data Model in this Paper

In this section, a definition is given of the data model $p(x_j, f_j|\mu, \phi, \theta)$ used throughout this paper. Under independent sampling of facet properties (x_j, f_j), the model on J samples is $\prod_{j=p}^{J} p(x_j, f_j|\mu, \phi, \theta)$. As mentioned in section 2.2, θ is separated into $(\theta_x, \theta_f) = (\Sigma_x, \Sigma_f)$, where capital Σ_x and Σ_f are used to denote covariance matrices defining the variability of the vector x_j around μ and f_j around ϕ, respectively. The

full likelihood for J samples is therefore

$$p(x_j, f_j|\mu, \phi, \theta) = \prod_{j=1}^{J} p(x_j|\mu, \Sigma_x)p(f_j|x_j, \phi, \Sigma_f).$$

In section 2.4.1 an attractive distribution for $p(x_j|\mu, \Sigma_x)$ is derived and in section 2.4.2 several useful feature distributions are discussed.

2.4.1 A Deformation Model

The deformation model derived in this section is meant for the situation in which an atlas image is used to define the template feature values ϕ given a fixed template orientation μ. The application to keep in mind is image registration. A dense group of facets is defined in the template image and the model given in this section quantifies reasonable deformations within that group of facets in new images from the class.

The set of template locations μ is fixed to lie on a d-dimensional grid in the d-dimensional atlas, and the deformation model is $p(x_j|\mu, \Sigma_x) = N(\mu, \Sigma_x = (\kappa P)^{-1})$ where $N(\cdot, \cdot)$ denotes the multivariate normal distribution. Throughout the following derivation, the subscript j denoting image will be ignored.

The goal is to derive a normal distribution whose covariance matrix $\Sigma_x = (\kappa P)^{-1}$ exhibits desirable deformation properties. For example, facets close together should be more highly correlated than those farther apart. Furthermore, the correlation structure should be isotropic and homogeneous, meaning that the correlation between two facets should be a function of the distance between them, but not have any higher dependence on their absolute locations. Another desirable property that will become apparent in subsequent chapters is the model's ability to conform to a hierarchical maximization scheme. Essentially, for this to be true, small groups of contiguous facets should have some kind of common representation in the model. Finally, it would be very useful to be able to derive the covariance matrix analytically and

therefore be able to produce exact samples from the model.

Many of these properties are implemented in the model defined in this section. A hierarchical normal distribution is developed that exhibits nearly isotropic, homogeneous covariance structure whose covariance matrix is derivable analytically and in which exact sampling is easily implemented. The hierarchical nature of the model facilitates the very successful hierarchical density maximization method discussed in Chapter 3.

The model is derived first in a one-dimension image for simplicity. (In this case, μ_i is a scalar and the vector μ is simply evenly spaced points.) Higher dimensional analogues are easily derived from the one-dimensional case and are discussed at the end of this section.

To derive the hierarchical normal distribution, the vector x is augmented to contain both observable and latent facets. The latent facets are simply placeholders in a graph structure on which the hierarchical normal distribution is placed. Figure 2.1 shows a graphical representation of the random variables used in this model. Observable facets are located on the bottom level of the graph and the remainder are latent.

Facet locations are now written $x = (x_1, ..., x_L)$ for L levels in the graph, so that x_L is the vector of observable facet locations. The mean of the normal distribution, μ, is augmented similarly, $\mu = (\mu_1, ..., \mu_L)$. Latent facet means $(\mu_1, \ldots, \mu_{L-1})$ are organized on grids similar to the observable facet means, however these grids become half as dense and half as numerous each level up in the hierarchy. The bottom level must have an even number of facets and the graph is terminated at the top with a single facet.

The random variable at each node in the graph is the one-dimensional location $x_{l,i}$ of facet i on level l. The conditional independence structure in the model is also given

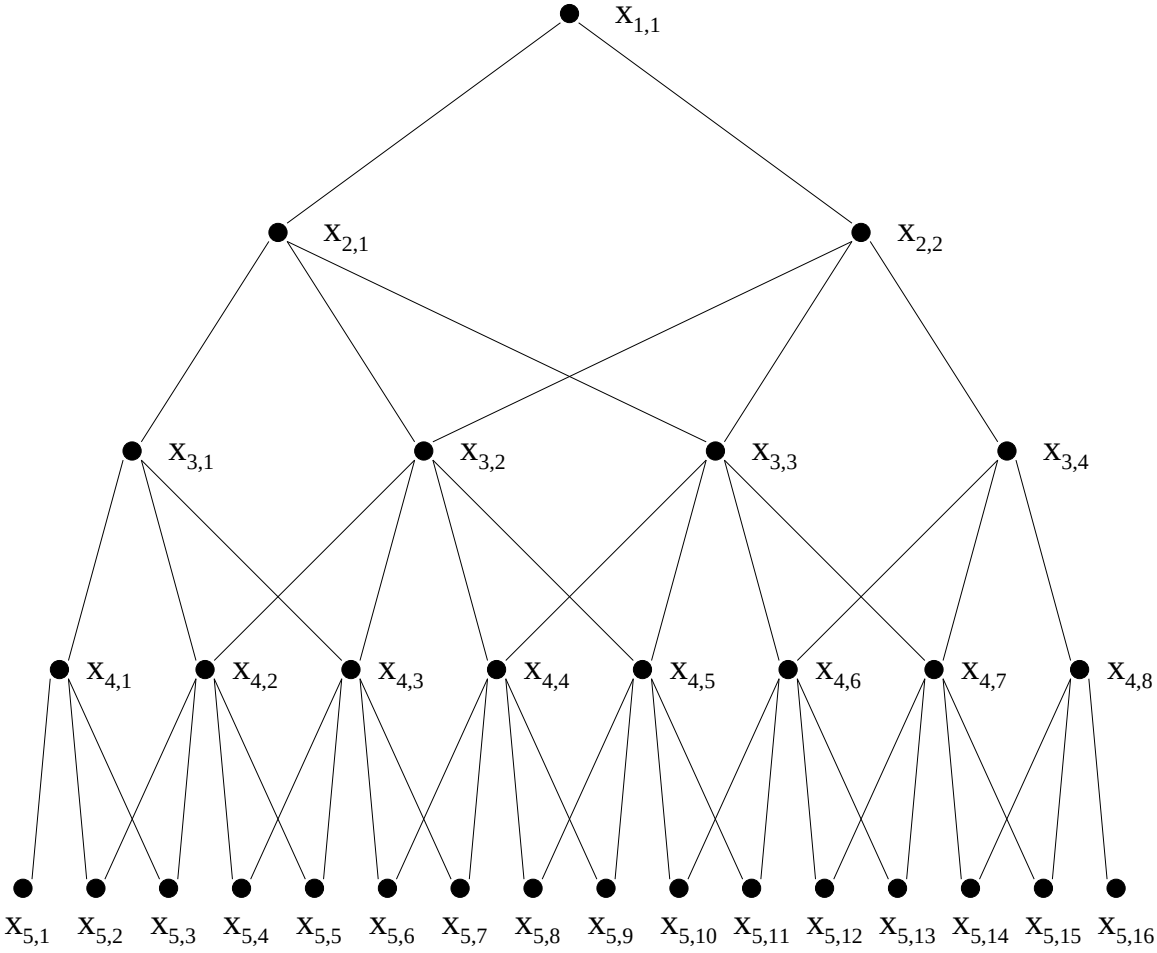


Figure 2.1: The conditional independence structure defined in the model equation (2.2). In the figure, $x_{l,i}$ denotes the location of facet i on level l. In the derivation of equation (2.2), cyclic boundary conditions are used which imply connections between, for example, $x_{3,1}$ and $x_{2,2}$ which are not shown in the figure.

in Figure 2.1. The lack of an arc connecting two nodes implies conditional independence between the two nodes given the rest of the nodes in the graph. Throughout the graph, parent facets are connected to the four children one level down whose means $\mu_{l,i}$ are closest to it.

The random vector $x = (x_1, ..., x_L)$ is distributed according to a hierarchical model

$$\begin{aligned} p(x_1|\cdot) &= N(\mu_1, (\kappa c_1)^{-1}) \\ p(x_2|x_1, \cdot) &= N(\mu_2 + A_2(x_1 - \mu_1), (\kappa c_2)^{-1} I) \\ p(x_3|x_2, \cdot) &= N(\mu_3 + A_3(x_2 - \mu_2), (\kappa c_3)^{-1} I) \\ &\vdots \\ p(x_L|x_{L-1}, \cdot) &= N(\mu_L + A_L(x_{L-1} - \mu_{L-1}), (\kappa c_L)^{-1} I) \end{aligned} \tag{2.2}$$

where conditioning on μ and κ is suppressed. In equation (2.2), $(c_2, \ldots, c_L)$ are scalar level weights (discussed below), I is the identity matrix, and A_l is a $2^{l-1} \times 2^{l-2}$ design matrix defining the connections in the graph. The first three design matrices are listed here for example.

$$A_2 = \begin{bmatrix} w_1 + w_2 \\ w_1 + w_2 \end{bmatrix},$$

$$A_3 = \begin{bmatrix} w_1 & w_2 \\ w_1 & w_2 \\ w_2 & w_1 \\ w_2 & w_1 \end{bmatrix},$$

$$A_4 = \begin{bmatrix} w_1 & 0 & 0 & w_2 \\ w_1 & w_2 & 0 & 0 \\ w_2 & w_1 & 0 & 0 \\ 0 & w_1 & w_2 & 0 \\ 0 & w_2 & w_1 & 0 \\ 0 & 0 & w_1 & w_2 \\ 0 & 0 & w_2 & w_1 \\ w_2 & 0 & 0 & w_1 \end{bmatrix},$$

where w_1 and w_2 are connection weights. These weights are constrained to be between 0 and 1 and to sum to unity so that the conditional deviation on a facet's location from its template location in (2.2) is equal to a weighted average of its parents' deviations. The weight w connecting facet $(\cdot)_{l,i}$ with facet $(\cdot)_{l+1,j}$ is defined in this paper to be inversely proportional to $|\mu_{l,i} - \mu_{l+1,j}|$ however other choices could be used to model different facet correlation structures. Note that for analytical convenience, cyclical boundary conditions have been used at the edges. In practice the edge facets could be connected only to their closest parents, however marginal variances and covariances involving edge facets will be altered relative to facets on the interior of the graph.

The level weights $(c_2, \ldots, c_L)$ can be chosen to ensure that the marginal variances of all facet locations $x_{l,i}$ are approximately equal. This property is very desirable since the model is then scalable, ie. marginal variances on the observable facets are always similar no matter how many levels of latent facets are used. Using this fact, the size of the graph can be determined by practical considerations such as the desired facet resolution and computational facilities.

To derive appropriate level weights ensuring that facet marginal variances are equal, the marginal covariance matrix on facet locations on one level is derived. The following simple identity (Gelman *et al.*, 1995) is used to derive this single level marginal covariance matrix. If vectors U and V are both multivariate normal such that

$$U|V \sim N(AV, \Sigma_{U|V})$$

$$V \sim N(\mu_V, \Sigma_V)$$

then the joint density on the vector $[U\ V]'$ is

$$\begin{bmatrix} U \\ V \end{bmatrix} \sim N\left(\begin{bmatrix} A\mu_V \\ \mu_V \end{bmatrix}, \begin{bmatrix} A\Sigma_V A' + \Sigma_{U|V} & \Sigma_{U|V} \\ \Sigma'_{U|V} & \Sigma_V \end{bmatrix} \right). \tag{2.3}$$

Due to the hierarchical form of the normal distribution (2.2), the marginal covariance matrix for facet locations on level l can be determined using the identity (2.3)

recursively:

$$\begin{aligned}
\Sigma_l &= A_l \Sigma_{l-1} A_l' + (\kappa c_l)^{-1} I \\
&= A_l \left(A_{l-1} \Sigma_{l-2} A_{l-1}' + (\kappa c_{l-1})^{-1} I \right) A_l' + (\kappa c_l)^{-1} I \\
&= A_l A_{l-1} \Sigma_{l-2} A_{l-1}' A_l' + (\kappa c_{l-1})^{-1} A_l A_l' + (\kappa c_l)^{-1} I \\
&\vdots \\
&= (\kappa c_1)^{-1} [A_l A_{l-1} \cdots A_2][A_2' \cdots A_{l-1}'] + \cdots + (\kappa c_l)^{-1} I \\
&= \kappa^{-1} \left\{ \sum_{j=1}^{l-1} c_j^{-1} \left[\prod_{j+1}^{i=l} A_i \right] \left[\prod_{i=j+1}^{l} A_i' \right] + c_l^{-1} I \right\}.
\end{aligned} \tag{2.4}$$

The notation $\prod_{j+1}^{i=l} A_i$ in the last row represents the matrix product $A_l A_{l-1} \cdots A_{j+1}$.

Setting $w_1 = .75$ and $w_2 = .25$ (according to the template facet spacing), the marginal variance on a facet location can be derived as a function of the level weights $(c_1, \ldots, c_L)$. For example, the marginal variance on facet $(\cdot)_{3,2}$ in figure 2.1 is

$$\kappa^{-1}(c_1^{-1} + 0.625 c_2^{-1} + c_3^{-1}).$$

The corresponding marginal variances for facets $(\cdot)_{4,4}$ and $(\cdot)_{5,8}$ are

$$\kappa^{-1}(c_1^{-1} + 0.532 c_2^{-1} + 0.625 c_3^{-1} + c_4^{-1}), \text{ and}$$
$$\kappa^{-1}(c_1^{-1} + 0.508 c_2^{-1} + 0.531 c_3^{-1} + 0.625 c_4^{-1} + c_5^{-1}),$$

respectively. To ensure that facet marginal variances are approximately equal across levels in the graph, the above three expressions are equated to generate the proper relationship among the level weights. The resulting relationship,

$$c_3 = 2.67 c_2, \; c_4 = 4.27 c_2, \text{ and } c_5 = 6.83 c_2, \tag{2.5}$$

ensures that all facet marginal variances are approximately equal to $\kappa^{-1}(c_1^{-1} + c_2^{-1})$. Variances on other facets on the same level in the graph are very close to the values given above. For example, under the constraint (2.5), level five facet marginal variances are within approximately $(\kappa c_2)^{-1}/10$.

This small calculation for a five level graph can be easily extended to larger graphs by using equation (2.4) to derive the higher level marginal covariance matrices.

Higher dimensional models are constructed analogously to the one-dimensional case. The template mean location vector μ is now placed on a d-dimensional grid, and the random variables on the graph in Figure 2.1 are d-dimensional locations. Deformations in all dimensions are modeled equivalently so that the marginal covariance matrix on each facet's d-dimensional location $x_{l,i}$ is proportional to the identity matrix I_d. The vector x is written as the facet's d-dimensional locations strung together, so that for a 5-level graph in 3-dimensions, x would have length $3 \times (8^4 + 8^3 + 8^2 + 8 + 1) = 14043$. The covariance matrix has a similarly increased size.

Parent/child connections are set up analogously to the one-dimensional case. For example, in two dimensions, each parent facet is connected to the $4 \times 4 = 16$ nearest facets on the next level down. Figure 2.2 shows a five-level two-dimensional graph viewed from the top. The sixteen children of one level 3 facet and of one level 4 facet are outlined. In three dimensions each parent has $4 \times 4 \times 4 = 64$ children.

In two-dimensions, the weight vector w has length 4, in 3-dimensions it has length 8, and its elements are again set inversely proportional to the Euclidean distance between parent and child mean locations. For example, under the grid geometry shown in Figure 2.2, each facet has one close parent, two medium distance parents, and one far parent.

The approximate marginal variance on a facet location in any direction can be calculated as in the one-dimensional case using equation (2.4). Furthermore, the relationship among level weights $(c_1, \ldots, c_L)$ similar to (2.5) can be derived to ensure approximately equal marginal variances on all facets in the graph.

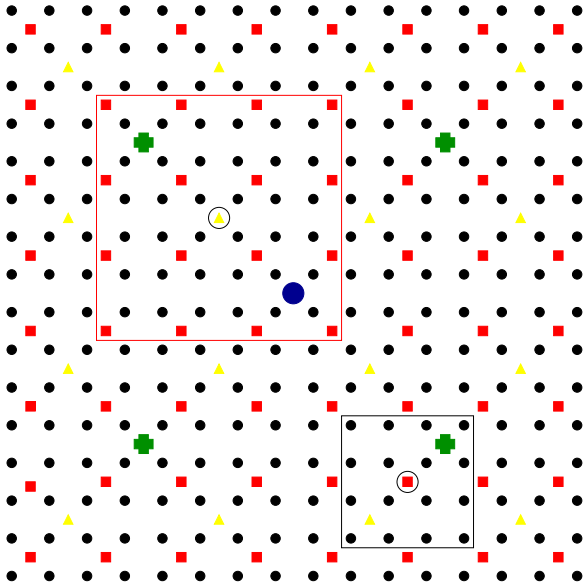


Figure 2.2: A two dimensional facet graph with five levels. The levels from top to bottom are represented by: blue circle, green plus, yellow diamond, red square, and black circle. Two sets of children are outlined on two levels below their circled parents.

2.4.2 Feature Models

In this section the feature model $p(f_j|x_j, \phi_j, \theta_f)$ on facets in one image j is discussed. For clarity the subscript j is dropped in the following discussion and the distribution is written simply $p(f|x, \phi, \theta_f)$. Remember that only the bottom level of facets in the hierarchical normal model is assumed to be observable, so the vector of features discussed here is only as long as the number of facets on the bottom level.

The feature model describes our beliefs about the appearance of facets found in the image class, where appearance is quantified by the notion of a feature function. There has been much work on deriving such functions to be useful in different applications. See, for example, Studholme *et al.* (1997).

For the purposes of this paper, a feature function $f_Q(x)$ is a function on d-dimensional space, $f_Q(x) : R^d \rightarrow R^n$. The subscript Q indicates that its functional form depends on an image $Q : R^d \rightarrow R^1$. Also note that the range of f_Q indicates

that it can be a vector-valued function in R^n.

The feature function is used to measure an image Q near the location x in such a way that the measurement is comparable with the template feature value ϕ. To be comparable, ϕ must have the same dimension as $f_Q(x)$, namely $\phi \in R^n$. A few examples will illuminate the meaning of the feature function.

The canonical feature function, $f_Q^1(x) : R^d \to R^1$, is the image intensity at x,

$$f_Q^1(x) = Q(x). \tag{2.6}$$

In this case each template feature value ϕ_i is one-dimensional and, under an appropriate feature distribution in the class could be thought of as the mean intensity value of the observed facet features f_i.

A useful extension of (2.6) is the function $f_Q^2(x) : R^d \to R^n$, whose vector value is the image intensity at several locations $(z_1, \ldots, z_n)$ in a small region centered at x (ie. $\frac{1}{n}\sum z_i = x$) and fixed with respect to x,

$$f_Q^2(x) = (Q(z_1), \ldots, Q(z_n)). \tag{2.7}$$

A second vector-valued feature function, $f_Q^3(x) : R^d \to R^d$, can be defined as the set of first spatial derivatives of Q at x,

$$f_Q^3(x) = (\frac{\partial}{\partial u_1}Q(x), \ldots, \frac{\partial}{\partial u_d}Q(x)) \tag{2.8}$$

where $x = (u_1, \ldots, u_d)$. Numerous extensions and combinations of this short list of feature functions might be found useful in application.

The feature distribution $p(f|x, \phi, \theta_f)$ models the variability of each facet's possible feature value f_i around the mean feature ϕ_i. In the following exposition, several feature distributions are defined that show promise in application. In Chapters 4 and 5, two of these distributions are used with image data. In all but the last of the following examples, variability is modeled using the normal distribution for its

parametric convenience, so that θ_f is a covariance matrix Σ_f, but more robust versions of these distributions could be entertained. One non-standard distribution has been found useful in some applications, and is described last in this section.

Using the first feature function $f_Q^1(x)$, a normal distribution can be defined which captures the notion that, when an observer searches for a facet location in a new image, they look for intensities in the image that are close to a mean intensity for the facet. For every facet i, its intensity is normally distributed with mean ϕ_i and variance $(\tau)^{-1}$. If all facets are treated independently, then the feature distribution on the vector of facet features f has the form

$$p(f|x,\phi) = (\tau/2\pi)^{p/2} \exp\left(-\frac{\tau}{2}\sum_i (f_i - \phi_i)^2\right). \tag{2.9}$$

However, it may be reasonable to assume that nearby facets have similar feature values. This can be modeled by generalizing Σ_f through a spatial normal model as investigated by Cressie (1993). For instance, in the flavor of an homogeneous Gaussian process, Σ_f could be parameterized such that the covariance between f_i and f_k is determined by $(x_i - x_k)$,

$$p(f|\phi,\Sigma_f(x)) = (2\pi)^{-p/2}|\Sigma_f(x)|^{-1/2}\exp\left(-\frac{1}{2}(f-\phi)'\Sigma_f(x)^{-1}(f-\phi)\right). \tag{2.10}$$

Using this feature density, the joint data model $p(x, f|\mu,\phi,\theta)$ can be written down, although it has a non-standard form since Σ_f is a function of x. A simplification of this model would parameterize Σ_f in terms of the template locations μ rather than x. This would ease inference on Σ_f and under small deformations would yield similar results as the model in (2.10).

