Bayesian Statistics in Computational Anatomy*

Christof Seiler
Department of Statistics
Stanford University

September 2016

Abstract

Computational anatomy is the science of anatomical shape examined by deforming a template organ into a subject organ. It compares and contrasts organ shapes to inspire personalized treatments or find group differences in case-control studies. Independently of the transformation model used, the task of finding deformations between organs is a statistical task concerned with estimating parameters. Recently it has become important to go beyond "best" estimates and quantify the variability of estimates. The variability is caused by noise in the image, model misspecification, or sampling variability in an observational study. Bayesian statistics provides a rigorous framework to build models that can quantify uncertainty. In this book chapter, we will review some of the basics of Bayesian statistics and related it to our own experience in applying Bayesian ideas in computational anatomy. We will divide the presentation into two parts. First, we formulate image registration using parametric Bayesian statistics and elaborate on some of the practical difficulties that we encountered in our own work. Second, we will give an example of nonparametric Bayesian statistics applied to clustering of deformation fields into parcels of contiguous voxels.

Keywords: Computational Anatomy, Image Registration, Bayesian Statistics, Bayesian Nonparametrics, Markov Chain Monte Carlo, Hamiltonian Monte Carlo, Gibbs Sampler

Contents

1 Introduction **2**

2 Parametric Bayesian Statistics **3**

2.1 Background 3

2.2 Example 5

2.3 Model 6

*Preprint of book chapter in Statistical Shape and Deformation Analysis: Methods, Implementations & Applications (G. Zheng, S. Li, and G. Szekely, eds.)

2.4 Markov Chain Monte Carlo 8
2.5 Hamiltonian Monte Carlo 9
2.6 Software Implementations 9
2.7 Convergence Diagnostics 9
2.8 Related Work 11

3 Nonparametric Bayesian Statistics **12**
3.1 Background 12
3.2 Example 13
3.3 Model 14
3.4 Gibbs Sampler 17
3.5 Software Implementations 17
3.6 Related Work 17

4 Conclusions and Open Problems **17**

1 Introduction

The goal of this book chapter is to show case two applications of Bayesian statistics we have successfully employed in our own research. The first application is on quantifying uncertainty in image registration using a parametric Bayesian model and an efficient sampling algorithm. The second application is on clustering of deformations into spatially contiguous regions using Bayesian nonparametrics.

Before introducing our statistical work, we would like to define what we understand by computational anatomy. The aim of computational anatomy is to describe the anatomy not by what it looks like but by how it deforms. Finding such deformations is usually referred to as image registration; throughout this text we will use computational anatomy and image registration interchangeably. Statistical estimation and analysis of deformations of a group of medical images can for instance be used for early diagnosis and prediction of disease. The emerging field of computational anatomy has many facets. It lays at the interface between medicine, geometry, computing and statistics. Ideally, mathematicians define a notion of shape and a deep theory that is rooted in geometry. The computing community takes these notions and theories and implements them on a computer. Statisticians then quantify the uncertainty when running these computer programs on real world data. Finally, medical doctors make treatment decisions based on the calculated statistics.

Going back in time, we can trace the origins of computational anatomy to Riemann and his "Habilitationsschrift" in 1854 where he linked manifolds to shape of solid figures. He conceived the analysis of shapes of a solid figure as an important part of geometry. In 1917, the biologist-mathematician D'Arcy Thompson wrote an entire chapter on the analysis of shapes using deformations in his book on "Growth and form" (Thompson,

1942). His key idea was to study the shapes not by what they look like, but by how they deform relative to each other. He showed how one could relate various species, different kinds of fish, monkeys and humans to each other by stretching, scaling, and, by some more complicated deformations like conformal maps. In 1966, Arnold wrote his groundbreaking paper on employing modern geometry to describe incompressible fluids (Arnold, 1966). From the nineties until today, Grenander, Miller, Trouvé, Younes, Holm (Grenander and Miller, 1998; Trouvé and Younes, 2011; Younes, 2010; Holm et al., 2009), and many others, modeled the human anatomy using fluid dynamics and build on Arnold's ideas to establish the foundations of computational anatomy.

Today, computational anatomy is surrounded by a vibrant community and several workshops during the MICCAI conference have been held over the past years on its mathematical foundations[1]. In 2015, an entire research program on the theoretical foundations of computational anatomy and its applications was organized at the University of Vienna[2].

2 Parametric Bayesian Statistics

Image registration is one of the major work horses of medical image analysis. One of the most important applications of image registration is to "normalize" brains to a common brain atlas. This is a crucial "preprocessing" step in most multimodal brain studies: For instance, in structural MRI studies, image registration is used to measure local brain morphology; or in functional and diffusion MRI studies, image registration is used to bring all subjects in the same anatomical space for comparison. Other imaging based communities have extended and adapted registration based method for their own application. For instance, orthopeadic research uses image registration to estimate implant designs and evaluate fracture risk of bones.

Image registration builds on assumptions and approximations. From a statistical viewpoint, image registration is the task of estimating the parameters that define non-linear deformations. In the first part of this book chapter, we review how estimation can be done using Bayesian statistics.

2.1 Background

Before going into the modeling and computational details of how Bayesian statistics can be used in image registration, we review the general concepts of Bayesian statistics. We start with a model

$$M = \{f(y \mid \theta) : \theta \in \Omega\}$$

and refer to function f as the likelihood and the variable θ as parameters. The parameters can take values in simple one dimensional spaces or high dimensional spaces; they can be discrete or continuous. This model describes the data as random samples from the likelihood function given a fixed parameter

$$Y_1, \ldots, Y_n \sim f(y \mid \theta).$$

[1] `http://www-sop.inria.fr/asclepios/events/MFCA13/`

[2] `http://www.mat.univie.ac.at/~shape2015/`

A Frequentist would stop here, a Bayesian needs to define a prior distribution on the parameters

$$\pi(\theta)$$

to complete the model. We can think of the prior as a way to incorporate our beliefs from previous experiments into our analysis. Combining everything using Bayes rule will give us the posterior distribution of the parameters after having observed data and given our prior beliefs

$$\pi(\theta|y) = \frac{\prod_{i=1}^{n} f(y_i|\theta)\pi(\theta)}{m(y)}.$$

The denominator $m(y)$ is called the marginal distribution,

$$m(y) = m(y_1, \ldots, y_n) = \int f(y_1, \ldots, y_n|\theta)\pi(\theta)d\theta = \int \prod_{i=1}^{n} f(y_i|\theta)\pi(\theta)d\theta,$$

and can be seen as the measurement of how well our model explains the data on average over all possible parameter values weighted by their prior probability. With the posterior distribution in hand, we can now compute posterior means

$$\bar{\theta} = \mathrm{E}(\theta \mid y) = \int \theta\, \pi(\theta \mid y)\, d\theta$$

or 95% credible intervals between

$$0.025 = \int_{-\infty}^{\theta_l} \theta\, \pi(\theta \mid y)\, d\theta \quad \text{and} \quad 0.975 = \int_{\theta_h}^{\infty} \theta\, \pi(\theta \mid y)\, d\theta,$$

where 95% of the posterior mass is between θ_l and θ_h. In most applications, we will not be able to calculate these integrals analytically but we will have to resort to computational approximations. Variational inference and Markov Chain Monte Carlo (MCMC) are the main computational approximations employed in practice. Even when $m(y)$ is unknown, we can use MCMC to draw samples from the posterior. Once can think of sampling as an alternative way of describing a distribution. If one were able to sample infinitely many times from a distribution, one would know everything about the distribution. However, in practice we will resort to finite sample approximations, and thus we will know the distribution up to approximation errors.

