Bayesian Models for Combining Data Across Domains and Domain Types in Predictive fMRI Data Analysis (Thesis Proposal)

Indrayana Rustandi
Computer Science Department
Carnegie Mellon University

March 26, 2007

Abstract

In the context of predictive fMRI data analysis, in particular classification-based analysis, the analysis is usually done separately for a particular subject in a specific study, or by combining data across subjects and studies after normalization to a common template by assuming that there are no inter-subject and inter-study variations. These approaches are suboptimal for a number of reasons. Based on findings in the cognitive neuroscience field, there is a reason to believe that data from other subjects and from different but similar studies exhibit similar patterns of activations, implying that there is some potential for leveraging data from other subjects. However, each subject's brain might still exhibit some variations in the activations compared to other subjects' brains, based on factors such as differences in anatomy, experience, or environment. Furthermore, current normalization techniques might blur and distort the data, causing some loss of information.

I propose to investigate and develop principled Bayesian methods to tackle these problems, and by doing so, enable combining data across domains and domain types, of which subjects and studies are instances. One goal is to improve predictive performance when compared with predictive analysis of each domain and domain type separately, especially in cases where there is a lot of commonality of activations across different domains and domain types. Another goal is to figure out the extent of commonality of various cognitive phenomena across domains and domain types.

1 Introduction

Functional Magnetic Resonance Imaging (fMRI) is a non-invasive technique to capture brain activations with a relatively high spatial resolution, in the order of several cubic millimeters. fMRI presents an opportunity to advance our understanding in how the brain works, but it also presents a challenge as to how to extract the information present in the data. It has been shown (reviewed in section 3.1) that machine learning techniques, in particular classification methods, can answer this challenge.

Despite the success, there are fundamental limitations of existing machine learning approaches when used to analyze fMRI data. In a typical fMRI study, there are in most cases more than one subject, each subject having a unique brain in terms of shape and size. In other words, the feature space is different for different subjects. This problem can be alleviated by registering all the subjects' brain to a common space. However, the uncertainty introduced by the registration is often ignored. Furthermore, it might still be the case that the patterns of activations for the same cognitive process might be different for different subjects because of the influence of factors such as each subject's unique experience and differences in the brain's vascular density for different subjects. As such, currently machine learning techniques are often applied independently for each individual subject, or they are applied to normalized data for all the subjects

without properly accounting for the uncertainty introduced by the normalization process and the inter-subject variations still present in the data.

In more than a few cases, there have been more than one fMRI study done to study a common cognitive aspect. For instance, there have been a couple of fMRI experiments done in our group to study semantic representations in the brain, with one study using words as the stimuli while pictures were used in the other study. In some cases, a research group runs the same study multiple times; an example is a study described in Wei et al. (2004), in which the fMRI activations corresponding to the auditory 2-back task were captured over eight sessions, with at least 3 weeks of time in between sessions. In other cases, several different research groups run similar studies, for instance a study described in Casey et al. (1998), in which an fMRI study on spatial working memory was done at four different institutions. Intuitively, there is some information common across these studies, mixed with variations introduced by, among others, different experimental conditions and different stimulus types. Current machine learning approaches are not flexible enough to handle these variations, so they are usually applied independently for each individual study, even for studies with the same subjects.

I have now mentioned about sharing data across subjects and studies. For this proposal, subjects and studies are examples of what I call *domain types*, while subject and study instances form the actual *domains*. I define a *domain* as a generator of data instances associated with a particular distribution. The type of the domain or the *domain type* is simply a way to categorize domains. Different domains are associated with different distributions. So for instance, an individual subject in an fMRI study can be considered as a distinct domain, since different subjects are assumed to have different generative distributions.

With that consideration, the main thesis of this proposal is

It is possible to invent machine learning and statistical techniques that can combine data from multiple domains and domain types to improve predictive performance and such that common patterns of activations can be distinguished from domain-specific or domain-type-specific patterns of activations.