Using the second feature function $f_Q^2(x)$, equation (2.7), one can model the notion that perhaps an observer locates features by matching intensities in a small region around the candidate location x to a template mean intensity profile represented by

ϕ. If an independent normal distribution is assumed on the n-dimensional feature vector around each facet, then the feature density on a whole set of features vectors

$$f = ((f_1)_1, \ldots, (f_1)_n, \ldots, (f_p)_1, \ldots, (f_p)_n)$$

is

$$p(f|x, \phi, \tau) = (\tau/2\pi)^{np/2} \exp\left(-\frac{\tau}{2} \sum_{i=1}^{p} \sum_{k=1}^{n} ((f_i)_k - (\phi_i)_k)^2 \right). \tag{2.11}$$

The same kind of vector-valued feature function can be used to incorporate the notion that an observer locates features by matching intensity profiles as above, but ignores the brightness or contrast of the profile. In this case the observer searches not for absolute intensities, but rather patterns in intensity. For example, one facet may always be found at a peak in intensity but the magnitude of that peak may not be important. Using the standardized feature vectors $(f_i)^*$ and $(\phi_i)^*$, where

$$(f_i)_k^* = \frac{(f_i)_k - \frac{1}{n}\sum_k (f_i)_k}{\sqrt{\frac{1}{n}\sum_k (f_i)_k^2 - (\sum_k (f_i)_k)^2}} \quad \text{and} \quad (\phi_i)_k^* = \frac{(\phi_i)_k - \frac{1}{n}\sum_k (\phi_i)_k}{\sqrt{\frac{1}{n}\sum_k (\phi_i)_k^2 - (\sum_k (\phi_i)_k)^2}}, \tag{2.12}$$

a density can be defined that is invariant to brightness and contrast. Under an independent normal distribution, this has the form

$$p(f|x, \phi, \tau) = (\tau/2\pi)^{np/2} \exp\left(-\frac{\tau}{2} \sum_{i=1}^{p} \sum_{k=1}^{n} ((f_i)_k^* - (\phi_i)_k^*)^2 \right) \tag{2.13}$$

Note that this feature model is indeterminate (and therefore improper) on the random variable f_i. However, under appropriate prior the posterior on τ is proper given data (x, f).

A final vector feature density is based on a standardized regression performed on the feature vector $((f_i)_1, \ldots, (f_i)_n)$ using the mean vector $((\phi_i)_1, \ldots, (\phi_i)_n)$ as a covariate. This feature density will be invariant not only to brightness and contrast,

but also to the linear trend relationship between f_i and ϕ_i. For each facet i, the linear regression can be represented in the form

$$(f_i)_k^* = (\beta_0)_i + (\beta_1)_i(\phi_i)_k^* + (\epsilon_i)_k \tag{2.14}$$

where $(\epsilon_i)_k \sim N(0, \tau^{-1})$. The amount of linear trend invariance in the model can be adjusted by a prior on the coefficients $(\beta_0)_i$ and $(\beta_1)_i$, after which the joint on $(f_i, (\beta_0)_i, (\beta_1)_i)$ could be marginalized over these coefficients. Or, a simple solution to allow complete linear trend invariance is to set the coefficients to their least squares estimates and calculate the resulting distribution on (f_i). For each facet i, the resulting feature density is

$$p(f_i|\phi_i, x_i, \tau) \propto \exp\left(-\frac{n\tau}{2}(1 - f_i^{*\prime}\phi_i^*\phi_i^{*\prime}f_i^*)\right), \tag{2.15}$$

and an independent density over all facets is a product of these distributions. The result of this heuristic plug-in method is a non-standard density for which posterior inference on τ is difficult, but for a given τ it has been found to be very effective in some registration applications. The cardiac SPECT application in Chapter 5 uses this density with great success.

2.5 Posterior and Predictive Distributions

Because of the simple normal forms of the shape and feature models given in section 2.4, under appropriate priors, posterior distributions on the hyperparameters μ, ϕ, κ and τ are quite straightforward. Similarly posterior predictive distributions seem to be as familiar, however the actual predictive distribution useful in application is more subtle and interesting, and will be discussed near the end of this section.

A full exposition of all possible posterior and predictive densities for the normal model class is not given here, see for example Gelman *et al.* (1995). Instead, only those distributions which are used in the later applications of Chapters 4 and 5 are discussed.

For the registration applications in this paper the template facet location vector μ and feature vector ϕ are always assumed known. A non-informative prior is used on κ and τ, $p(\kappa, \tau) \propto (\kappa\tau)^{-1}$. In the following discussion, the shape distribution $p(x|\mu, \kappa)$ considered is the model defined in section 2.4.1 and the feature model is always assumed to be the intensity distribution (2.9).

Data is in the form of manual observations of facet locations and their associated feature values (x, f) in J images from the class, given their locations and features (μ, ϕ) in an atlas image. The posterior distributions discussed in this section assume that every facet has been identified manually so that there is no missing data, however extension to the missing data case is relatively straightforward and is discussed in section 2.6.

The following discussion only deals with the observable facets on the bottom level of the facet graph, so the level indicator is dropped, i.e. $x_{L,i}$ is written x_i and the associated feature value is written f_i.

First, given manual observations of facet locations and features in J images, $\{(x_1, f_1), \ldots, (x_J, f_J)\} = (x, f)$, and using the intensity feature function (2.9), the posterior distribution on the hyperparameters κ and τ is the product of two independent Gamma distributions, $p(\kappa, \tau|x, f, \phi, \mu) = p(\kappa|x, \mu)p(\tau|f, \phi)$, where

$$p(\kappa|x, \mu) = \text{Gamma}\left(\frac{dpJ}{2}, \frac{1}{2}\sum_{j=1}^{J}(x_j - \mu)'P(x_j - \mu)\right) \tag{2.16}$$

and

$$p(\tau|f, \phi) = \text{Gamma}\left(\frac{pJ}{2}, \frac{1}{2}\sum_{j=1}^{J}(f_j - \phi)'(f_j - \phi)\right). \tag{2.17}$$

The sums in (2.16) and (2.17) are taken over facet identifications in the images $(1, \ldots, J)$, x_j and μ are vectors of length $d \times p$, and f_j and ϕ have length p. The precision matrix P is defined as $(\kappa\Sigma_L)^{-1}$ where Σ_L is given in equation (2.4). The

calculation of the second parameter in the posterior Gamma distribution on κ is discussed in section 2.6.

The joint posterior predictive distribution for facet locations $\tilde{x}$ and feature values $\tilde{f}$ in the image class is also a product of two independent distributions, $p(\tilde{x}, \tilde{f}|x, f) = p(\tilde{x}|\mu, x)p(\tilde{f}|\phi, f)$. The posterior predictive distribution on $\tilde{x}$ is written

$$p(\tilde{x}|\mu, x) = MVt_{dpJ}\left(\mu, \frac{1}{pJ}\left[(\mu - \bar{x})'P(\mu - \bar{x}) + \sum_{j=1}^{J}(x_j - \bar{x})'P(x_j - \bar{x})\right]' P\right), \tag{2.18}$$

where $\mathrm{MVt}_n(y, Z)$ is the multivariate t distribution on n degrees of freedom with center y and scale matrix Z, and $\bar{x}$ is the arithmetic mean vector $\bar{x} = \sum_{j=1}^{J} x_j$. Under the intensity feature density (2.9), the posterior predictive distribution on $\tilde{f}$ is a product of scalar t-distributions.

$$p(\tilde{f}|\phi, f) = \prod_i \mathrm{t}_{pJ}\left(\phi_i, \frac{1}{pJ}\sum_j\sum_i (f_{j,i} - \phi_i)^2\right). \tag{2.19}$$

The joint posterior predictive distribution $p(\tilde{x}, \tilde{f}|x, f, \mu, \phi) = p(\tilde{x}|x, \mu)p(\tilde{f}|f, \phi)$ on the facet properties $(\tilde{x}, \tilde{f})$ characterizes these properties in the image class as a whole. For instance, one could sample a posterior predicted set of facet properties and display them as $\tilde{f}$ vs. $\tilde{x}$ to give the illusion of a predicted image in the image class. It is only an illusion since this is not a realization of any image in the terminology of section 2.1 because no function on R^d has been sampled. However, with sufficient facets an image function could be approximated by, for instance, interpolation of $\tilde{f}$ between the facet locations $\tilde{x}_i$.

To automate the task of feature identification in a new image Q, for example to segment that new image, a different sort of predictive distribution is required. This distribution must incorporate the variability seen in the training facet properties

(x, f), while also paying heed to the fact that the predicted facet property pairs $(\tilde{x}, \tilde{f})$ must lie in the image Q.

In the formulation used so far on facet property pairs (x, f), this constitutes a constraint on every facet's property pair, since the pair must be allowable in the image Q. Namely, the constraint can be written as

$$f_i = f_Q(x_i) \quad \forall i, \tag{2.20}$$

where $f_Q(x_i)$ is the feature function defined in section 2.4.2. For example, using the vector feature function (2.7), for each facet i, the vector f_i must be the set of intensities in the image at the locations $((z_i)_1, \ldots, (z_i)_n)$ centered at x_i. The constraint (2.20) becomes

$$(f_i)_1 = Q((z_i)_1), \ldots, (f_i)_n = Q((z_i)_n) \quad \forall i.$$

However, under the joint density on (x, f), the event in (2.20) has zero probability. To understand this, refer to Figure 2.3 where a hypothetical joint density is displayed on a (scalar) facet location x and (scalar) feature f. (This could be considered a posterior predictive density or it could be just the data model given the hyperparameters.) The image is represented by the function $Q(x)$ on the graph and the constraint on (x, f) implied by this image is that the pair (x, f) must lie on this line. This is clearly a zero probability event under any of the joint models proposed thus far.

Therefore, to properly define a predictive density on facet locations in a new image, the image constraint is redefined so that the event it represents has non-zero probability and then a family of distributions is defined whose limit is the desired predictive density on facet locations. A formal discussion is given in Appendix A, but essentially the image constraint (2.20) is re-written

$$|f_i - f_Q(x_i)| < \epsilon \quad \forall i, \tag{2.21}$$

for the scalar feature case, or

$$|(f_i)_k - (f_Q(x_i))_k| < \epsilon \quad \forall i, k \tag{2.22}$$

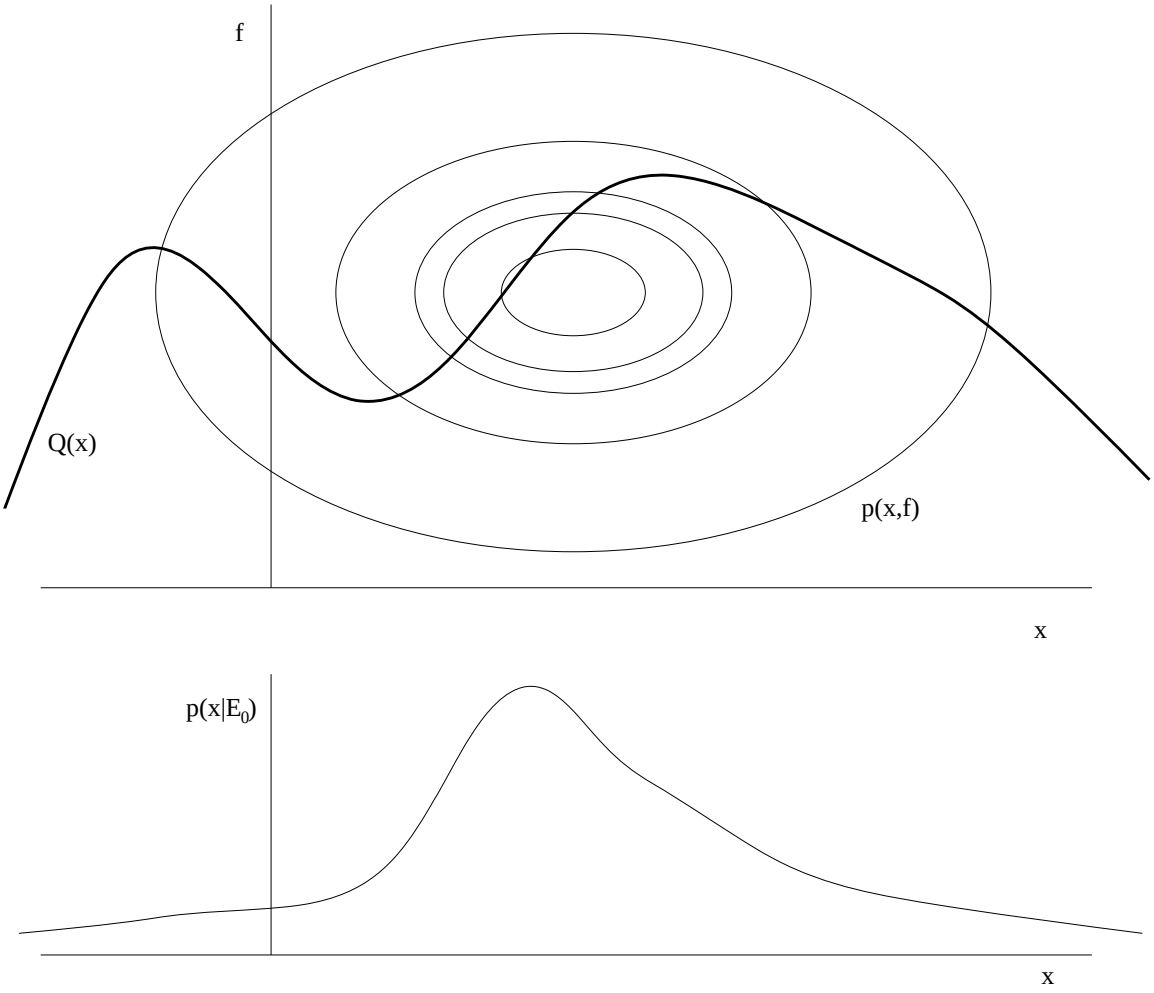


Figure 2.3: The joint density on the properties (x, f) of one facet and the constraint enforced by the one-dimensional image $Q(x)$ from the class. The contours in the top panel are iso-curves of $p(x, f)$. The bottom panel shows the shape of the predictive distribution $p(x|E_0)$ on the facet location in the image $Q(x)$ implied by the constraint E_0 applied to the joint in the top panel. See the text for the definition of E_0.

for a vector feature function. All the elements of the feature vector f_i must be within ϵ of the feature function on the image evaluated at x_i. If the (non-zero probability) event in equation (2.21) is written E_ϵ, then the resulting "conditional" distribution on x given the event E_ϵ is written $p(x|E_\epsilon, \cdot)$. Let E_ϵ in the limit as $\epsilon \to 0$ be written E_0. The limiting predictive distribution on facet locations in a new image from the class is shown in Appendix A to be

$$p(x|E_0, \cdot) \propto p(x, \{f_i = f_Q(x_i)\ \forall i\}|\cdot). \tag{2.23}$$

The distribution on x is proportional to the joint distribution on (x, f) in which each random variable f_i is set to the value of the feature function on the image Q at the location x_i. The density (2.23) can represent the constrained predictive density, in which case $(\cdot)$ represents the training data (x, f), or it can represent a constrained data model, in which case $(\cdot)$ represents the parameters (κ, τ). This result holds for scalar and vector valued feature functions by employing the constraints (2.21) and (2.22), respectively.

Since the conditional predictive distribution (2.23) on facet locations in a new image is known up to a normalizing constant, it can be maximized for point estimates and sampled for variability. These issues are developed in Chapter 3.

2.6 Missing Data Considerations

When an atlas image is used to implicitly define facet labels (see section 2.3), for the purpose of image registration, it is generally necessary to use many facets (on the order of the number of pixels in the image) to perform an adequate registration. Since it is impossible to have a human observer identify such a large number of facets in several images from the class, any posterior inference on the scale parameters θ_x and θ_f must be based on missing data. To calibrate the model, it is perhaps reasonable to gather from an observer locations of a subset of the facets.

Due to the normal form of the data model, under some conditions the appropriate posterior can be derived semi-analytically. First, in the registration problem the task involves always conditioning on the locations of facets in an atlas image where the atlas implicitly defines the template feature values ϕ from the fixed template locations μ. Given data (x, f), the relevant posterior on the scale parameters is therefore

$$p(\theta_x, \theta_f | x, f, \mu, \phi) \propto p(x, f | \mu, \phi, \theta_x, \theta_f) p(\theta_x, \theta_f). \tag{2.24}$$

The last factor does not condition on (μ, ϕ) since there is no dependence on these in the prior on (θ_x, θ_f).

In this section, the data considered is one observer viewing one image from the class. Generalizing these ideas to multiple image observations is straightforward if the observations in the multiple images are assumed independent, as is the case in section 2.5. The j-subscript is therefore dropped in this section and the data in one image is written (x, f).

The data model is again assumed to be

$$\begin{aligned} p(f | x, \phi, \tau) &= N(\phi, \tau^{-1} I_p) \\ p(x | \mu, \kappa) &= N(\mu, (\kappa P)^{-1}), \end{aligned} \tag{2.25}$$

where the shape model is written in terms of its precision matrix $P = (\kappa \Sigma_L)^{-1}$. For this missing data analysis, let the complete data (x, f) be split by the index i into the observed part $\{(x_i, f_i),\ i \in D\}$, and the unobserved part $\{(x_i, f_i),\ i \notin D\}$. Let the observed data be written (x_D, f_D) and the unobserved data be (x_{-D}, f_{-D}). Then the desired missing data posterior is

$$p(\kappa, \tau | x_D, f_D, \phi, \mu) \propto p(x_D, f_D | \kappa, \tau, \phi, \mu) p(\kappa, \tau) \tag{2.26}$$

where the data model can be written

$$p(x_D, f_D | \kappa, \tau, \phi, \mu) = p(f_D | x_D, \phi, \tau) p(x_D | \mu, \kappa). \tag{2.27}$$

Under the data model (2.25), the first factor is straightforward:

$$p(f_D|x_D, \phi, \tau) = N(\phi_D, \tau^{-1} I_q).$$

for q observed data points. The mean ϕ_D is the subset $\{\phi_i, i \in D\}$, of the complete data mean ϕ. This is a product of independent normal distributions on f_i over the subset of indices $i \in D$. Under non-informative prior $p(\tau) \propto \tau^{-1}$ the resulting posterior on τ is

$$p(\tau|f_D, \phi) = \text{Gamma}\left(\frac{q}{2}, \frac{q}{2}\sum_{i\in D}(f_i - \phi_i)^2\right).$$

To derive the second factor in (2.27), note that the factor can be re-written

$$p(x_D|\cdot) = \frac{p(x_D, x_{-D}|\cdot)}{p(x_{-D}|x_D, \cdot)} \tag{2.28}$$

where $(\cdot) = (\mu, \kappa)$. The numerator and denominator in this expression are both normal distributions. For illustration, the numerator and denominator are given explicitly

$$p(x_D, x_{-D}|\mu, \kappa) = (2\pi)^{dp/2}\kappa^{dp/2}|P|^{1/2}\exp\left(-\frac{\kappa}{2}\begin{bmatrix} x_D - \mu_D \\ x_{-D} - \mu_{-D}\end{bmatrix}' P \begin{bmatrix} x_D - \mu_D \\ x_{-D} - \mu_{-D}\end{bmatrix}\right)$$

and

$$p(x_{-D}|x_D, \mu, \kappa) = (2\pi)^{d(p-q)/2}\kappa^{d(p-q)/2}|P_{-D|D}|^{1/2} \times \exp\left(-\frac{\kappa}{2}\left[x_{-D} - E(x_{-D}|x_D)\right]' P_{-D|D}\left[x_{-D} - E(x_{-D}|x_D)\right]\right)$$

where $P_{-D|D}$ is the appropriate $(p-q) \times (p-q)$ conditional precision matrix. The expression (2.28) is valid for any x_{-D} and a convenient choice is to set x_{-D} equal to its conditional expectation $E(x_{-D}|x_D)$. The denominator then becomes (keeping the dependence on κ and x_D only)

$$p\left(x_{-D} = E(x_{-D}|x_D)|x_D, \cdot\right) \propto \kappa^{d(p-q)/2},$$

and the numerator is

$$p(x_D, x_{-D} = E(x_{-D}|x_D)|\cdot)$$

$$\propto \kappa^{dp/2} \exp\left\{ -\frac{\kappa}{2} \begin{bmatrix} x_D - \mu_D \\ E(x_{-D}|x_D) - \mu_{-D} \end{bmatrix}' P \begin{bmatrix} x_D - \mu_D \\ E(x_{-D}|x_D) - \mu_{-D} \end{bmatrix} \right\}.$$

Since $E(x_{-D}|x_D)$ is not a function of κ, under non-informative prior $p(\kappa) \propto \kappa^{-1}$ the posterior on κ given only the observed data x_D is

$$p(\kappa|x_D, \mu) = \text{Gamma}\left(\frac{dq}{2}, \frac{1}{2} \begin{bmatrix} x_D - \mu_D \\ E(x_{-D}|x_D) - \mu_{-D} \end{bmatrix}' P \begin{bmatrix} x_D - \mu_D \\ E(x_{-D}|x_D) - \mu_{-D} \end{bmatrix} \right).$$

The only difficulty is setting x_{-D} to the conditional expectation $E(x_{-D}|x_D)$. To evaluate this posterior, this can be done numerically using the ICM method (Besag, 1986) by setting x_D to the observed data and iterating over the full conditional distributions on all the unobserved elements x_{-D}. This method is applied in Chapter 4.