Sampling from high dimensional probability distributions is a crucial step towards the Bayesian treatment of computational anatomy. The goal is to setup a statistical model of computational anatomy, this includes constructing a prior for deformations and linking it through the likelihood to fixed and moving images. To infer the deformations parameters, we then sample many times from the posterior distribution that is a combination of the likelihood and prior distribution. We can then build uncertainty estimates form these samples. In computational anatomy, we need efficient samplers that work well in high dimensions. One promising candidate is the Hamiltonian Monte Carlo (HMC) method. It is a promising candidate for two reasons: it is efficient in high dimensions, and it provides a geometric structures very similar to the one encountered in computational anatomy.

Due to the computational complexity the focus has been mostly on finding efficient algorithms and implementations to obtain point estimates by optimizing an objective functions and finding its maximum value. However, computational anatomy is clearly a

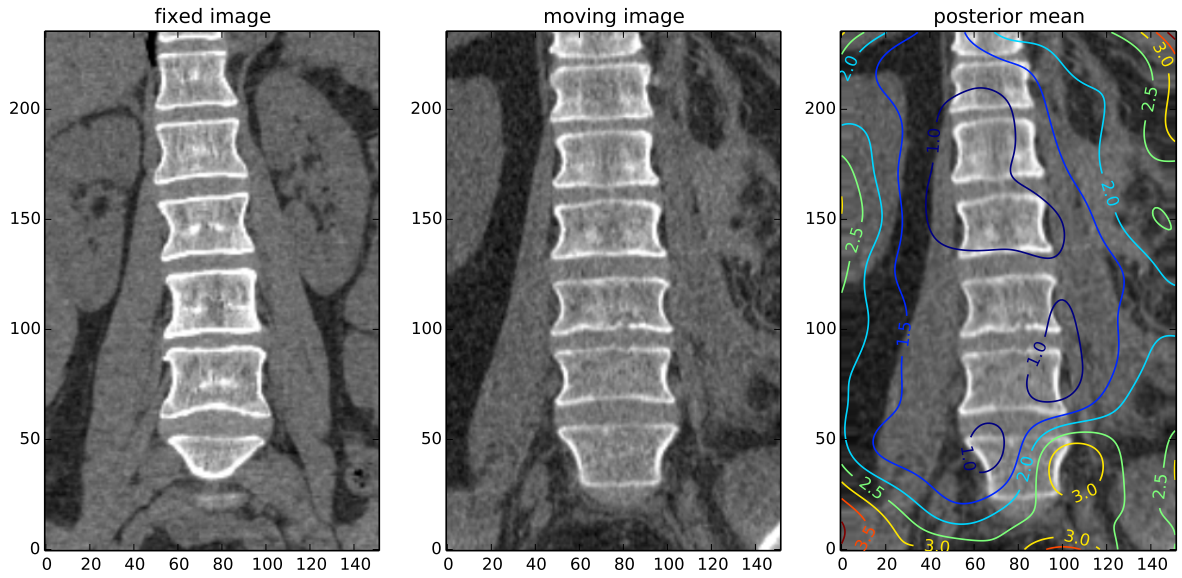

Figure 1: The contour plot (most right image) shows credible interval length of the displacement magnitude in millimeter when deforming the moving image (middle image) to the fixed image (most left image).

statistical problem considering that images are noisy and registered brains between subjects and template never match perfectly. With recent advances in computational power and Bayesian computations it has now become possible to add error bars to solutions. This provides the practitioner with additional information of the uncertainty associated to registration results. This is crucial in a clinical setting where it is important to obtain reproducible results.

2.2 Example

We start with an example from an ongoing back pain project. Figure 1 shows the uncertainty map when registering two participants in our back pain dataset. On the left, the *posterior mean* shows the expected deformation, overlaid with a contour map that shows 95% credible interval lengths derived from the posterior distribution of deformation parameters. We see that lumbar spine vertebrae L1 to L4 (L4 is the fifth vertebra from the top) deformations have an uncertainty around 1 millimeter. In contrast, the lumbar vertebra L5 deformation has an uncertainty up to 3 millimeter. This may indicate that the registration failed for L5. In three dimensional examples similar problems can occur due to anatomical abnormalities.

This illustrates the uncertainty estimation on the subject level between one fixed and one moving image. We generated this uncertainty map by sampling form a posterior distribution of deformations using Hamiltonian Monte Carlo (HMC). We will now carefully develop the model underlying this example and introduce the HMC sampler, which is a member of the Markov Chain Monte Carlo family.

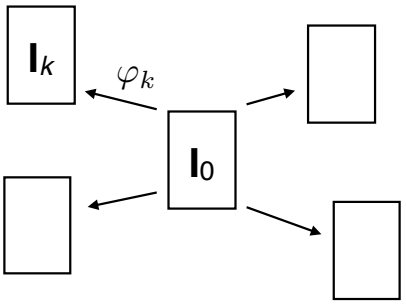

Figure 2: An template image $\mathbf{I_0}$ represents a typical anatomy. We model individual patients images $\mathbf{I_k}$ as non-linear deformations φ_k from the template. We will estimate the deformations from moving to fixed images with the goal of mapping all images into the same coordinate system, thus the inverse notation.

2.3 Model

We analyze geometric differences of the spine anatomy through geometric deformations. Our model describes patient images $\mathbf{I_k}$ as deformations φ_k from a common template image $\mathbf{I_0}$ (Figure 2). We model deformations with cubic multidimensional B-spline polynomial basis functions $\mathcal{B} : \mathbb{R}^3 \to \mathbb{R}^3$ (Unser, 1999). The model parameters are the weights associated to each basis function. If we pick C control points placed on a uniform grid over the template image $\mathbf{I_0}$, then we need to estimate a total of $q \in \mathbb{R}^{3C}$ weights in three dimensional volumetric CT images with voxels positions $x_i \in \mathbb{R}^3$. The estimated deformations ($\mathcal{N}_x$ is the support of the B-splines at spatial position x)

$$\varphi_k(x, q) = x + \sum_{x_i \in \mathcal{N}_x} q_i \, \mathcal{B}\left(x - x_i\right)$$

for each subject k can then be statistically analyzed.

There are two sources of uncertainty when estimating model parameters q from two images $\mathbf{I_0}$ and $\mathbf{I_k}$: image noise and model misspecification. The images are acquired with CT scanner which like other measurement devices introduce noise. The model misspecification is due to the crude assumption that the template can be deformed into a subject anatomy not accounting for shape difference that cannot be capture by such a simple model, e.g. missing part or entire vertebra in a patient. To capture these uncertainties, we assign a prior distribution to the parameters and sample form its posterior to compute posterior mean and posterior errors.