In other words, despite the challenges outlined above, I believe that we can develop methods that can account for data from multiple domains and domain types. These methods will be measured by their ability to make optimal predictions of some quantity of interest (e.g. class in classification) when presented with previously unobserved data, conditioned on the data that have been observed. In addition, the methods can also reveal common patterns vs patterns that are specific to specific domains or domain types. I will develop these methods in the course of the thesis work through the framework of Bayesian statistics, which affords us the ability to incorporate prior knowledge and deal with predictive inference in a principled probabilistic way.

What would be the benefits of being able to combine data across multiple domains in the case of fMRI data? fMRI data is high-dimensional but very few training examples relative to the number of dimensions, a not-so-ideal combination from the perspective of machine learning theory. By combining data across multiple domains, one benefit is that we can increase the number of training examples available, which can then improve the performance of these machine learning techniques. Another benefit would be the ability to extract components shared across one or several domains and distinguish them from components specific to particular instances of one or more domains. This is beneficial from the cognitive neuroscience perspective, because these methods can cognitive neuroscientists to integrate data from multiple subjects and studies so that common mechanisms for specific cognitive activities are revealed. By validating on their predictive performance, we can ascertain that these mechanisms can generalize to new instances, verifying that they are reliable and reproducible. Furthermore, the techniques that I will investigate can also be viewed as ways to do inductive transfer in the presence of complex task structures. Hence, the proposed work has the potential to advance the techniques for machine learning in general and transfer learning in particular.

In section 2, I give a brief overview of fMRI. Related work is reviewed in section 3. Section 4 describes some preliminary results in which the Gaussian Naïve Bayes is extended to incorporate inter-subject variations when combining data across subjects. Section 5 outlines the proposed work for the thesis. In section 6, I give the proposed schedule for undertaking the thesis work.

2 fMRI Overview

fMRI utilizes a strong magnetic field to detect fine-grained changes in the magnetic properties of the brain. In particular, fMRI is designed to take advantages of the changes in the magnetic properties of oxyhemoglobin and deoxyhemoglobin during neural activations compared to when neural activations are absent. Oxyhemoglobin (hemoglobin when it is carrying oxygen) is diamagnetic, while deoxyhemoglobin (hemoglobin when it is not carrying oxygen) is paramagnetic. At resting state, in the absence of any neural activations, there is a specific proportion between oxyhemoglobin and deoxyhemoglobin. When a neuron or a group of neurons activate, they elicit glucose consumption and supply of oxygen-carrying blood to the area around the activations. However, the amount of the oxygen consumed is less than the amount of the oxygen supplied, leading to a change in the proportion of oxyhemoglobin and deoxyhemoglobin compared to the proportion in the resting state. This causes a change in the magnetic properties around the location of neural activations, which is then captured by the fMRI scanner as a blood-oxygenation-level-dependent (BOLD) signal. For more on the relationship between neural activities and the BOLD signal, see Logothetis et al. (2001).

The BOLD signal is temporally blurred compared to the neural activations: neural activations lasting in the order of a few hundred milliseconds can give rise to a response in the BOLD signal in the order of a few (10-15) seconds. On the other hand, a relatively high spatial accuracy can be obtained in fMRI. Current state of the art fMRI scanners can capture data with a spatial resolution of $3 \times 3 \times 3 \ \mathrm{mm}^3$ for a volume element (called *voxel*), containing a population of several thousands of neurons. The resulting data is typically corrupted with noise from various sources. Some of this noise can be removed through some preprocessing steps, but some amount of noise will remain even in the preprocessed data.