In the preceding discussion the data consisted of locations of facets and their associated feature values (x_D, f_D) in one image. When similar observations of these facets are available in J images from the class, then the model equation (2.27) becomes

$$p((x_{D,1}, f_{D,1}), \ldots, (x_{D,J}, f_{D,J})|\kappa, \tau, \phi, \mu) = \prod_{j=1}^{J} p(f_{D,j}|x_{D,j}, \phi, \tau) p(x_{D,j}|\mu, \kappa).$$

The resulting missing data posteriors are

$$p(\tau|f_{D,\cdot}, \phi) = \text{Gamma}\left(\frac{qJ}{2}, \sum_{j=1}^{J} \sum_{i \in D} (f_{i,j} - \phi_i)^2 \right)$$

and

$$p(\kappa|x_{D,\cdot}, \mu) = \text{Gamma}\Big(\frac{dqJ}{2}, \frac{1}{2} \sum_{j=1}^{J} \begin{bmatrix} x_{D,j} - \mu_D \\ E(x_{-D,j}|x_{D,j}) - \mu_{-D,j} \end{bmatrix}' P \begin{bmatrix} x_{D,j} - \mu_D \\ E(x_{-D,j}|x_{D,j}) - \mu_{-D,j} \end{bmatrix} \Big).$$

Chapter 3

Investigations of Distributions

3.1 Density Maximization

In this section a method is given for maximizing the constrained joint density (2.23) to produce a point estimate of facet locations x in a new image given their template values ϕ and μ and the hyperparameters κ and τ. This point estimate of x is used in chapters 4 and 5 as an estimated registration between two images and as a starting point for MCMC sampling to demonstrate registration variability. The generated point estimate would also be an effective starting point for sampling the constrained posterior predictive density as long as a reasonable guess of the hyperparameters κ and τ were available. In these registration applications μ is always fixed and ϕ is set to the value of the atlas image evaluated at the locations μ.

The density to be maximized is re-written as a function of x

$$p(x|E_0,\mu,\phi,\kappa,\tau) = \frac{1}{Z}\exp\left(-\frac{\kappa}{2}(x-\mu)'P(x-\mu) - \frac{\tau}{2}\sum_i F(Q_0,\mu_{L,i},Q,x_{L,i})\right) \tag{3.1}$$

where Z is a normalizing constant independent of x, the summation is taken over

bottom level facets, and the function $F(Q_0, \mu_{L,i}, Q, x_{L,i})$ (for atlas image Q_0 and new image Q) depends on the feature density used in the joint model. For example, in the first feature density given in equation (2.9), this function would be

$$F(Q_0, \mu_{L,i}, Q_1, x_{L,i}) = \left(Q(x_{L,i}) - Q_0(\mu_{L,i})\right)^2 .$$

The form of the objective function (3.1) indicates that the only feature densities considered here are those which are products of marginal distributions over all facets on the bottom level. The spatial feature models mentioned in section 2.4.2 are not discussed further here.

For most of this section, the dependence of F on the fixed quantities Q_0, $\mu_{L,i}$, and Q_1 will be suppressed, writing it as simply $F(x_{L,i})$.

Of course, for $\tau = 0$, such that no image information is incorporated, the density (3.1) has unique maximum at $x = \mu$. The difficulty therefore comes from the sum of feature density functions $\sum_i F(x_{L,i})$. Each term in this sum is a non-linear function of $x_{L,i}$ which is highly multi-modal. The multi-modality of F can be reduced by an appropriate choice of feature density, however the problem is always present when applying the model in images of typical complexity.

The second concern in this optimization problem is the dimensionality of the objective function. For three-dimensional image registration applications like those explored in chapters 4 and 5, the number of facets necessary approaches the number of pixels in the image, typically between 5,000 and 500,000. This implies that the number of random variables in the facet graph is on the order of 15,000 to 1,500,000. Therefore, extreme care must be taken to derive efficient methods for maximizing this high-dimensional distribution.

Heuristically, an attractive course of action to maximize this high-dimensional and multi-modal distribution is to work from large scale features to small. When registering two images, certain large scale features are usually very similar in both

images and their locations can be found very easily. Then, once these large scale features are registered, it is natural to condition on these locations to find smaller scale features. As the scale is reduced the images become increasingly dissimilar and some small scale features may not be present in both images, but a hierarchical methodology should ensure robustness to these small scale variations.

The hierarchical structure of the shape model $p(x|\mu,\kappa)$ lends itself well to this type of top-down conditioning. Essentially, small models containing only a set of top levels of the facet graph are maximized first, then the next larger model containing one more level in the graph is maximized conditional on the estimate of the smaller model's mode. Levels are added one by one and these incrementally larger models are maximized conditional on the mode estimated in the previous step. For each sub-model containing, for instance, facets on levels $1,\ldots,l$, the bottom level l is treated as the image level for the sub-model and the original sum of log feature density contributions, $\sum_i F(x_{L,i})$ in (3.1), is approximated by a smaller sum $\sum_i F_l(x_{l,i})$ taken over facets on level l. The shape component in the sub-model is the marginal normal distribution on levels $1,\ldots,l$ given in equation (2.2). This sub-model approach significantly reduces the dimensionality of each maximization step and implements the pragmatic top-down approach.

Maximization proceeds as follows. First the d-dimensional objective function for the sub-model containing only the single top level facet is maximized,

$$p(x_{1,1}|\mu_{1,1},\phi_{1,1},\kappa,\tau) \propto \exp\left(-\frac{\kappa c_1}{2}(x_{1,1}-\mu_{1,1})'(x_{1,1}-\mu_{1,1}) - \frac{\tau}{2}F_1(x_{1,1})\right). \qquad (3.2)$$

The function $F_1(x_{1,1})$ used for this one level model is analogous to the original model's $F(x_{L,i})$ but is modified to reduce the sub-model's multi-modality. The issue of formulating effective F_l functions is discussed below.

The Nelder-Mead simplex method (Nelder and Mead, 1965) is effective for numerically maximizing functions like eq. (3.2) since it does not require the calculation

of derivatives with respect to x and is somewhat robust to a multi-modal objective function. This method is used for all the required numerical maximizations in this section.

Given the estimated maximum density location $x_{1,1}$ from (3.2) the second largest sub-model which includes levels 1 and 2 can be maximized. First, the conditional distribution of level 2 facets given the top level facet is maximized numerically,

$$p(x_2|x_1,\cdot) \propto \exp\Big(-\frac{\kappa c_2}{2}(x_2-\mu_2-A_2(x_1-\mu_1))'(x_2-\mu_2-A_2(x_1-\mu_1)) \\ -\frac{\tau}{2}\sum_i F_2(x_{2,i})\Big), \tag{3.3}$$

where $(\cdot)=(\mu,\phi,\kappa,\tau)$. The length of x_2 is $d2^d$, and the fixed $d2^d \times d$ design matrix A_2 is defined in section 2.4.1. Note that this is a product of independent (d-dimensional) distributions on the elements $x_{2,i}$ of x_2 and therefore can be maximized by sequentially maximizing the marginal distributions on $x_{2,i}$. At this point, ICM iterations can be used to maximize the joint density on (x_1, x_2). The conditional density on x_1 given x_2 is normal and can be maximized analytically. The form of this conditional and all other full-conditionals needed for ICM steps in the larger sub-models are given in section 3.2.

This procedure of maximizing sub-models on the top l levels continues as l is increased. For each sub-model, first the facets on the bottom level l are numerically maximized conditional on the results of the last sub-model, then all facets above this level can be analytically set to their full conditional means inside a set of ICM cycles. When $l = L$ all facets in the true model are included in the maximization step and, under the condition that $F_L(x_{L,i})$ equals the true model's $F(x_{L,i})$, the true density, eq. (3.1), is evaluated in this final step. The sub-model procedure described thus far can be viewed as a method for generating a good guess of the mode of this distribution before starting the ICM cycles on its implied full conditionals.

The sub-model feature density functions $F_1(x_1), \ldots, F_L(x_L)$ are now discussed. First, as $l \to L$, these functions $F_l(x_{l,i})$ should approach the feature density function in the true model so that $F_L = F$. One obvious way of doing this is to use the same functional form in each sub-model. In this case, maximizing each sub-model represents predicting the locations of a smaller set of facets in the image Q. (If using one of the vector feature function densities (2.13 or 2.15) then the size of the region surrounding each bottom-level facet in a particular sub-model, represented by the vector z_i in the notation of section 2.4.2, could be scaled appropriately to account for the larger spacing between bottom level facets in the sub-model.)

However, setting all the F_l functions equal has one considerable drawback: the multi-modality of each function $F_l(x_{l,i})$ is not reduced and therefore the joint density on $(x_1, \ldots, x_l)$ under the sub-model on these levels has roughly as many modes (per dimension) as does the true joint density on the whole graph $(x_1, \ldots, x_L)$.

The well-investigated tools of multi-scale image analysis (Lindeberg, 1990) can help reduce the multi-modality problem. In multi-scale image analysis, the d-dimensional image is embedded in a $(d+1)$-dimensional "scale-space" where the added scale dimension is denoted σ. The scale-space of the d-dimensional image Q is created by calculating the family of convolutions of the image with a d-dimensional zero-mean spherical Gaussian kernel G of standard deviation σ,

$$Q^\sigma(x) = \int_{R^d} Q(x) G(\alpha - x, \sigma) d\alpha, \tag{3.4}$$

and the family of generated images is indexed by σ. There are several reasons for using a Gaussian kernel to generate this family, see for instance Lindeberg (1990), but the most important reason for this application is that the number of modes in the generated image $Q^\sigma(x)$ is a monotonically decreasing function of σ. As σ increases, the amount of local information contained in Q^σ decreases.

Scale-space therefore offers a promising method of defining the sub-model feature

density functions $F_l(x)$. Using the explicit notation defined at the beginning of the section, let $F_l(Q_0, \mu_{l,i}, Q, x_{l,i}) = F_l(Q_0^{\sigma_l}, \mu_{l,i}, Q^{\sigma_l}, x_{l,i})$, where $\sigma_1 > \sigma_2 > \cdots > \sigma_L = 0$. The bottom level feature density function is the original function defined in (3.1) and each sub-model feature function is a smoothed (less multi-modal) version of the original. This set of functions is in keeping with the heuristic plan of top-down maximization because estimating the mode in the smallest (highest-level) sub-models represents locating large scale features in the image Q, and as larger and larger sub-models are maximized they approach the true density (3.1) exactly at the Lth sub-model.

The one remaining concern is the choice of $\sigma_1, \ldots, \sigma_{L-1}$. Collins *et al.* (1994) approach a similar optimization problem from the perspective of spatially sampling the image at different resolutions. They point out that a Gaussian convolution is essentially a low-pass spatial filter with cut-off frequency approximately equal to the standard deviation of the kernel. Their methodology would suggest to use a scale on level l equal to the atlas grid spacing on that level since this is approximately the inverse of the Nyquist frequency of the spatial sampling for that level (Yaglom, 1962). This choice of scales also seems to be reasonable in this paper's model for the following reason.

Consider that after a sub-model on levels $(1, \ldots, l)$ has been maximized, that the rest of the facets in the true model on the levels between $l+1$ and the bottom level L are set to their conditional maxima under the shape model $p(x_{l+1}, \ldots, x_L | x_l, \mu, \kappa)$. This is a reasonable current guess for the maximum density locations of all the facets in the graph under the true model. Next consider that higher order parent/child connections are ignored, namely that all parent/child weights w_i for $i > 1$ are set to 0. In this case the facet graph can be drawn as a tree and the maximum density locations of the 2^d remaining children of any facet are completely defined by the

location of that one parent facet. (This would be a reasonable approximation to the true density if w_1 were much larger than the higher-order weights.)

Then the location of a facet on level l, the bottom level in the sub-model just considered, affects a compact set of descendents on the bottom level L in the facet graph. For instance in three dimensions, this set is cubic, centered at the location of the level l facet, and has size $2^{L-l}s \times 2^{L-l}s \times 2^{L-l}s$ for a bottom level template grid spacing s. If the higher order parent/child connections are not ignored then the set of affected bottom level facets is larger, with facets near the edge of the cube being less affected than those near the center. (This decreasing envelope of affection is only true for decreasing parent/child weights as a function of parent/child template distance, but this condition is always true in this paper.)

Next, notice that the scale-space smoothing of the feature density functions $F_l(Q_0^{\sigma_l}, \mu_{l,i}, Q^{\sigma_l}, x_{l,i})$ changes the two image inputs from the original images Q_0 and Q by performing a weighted spatial average on each image, and the extent of the non-negligible weights in the spatial average is defined by the standard deviation σ_l of the Gaussian kernel. It is reasonable to roughly match the size of the scale-space Gaussian kernel used on this level to the size of the region of affected bottom level facets discussed in the last paragraph. This would suggest that $\sigma_l \approx 2^{L-l}s$. In practice σ_L might be set to zero so that the bottom level of facets represents the original image without any loss of information.

Remember that this set of semi-heuristic approximations to the true model (3.1) approaches and becomes the true model for the approximation on level L. This does not guarantee that the method will converge to the global maximum of (3.1), but the top-down approach has proven itself in practice to be quite robust to local maxima.

At every step during the top-down maximization the value of the true (unnormalized) density (3.1) can be evaluated by plugging in the current estimated facet

locations. After the sub-model on levels $(1, \ldots, l)$ has been maximized a reasonable estimate of the achievable unnormalized density, after the whole top-down procedure is complete, can be calculated by analytically setting the facets on levels $(l+1, \ldots, L)$ to their conditional means under the shape model $p(x|\mu, \kappa)$ and calculating equation (3.1) for this graph configuration.

3.2 Sampling Scheme

3.2.1 Joint Sampling of Facet Properties

Drawing a sample from the joint distribution on facet locations x and features f is generally straightforward. Consider first drawing (x, f) from the data model $p(x, f|\mu, \phi, \kappa, \tau)$. Under both feature models used in this paper, (2.9) and (2.15), the data model is the product of two components $p(x|\mu, \kappa)p(f|\phi, \tau)$ that can be sampled separately for a full sample (x, f).

Conditional on ϕ and τ, the intensity feature model (2.9) is a product of normal distributions over the bottom level facets with mean ϕ and variance τ. On the other hand, the standardized regression feature model (2.15) is not a standard form that would have to be sampled by Monte Carlo methods. It is also indeterminate on the original feature values f and must therefore be constrained, for example by insisting that the mean and variance of each facet's feature vectors are 0 and 1, respectively.

Since the shape model $p(x|\mu, \kappa)$ given in section 2.4.1 is a hierarchical normal distribution, it is easily sampled using the method of composition. The top facet is sampled from its marginal distribution, then the second level is drawn from its conditional distribution given the top facet sample, and so forth. Equation (2.2) defines all the necessary conditional distributions. Each facet on a particular level can be sampled independently since, conditional on the level above, the distribution on all facet locations is a product of independent normals.

Figure 3.1 displays one sample of the 32×32 bottom level facets from a two-dimensional, 5 level graph under the shape model defined by equation (2.2). The methods of section 2.4.1 were used to define the level weights and parent/child connection weights, and the overall precision parameter κ was manually set to discourage facet overlap and facilitate viewing of the shape of the sample. One vertical line of facets is shown in red to highlight the correlation between nearby facets.

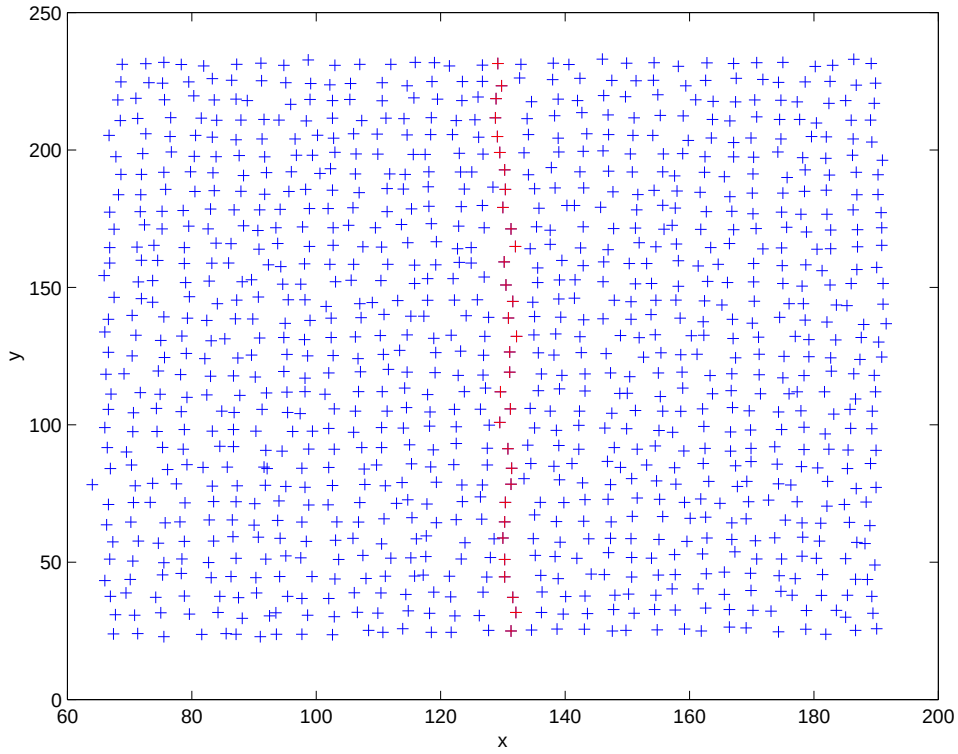


Figure 3.1: A sample from the prior on 32×32 2D facets (6 level graph)

The posterior predictive joint density $p(\tilde{x}, \tilde{f}|(x, f))$ after observing J samples $\{(x_1, f_1), \ldots, (x_J, f_J)\} = (x, f)$ can be sampled using composition by first drawing κ and τ from their posterior distributions (2.16) and (2.17) and then, conditional on these parameters, drawing from the normal data model.

3.2.2 Constrained Predictive Density

When sampling facet locations in a new image from the class, conditional on either the hyperparameters κ and τ or conditional on a set of manually identified facet locations and features from other images from the class, the predictive distribution is only know up to a normalizing constant. See section 2.5 for a full discussion. However, because of the form of the predictive distribution, full conditional distributions on all facet locations, given the locations of their parents and children, are easily written down.

Consider the constrained data model distribution given in equation (3.1). Since the sum of feature functions is taken over only the bottom level of facets in the graph, it is easily seen that the full conditional distributions on the locations of facets above the bottom level in the graph are normal given their parents and children. However, the full conditionals on the bottom level facet locations given their parents are non-standard due to the nonlinear function F.

One approach to sampling this density is to proceed facet by facet and sample from each facet location's d-dimensional full conditional distribution in the style of Gibbs sampling (Geman and Geman, 1984). To sample from the full conditional distribution on bottom level facets, one must use a Monte Carlo technique due to its non-standard form. The Metropolis algorithm with a spherical normal proposal density has worked well in practice for these full conditionals. Analytical Gibbs sampling can be used on locations of facets above the bottom level since these full conditionals are known to be d-dimensional normals.

The full conditional mean on facet locations above the bottom level is

$$\mu^c_{l,i} = \mu_{l,i} + (\sigma^2)^c \left[\sum_{k \in C_{l,i}} c_{l+1} w_{i,k} (x_{l+1,k} - \mu_{l+1,k}) + \sum_{k \in P_{l,i}} c_l w_{i,k} (x_{l-1,k} - \mu_{l-1,k}) \right] \tag{3.5}$$

where $C_{l,i}$ and $P_{l,i}$ are the sets of children and parents, respectively, of facet $(\cdot)_{l,i}$ and $w_{i,k}$ is the parent/child connection weight between facets i and k. The full conditional variance $(\sigma^2)^c$ is equal to

$$(\sigma^2)^c = \left[\sum_{k \in C_{l,i}} c_{l+1} w_{i,k} + \sum_{k \in P_{l,i}} c_l w_{i,k} \right]^{-1}. \tag{3.6}$$

If this facet-by-facet type of Gibbs sampling converges slowly due to excessive correlation between parent and child facets, then one can use the hierarchical structure of the shape component to derive a more efficient sampling algorithm. If the facets are split into two groups, those on the bottom level and those on all the other upper levels, then the full conditional distribution on the locations of all the facets in the upper group given the lower group is known analytically and can be sampled exactly. (This type of sampling does not improve the efficiency of sampling on the bottom level facets since these full conditionals are non-standard.). A Gibbs sampling scheme can be employed which iterates between the full conditional on the bottom level facet locations given the upper levels, and vice versa.

The joint distribution on all levels above the bottom conditional on the bottom level can be factored in terms of its "backward" conditionals

$$p(x_1, \ldots, x_{L-1} | x_L, \cdot) = p(x_1 | x_2, \cdot) p(x_2 | x_3, \cdot) \cdots p(x_{L-1} | x_L, \cdot) \tag{3.7}$$

where $(\cdot) = \mu, \phi, \kappa, \tau$. Each of these backwards conditionals are known analytically,

$$p(x_l | x_{l+1}, \cdot) \propto p(x_{l+1} | x_l, \cdot) p(x_l | \cdot),$$

where $p(x_{l+1} | x_l, \cdot)$ is given in equation (2.2) and $p(x_l | \cdot)$ is normal with mean μ_l and covariance matrix given in equation (2.4). The resulting backward conditional distribution is

$$p(x_l | x_{l+1}, \cdot) = N(\mu_{l|l+1}, \Sigma_{l|l+1})$$

where the conditional covariance matrix is

$$\Sigma_{l|l+1} = (\kappa c_{l+1} A'_{l+1} A_{l+1} + \Sigma_l^{-1})^{-1}$$

and the conditional mean is

$$\mu_{l|l+1} = \Sigma_{l|l+1} A'_{l+1} (x_{l+1} - \mu_{l+1}),$$

where the design matrix A_l is defined in section 2.4.1. The joint distribution on all levels above the bottom can be sampled exactly using the method of composition on these known backward conditionals. This result improves convergence of the Markov chain to the true joint distribution in terms of the number of iterations required, but each iteration is expensive to compute due to the matrix inversion required in the backward conditional covariance matrix. In fact, for large facet graphs, the size of this matrix could render this method infeasible.