To control the spatial smoothness of the deformations we put a normal prior distribution

$$q \sim N(0, (\lambda_1 \Lambda)^{-1}) = \frac{1}{Z} \exp\left(\lambda_1 \sum_{i=1}^{N} \| \operatorname{Jac} \varphi_k(x_i, q)\|_{\ell_2}^2\right) = \frac{1}{Z}\left(-\lambda_1 \operatorname{Reg}(q)\right)$$

on the deformation parameters. The term $\operatorname{Reg}(q)$ can be seen as a form of regularization, penalizing large first order derivatives of the deformations; lower values represent more regular deformations. The block precision matrix Λ is defined by the element-wise partial derivatives of the $\mathcal{B}$-spline basis matrix ($\otimes$ denotes the Kronecker product)

$$\Lambda = I_3 \otimes \sum_{i=1}^{3} \left(\frac{\partial \mathcal{B}}{\partial x^i}^{\mathsf{T}} \frac{\partial \mathcal{B}}{\partial x^i}\right)$$

and the Jac is the Jacobian matrix. For more details see Andersson et al. (2007).

We then make the link to the imaging data through the likelihood term

$$(\mathbf{I_k} \mid q) \sim \frac{1}{Z}\left(-\sum_{i=1}^{N}(\mathbf{I}_k \circ \varphi_k(x_i, q) - \mathbf{I_0}(x_i))^2\right) = \frac{1}{Z}\exp\left(-\operatorname{Dist}(\mathbf{I_0}, \mathbf{I_k}, q)\right)$$

measuring the dissimilarity $\operatorname{Dist}(\mathbf{I_0}, \mathbf{I_k}, q)$ between template image $\mathbf{I_0}$ and subject image $\mathbf{I_k}$ after applying the deformation φ_k; lower values represent a better match. Combining the prior and likelihood term into the posterior distribution yields

$$(q \mid \mathbf{I_k}) \sim \frac{1}{Z}\exp\left(-\operatorname{Dist}(\mathbf{I_0}, \mathbf{I_k}, q) - \lambda_1 \operatorname{Reg}(q)\right),$$

which completes the Bayesian model for geometric deformations.

To make inference about deformation parameters, we can now compute different functionals of the posterior distribution. For example, the posterior mean

$$\theta = \mathbb{E}(q \mid \mathbf{I_k}) = \int q\, \pi(q \mid \mathbf{I_k})\, dq.$$

This $\mathbb{R}^{3C}$ dimensional integral is intractable analytically and to solve it one must resort to numerical methods. The first idea of evaluating the posterior at evenly distributed grid points is infeasible due to the large number of grid points. A clever alternative are Markov Chain Monte Carlo (MCMC) sampling algorithms. The main idea for sample-based estimators of integrals is to use the sample mean estimator

$$\widehat{\theta} = \frac{1}{T}\sum_{t=1}^{T} q_t$$

with draws from the posterior distribution

$$q_1, \ldots, q_T \sim (q \mid \mathbf{I_k}).$$

A wide range of MCMC samplers exists. In the next section, we will introduce the basic principle and elaborate on the more advanced Hamiltonian Monte Carlo.

In addition to the sample mean, we can also compute credible intervals from samples by calculating sample quantiles. For Figure 1, we computed element-wise quantiles of the parameter vector

$$q = \left(q^1, \ldots, q^{3C}\right)^{\mathsf{T}}$$

to obtain a 95% confidence interval

$$\hat{\theta}_h = \left(q^1_{(\alpha)}, \ldots, q^{3C}_{(\alpha)}\right)^{\mathsf{T}} \quad \text{and} \quad \hat{\theta}_h = \left(q^1_{(1-\alpha)}, \ldots, q^{3C}_{(1-\alpha)}\right)^{\mathsf{T}}$$

by ordering parameters from smallest to largest indicated by $_{(.)}$, where $_{(\alpha)}$ is the αth smallest value. We can then compute deformations by plug-in and compute the desired credible interval lengths as the difference between the 2.5% and 97.5% percentiles

$$|\varphi(x, \hat{\theta}_h) - \varphi(x, \hat{\theta}_l)|.$$

2.4 Markov Chain Monte Carlo

As the name suggests, MCMC is composed of a Markov chain and a Monte Carlo simulation. For lower dimensional integrals Monte Carlo simulations without Markov chains are possible. This works by drawing independent samples from $\pi(\theta \mid \mathbf{I_k})$

$$\hat{\theta} = \frac{1}{T}\sum_{i=1}^{T} q_i, \quad q_1, \ldots, q_T \overset{indep.}{\sim} \pi(q \mid \mathbf{I_k}).$$

However Monte Carlo simulations breaks down for problems of three or more dimensions. In this case, we resort to constructing a Markov chain that generates dependent samples

$$\hat{\theta} = \frac{1}{T}\sum_{i=1}^{T} q_i, \quad q_1, \ldots, q_T \overset{dep.}{\sim} \pi(q \mid \mathbf{I_k}).$$

Metropolis algorithm is the simplest MCMC algorithm. Consider a finite sample space. Think of it as a state space, where each outcome corresponds to a state. Metropolis can sample from an unormalized probability $\pi(x)$ on finite state space $\mathcal{X}$. Define a Markov transition matrix $J(x, y)$ that assigns nonzero probabilities of moving from x to y and y to x. Metropolis changes $J(x, y)$ to a new matrix $K(x, y)$ that corresponds to a possibly unnormalized version of $\pi(x)$.

The algorithm contains the following steps:

- Pick initial point in sample space x_0
- Pick potential next move from $J(x, y)$ with $J(x, y) > 0$ and $J(y, x) > 0$
- Evaluate
$$A(x, y) = \frac{\pi(y)J(y, x)}{\pi(x)J(x, y)}$$
- If $A(x, y) \geq 1$ move to y
- If $A(x, y) < 1$ flip a coing with this success probability
  - and move to y if success
  - otherwise stay at x

We can write this in matrix form

$$K(x, y) = \begin{cases} J(x, y) & \text{if } x \neq y, A(x, y) > 1 \\ J(x, y)A(x, y) & \text{if } x \neq y, A(x, y) < 1 \\ J(x, y) + \sum_{z:A(x,z)<1} J(x, z)(1 - A(x, z)) & \text{if } x = y \end{cases}$$

Then we can use the Fundamental Theorem of Markov Chains to prove that

$$K^n(x, y) \to \pi(y) \quad \text{for each } x, y \in \mathcal{X}$$

or in other words, the matrix $K^n = K^1 K^2 \cdots K^n$ converges to a matrix

$$\pi K = \lambda\pi \quad \leftrightarrow \quad \pi K = \pi$$

with one left eigenvector π and one eigenvalue $\lambda = 1$.

To sample from π, apply J to the left of the current sample position x_t. This will give us the next sample x_{t+1}. A nice introduction to the subject of MCMC is Diaconis (2009).

2.5 Hamiltonian Monte Carlo

As the name suggest, HMC involves defining a Hamiltonian function $H : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ with a potential energy term $V(q)$ which is equal to the $-\log$ transformed posterior distribution and a kinetic energy term $K(p)$ which will have the quadratic form $\frac{1}{2}\, p^\intercal G p$. The sum of both terms

$$H(q, p) = V(q) + K(p)$$

is the Hamiltonian function. It can be shown that the following algorithms produces samples from the distribution of the random variable $(q \mid \mathbf{I_k})$:

- Fix a starting position q_0
- Draw p_0 from a Gaussian $N(0, G^{-1})$
- Solve the Hamiltonian system

  $$\frac{dq}{dt} = \frac{\partial H}{\partial p} \quad \text{and} \quad \frac{dp}{dt} = -\frac{\partial H}{\partial q}$$

  for a predefined amount of time and record end point q_1
- Repeat the previous steps $T_0 + T$ times yielding $q_1, \ldots, q_{T_0}, \ldots, q_T$
- Then $q_{T_0}, \ldots, q_T \sim (q \mid \mathbf{I_k})$ are samples from the posterior. Note that samples before T_0 will be discarded because they could be biased due to a bad starting point.