Just like there are variations in weights and heights across the population, there are variations caused by the differences in brain volumes across different individuals. This gives rise to different feature spaces for different human subjects. Methods are available to map fMRI data from different brains into a common brain template. However, these methods might introduce distortions in the data, caused by the necessary inter/extrapolations from the original voxels to the voxels in the common brain template. Furthermore, the BOLD signal also depends highly on the density of the blood vessels, and there might be differences in the vascular density at a specific location in different brains. Lastly, even though, there are common functional localities across different brains, we also need to consider that different people encounter different life experiences. It is not yet known how these different experiences reflect in the patterns of activations for a specific cognitive phenomenon, for instance, how the patterns of activations representing the semantic category food differ in native English speakers vs English-as-second-language speakers. These are some of the main challenges that need to be addressed in order to be able to effectively extract information from fMRI data across subjects in various settings.

3 Related Work

Now I will review works related to the thesis proposed here. I will break down these works into works in classification of fMRI data, approaches for group analysis in fMRI, approaches for multitask learning or inductive transfer, and a Bayesian approach to obtain a extract lower-dimensional factors out of high-dimensional data.

3.1 Classification of fMRI Data

Classification of fMRI data refers to the problem of classifying mental states using fMRI data. There have been quite a few publications regarding the application of classification for the analysis of fMRI data. I will not exhaustively cover every work that has been published and will instead focus on those works that have dealt with combining data across subjects in some sense. In this regard, Wang et al. (2004) combined data for a sentence-picture matching study across subjects by first normalizing the data into region-of-interest (ROI) supervoxels, i.e. the average of the voxels in the corresponding ROI, and pooling the data for all the

subjects and use them as training examples for a single Gaussian Naïve Bayes (GNB) classifier. In section 4, I will compare their method with an extension of the GNB classifier that is designed to account for inter-subject variability when using data from multiple subjects. In the context of classifying whether subjects were lying or not, Davatzikos et al. (2005) also pooled data from multiple subjects after normalization to a standard brain template (the Montreal Neurological Institute or MNI template), which in turn became the training examples of a support vector machine with a Gaussian kernel. The pooling was also done by Mourão-Miranda et al. (2006), which considered the effects of temporal compression and space selection on single-subject and multi-subject classification of fMRI data using the linear support vector machine.

While there are a few publications that consider combining data across subjects, there has not been any publications regarding combining data across studies. In our group, there have been (unpublished) investigations of how well naïve pooling across studies works for fMRI datasets studying semantic categories (with different stimulus types for different studies) after normalization into the MNI template. In some cases, better-than-random classification accuracies can be obtained. This indicates that indeed sharing across studies is possible, and a principled way to do that will contribute significantly to the field.

As noted in Haynes and Rees (2006) and Norman et al. (2006), the ability to find out spatial mappings across subjects that are tuned to the needs of classification algorithm is important for the advance of the field. However, at this point, no methods providing that ability has been proposed.

3.2 Group Analysis in fMRI

In conventional fMRI data analysis, analysis over multiple subjects is called group analysis. The main focus is to obtain reliable population inference by accounting for variations across subjects. In the frequentist literature, this can be done using *mixed-effects models*, first suggested by Woods (1996). A mixed-effect model consists of some *fixed effects* and some *random effects*. The fixed effects represent some quantities for the whole population, while the random effects represent variations from the population-level quantities. The random effects are usually modeled as Gaussian with mean zero and variance components that need to be estimated. A widely used method to obtain estimates for mixed-effects models is the *summary statistics* method (Holmes and Friston (1998), generalized in Beckmann et al. (2003)). Instead of estimating all the effects simultaneously using the maximum likelihood principle, the summary statistics method breaks down the process by first estimating the subject-specific effects and then use these estimates to obtain the population effects. Penny et al. (2003) showed that the summary statistics approach yields the same results as the maximum likelihood approach without requiring as much computational resources. Another way to estimate parameters of a mixed-effects model is the restricted maximum likelihood (ReML) method (Harville (1977)), which was advocated by Worsley et al. (2002) as giving less biased variance estimates than maximum likelihood. Friston et al. (2002a) formulated group analysis within the framework of hierarchical Bayes modeling and used the parametric empirical Bayes method proposed in Friston et al. (2002b) to obtain fixed-effects and random-effects estimates. The use of the Bayesian procedure was tied in with the desire to obtain maps of posterior probabilities of activations. Lazar et al. (2002) surveyed other methods from the field of meta analysis—the field concerned with combining information from a group of studies—that can be applicable in the context of fMRI data analysis to pool information over multiple subjects.