Finally, to sample from the posterior predictive distribution on facet locations in a new image given a set of manual facet identifications from images in the class, the most straightforward course of action is to first sample κ and τ from their posterior distributions and then, conditional on these draws, use one of the methods above to sample from the constrained data distribution (3.1).

Chapter 4

Application: Neonatal Mouse Brain Magnetic Resonance Microscopy

This chapter deals with a study being undertaken at the Duke University Medical Center for in vivo Microscopy to quantify hippocampal shape and volume changes in apoE-deficient mice. A deficiency in apolipoprotein E is a major risk factor for Alzheimer's disease, and the aim of the study is to determine if these mice display the 20-30% reduction in hippocampal volume that has been noted in some human Alzheimer's patients. To quantify any hippocampal changes in these mice, the study will follow several apoE-deficient mice and an age matched control group of C57 black mice by taking regular MR scans throughout their lifetimes. Each scan will be segmented to quantify morphological changes in the hippocampus.

The methods of this paper may be useful for quantifying the hippocampal shape change by automatically segmenting the region in the control and apoE mice groups, an attractive prospect because it is reproducible and not prone to the variability of technician expertise. To demonstrate how the model can be used for segmentation, two mice brains, one apoE and one control, were imaged and the facet model was applied to register them. From this registration a manual hippocampal segmentation performed on the apoE mouse could be transfered to the control mouse and tested

for accuracy against a manual segmentation of the control mouse hippocampus.

For this *in vitro* experiment, the mice were perfused transcardially with 0.9% (wt/vol) saline followed by fixative (buffered FORMA-SCENT fixative, 10% w/vol formalin, PH 6.9-7.1) and the brains were removed from the skull and embedded in fomblin (perfluoropoly ether) to keep them from dehydrating and to limit susceptibility effects at the surface. (Susceptibility artifacts in magnetic resonance imaging have the appearance of washed out regions near interfaces of two materials with different magnetic properties.) The cuvet holding the brain was oriented in a 1 cm solenoid coil inside a Brucker CSI Magnetic Resonance Imaging (MRI) instrument using shielded gradients capable of 90 G/cm. The cuvet was registered physically inside the instrument so that very little rotational misregistration between images was encountered in the resulting images. Each $T2^*$ image was acquired in approximately 2.7 hours with flip angle 35°, 150ms relaxation time, and 9ms echo time. The three-dimensional images consisted of 128 slices of 256×256 pixels, where each three-dimensional voxel had dimensions .039mm×.039mm×.16mm. (The voxel anisotropy must be incorporated into the model since facet locations are modeled in real space rather than the arbitrarily sized image cube.)

The manual and automatically generated hippocampus segmentations in this section are represented by an image of the same dimension as the original image, where each pixel intensity represents whether or not the pixel belongs to the hippocampus. To automatically generate a segmentation image of a new brain, the facet model is applied to register the new image to a manually segmented atlas image, thereby determining the locations in the new image of all the "hippocampus" facets, known from the atlas segmentation image. The segmentation image for the new brain is generated by simply defining that the pixels near each hippocampus facet belongs to the hippocampus region, where the meaning of "near" will be defined in section 4.3.

The control mouse image was arbitrarily chosen as the atlas and facets locations were predicted in the apoE mouse image. Image intensities were rescaled to match the tenth and ninetieth quantiles of their empirical intensity distributions inside a three-dimensional rectangle manually defined to cover most of the interior of the brain in each image. Several choices of the rescaling rectangle were tested and it was noted that the rescaling factors were insensitive to reasonable choices of this region. This rescaling was done to facilitate the use of the intensity feature model, equation (2.9), which assumes that, after rescaling, homologous features have similar intensities in both images. This simple feature model performed well in this application and was computationally efficient.

In this three-dimensional application the facet graph was set to have 6 levels for a total of 37449 facets. The template grid μ was stretched in all three dimensions to cover the hippocampus region more effectively. The resulting bottom level template facet spacing was .15mm$\times$.26mm$\times$.12mm. The template feature values ϕ were fixed to be equal to the atlas image evaluated at the template locations.

The deformation model of section 2.4.1 was used and the associated level weights were set using the methods of that section to ensure that all facets had equal marginal variance under the shape model. The top level weight c_1 was set to a very low value to ensure model insensitivity to translations of all facets as a group, since these two images were translationally misregistered in all three dimensions. Parent/child connection weights were set to be inversely proportional to the atlas distance from parent to child and to sum to one over each set of parents.

In section 4.1 the hyperparameter estimation methods of section 2.5 are highlighted. To estimate these parameters, fifty randomly chosen facets were displayed on the template image and their homologous points were manually located in the new image.

In sections 4.2 and 4.4 the constrained data model density (2.23), given the estimated hyperparameters, is maximized and sampled using the methods of section 3.1 and 3.2. And finally, in section 4.3 segmentation images are predicted and the results are compared with manual segmentations.

Throughout the discussion the control mouse image will be called the "atlas" and the apoE mouse image the "new" image. Locations in these anatomical images will be referred to by their geographical location, eg. N for the top of the image and E for the right side, for the reader not comfortable with medical imaging terminology.

4.1 Parameter Estimation

To estimate the hyperparameters κ and τ, data was gathered in the form of manual facet locations in the new image given their locations in the atlas image. Once the template facet locations μ were set to cover the hippocampus in the atlas image, fifty facets were chosen at random from the bottom level of facet template means μ_6 in the 6 level graph. These facets were displayed individually on the atlas image (by displaying the slice corresponding to the facet's z index and highlighting the pixel in that slice corresponding to the facet's x, y template location) and an observer manually located the homologous three-dimensional locations in the new image. The observer used an image viewer with which he could chose an appropriate image slice and locate any x, y pixel in that slice.

The intensity feature density (2.9) was used for this application. This density assumes that facet intensities in the new image at their correct locations are normally distributed with mean equal to the atlas image evaluated at the facet template location. To check this normality assumption the sorted observed feature differences $(f_i - \phi_i)$ were plotted versus quantiles of the standard normal distribution. The top panel of figure 4.1 shows the quantile-quantile plot for all fifty data points. There are

two gross outliers at the bottom of the plot, and one possible outlier at the top. On closer inspection of the facet placement for these outliers, it was noted that the top facet was incorrectly placed by the observer in a region of high deviation in intensity which resulted in the lack of fit to the model. The bottom two facets were in regions of the atlas that had no homologous point in the new image. Both were near the base of the brain where the images differ widely due to the brain extraction procedure. These results indicate that the ever present effect of the lack of homology between two images could be taken account of by a more robust feature model, however for this application these three outliers were simply discarded for parameter estimation. The bottom panel of figure 4.1 displays the 47 non-outliers. These seem to follow the normal model extremely well, and the resulting posterior distribution on τ was calculated to be $p(\tau|(x,f),\phi)$ =Gamma$(23.5, 75,000)$. Under squared error loss the Bayes estimate is the posterior mean $\hat{\tau} = 0.00038$.

Figure 4.2 displays results from the ICM procedure for estimating κ outlined in section 2.6. The three outliers seen in the estimation of τ were also left out for this estimation procedure. The $q = 47$ observed facet locations x_D were fixed and ICM cycles were used to set the rest of the facets in the graph x_{-D} to their conditional expectations $E(x_{-D}|x_D)$ given the observed facet locations. The vertical axis in figure 4.2 is the square root of the renormalized log shape density, $\log\ p(x = (x_D, \hat{E}_r(x_{-D}|x_D))|\mu,\kappa)$, for the current estimate of the conditional expectation $\hat{E}_r(x_{-D}|x_D)$. The log density has been renormalized before plotting since, under the shape model, all facets have approximately the same marginal variance and the log shape density can be renormalized to give an estimate of that marginal variance. At iteration r of the ICM cycles, the estimated marginal variance is

$$\hat{\sigma}_r^2 = \frac{c_1^{-1} + c_2^{-1}}{q} \begin{bmatrix} x_D - \mu_D \\ \hat{E}_r(x_{-D}|x_D) - \mu_{-D} \end{bmatrix}' P \begin{bmatrix} x_D - \mu_D \\ \hat{E}_r(x_{-D}|x_D) - \mu_{-D} \end{bmatrix}$$

where P is the joint precision matrix on all facets in the graph, defined implic-

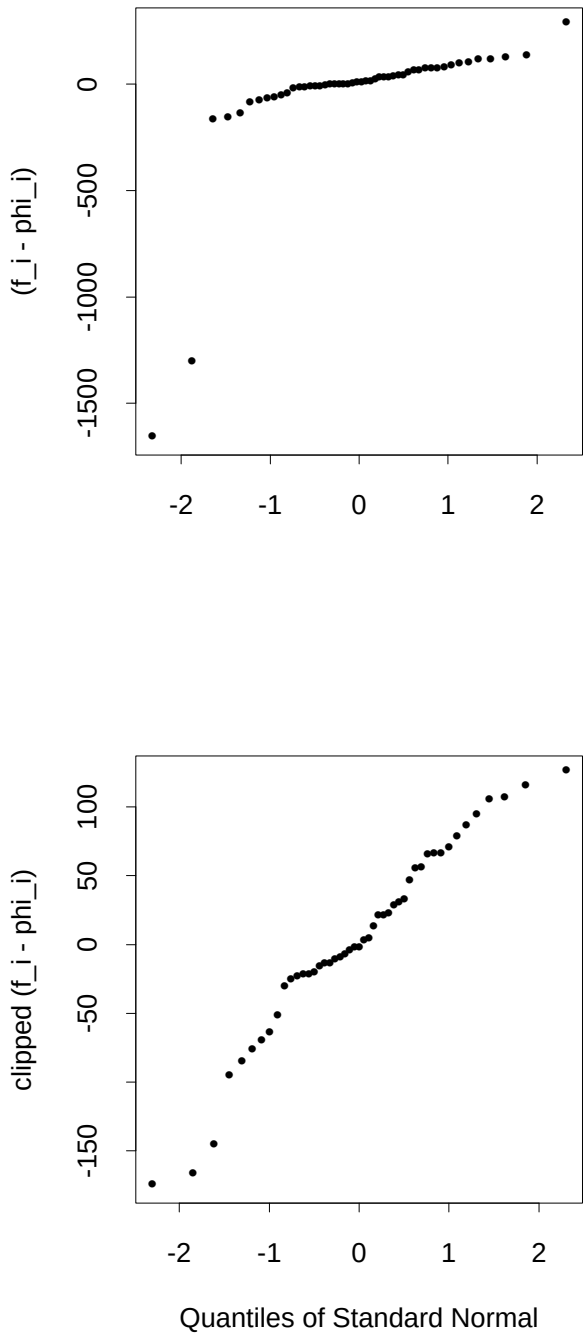


Figure 4.1: Sorted feature differences $f_i - \phi_i$ for fifty manually located facets. All fifty data points are shown in the top panel and the three outliers are removed in the bottom panel. The atlas and new images were rescaled as discussed in the text before producing these plots.

itly in equation (2.2) (with κ set to one in that definition). The estimated missing data conditional expectation $\hat{E}_r(x_{-D}|x_D)$ converges to the true conditional expectation $E(x_{-D}|x_D)$ as the ICM iteration $r \to \infty$. The resulting posterior on κ is $\text{Gamma}(\frac{dq}{2}, \frac{q}{2(c_1^{-1}+c_2^{-1})}\hat{\sigma}_\infty^2) \approx \text{Gamma}(70.5, 1.96)$.

4.2 Registration Results

A point estimate of a registration between the atlas image and the new image can be computed by maximizing the constrained data model (2.23) using the methods of section 3.1. For this point estimate, the Bayes estimates of the hyper-parameters $\hat{\kappa}$ and $\hat{\tau}$ determined in section 4.1 were used. In this section qualitative results of the maximization are given by displaying the predicted facet locations in the new image and comparing them with their locations in the template.

Figure 4.3 displays the unnormalized log density achieved during maximization. The top curve is the unnormalized log shape component

$$-\frac{\kappa}{2}(x-\mu)'P(x-\mu),$$

the bottom curve is the unnormalized feature density

$$-\frac{\tau}{2}\sum_i (f_i - \phi_i)^2, \tag{4.1}$$

and the horizontal axis is the maximization step. As maximization proceeded down the graph in each submodel, the full conditional distributions were maximized numerically for facets on the current level and analytically for those higher up. Five ICM cycles were performed on the full conditional distributions of all facets in the submodel. Given six levels in the graph, there were 6×5 steps in the maximization. After each level was complete the estimated achievable log density was calculated according to the methods of section 3.1. These joint density values are recorded in

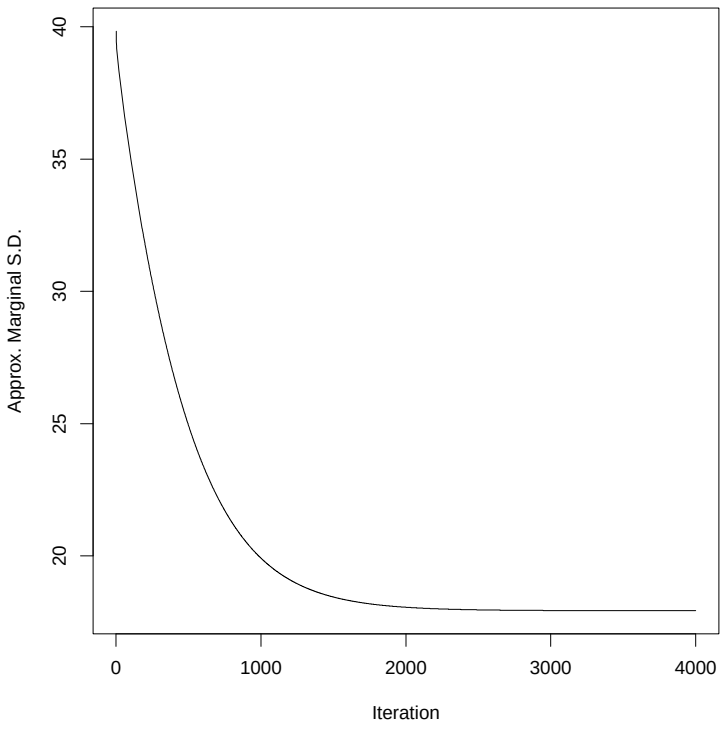


Figure 4.2: ICM cycles to estimate κ. Plotted at each iteration is the current estimate of $\sqrt{\kappa(c_1^{-1} + c_2^{-1})}$, the approximate marginal standard deviation for any facet in the graph.

the figure. During maximization the deviation seen in feature log density was approximately ten times the deviation in shape log density. This weighting of the two components of the model is on par with previous experience using models of this kind in earlier applications (McCulloch *et al.*, 1997).

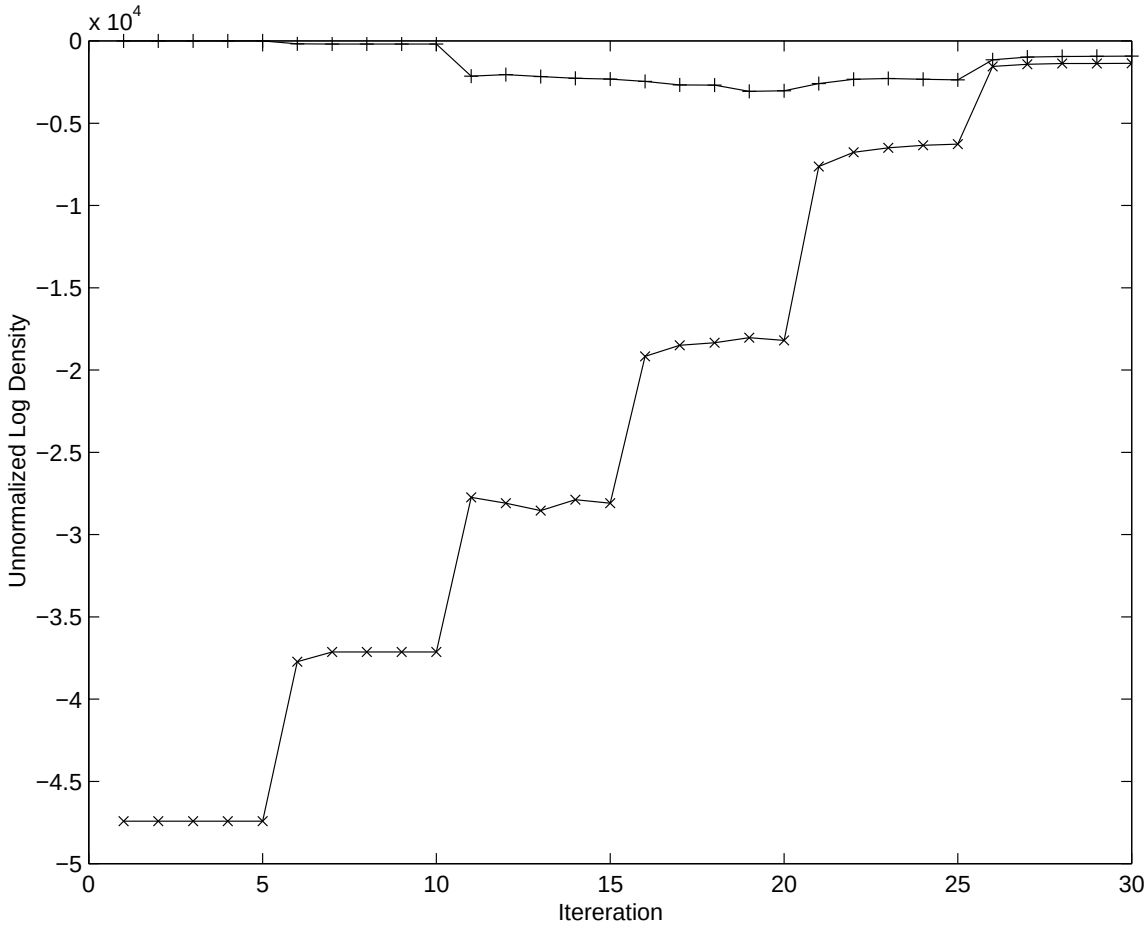


Figure 4.3: Unnormalized log density during maximization procedure. The feature and shape components are represented by $\times$ and $+$, respectively.

Figure 4.4 (a) displays the coronal view of the atlas image slice 44. Superimposed on this image are the first two indices of the three dimensional template locations $\mu_{6,i}$ for all bottom level facets i whose third (slice) index of $\mu_{6,i}$ fell in the range $(44, 45)$. Panels (b), (c), and (d) display the maximum density point estimate of the facet locations $x_{6,i}$ superimposed on coronal slices 39, 40, and 41. In these sub-figures,

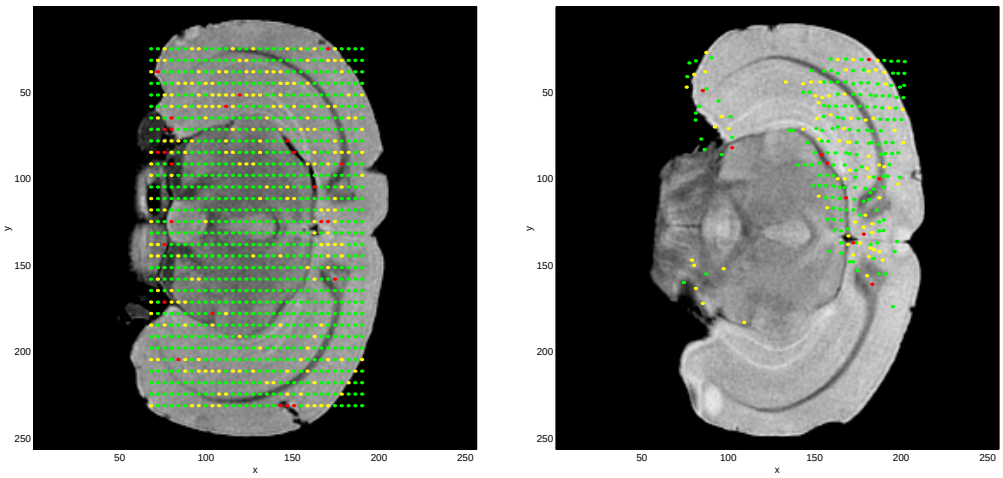
slice s is displayed as the background for all facets i whose third indices of $x_{6,i}$ are in the range $(s, s+1)$ (and whose template third indices fell in the template slice shown in panel (a)). It is necessary to show several slices in the new image since the facet locations $x_{6,i}$ are defined in three dimensions and can therefore be predicted out of slice. In fact, because of the gross translational mis-registration between the images, facets were predicted approximately 5 slices down in the new image compared with the atlas image. The model also detected that the two images were misregistered by a small rotation in the NE/SW direction and therefore predicted that generally NE facets should be two slices higher than SE facets. The model is not strictly invariant to rotation, but did approximate this invariance adequately by adjusting facets appropriately in the higher levels of the graph. On a second inspection of the two images, this rotational misregistration was deemed to be a true artifact. Interestingly, the fit of the model pointed out this feature that was not obvious on casual inspection of the two images before analysis.

Facets are color-coded in figure 4.4 to showcase one property of the model. The facets are color-coded by their marginal feature density values (eq. 4.1): green for high density, yellow for medium, and red for low feature density. The color cutoffs were chosen arbitrarily from the empirical distribution on feature density contribution for facets in this atlas slice. Figure 4.5 gives this distribution along with the ranges chosen for display in figure 4.4. The red facets show regions of poor intensity match between the two rescaled images. Note the three facets in the SE corner which fall on a tear in the atlas image brain. These facets were predicted to lie in a reasonable region in the new brain even though there is of course no matching tear in this image. To predict these facet locations correctly, the model drew strength from the facets nearby which do have homologous and therefore high feature density locations in the new image. This mechanism of drawing strength occurs in several places in this

slice, including an important region near the NE corner on the inner boundary of the hippocampus (See figure 4.7 for a depiction of the hippocampus). Two red facets in a ridge of low atlas intensity were well predicted in the new image even though the hippocampal boundary is not as well defined in the new image. On the other hand, several facets near the base of the atlas brain in the NW corner of panel (a) lie in an extended region of the atlas that has no homologous region in the new image. Marginal feature densities in this region are low (red) and there is a large amount of deformation in facet locations. The model necessarily fits poorly in this region since the two images are so different there.