The MCMC samples are correlated and standard Monte Carlo errors estimates do not apply. It is common to use trace plots to choose the appropriate T, which we found to be $T = 500$ in our spine registration problem. We refer to Seiler et al. (2014); Holmes et al. (2014) for a theoretical investigation on this topic. The book chapter by Neal (2011) is an excellent source for more background material and illustrative toy examples on HMC.

2.6 Software Implementations

The main HMC sampling algorithm is easy to implements and contains only few steps. The tricky part is solving the Hamiltonian system. This is usually done using an Euler numerical scheme. However the fine tuning of parameters is not straight forward and software has been implemented to automize this step (Carpenter et al., 2016).

A specific implementation of HMC for medical images can be found on the GitHub repository of the first author.[3]

2.7 Convergence Diagnostics

Some statisticians[4] argue that diagnostics for MCMC only finds "obvious, gross, embarrassing problems that jump out of simple plots". One of the concerns has to do with

[3] Code available: `https://christofseiler.github.io/BayesianImageRegistration`

[4] `http://users.stat.umn.edu/~geyer/mcmc/diag.html`

bad starting points. Consider an event B having high probability under the equilibrium distribution, that is, the distribution that is unchanged if we ran the MCMC sampler for a long time. Suppose we are unlucky and start the sampler at a bad starting point and it would take a long time, say longer than the age of the universe to reach B, then the chance of diagnosing this problem will be highly improbable.

Keeping this limitation in mind, convergence diagnostics of MCMC samplers are the only way to quantify the quality of our sample. The following description builds heavily on (Robert and Casella, 2009, Chapter 8). The `R` package `coda` offers implementations for the most popular diagnostic tools.

We consider two types of convergence of MCMC methods:

1. Convergence to the stationary distribution: Check that distribution of the chain x_t is the stationary distribution f. In practice this is impossible to test with just one chain. Several chains have to be run to test this and it can actually never be tested exactly. What is tested is how independent the chains are at time t when started at different starting positions.

2. Convergence of averages: Once stationarity is established, we can focus on evaluating the Monte Carlo error. However, in contrast to the usual Monte Carlo error, we have additional problems to take into account: the samples are dependent across time. The stronger this dependence the slower we explore the posterior distribution.

To be more concrete, define an estimator

$$\hat{\theta} = \frac{1}{T} \sum_{t=1}^{T} h(x_t)$$

of the parameter

$$\theta = \mathrm{E}(h(X)).$$

The variance of this estimator $\mathrm{Var}(\hat{\theta}_{\mathrm{MC}})$ in case of identically and independent Monte Carlo samples is given by the central limit theorem. The variance of this estimator $\mathrm{Var}(\hat{\theta}_{\mathrm{MCMC}})$ in case of dependent Markov chain samples gets worse by a factor that depends on the amount of correlation between draws. This can be measured by the effective sample size

$$T_{\mathrm{MC}} = \frac{T_{\mathrm{MCMC}}}{\kappa(h)}$$

with the autocorrelation time

$$\kappa(h) = 1 + 2 \sum_{t=1}^{\infty} \mathrm{corr}\left(h(x_0), h(x_t)\right).$$

Intuitively, the larger $\kappa(h)$ the more dependent are draws and we need to increase the sample size to reach the Monte Carlo error. If draws are independent then we have $\kappa(h) = 1$ and the Monte Carlo and MCMC samples are equivalent. Diagnostics tools can now be build on this concept.

For more involved samplers, e.g. the HMC sampler, diagnostics becomes even more complicated. Recently, we have reformulated HMC in the language of Riemannian geometry

and applied theorems from Joulin and Ollivier (2010) on Markov chain curvature that inspired a new diagnostic tool for HMC (Seiler et al., 2014; Holmes et al., 2014). The idea of Markov chain curvature is similar to measuring the autocorrelation time of the Markov chain, in fact, the curvature is inverse proportional to the correlation between draws from the Markov chain. Intuitively, a random walk moves faster through low density regions of the space, and slower through high density regions.

2.8 Related Work

We will distinguish between two types of deformations: *small deformations* and *large deformations*. To keep this book chapter focused on uncertainty quantification we will not go into the difference between the two types expect to say that one can think of large deformations as compositions of small deformations.

First, we focus on the *small deformations* setting. Allassonnière et al. (2007, 2010) describe a maximum a posterior estimation (MAP) procedure based on an Expectation-Maximization (EM) algorithm to estimate the template from a set of subject images. Similarly Van Leemput (2009) estimated the MAP template using a combination of pseudo-likelihood technique (Besag, 1975), EM algorithm, and Laplace's method. In contrast, Risholm et al. (2010b,a) sample from the posterior distribution of pairwise registration parameters using Metropolis-Hastings which is part of the family of MCMC algorithms. In addition to MCMC, Risholm et al. (2013) marginalize over hyperparameters (modeling noise in the image intensities) using Laplace's method. Simpson et al. (2012) use mean-field variational Bayes to approximate the full posterior distribution. This is an optimization-based alternative to MCMC and much faster in practice. However it is unclear how accurate the posterior distribution is approximated with this method. In contrast, MCMC-type methods enjoy the property that they are sampling exactly from the true posterior distribution given that the sampler is run for long enough. Quantifying running times is however very tricky and usually only possible for simplified toy examples. In Simpson et al. (2015) an additional Laplace's method is used for hyperparameters analog to Risholm et al. (2013). Heinrich et al. (2016) propose an alternative approach using dynamic programming to find the MAP.

Now, we focus on the *large deformations* setting. Zhang et al. (2013) use Hamiltonian Monte Carlo (HMC) to sample diffeomorphic deformations. The HMC algorithm is part of a Monte Carlo EM algorithm to estimate an image template. This is quite computational expensive and Zhang and Fletcher (2015) provide a fast algebraic approximation that is useful in a sampling scheme. Wassermann et al. (2014) approximate the full Bayesian posterior distribution by a variational formulation. Their method can be viewed as a combination of the Laplace's method (describing stochastic differential equations by Gaussian processes) and variational Bayes (minimizing the Kullback-Leibler divergence). Yang and Niethammer (2015) propose a low rank approximation of the Hessian matrix at the mode of the posterior distribution.

3 Nonparametric Bayesian Statistics

3.1 Background

The following introduction to Bayesian Nonparametrics (BN) builds strongly on the lecture notes by Larry Wasserman[5]. A good reference on the theoretical background are the lecture notes by van der Vaart[6].

We replace the finite dimensional model from the previous section

$$\{f(y|\theta) : \theta \in \Theta\}$$

with an infinite dimensional model and constrain the second derivate of possible functions to be finite

$$\mathcal{F} = \left\{ f : \int (f''(y))^2 dy < \infty \right\}.$$

Other constrains are possible but for illustrative purposes we focus on this model. In order to translate parametric Bayesian ideas to the nonparametric setting, we need to address some questions. First, we will need to put a prior π on an infinite dimensional space. For example, suppose we observe

$$X_1, \ldots, X_n \sim F$$

with unknown distribution F with density f. We put prior π on set of all distributions $\mathcal{F}$. In many cases, we cannot explicitly write down a formula for π. This follows from technical arguments about the infinite dimensional set $\mathcal{F}$ that we will not cover in this short introduction. How can we describe a distribution π in another way than writing it down? If we know how to draw from π we can get many samples and then even without knowing the formula for π we can plot and summarize it in any way we want. The idea is to find an algorithm to sample from this model

$$\begin{aligned} F &\sim \pi \\ X_1, \ldots, X_n \mid F &\sim F. \end{aligned}$$

The usual frequentist estimate of F is the empirical distribution function

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x).$$

To estimate F from a Bayesian perspective we put a prior on π on the set of all $\mathcal{F}$. Such a prior was invented by Ferguson (1973). The prior has two parameter: F_0 and α denoted by $\mathrm{DP}(\alpha, F_0)$. F_0 is a distribution function and should be thought of as a prior guess of F. The number α controls how tightly concentrated the prior is around F_0. The model is

$$\begin{aligned} F &\sim \mathrm{DP}(\alpha, F_0) \\ X_1, \ldots, X_n \mid F &\sim F. \end{aligned}$$

But how to draw samples from this model? First to draw samples from the prior $\mathrm{DP}(\alpha, F_0)$, we follow four steps:

[5]Lecture notes: `http://www.stat.cmu.edu/~larry/=sml/nonparbayes.pdf`

[6]Lecture notes: `http://www.math.leidenuniv.nl/~avdvaart/BNP/`

1. Draw $s_1, s_2, \ldots$ independently from F_0
2. Draw $V_1, V_2, \cdots \sim \text{Beta}(1, \alpha)$
3. Stick breaking process: Let $w_1 = V_1$ and $w_j = V_j \prod_{i=1}^{j-1}(1 - V_i)$ for $j = 2, 3, \ldots$
   - Imagine a stick of unit length
   - Then w_1 is obtained by breaking the stick at the random point V_1
   - The stick has now length $1 - V_1$
   - The second weight w_2 is obtained by breaking a proportion V_2 from the remaining stick
   - The process continues and generates the whole sequence of weights $w_1, w_2, \ldots$
4. Let F be the discrete distribution that puts mass w_j at s_j, that is, $F = \sum_{j=1}^{\infty} w_j \delta_{s_j}$ where δ_{s_j} is a point mass at s_j

After we observe the data $X = (X_1, \ldots, X_n)$, we are interested in the posterior distribution. The same idea applies here, instead of writing down a formula we describe an algorithm to sample for the posterior distribution. To sample from the posterior, we need the following theorem. Let F_n be the empirical distribution.

Theorem 3.1 (Ferguson (1973))**.** *Let $X_1, \ldots, X_n \sim F$. Let F have prior $\pi = \text{DP}(\alpha, F_0)$. Then the posterior π for F given $X_1, \ldots, X_n$ is $\text{DP}(\alpha + n, \bar{F}_n)$ where*

$$\bar{F}_n = \frac{n}{n+\alpha} F_n + \frac{\alpha}{n+\alpha} F_0.$$

Since the posterior is again a Dirichlet process, we can sample from it the same way as we did for the prior. We only replace α with $\alpha + n$ and F_0 with $\bar{F}_n$. Thus the posterior $\bar{F}_n$ is a convex combination of the empirical distribution F_n and the prior guess F_0. To explore the posterior distribution, we could draw many random distribution functions form the posterior. We could then numerically construct two functions L_n and U_n such that

$$\pi(L_n(x) \leq F(x) \leq U_n(x) \text{ for all } x \mid X_1, \ldots, X_n) = 1 - \alpha$$

This is a Bayesian credible interval for F. When n is large then $\bar{F}_n \approx F_n$.

3.2 Example

Unfortunately, most people will be affect by lumber back pain (LBP) during the coarse of their lives. We classify the pain into acute and chronicle pain. The acute pain is most commonly caused by muscle strain or ligament sprain. The chronicle pain is most commonly cause by a disc tear, facet joint disorder, or sacroiliac joint disfunction.

In addition to the individual suffering of every patient, the society as a whole pays a big price for the treatment of LBP. For instance, the direct costs of LBP to be at 2.6 billion Euro, 6.1% of the total healthcare expenditure in Switzerland Wieser et al. (2011), which results in a total economic burden between 1.6 and 2.3% of Swiss GDP.

Despite the enormous burden on individual patients and the society, the geometric variability of deformations of the spine are still unexplored. For example, it has not been

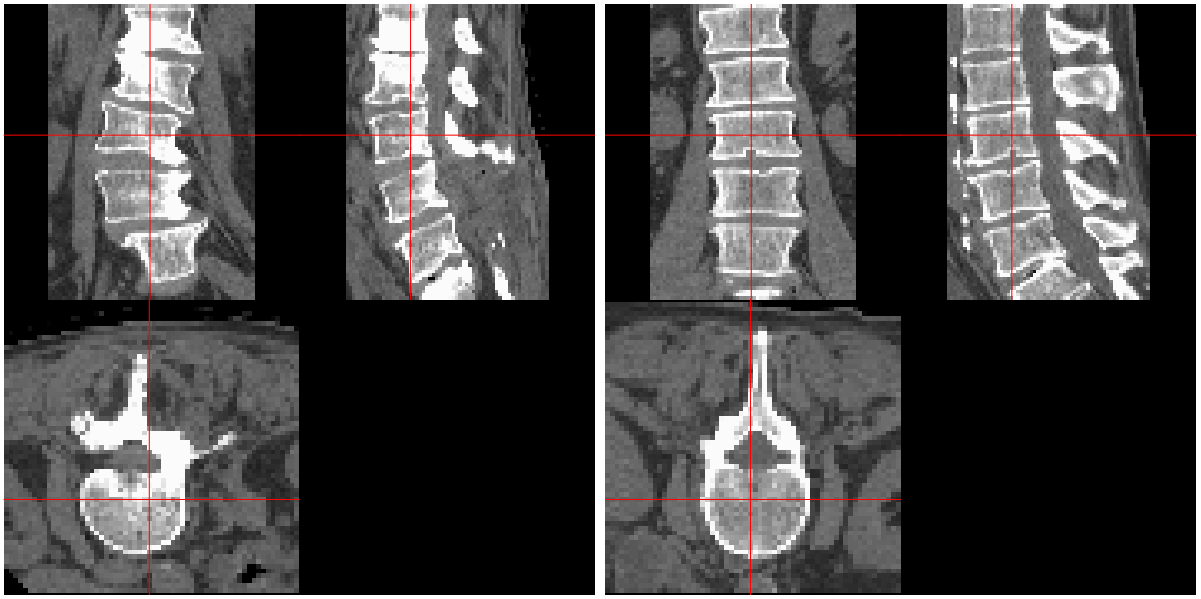

Figure 3: Left: Lumbar back pain patient with scoliosis. Right: Abdominal pain patient.

reported whether scoliosis patients (sidewise curvature of the spine) suffer more often from LBP than normal patients (Figure 3).