Figure 4.6 highlights how the model can be used to automatically segment a region of interest in a new image given a manual segmentation of the atlas image. In the top panel, an observer has highlighted those facets in the atlas image slice 39 he believes to be "hippocampus" facets. He could similarly highlight hippocampus facets in all other relevant slices. The bottom panels give the maximum density predictions in the new image of the facets highlighted in the top panel. Notice that in the atlas image the north half of the hippocampus seems to protrude further to the west than the south half does, whereas in the new image, the south half seems to protrude farther to the west. This is largely an artifact of the rotational misregistration between the images and the model correctly fit that misregistration by placing the north facet group in a higher slice than most of the south facet group. Remember that facets are predicted in R^3 whereas the images are a finite set of voxels. Therefore the threshold between facets displayed in panel (b) and those displayed in panel (c) is an arbitrary function of the orientation of the slices in the image dataset.

Apart from the predicted rotational and translational misregistrations, the model also predicted the more subtle shape differences in the two hippocampi. This is easily seen by the deformation from the atlas grid $\mu_{6,i}$ to the predicted locations $x_{6,i}$. All



(a) Atlas image coronal slice 44

(b) Observed image coronal slice 41

(c) Observed image coronal slice 39

(d) Observed image coronal slice 40

Figure 4.4: All facets in atlas slice 44 and their predicted locations in the new image. Feature match $f_i - \phi_i$ is color-coded by the rule given in Figure 4.5.

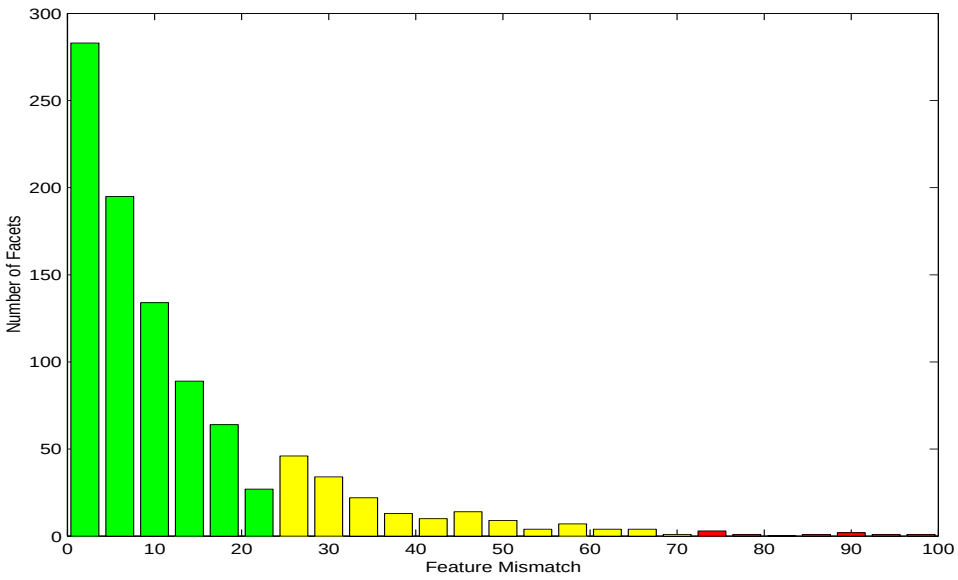


Figure 4.5: The log feature density contribution to the mode estimate shown in figure 4.4. The square-root of $(f_i - \phi_i)^2$ is given on the horizontal axis for all facets in atlas slice 44. The colored regions of the histogram correspond to the color coding used in figure 4.4.

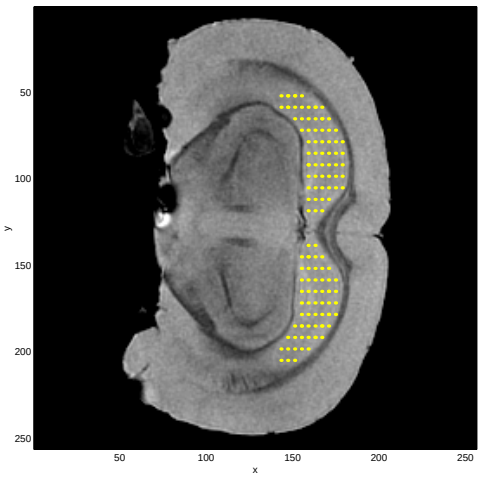
hippocampus facet locations seem to be predicted exceptionally well.

The maximization procedure in this section took approximately five minutes of CPU time on a mid-priced UNIX workstation.

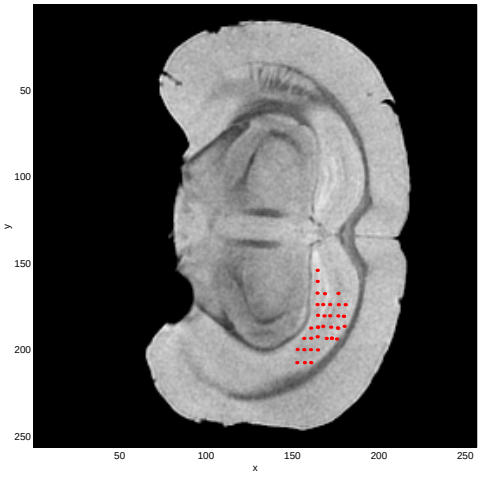
4.3 Segmentation Analysis

Figure 4.7 shows slice 39 of the atlas image with two manual segmentations superimposed on it. For this experiment, the hippocampi in the atlas image and the new image were both segmented by two observers. Large variability was seen in the two segmentations so to incorporate this variability, two regions were dealt with separately: the region where both observers agreed the pixel belonged to the hippocampus (green in figure 4.4), and the region where at least one observer believed the pixel was hippocampus (yellow or blue in figure 4.4).

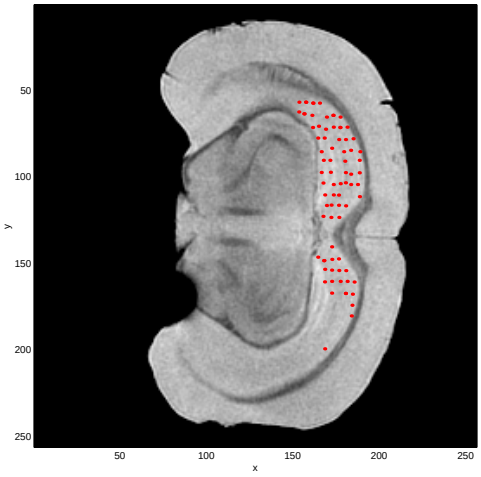
The goal of this experiment was to use the previous section's predicted set of facet



(a) Atlas image coronal slice 39



(b) Observed image coronal slice 35



(c) Observed image coronal slice 36

Figure 4.6: Hippocampus facets in atlas slice 39 and their predicted locations in the new image.

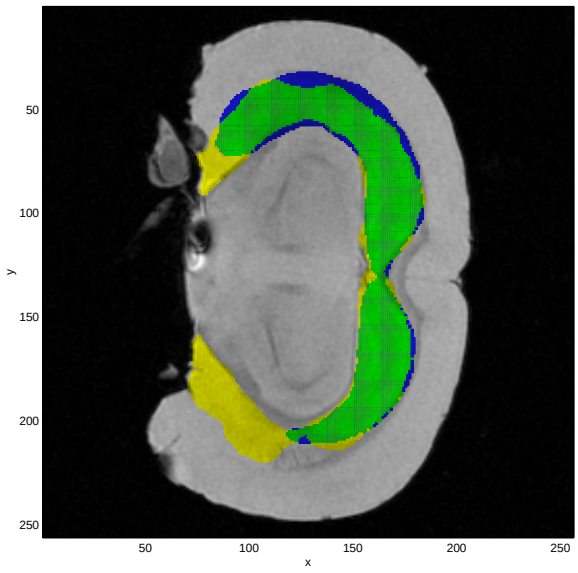


Figure 4.7: Slice 39 of atlas image with two manual segmentations superimposed on it: one shown in blue, and one in yellow. The region of agreement is shown in green.

locations to predict three-dimensional segmentation images like those displayed in the colored regions in figure 4.7. A very simple methodology was used: for every facet labelled "hippocampus" by a manual segmentation of one image, every pixel in a small region around the facet's location in the other image was set to the hippocampus pixel identifier, thus producing a segmentation of the other image. To be able to easily define the size of the small fill-region, a methodology was employed which is backward to what one might expect: the atlas image segmentation was predicted given a manual segmentation of the new image. In this way, the fill-region was naturally defined by the grid spacing of the atlas facet locations μ. Only the bottom level of facets in the graph were used to predict segmentations, all higher-level facets were ignored.

Given the predicted facet location calculated in section 4.2, each bottom level facet was given a label using the two manual segmentations of the new image. For each bottom level facet, if the predicted facet location $x_{6,i}$ in the new image fell in one observer's segmentation but not the other, it was labeled "1". If it fell in both observers' segmentations it was labeled "2". Otherwise it was labeled "0". A predicted segmentation of the atlas image could now be generated by filling in small regions around each facet's template position $\mu_{6,i}$ with the facet's label.

This predicted segmentation image of the atlas image therefore had three regions: "0" for the agreed non-hippocampus region, "1" for disagreement between the observers, and "2" for the agreed hippocampus region. To judge performance of the automated segmentation, the manual segmentation images of the atlas image were similarly combined to produce three-region segmentations.

Figure 4.8 compares three slices of the predicted segmentation with the manual segmentation images. Each panel is a difference image of the manually segmented region minus the automatically segmented region, where the region compared is the set of pixels either observer deemed as hippocampus (regions "1" and "2" combined).

Gray denotes agreement between the predicted and manual segmentation, and white and black denote disagreement. First, notice that the automated region is much chunkier than the manual region since there were significantly fewer bottom level facets than voxels in the manual segmentations. An added level of the graph would reduce this effect since the fill region around each facet would be one eighth the volume used here. Two of these slices seem to match very well and one does not. The difficult variability in the middle part of the hippocampus, shown in the middle panel, is not predicted well by the model because of the large difference in shape between the atlas and new image.

Table 4.1 gives quantitative results comparing the predicted and manual segmentations. The measure used to compare these regions is the percentage overlap, used by Collins *et al.* (1995), defined as the number of pixels in agreement between the two segmentations divided by the total number of pixels in one region, written as a percentage. The percentage must be normalized by each region's volume in turn and results examined in case one region is entirely contained in the other. (This situation would give a perfect score if the percentage were only normalized by the smaller region.) The left column in the table compares the automatic and manual segmentations for the observer agreed regions and the right column compares the larger regions which contain the disputed pixels. These results may seem discouraging, however when the two manual segmentations of the same image are compared, the overlap results are only 89.0 and 87.2. Since the automated procedure depends on the obviously variable manual segmentations of the new image for facet labels, then is scored against the variable manual segmentations of the atlas image, it is expected that the score should be significantly lower than the 89/87 manual segmentation score.

When segmentations of only the upper part of the hippocampus, on the east side

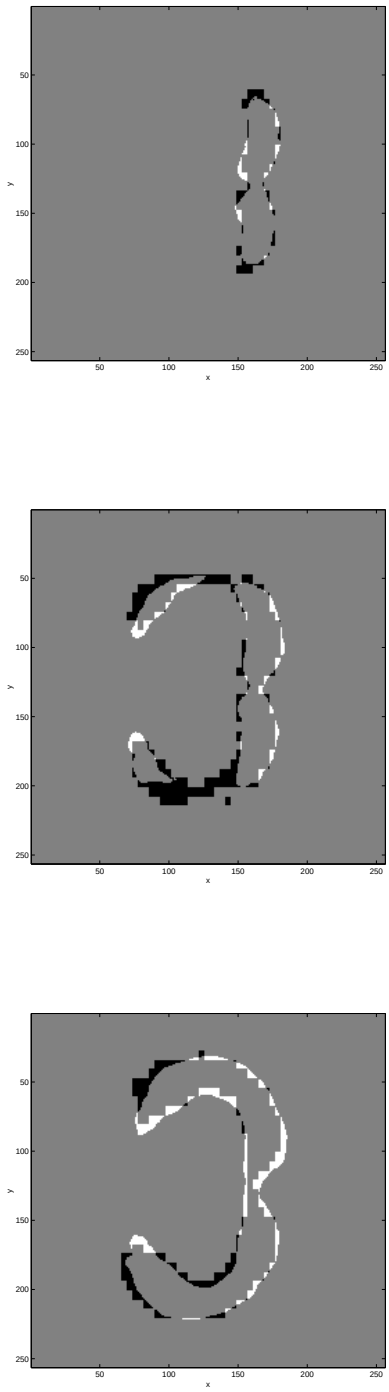


Figure 4.8: Difference images comparing three slices of the manual segmentation with the predicted segmentation.

Normalized by	Both observers region	Either observer region
predicted volume	80.2	84.2
manual volume	77.5	76.3

Table 4.1: Percent overlap results comparing predicted segmentations with manual segmentations. When the two manual segmentations are compared, percent overlap is 89.0 and 87.2.

of the images, are compared, the results are significantly better. The scores in Table 4.2 were calculated by simply cutting off the W side of all the segmentations below $x = 135$ (see figure 4.7). The corresponding scores comparing the two manual segmentations of this upper region are 90.3% and 90.2%. These results show that much of the error in the whole hippocampus segmentation occurs in the highly variable lower portion (west side of the images).

Normalized by	Both observers region	Either observer region
predicted volume	86.1	87.8
manual volume	81.6	81.0

Table 4.2: Percent overlap results comparing predicted segmentations with manual segmentations for only the upper part of the hippocampus, ie. $x > 135$. Comparing the two manual segmentations, percent overlap is 90.3 and 90.2.

4.4 Predictive Variability Inference

The methods of section 3.2 were used to sample from the constrained data model (2.23) on facet locations in the new image given the facet locations in the atlas image, where the hyperparameters were set to their posterior means calculated in section 4.1.

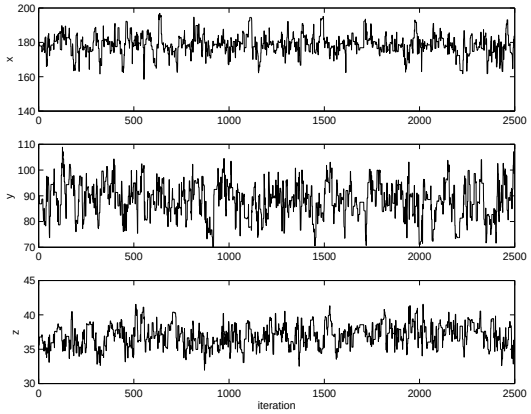
A Markov chain was constructed by sampling from the three-dimensional full conditional distributions on each facet given all of its neighboring facets in the graph. A hybrid Gibbs/Metropolis chain was used since all facets above the bottom level have known full conditional distributions but facets on the bottom level have non-standard

distributions due to the image constraint. In the Metropolis step on every bottom level facet, the proposal density was a spherical normal distribution with mean zero and adjustable variance. The proposal variance was determined by monitoring the acceptance rates of several bottom level facets and ensuring that each rate was in the range $(.25, .5)$. The Markov chain was started at the estimated maximum density location determined in section 4.2 and was run for 2500 iterations. For this model with 37449×3 random variables, each iteration took approximately .25 seconds on a mid-priced UNIX workstation. At each iteration, locations were recorded for five bottom level facets and one parent of one of these five.

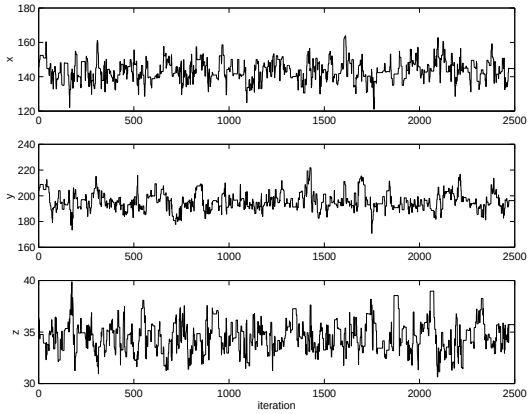
The atlas locations of the five bottom level facets are shown in figure 4.10 (a) and trace plots are shown in figure 4.9 for two of the five and the one parent facet. Notice that the marginal trace of the one recorded parent facet (figure 4.9 (c)) has a different appearance than the other two shown in that figure since sampling its full-conditional is a Gibbs step for which every sample is accepted. A burn-in period of approximately 500 iterations is noticeable in the parent trace (c), but virtually no burn-in period was noticed for any other facets recorded.

To be conservative about the burn-in period, only the second half of the recorded MCMC traces were used to create kernel density estimates of the five tri-variate marginal distributions on the bottom level facet locations. In fact, for each facet, to be able to view its three-dimensional marginal density, a conditional density estimate was calculated by thresholding the samples and keeping only those which fell in a particular image slice. An optimally-chosen (Terrell, 1990) normal kernel was used for the density estimate. The resulting marginal densities for the five facets, conditional on their being in slice 35, are shown in figure 4.10. For each facet the half-maximum density contour is displayed.

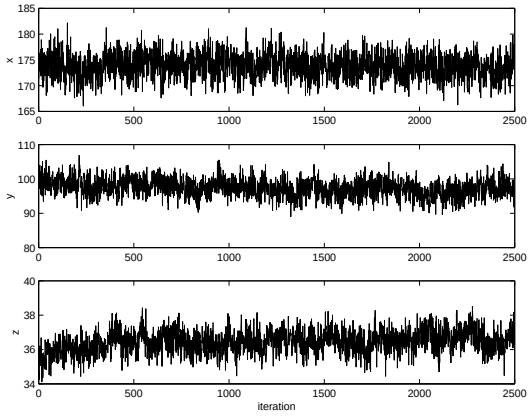
Remember that, under the shape distribution $p(x|\mu, \kappa)$, all of these densities would



(a) Mid-hippocampus facet



(b) Boundary-hippocampus facet



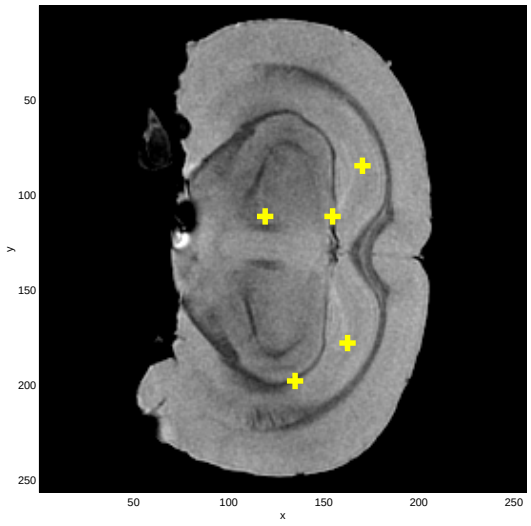
(c) Parent of Mid-hippocampus facet

Figure 4.9: MCMC samples for two bottom level facets and the parent one parent (in three dimensions)

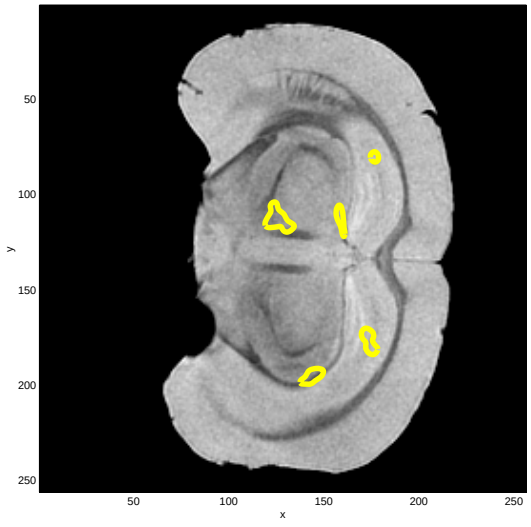
be spherical normal. It is the contribution of the feature density which causes the contours to elongate along isolines of intensity in the new image. This effect is seen most strongly in the two hippocampus boundary facets highlighted.

The mid-hippocampus facet to the south of the image seems to show reasonable variability, but the north-most middle hippocampus facet shows a very tight contour and seems to be predicted too high. The reason for this is as follows: remember from the density maximization, section 4.2, that this facet is more likely in slice 36 rather than the slice 35 displayed. This small and high contour is an artifact of the threshold conditioning used to display the facet density in slice 35, since the full trivariate density is elongated along the direction from the south in slice 36 to the north in slice 35. The large variability seen in the furthest west facet is most likely due to the lack of homology between the two brains in the lower brain.

Figure 4.11 shows three views (coronal, transverse, and sagittal) of the marginal density estimate for the south-most facet in figure 4.10. The left column shows the atlas brain centered at the facet's atlas location, and the right column shows the new brain centered at the maximum density predicted location. All contours displayed are conditional densities computed by thresholding in the direction perpendicular to the page. In all views the contours enclose reasonable predictive regions for this facet's location.



(a) Atlas image coronal slice 39



(b) Observed coronal slice 35

Figure 4.10: Kernel density estimates of marginal predictive densities on five bottom level facets from MCMC samples.