To explore this issue, we will investigate geometric differences between LBP and abdominal patients. We expect to see regional differences between the two patient groups. We propose to estimate regions using a BN clustering method that allows to incorporate geometric prior information. This clustering algorithm will find spatially contiguous voxel clusters (Figure 4) without knowing in advance the number of clusters.

3.3 Model

Deformation fields can be found by registration methods as described in the first part of this book chapter. Additionally in this part we will assume that deformations are diffeomorphic, which means that they are differentiable and their inverses are differentiable. The input to our clustering algorithm are deformation fields encoded as Stationary Velocity Fields (SVF), which can be obtain through various registration algorithms (Ashburner, 2007; Hernandez et al., 2007; Vercauteren et al., 2009; Lorenzi et al., 2013). The SVF v is the unique solutions to the Ordinary Differential Equation (ODE) $\partial\phi(x,t)/\partial t = v(\varphi(x,t))$ with initial condition $\varphi(x,0) = \text{identity}$. The reason that ODE's are useful for image registration is that we can generate a diffeomorphic mapping of a patient image $\mathbf{I_k}$ to a template image $\mathbf{I_0}$ with $\mathbf{I_0}(x) = \mathbf{I_k}(\varphi_k(x))$, spatial position $x \in \mathbb{R}^3$, intensity image $\mathbf{I} : \mathbb{R}^3 \mapsto \mathbb{R}$, and diffeomorphic mapping $\varphi : \mathbb{R}^3 \mapsto \mathbb{R}^3$. This assumption makes sense for spines in the absence of fractures and collapse of tissue.

We model the observed velocity fields as a linear combination of linear transformations

$$v(x) = \sum_{i=1}^{p} w_i(x) \begin{bmatrix} L_i & q_i \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ 1 \end{bmatrix} + \varepsilon(x)$$

with an affine part L_i, translational part q_i, and additive independent and identically

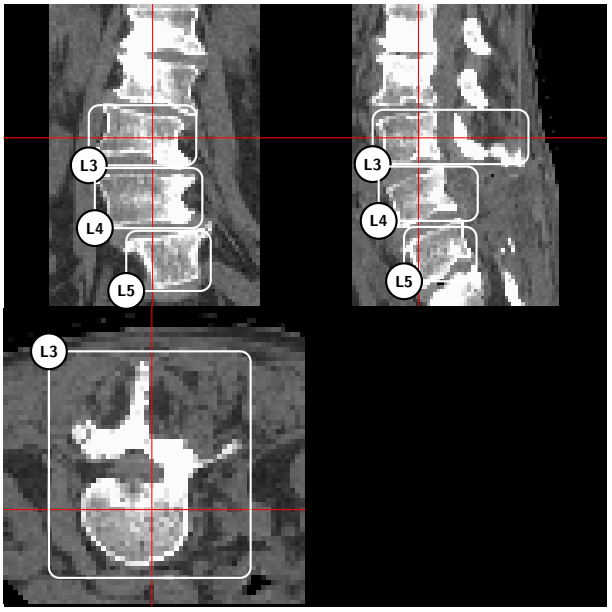

Figure 4: Parcellations in computational anatomy.

distributed voxelwise Gaussian noise $\varepsilon(x)$. Our goal is to infer both the number of parcels p and shape $w_i(x)$ with the assumption that $w_i(x)$ are non-overlapping binary weight images.

We formulate this in terms of a BN model by vectorizing both the image matrix and the linear transformations, and by introducing the binary matrix $\mathbf{W}$ that assigns n voxels to p parcels

$$\begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = \mathbf{W} \begin{bmatrix} \text{Vectorize}(L_1) \\ q_1 \\ \vdots \\ \text{Vectorize}(L_p) \\ q_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Each column represents one weight image $w_i(x)$. The columns can grow in size as more more observations become available (higher resolution images). The parameters of interest in this BN models are $\mathbf{W}$, L_i, and q_i. The BN part of this model is the matrix $\mathbf{W}$ because it is not fixed in column size.

Creating parcels in an image is similar to clustering voxels. We can formulate clustering as a density estimation problem by using a mixture model to approximate densities. Each component of the mixture model defines a cluster. Positions that are close to the mode of one component are assigned the same label. In BN we can perform density estimation using an extension of the Dirichlet process prior to the Dirichlet process mixture model. We now give a short introduction to density estimation in BN.

Consider that we observe $X_1, \ldots, X_n \sim F$ from a distribution F with density f. Without loss of generality we can assume that $X_i \in \mathbb{R}$. Our goal is to estimate the unknown density function f. The Dirichlet process is not an appropriate prior for this problem because it produces discrete distributions and densities are continuous. An intuitive way to construct the nonparametric estimation procedure is by starting with a parametric model and letting the number of parameters go to infinity. Consider the Gaussian mixture

model

$$f(x) = \sum_{j=1}^{k} w_j f(x; \theta_j),$$

where $f(x; \theta_j)$ is normal and each component is parametrized with its mean and variance $\theta_j = (\mu_j, \sigma_j^2)$. In this model, we would have to estimate the number of components k, weights w_j, and parameters μ_j, σ_j^2. In the Bayesian approach, we have to put priors on k, w_j, and μ_j, σ_j. One option is to separate the estimation task into two parts by comparing the quality of a fixed set of models $k = 1, \ldots, K$. Recently, it became popular to use an infinite mixture model

$$f(x) = \sum_{j=1}^{\infty} w_j f(x; \theta_j),$$

which trades the finite model comparison problem into a more continuous problem with possible infinitely many components k. Nevertheless, as we will see, we still need to pick a parameters that controls the number of components k indirectly. As a prior for the parameters we could take $\theta_1, \theta_2, \ldots$ to be drawn from some F_0 and we could take $w_1, w_2, \ldots$ to be drawn from the stick breaking prior. This is known as the Dirichlet process mixture model and is an extension of the Dirichlet process prior $F \sim \mathrm{DP}(\alpha, F_0)$ with the difference that we replace the point mass distribution δ_{θ_j} in the original form $F = \sum_{j=1}^{\infty} w_j \delta_{\theta_j}$ by smooth densities $f(x; \theta_j)$. Combining everything, the model is

$$\begin{aligned} F &\sim \mathrm{DP}(\alpha, F_0) \\ \theta_1, \ldots, \theta_n \mid F &\sim F \\ X_j \mid \theta_j &\sim f(x; \theta_j), \quad j = 1, \ldots, n. \end{aligned}$$

It is important to note that the discreteness of F automatically creates a clustering of the parameters θ_j's. This can be considered an implicit prior for the number of components k. We can control the number of k indirectly by choosing an appropriate concentration parameter α. However, there is no free lunch and choosing α usually involves additional priors (Escobar and West, 1995, 1998).

To complete the model, we also define Gaussian priors on transformation parameters L_i and q_i. We decompose locally linear transformations

$$A_i x + b_i = \exp\left(\begin{bmatrix} L_i & q_i \\ 0 & 0 \end{bmatrix}\right) \begin{bmatrix} x \\ 1 \end{bmatrix}$$

using the Jordan/Schur decomposition

$$\begin{aligned} L_i &= \frac{1}{2}(L_i - L_i^\mathsf{T}) + \frac{1}{2}(L_i + L_i^\mathsf{T}) \\ L_i &= \text{rotation} + \text{scaling} = \theta \begin{bmatrix} 0 & -r_3 & r_2 \\ r_3 & 0 & -r_1 \\ -r_2 & r_1 & 0 \end{bmatrix} + \mathrm{diag}(s) \end{aligned}$$

to obtain a rotation axis $\begin{bmatrix} r_1 & r_2 & r_3 \end{bmatrix}^\mathsf{T}$ and a rotation angle θ. This rotation axis and rotation angle are better interpretable than a transformation matrix and a translation vector and allow us to define subjective priors. For instance, we may want to define a prior on the angle to favors deformations centered at 0° with standard deviation 30°.

3.4 Gibbs Sampler

If we are only interested in cluster assignments $\mathbf{W}$ and ignore the transformation parameter, we can integrate out L_i and q_i, and sample from the remaining integral using the distant dependent Chinese Restaurant Process (Blei and Frazier, 2011). This process is a Gibbs samplers. Gibbs samplers are convenient whenever we wish to sample from a posterior that can be decomposed into conditional distributions for which fast ways of sampling are available. For instance, to draw sample from this joint distribution

$$\theta_1, \ldots, \theta_T \sim \pi(\theta^1, \theta^2 \mid y)$$

we can iterate between their respective conditional distributions,

- Step 1: $\theta_i^1 \sim \pi(\theta_1 \mid y, \theta_{i-1}^2)$
- Step 2: $\theta_i^2 \sim \pi(\theta_2 \mid y, \theta_i^1)$

and repeating it many times to obtain samples from the joint distribution

$$\left(\theta_1^1, \theta_1^2\right), \ldots, \left(\theta_T^1, \theta_T^2\right) \sim \pi(\theta^1, \theta^2 \mid y).$$

We use the distant dependent Chinese Restaurant Process to draw samples form the marginal distribution for LBP versus abdominal pain dataset as illustrated in Figure 5. Details on the technical implementation can be found in our recent conference article (Seiler et al., 2013).

3.5 Software Implementations

An implementation of a variety of BN tools is available in the `R` package `DPpackage`. Our specific implementation for medical images can be found on the GitHub repository of the first author.[7]

3.6 Related Work

The usage of BN in computational anatomy is still in its infancy. Related work in the medical context are the detection of spatial activation patterns in fMRI (Kim et al., 2006) or tractography segmentation (Wang et al., 2011) using the Dirichlet processes.

4 Conclusions and Open Problems

In this book chapter, we reviewed general concepts in Bayesian statistics and reported our experience with applying them to problems in computational anatomy. In the parametric part, our treatment focused on the *small deformation* framework. The translation of our work to *large deformation* framework, especially the translation of diagnostics tools for MCMC is currently open. A successful treatment of diagnostics for *large deformation*

[7]Code available: `https://github.com/ChristofSeiler/BayesianNonparametrics`

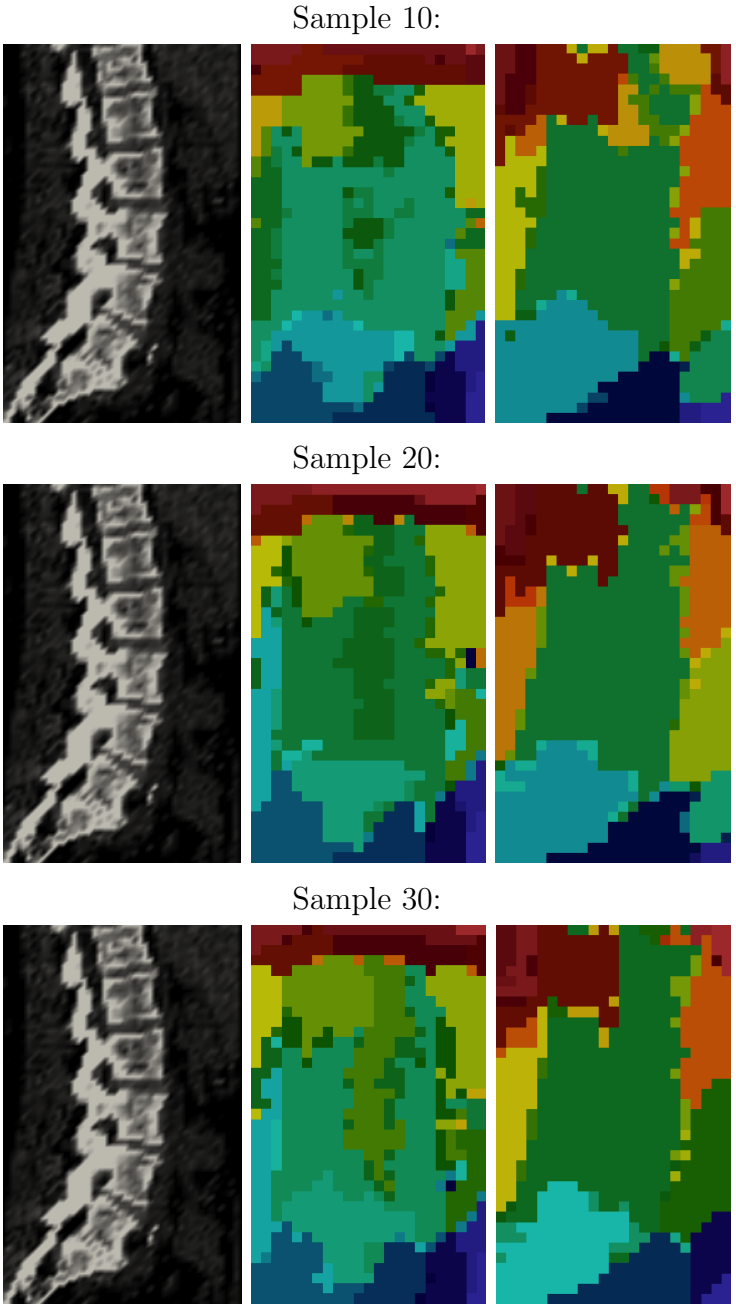

Figure 5: Colors are clusters. Left row: Template spine image. Middle row: Back pain patients. Right row: Abdominal pain patients.

will most likely require an even stronger interplay between geometry and probability. As reported in the nonparametric part, we have found only sparse literature on applying BP ideas to computational anatomy problems.

Besides the theoretical developments, it will be paramount to provide the community with efficient software implementations in the form of `R` packages for reproducible research. The recent growing community around the `STAN` software (Carpenter et al., 2016) implementing HMC will hopefully facilitate a more routine usage of Bayesian statistics in computational anatomy.

References

Allassonnière, S., Y. Amit, and A. Trouvé
2007. Towards a coherent statistical framework for dense deformable template estimation. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 69(1):3–29.

Allassonnière, S., E. Kuhn, and A. Trouvé
2010. Construction of Bayesian deformable models via a stochastic approximation algorithm: a convergence study. *Bernoulli*, 16(3):641–678.

Andersson, J. L. R., M. Jenkinson, and S. Smith
2007. Non-linear optimisation. Technical Report TR07JA1, FMRIB Analysis Group of the University of Oxford.