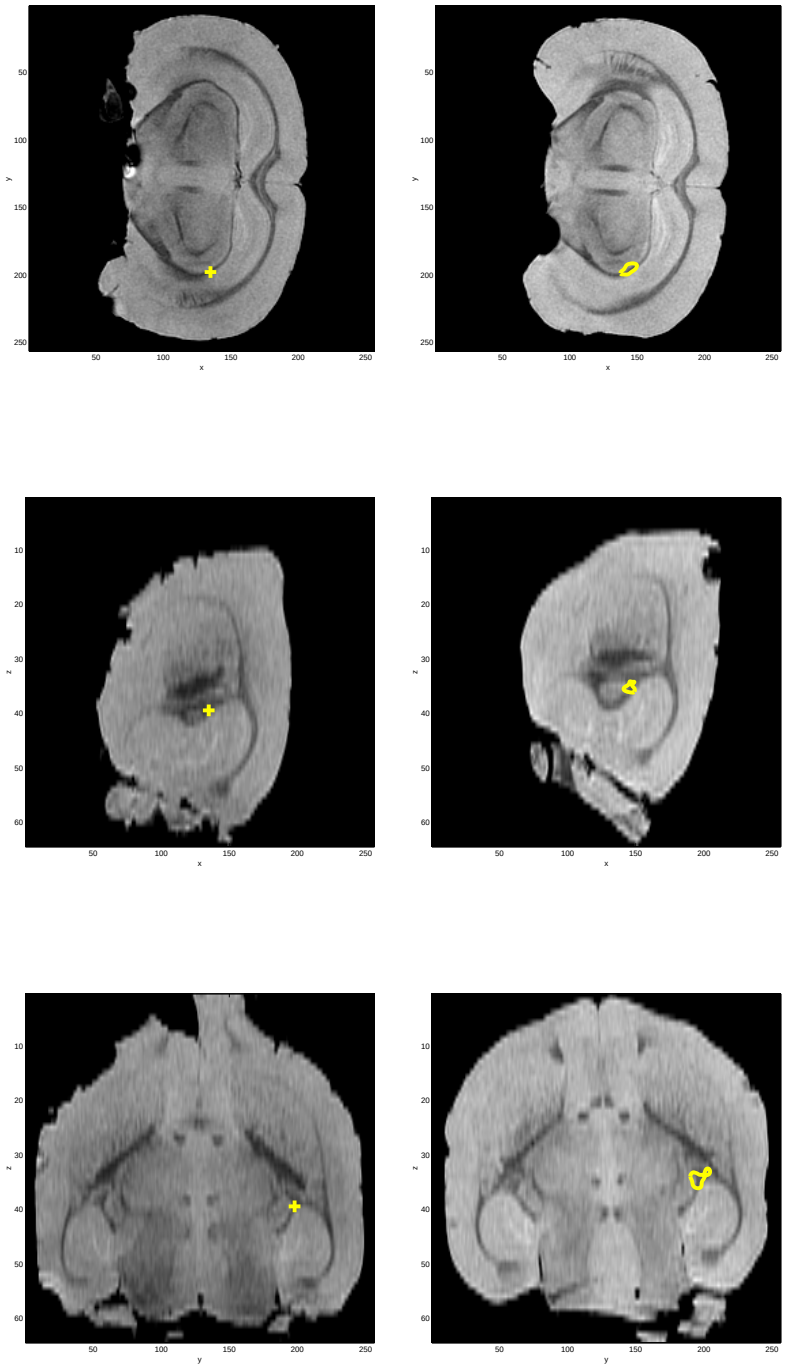


Figure 4.11: Three views of the tri-variate marginal density on a facet location in the new image. The atlas is displayed in the left column in three views centered at the facet template location. Predictive density contours are displayed in the right column on the new image.

Chapter 5

Application: Time Gated Cardiac Single Photon Emission Computed Tomography

5.1 Introduction and General Framework

Single Photon Emission Computed Tomography (SPECT) is used widely in the US for the diagnosis of coronary artery disease. Although other modalities have shown promise for this purpose, SPECT provides a relatively low-cost and non-invasive tool for visualizing normal and abnormal heart function. Heart function is measured by imaging the spatial distribution of a radioactive tracer that has been injected into the patient's bloodstream. An infarct can be detected by a dark spot in the image because less blood, and therefore less tracer, travels to the damaged area. To acquire an image, a camera counting gamma-ray photons ejected from the body is oriented at several angles of rotation around the body. At each angle the camera acquires one view (or projection) of the photon ejection profile in the body. By combining a set of projections at different angles around the body, a three-dimensional reconstruction can be calculated that displays the body in real space. This reconstructed image is the final product that clinicians view for diagnosis purposes.

Even though cardiac SPECT has become a clinically routine technique, much

improvement is still possible in acquisition and processing techniques. One limitation of ordinary cardiac SPECT is that to produce a reasonable signal-to-noise ratio (SNR) in the reconstructed image, the camera must be fixed at each angle around the body for several seconds, during which the heart beats several times. This produces motion artifact: any part of the heart which moves significantly during the beat cycle is blurred out because of its motion relative to the camera. The region most severely affected by motion artifact is the wall of the left ventrical (LV) due to its large motion while beating. Abnormalities in the LV can only be detected if they have size on the order of (or larger than) the magnitude of its beating motion.

One promising direction of research to correct for this artifact is gated myocardial perfusion imaging (Faber *et al.*, 1991). In this technique, rather than acquiring one long image of the heart with high signal-to-noise but poor motion artifact, several images are acquired by binning the photon counts into gates triggered from the patient's own electro-cardiogram (ECG). By gating these images using the ECG, each image acquired becomes a view of the heart in one pose in its beat cycle. Each gated image now has very little motion artifact when compared with standard non-gated SPECT. The trade-off is that, for equal total scan-times, the signal-to-noise ratio compared with an ungated image is reduced by a factor equal to the number of time gates imaged.

This ECG gating is a major advance in itself since it gives the physician the option of looking for smaller abnormalities in these gated images, but the low SNR of each gate makes detection of these infarcts difficult. In this chapter it will be shown that the facet model can be used very effectively to combine the information from all the gates into one "composite" image that displays the heart in any pose with low motion artifact and high SNR.

The advantage of the facet model is that anatomical locations on the heart can be

followed through the gated images. Once these locations are found in each time gate, then a reasonable composite image can be formed by averaging, for each anatomical location, its image value in each time gate.

To produce this "facet-composite" image, one gate is chosen to be the atlas image (in the terminology of previous chapters) and the facet model is applied to register every other gate to the atlas. Any gate can be chosen as the atlas, and the facet-composite image will display the heart in the pose of that chosen gate. The template facet locations μ are set so that the bottom level facets in the graph are at the center of each voxel in the atlas image and the predictive distribution on facet locations (eq. 2.23) is maximized for each other gate in turn. Each bottom level facet then has a predicted location $(\hat{x}_{L,i})_j$ in each time gate j. The facet composite image is formed by displaying, for every facet, its average intensity value over the J gates located at its atlas position. Each pixel in the atlas has an associated bottom level facet, so the value at pixel i in the facet composite image is

$$V_i = \frac{1}{J} Q_0\left(\mu_{L,i}\right) + \frac{1}{J-1} \sum_{j=1}^{J-1} Q_j\left(\left(\hat{x}_{L,i}\right)_j\right),$$

where Q_0 is the atlas gated image and $(Q_1, \ldots, Q_{J-1})$ are the other gates. All facets above the bottom level are treated as latent random variables in the shape model and ignored when creating the composite image.

5.2 Results

A study was conducted using six patient images from the Duke University Medical Center. The patients were injected with Tc-99m-labeled (140 keV gamma ray) Sestamibi, a myocardial tracer commonly used in Nuclear Medicine, after they exercised for a short time. The patients were scanned with ECG gated SPECT on a three headed scanner ("Triad", Trionix, Inc., Twinsburg, OH) with each head using a low

energy, general purpose collimator. The camera acquired data in eight equally distributed time gates, triggered by the patient's own ECG signal, where, in the final images, Gate 2 corresponded to the systolic (compressed) cardiac phase and Gate 7 corresponded to the diastolic (relaxed) cardiac phase. The 90 total acquired angular samples for all gated data sets were reconstructed with a filtered back-projection algorithm with post reconstruction smoothing employing a Hann filter (0.60 cy/cm cutoff from the Nyquist frequency). The images were then independently cropped to include the reconstructed heart volume, and then reoriented so that the heart view was the short axis view. (The other two orthogonal views are referred to as the vertical and horizontal long axis view.) The resulting three-dimensional images consisted of $32\times32\times20$ voxels of size $7.12\text{mm}\times7.12\text{mm}\times7.12\text{mm}$.

Figure 5.1 shows a mid-heart slice of the short axis view of one patient at all eight time gates. The curve shown to the right is a typical volume curve (Opie, 1991) for a heart as it beats through its cycle. For all analysis in this section, the seventh gate, in the diastolic cardiac phase, was used as the atlas image.

A five level facet graph was used to model the deformation so that the bottom level had $16\times16\times16$ facets. In all patients this number of bottom level facets was sufficient to cover the whole atlas heart with one facet per voxel. Level weights were set to ensure equal facet marginal variances and parent/child weights were set inversely proportional to the atlas separation, as discussed in section 2.4.1.

The standardized regression feature model (2.15) was used because the nature of the images made it difficult to rescale the intensities in a sensible way. Since these images consist of extended regions of very low intensity and extended regions of very high intensity, any rescaling based on the tenth and ninetieth quantiles of intensity inside a closed region of interest (as done in Chapter 4) is very sensitive to the choice of the rescaling region. The standardized regression feature function is

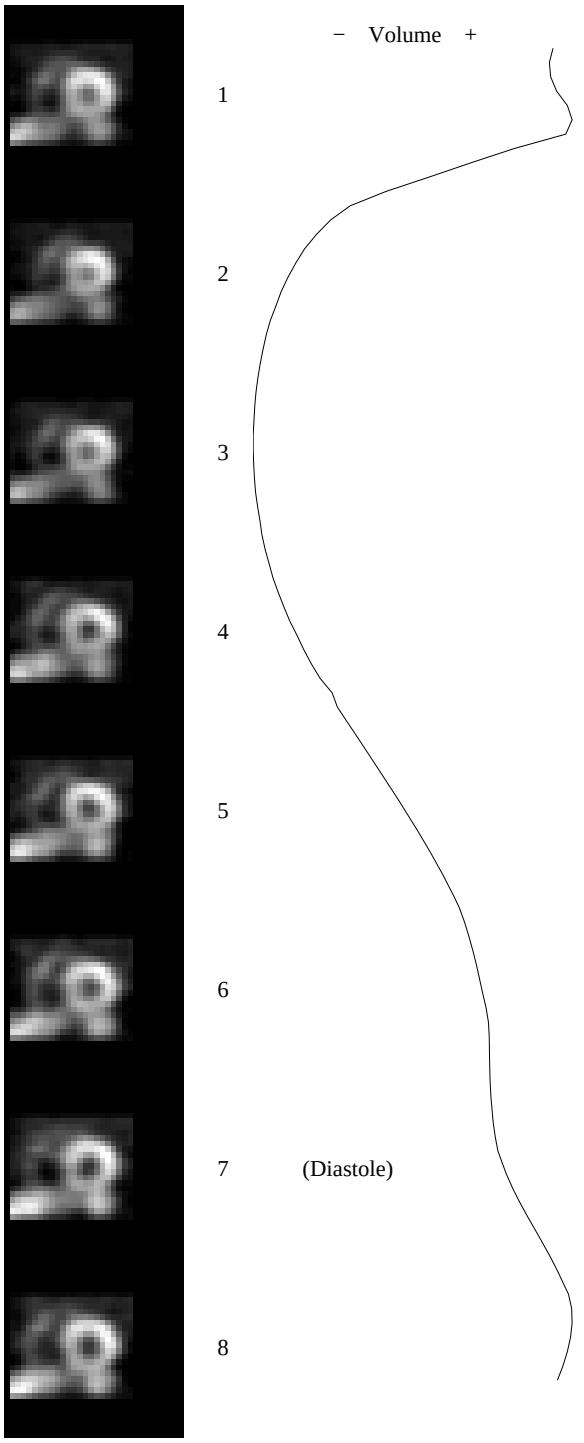


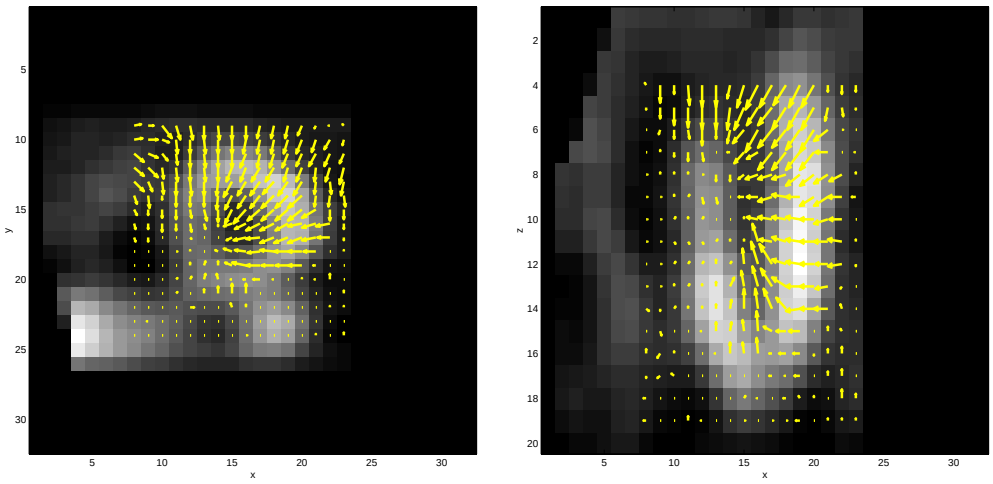
Figure 5.1: Eight Gates of the short axis view of one patient and a typical heart volume curve. Gate 7 was used as the atlas in all cases.

insensitive to the absolute intensity near each facet, and therefore works extremely well in this application. For each bottom level facet $(\cdot)_{L,i}$ the regression design points $((z_{L,i})_1, \ldots, (z_{L,i})_n)$ were oriented as 72 quasi-random spherical normal samples in three-dimensions with mean equal to the facet's location and standard deviation equal to the bottom level atlas grid spacing, 1 pixel. (This number of samples in three dimensions is suggested by Shaw (1988) according to space-filling optimality criteria.)

Predicted facet locations in all gates were estimated by maximizing the constrained facet data model, eq. 2.23, using the diastolic gate 7 as the atlas image. The hyperparameters τ and κ were set manually by observing a few gate registrations in one patient and the same values were used for all patients.

Figure 5.2 shows sample results of the registration process in one patient. The predicted facet motion from the diastolic phase gate 7 to the systolic gate 2 is represented by a set of arrows. The background in both panels is the atlas image and the arrows point from the gate 7 positions $\mu_{L,i}$ to the predicted gate 2 positions $x_{L,i}$. The left panel shows the short axis view of facets in atlas slice 12 of 20 and the right shows the horizontal long axis view of facets in atlas slice 16 of 32. Remember that facets can be predicted to be out of slice, so some arrows may actually point into or out of the page. The region of largest motion is in the top right of both panels and corresponds well to expected LV wall motion.

Figures 5.3 and 5.4 display vertical and horizontal long axis views in one patient. In both figures, each row is one slice perpendicular to the direction of the view. In the left column the original gate 7 data is displayed, in the middle column is the facet-wise composite image, and in the right column is the ungated SPECT image (produced from these datasets by a pixel-wise sum over all gates). It is clear that the facet-composite image recovers the same level of SNR as the ungated column while



(a) Short axis view

(b) Horizontal long axis view

Figure 5.2: Facet motion when matching gate 7 (diastolic phase) to gate 2 (systolic phase). The left panel shows the short axis view slice 12 of 20 and the right panel is the horizontal long axis view slice 16 of 32. In both panels the background image is the atlas image, diastolic gate 7. The tail of the arrow represents the atlas facet position μ in gate 7 and the head is the predicted facet position x in gate 2. Facets not displayed were predicted to move a negligible distance. The facet motion closely parallels the expected left ventricular wall motion.

reducing the motion artifact by adhering to the pose of the gate 7 raw image.

In an effort to quantitatively measure the performance of the algorithm in this application, first a trained radiologist was asked to pick from each patient's gate 7 image a homologous slice near the mid-heart in the short axis view. Next, in each patient, the observer chose a horizontal line through the center of the heart in this short axis slice. Given this manually chosen one dimensional line in the each patient's three-dimensional data sets, an intensity profile could be plotted through the heart. Figure 5.5 shows a typical profile for one patient from the six. The same line was used to generate intensity profiles in the facet-wise composite image and the ungated image. The presence of motion artifact in the ungated image causes the two intensity peaks on either side of the LV to be closer together and fatter than they are in the gated image, since gate 7 is at a point in the heart beat cycle where the heart is relaxed and has a fairly large volume. The two peaks in each profile were fit to Gaussian kernels using the least squares method and the fitted means and full-width-half-maxima (FWHM) were recorded.

In Table 5.1, the three image methods (gated, facet-wise average, and pixel-wise average) are compared by the separation of the fitted peaks on either side of the LV. In all cases except the last the facet-wise composite profile adhered more closely to the gate 7 pose than did the ungated profile. The average percent difference in centroid separation in the facet composite profile was approximately 5% compared with 9% for the ungated profile. This indicates that a good portion of the motion artifact has been eliminated in the facet-wise composite.

Table 5.2 gives similar results of a comparison of the fitted FWHM in the three image methods. In each cell of the table values are given for the left and right walls in the profile. On average wall thickening was increased in the facet-composite profile by 32%/4% (left/right) compared with 36%/14% in the ungated profile.

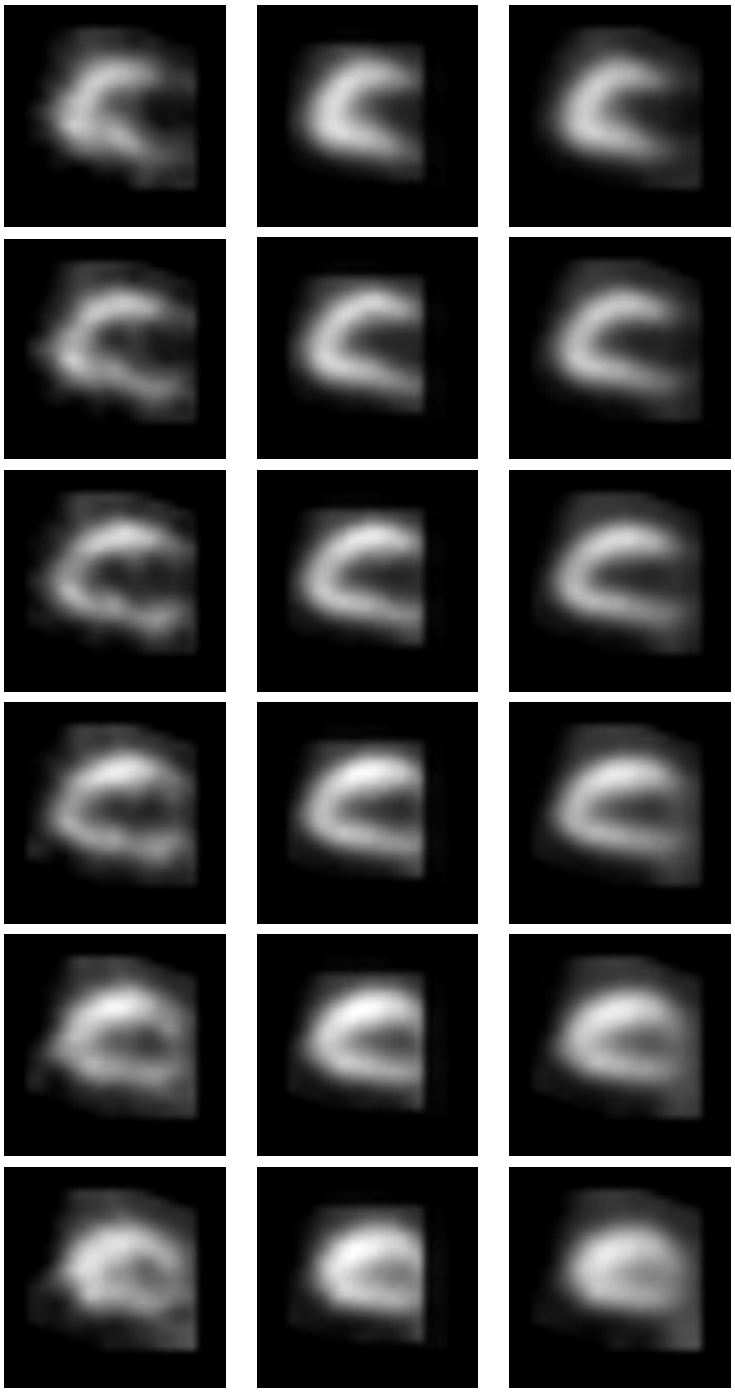


Figure 5.3: Six planes of the vertical long axis view by 3 methods. The left column is the raw gate 7 data, the middle column is the facet-wise composite, and the right column is the ungated image. Note that the facet-wise composite image is only defined in the cubic region where facets were located in the atlas, hence the large black borders in the middle column (these images have been interpolated and re-pixilated from their original 20×32 to 256×256 for easier viewing).

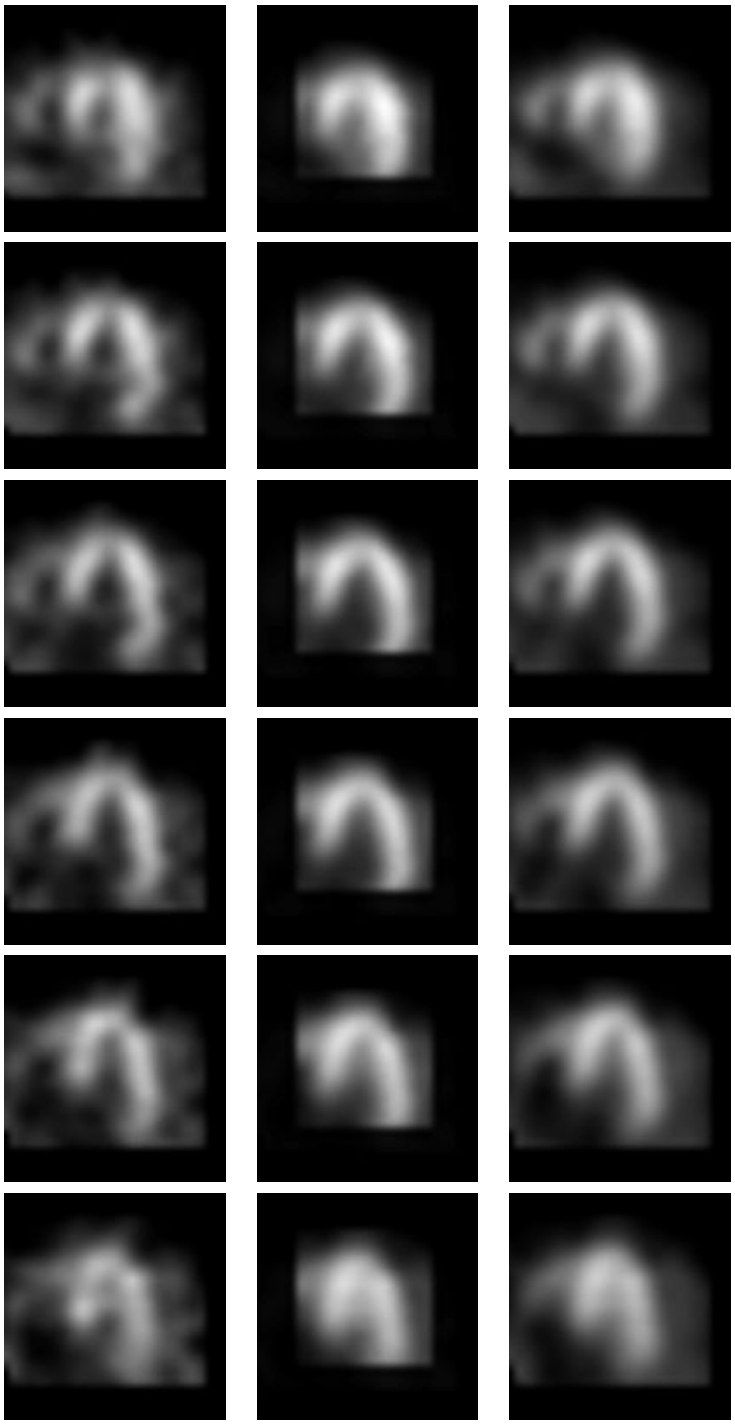


Figure 5.4: Six planes of the horizontal long axis view by 3 methods. The left column is the raw gate 7 data, the middle column is the facet-wise composite, and the right column is the ungated image. Note that the facet-wise composite image is only defined in the cubic region where facets were located in the atlas, hence the large black borders in the middle column (these images have been interpolated and re-pixilated from their original 20×32 to 256×256 for easier viewing).

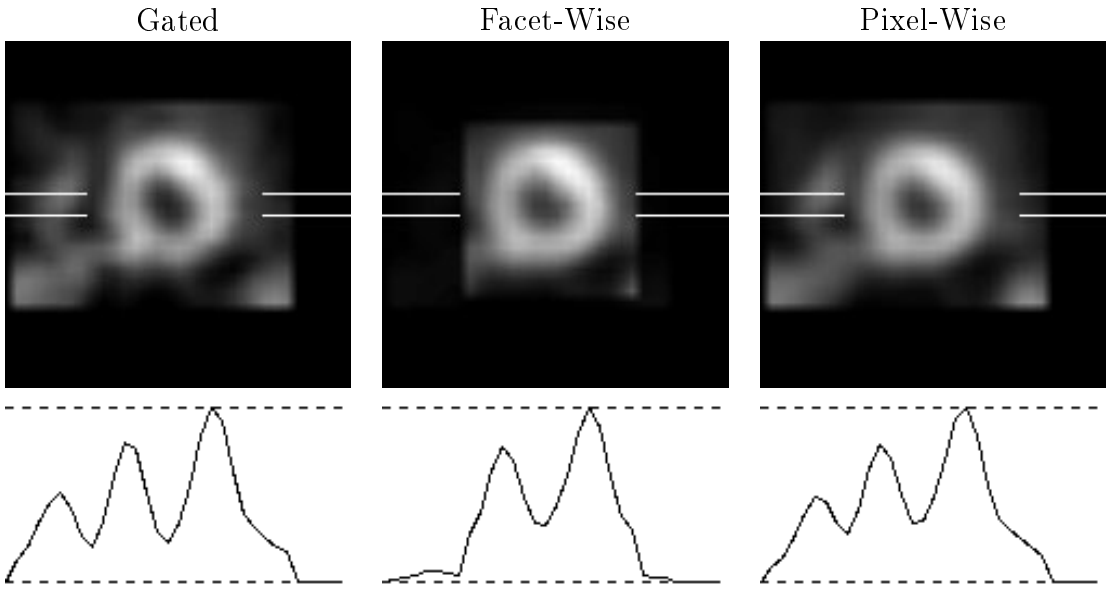


Figure 5.5: Three methods for producing a short axis view in one patient. For each of the 6 patients, a 1 dimensional horizontal intensity profile was measured. The orientation of this slice was determined by a trained radiologist to be located centrally in the short axis view and in a similar slice in each patient. Typical intensity profiles from the three methods are shown on the bottom panels.

Patient	Gate 7	Facet-wise (FW)	Pixel-wise(PW)	FW-PW
A	6.36	5.72	5.38	.34
B	7.41	7.25	6.56	.69
C	6.35	5.75	5.52	.23
D	6.63	6.18	6.01	.17
E	10.14	9.64	9.31	.33
F	7.69	7.93	7.70	.23
mean	7.43	7.07	6.75	.32

Table 5.1: The separation in pixels (pixel edge length 7.12mm) between intensity peaks in the left and right walls of left ventricle measured on a 1 dimensional horizontal slice located centrally in the short axis view and in a similar slice in each patient.

Patient	Gate 7	Facet-wise (FW)	Pixel-wise(PW)	FW-PW
A	3.99/3.37	8.40/3.92	10.20/4.23	-1.80/-0.31
B	9.56/11.54	10.25/11.1	11.16/12.07	-0.91/-0.97
C	4.47/5.56	7.11/5.74	5.60/6.16	1.51/-0.42
D	5.87/7.76	8.64/7.32	7.32/7.48	1.32/-0.16
E	4.07/5.55	4.66/5.96	4.72/5.88	-0.06/0.08
F	5.36/5.52	4.91/6.71	6.02/9.15	-1.11/-2.44
mean	5.54/6.55	7.33/6.79	7.51/7.49	-0.18/-0.70

Table 5.2: The Left Ventriclular wall widths (left/right) in pixels (pixel edge length 7.12mm) determined by the full-width-half-maximum (FWHM) of the extracted one-dimensional intensity profiles.

Finally, Figure 5.6 displays a representation of the three-dimensional cardiac images called a "bullseye" plot. Essentially the heart is numerically flattened out from base to apex so that each quadrant of the image represents one quarter of the heart and moving radially out from the center corresponds to moving up the wall of the heart from apex to base. These plots are used extensively as a diagnosis tool because the most important parts of the whole three dimensional image set are reduced into one image for the physician to view.

The right three panels of figure 5.6 are the three image methods for a normal patient drawn from the original six, and the left panels show the only infarcted patient in the set. The muscle damage is represented in each left panel by a dark non-circular region near the center of each image. While the right panels have insufficient relief to see any significant difference between the three analysis methods, the infarct in the left panels exposes the motion artifact that plagues the pixel-wise average method. The true extent of the infarct shown in the gated image is well modeled in the facet-wise composite, but the pixel-wise average considerably distorts and blurs the region. Additional artifacts are also present toward the SW of the pixel-wise image which are not represented in either of the other two images. The facet-wise composite image is faithful to the structure of the gated image while increasing the signal-to-noise ratio

by averaging over the eight gates in the dataset.

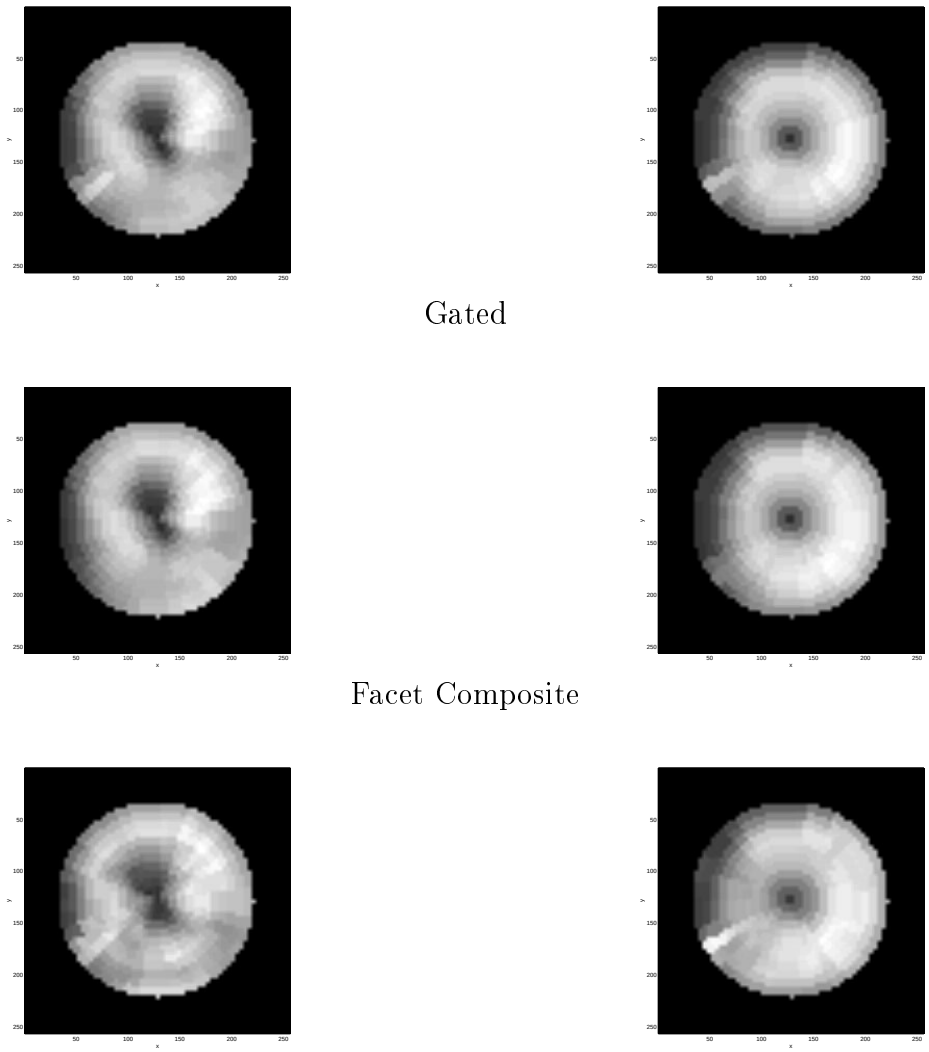


Figure 5.6: Bullseye representations. The left panels show an infarcted patient by three image methods and the right panels show a healthy patient.

Chapter 6

Summary and Extensions

6.1 Important Results

The models outlined in this paper show tremendous promise as a coherent system for automated feature-based image analysis. Feature registration of two images is accomplished by predicting the locations of a large number of facets in one of the images given their locations in the other. Furthermore, an immediate outcome of this registration is the automated segmentation of one image given a manual segmentation of the other.

Because the system is a full statistical model of how an observer locates features in an image, the resulting procedures can be characterized far more comprehensively than any of the more heuristic numerical algorithms put forth previously. The tuning parameters, which in a simple numerical algorithm would have to be set manually, are parameters in the statistical model which can be estimated from appropriate training data. This system therefore provides the true objectivity so fervently pursued in automatic procedures. The simple form of the statistical model proposed in this paper allows for straightforward estimation of these parameters.

Furthermore, the unavoidable issue of observer variability that is ignored in a

simple numerical optimization algorithm can be addressed in a statistical model. For example, when considering real images there is never a single "correct" segmentation; every intelligent observer is correct. Any system for automating the segmentation process must therefore integrate this fact.

The straightforward joint modeling of facet locations and feature values to estimate hyperparameters leads to a predictive distribution on facet locations in a new image which incorporates several attractive properties. Its hierarchical structure permits the practical top-down approach for density maximization by conditioning on large scale image feature locations before addressing the local variability. The result is a very robust and fast procedure even in large three-dimensional images. Likewise, the hierarchical detachment of large scale "unobservable" facets from those "observable" facets on the bottom level of the graph permits fast Gibbs sampling on all the upper level facet locations. Registration variability estimates are therefore feasible in a modest amount of time.

Both registration analyses given in chapters 4 and 5 were extremely successful. The facet-composite cardiac SPECT images showed marked improvement in signal-to-noise compared to the single image gates while significantly reducing the motion artifact seen in the regular pixel-wise average images. In the mouse brain MR registration analysis, a method was highlighted for displaying regions of poor model fit which could be used as a detector of abnormality relative to the atlas image. In addition, a straightforward extension of the model was used to produce a reasonable hippocampus segmentation of one brain given a segmentation of the other.

6.2 Extensions

The models presented in this paper are very flexible and many extensions can be envisioned. This section provides a few suggestions for tailoring the methods to

different applications, as well as discussion regarding the use of manual segmentation data to estimate the model hyperparameters.

6.2.1 Model Development

First, the shape model given in section 2.4.1 is extensible to incorporate more prior information about the kinds of deformations expected in the class. For example, simple changes could be implemented such as favoring deformations along a particular plane by allowing different variability parameters by image dimension: $\kappa \rightarrow (\kappa_1, \ldots, \kappa_d)$. This generalization allows the model to be applied in, for example, stereopsis applications (Dhond, 1989) in robot vision and remote sensing, where the aim is to generate three-dimensional landscapes from two different two-dimensional views. Since relative deformation between the views is constrained to be in the direction connecting the cameras, the facet model could be applied by simply constraining facet deformations accordingly. This is an extreme case of adjusting the new d-dimensional κ since each facet location can actually be represented by a one-dimensional random variable rather than the original two for general deformations.

The shape model can also be modified to incorporate prior knowledge about objects in an image class that are expected to deform together within the class. This form of prior information is easily modeled by increasing the parent/child weights between constituent facets in each object, and decreasing those connection across the object's boundary. One extreme approach has been investigated by Laading *et al.* (1997) in which certain parent/child connections were set to zero, transforming the facet graph into a tree structure.

More complex facet correlation structures could be modeled by treating all the parent/child weights as unknown parameters and simulating from their (non-standard) joint posterior distribution given manual training data. An informative prior would

have to be specified on the parent/child weights since the amount of data would invariably be sparse relative to the number of additional parameters. In fact, data in the form of manual segmentations rather than individual facet locations would most likely have to be used. This type of data is discussed below.

In applications like the cardiac SPECT images in chapter 5, where several images are taken in a sequence in time, modeling the whole four-dimensional dataset would be fruitful. By adding the fourth time dimension, structural constraints can be added to the allowable deformations under the shape model that are inspired by past heart motion studies. There have been several such studies (McEachen and Duncan 1997, Clarysse *et al.* 1996, and Park *et al.* 1996) that might be useful for deriving a structural prior template of sensible heart motion.

The feature models used in this paper can also be generalized by incorporating different image measures to be effective in specific applications. A study of different image measure possibilities is given by Studholme *et al.* (1997). The spatial feature models alluded to in section 2.4.2 could be investigated, although their added complexity might outweigh the precision gained in the posterior on feature hyperparameters.

6.2.2 Parameter Estimation Using Region Data

The final issue discussed here is the matter of data acquisition. The fifty data points in the mouse MR application were acquired because this form of data was most suitable for estimating the parameters in this model. This tedious chore might be avoided if a method can be found for using the already abundant data in the form of manual region segmentations. This task is just as tedious, however it is constantly performed in the clinic since automatic algorithms have yet to prove themselves.

Region data can be treated as missing data on facet locations. Consider that two

images have been segmented manually. Choosing one arbitrarily as the atlas image, the atlas segmentation essentially assigns labels to each facet. Each facet i for which $\mu_{L,i}$ falls within a particular atlas segmentation region A_t is a t-type facet. For a set of region types T segmented in the atlas image there is the same number of facet types $t \in T$. Then, given a similar set of segmentation regions $\{B_t, t \in T\}$ in the new image, the data can be represented as

$$x_{L,i} \in B_{t_i} \text{ for all } i \tag{6.1}$$

where facet i has atlas segmentation type t_i. This means that, for every facet on the bottom level of the graph, its location must lie in the region of the new image covered by the manual segmentation of its type in that image. In keeping with the protocol of having only bottom level facets observable, segmentation constraints should not be enforced on higher level unobservable facets.

The difficulty using this data to estimate the model parameters is that the image function Q inside each region must also be incorporated to indicate the associated feature value with each facet's location. This constraint has the form

$$f_{L,i} = f_Q(x_{L,i}) \text{ for all } i$$

where the feature function f_Q is discussed in section 2.4.2. If the constrained data model (2.23) is used to enforce this constraint, then the applicable data model is

$$p(x|\kappa, \tau, \mu, \phi) = \frac{1}{Z(\kappa, \tau, \mu, \phi)} \exp\left(-\frac{\kappa}{2}(x-\mu)'P(x-\mu) - \frac{\tau}{2}\sum_i F(x_{L,i})\right) \tag{6.2}$$

and the data on x is the constraint given in (6.1). The function F is defined in section 3.1. Since the unknown normalizing constant Z is a function of κ and τ, the posterior on these parameters is unknown. However the conditional distribution on x is an exponential family with sufficient statistics

$$(x-\mu)'P(x-\mu) \quad \text{and} \quad \sum_i F(x_{L,i}). \tag{6.3}$$

A maximum likelihood estimate (MLE) of κ and τ can therefore be calculated by examining the observed sufficient statistics (from the region data) and the expected statistics under the distribution (6.2) given κ and τ. The parameter values which produce expected sufficient statistics matching the observed sufficient statistics are the MLE given the data (Bickel and Doksum, 1977).

Unfortunately, the observed sufficient statistics are not immediately calculable from the region data, nor are the expected sufficient statistics known analytically. A missing data framework (Geman and McClure, 1987) must be employed in which two streams of Markov chains are run conditional on a set of parameters κ and τ: one enforcing the constraint (6.1) and one ignoring it. For both Markov chains the two sufficient statistics (6.3) are recorded. The Monte Carlo mean estimate from the constrained Markov chain gives an estimate of the "expected" observed sufficient statistics under the model and the unconstrained Markov chain gives an estimate of the expected sufficient statistics under the model. Similar pairs of chains are run for several choices of κ and τ and the pair that produces the closest match of expected sufficient statistics is the MLE of κ and τ given the region segmentations.

Other than the obvious computational burden of this parameter estimation technique, there are also some technical difficulties that must be overcome. First, the choice of initial condition for the iterative sampling is not straightforward. In the constrained Markov chain, facets must be placed in their proper regions according to the manual segmentation while maintaining a joint configuration close to the mode of the distribution. Second, a sufficient statistic match criterion must be chosen since there is a separate statistic for every parameter in the model. Furthermore, the MLE estimated using two segmentations in a class of images is only strictly applicable for those two images. Some thought must be given to reasonable methods for combining estimates from multiple image segmentations.

Appendix A

Derivation of Equation 2.23

In this appendix, a discussion is given of the derivation of the conditional distribution (2.23). The one-dimensional case, corresponding to one facet in a one-dimensional image, is considered first.

Given a joint probability density $f(x, y)$ on two continuous random variables X and Y, the limiting probability density on X under the constraint that $|Y - L(X)| < \epsilon$ as $\epsilon \to 0$, for a continuous function L is proportional to the joint density evaluated at $y = L(x)$, namely $f(x, y = L(x))$.