Arnold, V. I.
1966. Sur la géométrie différentielle des groupes de Lie de dimension infinie et ses applications à l'hydrodynamique des fluides parfaits. *Ann. Inst. Fourier (Grenoble)*, 16(fasc. 1):319–361.

Ashburner, J.
2007. A fast diffeomorphic image registration algorithm. *NeuroImage*, 38(1):95–113.

Besag, J.
1975. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 24(3):179–195.

Blei, D. M. and P. I. Frazier
2011. Distance dependent Chinese restaurant processes. *Journal of Machine Learning Research*, 12(Aug):2461–2488.

Carpenter, B., A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell
2016. Stan: A probabilistic programming language. *Journal of Statistical Software (in press)*.

Diaconis, P.
2009. The Markov chain Monte Carlo revolution. *Bull. Amer. Math. Soc. (N.S.)*, 46(2):179–205.

Escobar, M. D. and M. West
1995. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588.

Escobar, M. D. and M. West
1998. Computing nonparametric hierarchical models. In *Practical Nonparametric and Semiparametric Bayesian Statistics*, Pp. 1–22. Springer.

Ferguson, T. S.
1973. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, Pp. 209–230.

Grenander, U. and M. I. Miller
1998. Computational anatomy: An emerging discipline. *Quart. Appl. Math.*, 56(4):617–694. Current and future challenges in the applications of mathematics (Providence, RI, 1997).

Heinrich, M. P., I. J. Simpson, B. W. Papież, M. Brady, and J. A. Schnabel
2016. Deformable image registration by combining uncertainty estimates from supervoxel belief propagation. *Medical image analysis*, 27:57–71.

Hernandez, M., M. N. Bossa, and S. Olmos
2007. Registration of anatomical images using geodesic paths of diffeomorphisms parameterized with stationary vector fields. In *ICCV 2007*, Pp. 1–8. IEEE.

Holm, D. D., T. Schmah, and C. Stoica
2009. *Geometric mechanics and symmetry*, volume 12 of *Oxford Texts in Applied and Engineering Mathematics*. Oxford University Press, Oxford. From finite to infinite dimensions, With solutions to selected exercises by David C. P. Ellis.

Holmes, S., S. Rubinstein-Salzedo, and C. Seiler
2014. Curvature and concentration of Hamiltonian Monte Carlo in high dimensions. *Preprint arXiv:1407.1114*.

Joulin, A. and Y. Ollivier
2010. Curvature, concentration and error estimates for Markov chain Monte Carlo. *Ann. Probab.*, 38(6):2418–2442.

Kim, S., P. Smyth, and H. Stern
2006. A nonparametric Bayesian approach to detecting spatial activation patterns in fMRI data. In *MICCAI 2006, Part II*, LNCS, Pp. 217–224. Springer Heidelberg.

Lorenzi, M., N. Ayache, G. B. Frisoni, and X. Pennec
2013. LCC-demons: A robust and accurate diffeomorphic registration algorithm. *NeuroImage*.

Neal, R. M.
2011. MCMC using Hamiltonian dynamics. In *Handbook of Markov chain Monte Carlo*, Chapman & Hall/CRC Handb. Mod. Stat. Methods, Pp. 113–162. CRC Press, Boca Raton, FL.

Risholm, P., F. Janoos, I. Norton, A. J. Golby, and W. M. W. III
2013. Bayesian characterization of uncertainty in intra-subject non-rigid registration. *Med Image Anal*, 17(5):538–55.

Risholm, P., S. Pieper, E. Samset, and W. M. Wells III
2010a. Summarizing and visualizing uncertainty in non-rigid registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Pp. 554–561. Springer.

Risholm, P., E. Samset, and W. Wells, III
2010b. Bayesian estimation of deformation and elastic parameters in non-rigid registration. In *Biomedical Image Registration*, B. Fischer, B. M. Dawant, and C. Lorenz, eds., volume 6204 of *Lecture Notes in Computer Science*, Pp. 104–115. Springer Berlin Heidelberg.

Robert, C. and G. Casella
2009. *Introducing Monte Carlo Methods with R*. Springer Science & Business Media.

Seiler, C., X. Pennec, and S. Holmes
2013. Random spatial structure of geometric deformations and Bayesian nonparametrics. In *Geometric Science of Information*, volume 8085 of *LNCS*, Pp. 120–127. Springer.

Seiler, C., S. Rubinstein-Salzedo, and S. Holmes
2014. Positive curvature and Hamiltonian Monte Carlo. In *Advances in Neural Information Processing Systems (NIPS)*, Pp. 586–594.

Simpson, I., M. Cardoso, M. Modat, D. Cash, M. Woolrich, J. Andersson, J. Schnabel, S. Ourselin, A. D. N. Initiative, et al.
2015. Probabilistic non-linear registration with spatially adaptive regularisation. *Medical image analysis*, 26(1):203–216.

Simpson, I. J. A., J. A. Schnabel, A. R. Groves, J. L. R. Andersson, and M. W. Woolrich
2012. Probabilistic inference of regularisation in non-rigid registration. *NeuroImage*, 59(3):2438–2451.

Thompson, D. W.
1942. *On growth and form.* Cambridge Univ. Press.

Trouvé, A. and L. Younes
2011. Shape spaces. In *Handbook of Mathematical Methods in Imaging*, Pp. 1309–1362. Springer.

Unser, M.
1999. Splines: A perfect fit for signal and image processing. *Signal Processing Magazine, IEEE*, 16(6):22–38.

Van Leemput, K.
2009. Encoding probabilistic brain atlases using Bayesian inference. *IEEE Transactions on Medical Imaging*, 28(6):822–837.

Vercauteren, T., X. Pennec, A. Perchant, and N. Ayache
2009. Diffeomorphic demons: Efficient Non-Parametric image registration. *NeuroImage*, 45(1 Suppl):S61–S72.

Wang, X., E. E. Grimson, and C.-F. F. Westin
2011. Tractography segmentation using a hierarchical Dirichlet processes mixture model. *NeuroImage*, 54(1):290–302.

Wassermann, D., M. Toews, M. Niethammer, and W. Wells III
2014. Probabilistic diffeomorphic registration: Representing uncertainty. In *Biomedical Image Registration*, Pp. 72–82. Springer.

Wieser, S., B. Horisberger, S. Schmidhauser, C. Eisenring, U. Brügger, A. Ruckstuhl, J. Dietrich, A. F. Mannion, A. Elfering, O. Tamcan, and U. Müller
2011. Cost of low back pain in switzerland in 2005. *The European Journal of Health Economics*, 12(5):455–467.

Yang, X. and M. Niethammer
2015. Uncertainty quantification for LDDMM using a low-rank Hessian approximation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Pp. 289–296. Springer.

Younes, L.
2010. *Shapes and diffeomorphisms*, volume 171. Springer Science & Business Media.

Zhang, M. and P. T. Fletcher
2015. Finite-dimensional Lie algebras for fast diffeomorphic image registration. In *International Conference on Information Processing in Medical Imaging*, Pp. 249–260. Springer.

Zhang, M., N. Singh, and P. T. Fletcher
2013. Bayesian estimation of regularization and atlas building in diffeomorphic image registration. In *Information Processing in Medical Imaging (IPMI)*, LNCS, Pp. 37–48. Springer.