If the function L is defined to be constant, $L(x) = y_0$, then the resulting limiting conditional probability density on x is the regular $g_1(x|y = y_0)$. It is instructive to derive this density from the underlying c.d.f.'s. The desired c.d.f. is the following probability

$$P(X < x|\ |Y - y_0| < \epsilon) = \frac{P(X < x \cap |Y - y_0| < \epsilon)}{P(|Y - y_0| < \epsilon)} \tag{A.1}$$

in the limit as $\epsilon \to 0$. The event written in the numerator is shown in figure A.1 (a). Using the conventional definition of the joint c.d.f. $F(x, y) = P(X < x \cap Y < y)$, equation (A.1) can be re-written

$$P(X < x|\ |Y - y_0| < \epsilon) = \frac{[F(x, y_0 + \epsilon) - F(x, y_0 - \epsilon)]/\epsilon}{[F_2(y_0 + \epsilon) - F_2(y_0 - \epsilon)]/\epsilon} \tag{A.2}$$

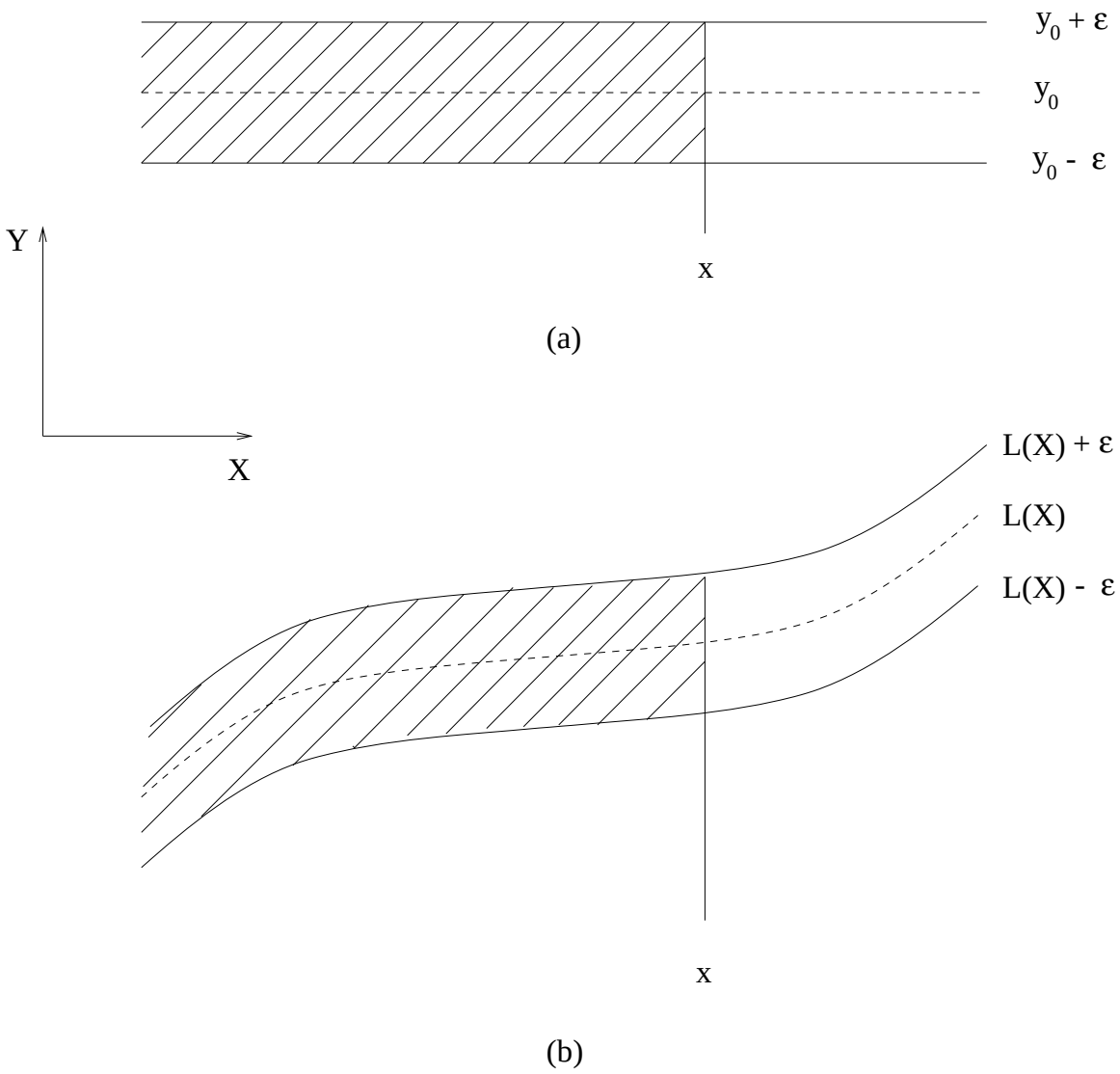


Figure A.1: The cross-hatched event $X < x \cap |Y - L(X)| < \epsilon$ is displayed in part (a) for $L(x) = y_0$ and in part (b) for general $L(x)$.

where $F_2(y) = P(Y < y)$. Both the numerator and denominator have been divided by ϵ so that we may re-write the equation using probability densities,

$$\lim_{\epsilon \to 0} P(X < x | \, |Y - y_0| < \epsilon) = \frac{\left[\frac{\partial}{\partial y} F(x,y)\right]_{y_0}}{f_2(y_0)}. \tag{A.3}$$

Finally we arrive at the expected result

$$\frac{\partial}{\partial x} \lim_{\epsilon \to 0} P(X < x | \, |y - y_0| < \epsilon) = \frac{f(x,y_0)}{f_2(y_0)} = g_1(x|y = y_0). \tag{A.4}$$

Now, for a general continuous function L, the desired conditional c.d.f. is

$$P(X < x | \, |Y - L(X)| < \epsilon) = \frac{P(X < x \cap |Y - L(X)| < \epsilon)}{P(|Y - L(X)| < \epsilon)}. \tag{A.5}$$

in the limit as $\epsilon \to 0$. The constraint here, shown in figure A.1 (b), is the region of width 2ϵ bounding the function L above and below for its entire domain. (This is the desired region in the case of equation (2.23) since the support of the r.v. representing facet location is not affected by the image constraint, it is still the whole real line.)

The denominator in equation (A.5) can be rewritten in terms of the joint density $f(x,y)$,

$$\begin{aligned} P(|Y - L(X)| < \epsilon) &= \int_{-\infty}^{\infty} \int_{L(x)-\epsilon}^{L(x)+\epsilon} f(x,y) dy dx \\ &= \int_{-\infty}^{\infty} f_1(x) \int_{L(x)-\epsilon}^{L(x)+\epsilon} g_2(y|x) dy dx \\ &= \int_{-\infty}^{\infty} f_1(x) \left[G_2(L(x) + \epsilon | x) - G_2(L(x) - \epsilon | x) \right] dx \end{aligned} \tag{A.6}$$

and, similarly, the numerator in equation (A.5) can be rewritten

$$\begin{aligned} P(X < x \cap |Y - L(X)| < \epsilon) &= \int_{-\infty}^{x} \int_{L(x)-\epsilon}^{L(x)+\epsilon} f(t,y) dy dt \\ &= \int_{-\infty}^{x} f_1(t) \left[G_2(L(t) + \epsilon | t) - G_2(L(t) - \epsilon | t) \right] dt. \end{aligned} \tag{A.7}$$

Finally, the desired probability is

$$\lim_{\epsilon \to 0} P(X < x | \; |Y - L(X)| < \epsilon)$$

$$= \lim_{\epsilon \to 0} \frac{\int_{-\infty}^{x} f_1(t) \left[G_2(L(t) + \epsilon | t) - G_2(L(t) - \epsilon | t) \right] / \epsilon \quad dt}{\int_{-\infty}^{+\infty} f_1(t) \left[G_2(L(t) + \epsilon | t) - G_2(L(t) - \epsilon | t) \right] / \epsilon \quad dt} \tag{A.8}$$

$$= \frac{\int_{-\infty}^{t} f_1(t) g_2(L(t)|t) dt}{\int_{-\infty}^{+\infty} f_1(t) g_2(L(t)|t) dt}.$$

The denominator in equation (A.8) is not a function of x which allows us to write the desired density on x under the appropriate constraint as

$$\frac{\partial}{\partial x} \lim_{\epsilon \to 0} P(X < x | \; |Y - L(X)| < \epsilon) = \frac{1}{Z} f(x, L(x)). \tag{A.9}$$

This progression extends to higher dimensional X and Y. In practice, X represents the vector of facet locations and Y represents the vector of facet feature values. Therefore, let X be a $(p \times d)$-dimensional vector

$$X = (X_{1,1}, \ldots, X_{1,d}, \ldots, X_{p,1}, \ldots, X_{p,n})$$

and Y be a $(p \times n)$-dimensional vector (to account for the possibility of a n-dimensional feature vector associated with every facet)

$$Y = (Y_{1,1}, \ldots, Y_{1,n}, \ldots, Y_{p,1}, \ldots, Y_{p,n}).$$

The function $L(x)$ represents the image function $Q(x) : R^d \to R$. Then the analogous constraint to the one-dimensional case is

$$|Y_{i,k} - L\left(z_k(x_i)\right)| < \epsilon \quad \text{for } i = 1, \ldots, p \text{ and } k = 1, \ldots, n$$

where the deterministic function $z_k(x_i)$ is used to create the n-dimensional vector $(z_1(x_i), \ldots, z_n(x_i))$ for instance to represent a set of n points in a region centered at x_i.

The integral in (A.6) becomes

$$P(|Y_{i,k} - L\left(z_k(x_i)\right)| < \epsilon \quad \forall i,k)$$

$$= \int_{R^{p\times d}} f(x_{1,1},\ldots,x_{p,d}) \int_{L(z_1(x_1))-\epsilon}^{L(z_1(x_1))+\epsilon} \cdots \int_{L(z_n(x_p))-\epsilon}^{L(z_n(x_p))+\epsilon} g_2(y_{1,1},\ldots,y_{p,n}|x_{1,1},\ldots,x_{p,d}) \, dy_{1,1}\cdots dy_{p,n} dx_{1,1}\cdots dx_{p,d}$$

and the integral in (A.7) is similar, with appropriately changed integration limits. The limiting joint density on the whole vector x as $\epsilon \to 0$ becomes

$$\frac{1}{Z} f\left(x_{1,1},\ldots,x_{p,d}, L(z_1(x_1)),\ldots,L(z_n(x_p))\right).$$

Bibliography

Amit, Y., Grenander, U. and Piccioni, M. (1991) Structural image restoration through deformable templates. *Journal of the American Statistical Association*, **86,** no. 414, 376–387.

Amit, Y. and Kong, A. (1994) Graphical templates for model registration. Technical Report. Department of Statistics, Chicago, IL.

Bajcsy, R. and Kovacic, S. (1989) Multiresolution elastic matching. *Computer Vision, Graphics, and Image Processing*, **46,** 1–21.

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, series* B, **36,** no. 2, 192–236.

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

Bickel, Peter J. and Doksum, K. A. (1977) *Mathematical statistics : basic ideas and selected topics*. Holden-Day.

Bookstein, F. (1991) *Morphometric Tools for Landmark Data*. Cambridge: Cambridge University Press.

Bookstein, F. L. (1991) Thin-plate splines and the atlas problem for biomedical images. In *Lecture Notes in Computer Science*, vol. 511, pp. 326–342. Springer-Verlag.

Bowsher, J. E., Johnson, V. E., Turkington, T. G., Jaszczak, R. J., Floyd Jr., C. E. and Coleman, R. E. (1995) Bayesian reconstruction and use of anatomical a priori information for emission tomography. Technical Report. Institute of Statistics and Decision Sciences.

Christensen, G. E., Rabbitt, R. D. and Miller, M. I. (1993) A deformable neuroanatomy textbook based on viscous fluid mechanics. In *Proceeding of the 1993 Conference on Information Sciences and Systems*, pp. 211–216.

Clarysse, P., Friboulet, D. and Magnin, I. E. (1996) Tracking geometrical descriptors on 3-D deformable surfaces: application to the left-ventricular surface of the heart. *IEEE Transactions on Medical Imaging*, **16,** no. 4, 392–404.

Clifford, P. and Nicholls, G. (1994) A metropolis sampler for polygonal image reconstruction. Technical Report. Department of Statistics, Oxford University.

Collins, D., Holmes, C., Peters, T. and Evans, A. (1995) Automatic 3-D model-based neuroanatomical segmentation. *Human Brain Mapping*, **3,** 190–208.

Collins, D. L. and Evans, A. C. (1996) Animal: validation and applications of non-linear registration-based segmentation. *International Journal of Computer Vision*, 1271–1294.

Collins, D. L., Neelin, P., Peters, T. M. and Evans, A. C. (1994) Automatic 3D intersubject registration of MR volumetric data in standardized talairach space. *Journal of Computer Assisted Tomography*, **18,** no. 2, 192–205.

Cootes, T. F., Hill, A., Taylor, C. J. and Haslam, J. (1994) The use of active shape models for locating structures in medical images. *Image and Vision Computing*, **12,** no. 6, 355–366.

Cootes, T. F., Taylor, C. J. and Lantis, A. (1994) Active shape models: evaluation of a multi-resolution method for improving image search. vol. 1, pp. 327–336. BMVA Press.

Cressie, N. A. (1993) *Statistics for Spatial Data.* Wiley.

Davatzikos, C. and Prince, J. (1995) An active contour model for mapping the cortex. *IEEE Transactions on Medical Imaging*, **14,** no. 1, 65–80.

Davatzikos, C. and Prince, J. (1996) Image registration based on boundary mapping. *IEEE Transactions on Medical Imaging*, **15,** no. 4, 212–215.

Davatzikos, C., Vaillant, M., Resnick, S., Prince, J., Letovsky, S. and Bryan, N. (1996) A computerized approach for morphological analysis of the corpus callosum. *Journ Comp. Ass. Tom.*, **20,** no. 1, 88–97.

Dhond, U. R. (1989) Structure from stereo–a review. *IEEE Transactions on Systems, Man, and Cybernetics*, **19,** no. 6, 1489–1492.

Dryden, I. L. (1997) General shape and registration analysis. Technical Report. Department of Statistics, University of Leeds.

Dryden, I. L. and Mardia, K. V. (1996) Statistical shape analysis. Technical Report. University of Leeds.

Eberly, D. H. (1994) Geometric methods for analysis of ridges in n-dimensional images. Ph.D. Thesis. University of North Carolina at Chapel Hill.

Faber, T. L., Akers, M. S., Peshock, R. M. and Corbet, J. R. (1991) Three-dimensional motion and perfusion quantification in gated single-photon emission computed tomograms. *Journal of Nuclear Medicine*, **32,** 2311–2317.

Gee, J. C., Reivich, M. and Bajcsy, R. (1993) Elastically deforming 3D atlas to match anatomical brain images. *Journal of Computer Assisted Tomography*, **17,** no. 2, 225–236.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1995) *Bayesian Data Analysis*. Chapman and Hall.

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,** no. 6, 721–741.

Geman, S. and McClure, D.E. (1987) Statistical methods for tomographic image reconstruction. *Bulletin of the International Statistical Institute*, **52,** no. 4, 5–21.

Godsill, S. J. and Kokaram, A. C. (1997) Restoration of image sequences using a causal spatio-temporal model. In *The Art and Science of Bayesian Image Analysis* (eds K. V. Mardia, C. A. Gill and R. G. Aykroyd), pp. 189–194. Leeds University Press.

Green, P.J. (1995) Reversible jump MCMC computation and Bayesian model determination. *Biometrika*, **82,** no. 4, 711–732.

Greitz, T., Bohm, C., Holte, S. and Eriksson, L. (1991) A computerized brain atlas: construction, anatomical content, and some applications. *Journal of Computer Assisted Tomography*, **15,** no. 1, 26–38.

Grenander, U. and Miller, M. I. (1994) Representations of knowledge in complex systems. *Journal of the Royal Statistical Society: Series* B, **56,** no. 4, 549–603.

ter Haar Romeny, B. M., Florack, L. M. J., Koenderink, J. J. and Viergever, M. A. (1991) Scale-space: its natural operators and differential invariants. In *Lecture Notes in Computer Science, Berlin, Germany*, vol. 511, pp. 239–255. Springer-Verlag.

Higdon, D. M., Bowsher, J. E., Johnson, V. E., Turkington, T. G., Gilland, D. R. and Jaszczak, R. J. (1997) Fully Bayesian estimation of Gibbs hyperparameters

for emission computed tomography data. Technical Report. Institute of Statistics and Decision Sciences, Duke University.

Hoffman, E. H., Cutler, P. D., Digby, W. M. and Mazziotta, J.C. (1990) 3-D phantom to simulate cerebral blood flow and metabolic images for PET. *IEEE Transactions on Nuclear Science*, **37,** 616–620.

Hurn, M. and Rue, H (1997) High-level image priors in confocal microscopy applications. In *The Art and Science of Bayesian Image Analysis* (eds K. V. Mardia, C. A. Gill and R. G. Aykroyd), pp. 36–43. Leeds University Press.

Johnson, V. E. (1994) A model for segmentation and analysis of noisy images. *Journal of the American Statistical Association*, **89,** no. 425, 230–241.

Kass, M., Witkin, A. and Terzopoulos, D. (1988) Snakes: axtive contour models. *International Journal of Computer Vision*, 321–331.

Koenderink, J. J. (1984) The structure of images. *Biological Cybernetics*, **50,** 363–370.

Laading, J. K., McCulloch, C. C. and Johnson, V. E. (1997) A hierarchical object deformation model applied to the digital chest radiograph. In *The American Statistical Association Proceedings of the Section on Bayesian Statistical Science, Anaheim, California.*

Le Goualher, G. and Barillot, C. Three-dimensional segmentation and representation of cortical sulci using active ribbons. In *International Journal of Pattern Recognition and Artificial Intelligence* (in press).

Lele, S. and Cole, T. M. (1996) A new test for shape differences when variance-covariance matrices are unequal. *Journal of Human Evolution*, **31,** 193–212.

Lifshitz, L. and Pizer, S. (1990) A multiresolution hierarchical approach to image segmentation based on intensity extrema. *IEEE Trans. PAMI*, **12,** no. 6, 529–640.

Lindeberg, T. (1990) Scale-space for discrete signals. *IEEE Transactions on Pattern Analysis and Machince Intelligence*, **12,** 234–254.

Lindeberg, T. (1994) *Scale-space Theory.* Boston, MA: Kluwer Academic Publishers.

Mardia, K. V. and Dryden, I. L. (1989) Shape distributions for landmark data. *Advances in Applied Probability*, **21,** 742–755.

Mardia, K. V., McCulloch, C. C., Dryden, I. L. and Johnson, V. E. (1997) Automatic scale-space method of landmark detection. Technical Report. University of Leeds.

Mazziotta, J., Toga, A., Evans, A., Fox, P. and Lancaster, J. (1995) A Probalistic atlas of the human brain: theory and rationale for its development. *Neuroimage*, **2,** 89–101.

McCulloch, C. C., Laading, J. K. and Johnson, V. E. (1997) Image feature identification via Bayesian hierarchical models. In *The American Statistical Association Proceedings of the Section on Bayesian Statistical Science, Anaheim, California.*

McCulloch, C. C., Laading, J. K., Wilson, A. G. and Johnson, V. E. (1996) A shape-based framework for automated image segmentation. In *The American Statistical Association Proceedings of the Section on Bayesian Statistical Science, Chicago, Illinois*, pp. 1–6.

McEachen, J. C. and Duncan, J. S. (1997) Shape-based tracking of left ventricular wall motion. *IEEE Transactions on Medical Imaging*, **16,** no. 3, 270–283.

McInerney, T. and Terzopoulos, D. (1996) Deformable models in medical image analysis: a survey. *Medical Image Analysis*, **1,** no. 2, 1–25.

Nelder, J. and Mead, R. (1965) A simplex method for function minimization. *The Computer Journal*, **7,** 308–313.

Opie, L.H. (1991) (ed.) *The Heart: Physiology and Metabolism*, 2 edn, ch. 13. Raven Press.

Park, J., Metaxas, D., Young, A. and Axel, L. (1996) Deformable models with parameter functions for cardiac motion analysis from tagged MRI data. *IEEE Transactions on Medical Imaging*, **15,** no. 3.

Phillips, D. B. and Smith, A. F. M. (1994) Bayesian Faces via Hierarchical Template Modeling. *Journal of the American Statistical Association*, **89,** no. 428, 1151–1163.

Ripley, B. (1991) The use of spatial models as image priors. In *Spatial Statistics and Imaging* (ed. A. Possolo), vol. 20, pp. 309–340. IMS Lecture Notes, Monograph Series, Institute of Mathematical Statistics.

Sandor, S. and Leahy, R. (1997) Surface-based labeling of cortical anatomy using a deformable atlas. *IEEE Transactions on Medical Imaging*, **16,** no. 1, 41–54.

Shaw, J. (1988) A quasirandom approach to integration in Bayesian statistics. *The Annals of Statistics*, **16,** no. 2, 895–913.

Sibson, R. (1978) Studies in the robustness of multidimensional scaling: Procrustes statistics. *Journal of the Royal Statistical Society*, **40,** no. 2, 234–238.

Stoyan, D. (1994) *Fractals, random shapes, and point fields : methods of geometrical statistics.* Wiley and Sons.

Studholme, C., Hill, D. L. G. and Hawkes, D. J. (1997) Automated three-dimensional registration of magnetic resonance and positron emission tomography brain images by multiresolution optimization of voxel similarity measures. *Medical Physics*, **24,** no. 1.

Talairach, J. and Tournoux, P. (1988) *Co-planar stereotaxic atlas of the human brain.* Thieme.

Terrell, G. (1990) The Maximal Smoothing Principle in Density Estimation. *Journal of the American Statistical Association*, **85,** no. 410, 470–477.

Thompson, P. and Toga, A. (1996) A surface-based technique for warping three-dimensional images of the brain. *IEEE Transactions on Medical Imaging*, **15,** no. 4, 402–417.

Titterington, D. M. (1997) Some aspects of Bayesian image analysis. In *The Art and Science of Bayesian Image Analysis* (eds K. V. Mardia, C. A. Gill and R. G. Aykroyd), pp. 153–160. Leeds University Press.

Wilson, A. G (1995) Statistical methods for shapes and deformations. Ph.D. Thesis. Duke University.

Woods, R. P., Cherry, S. R. and Mazziotta, J.C. (1992) Rapid automated algorithm for aligning and reslicing PET images. *Journal of Computer Assisted Tomography*, **16,** no. 4, 620–633.

Yaglom, A. M. (1962) *An Introduction to the theory of Stationary Random Functions.* Englewood Cliffs, New Jersey: Prentice-Hall.

Biography

B.Sc.E. Engineering Physics, specialization Materials Science, Queen's University at Kingston, Ontario, Canada, 1994
M.Sc. Statistical Consulting, Queen's University at Kingston, Ontario, Canada, 1995
Ph.D., Duke University, 1998
Accepted Assistant Professor position, Johns Hopkins School of Public Health Department of Biostatistics, 1998