Note that the main focus of the works outlined above is to obtain more reliable activations in the context of univariate hypothesis testing. Hence, the objective is inherently different from the predictive objective underlying the proposed thesis. Despite this difference, ideas used in the above works can be incorporated into some of the methods that I propose to investigate, especially those involving modeling across subjects.

3.3 Multitask Learning/Inductive Transfer

In the context of multitask learning (Caruana (1997)), domains defined in this proposal are equivalent to *tasks*. Hence, machine learning techniques that work for data across domains can also be viewed as techniques for multitask learning or inductive transfer. There have been a few methods proposed to do multitask learning in the probabilistic or Bayesian setting. Bakker and Heskes (2003) discussed probabilistic

ways to differentiate task similarities in the context of multitask learning using neural networks. Yu et al. (2005) proposed a way to learn Gaussian processes from multiple related tasks by specifying a common multivariate Gaussian prior instantiations of related tasks taking the form of functions. Zhang et al. (2006) used a probabilistic model based on independent components analysis (Hyvärinen et al. (2001)) to model interactions between tasks. Rosenstein et al. (2005) extended the naïve Bayes model for multinomial data using a hierarchical Bayes model and apply it to meeting acceptance data. Marx et al. (2005) extended the logistic regression model for predicting a new task by using a Gaussian prior on the logistic regression coefficients across tasks and learning the parameters for this prior for the new task by using maximum likelihood over the coefficients for the related tasks. Xue et al. (2007) considered connecting logistic regression coefficients across related tasks, and enabling task clustering using the Dirichlet process mixture model. The Dirichlet process mixture model was also used by Roy and Kaelbling (2007) to enable clustering across tasks in a naïve Bayes model. A direct analogue in the Bayesian statistics field for classification across multiple tasks is the hierarchical logistic regression (see for instance chapter 16 of Gelman et al. (2003)), which uses hierarchical models to couple logistic regression coefficients across related groups of data.

From the above, we see that there are a few methods for considering clustering among tasks. Some of the ideas might be potentially valuable for the analysis of fMRI data, but it remains to seen whether such ideas are applicable when tasks are organized into domain type structure, such as those considered in this proposal. Also, one thing that the above methods have in common is the assumption of a common feature space across tasks. As mentioned in section 3.1, when doing predictive analysis (e.g. classification) of fMRI data, it might be desirable to deal directly with data with differing feature spaces without having to normalize them to a common space.

3.4 Sparse Bayesian Factor Regression

Factor analysis assumes that the set of variables in the data can be grouped into a smaller set of factors. Each original variable provides some contribution to each factor, quantified by what is called the factor loading matrix. West (2003) takes this idea and, adopting the Bayesian viewpoint, extends it for regression problems for high-dimensional data. To deal with the high dimensionality of the data, he adopts a prior that inherently assumes that the factor loading matrix is sparse, i.e. only a small number of variables contribute to each factor. The model (with extensions) has been applied to gene expression data (see, for instance, Carvalho et al. (2005) and Lucas et al. (2006)), and it has the potential for the predictive analysis of fMRI data analysis by breaking down joint contribution of voxels into factors.

The model proposed by West (2003) requires the specification of the number of factors in advance. Griffiths and Ghahramani (2006) propose a prior, called the Indian buffet process prio,r to model the possibility of infinite number of factors in the style of the Dirichlet process prior. With this prior, the number of factors need not be specified in advance but can adapt to the data.

One assumption of these models is that it deals with data coming from one single domain and domain type. That means in the context of fMRI it can be applied to a single subject in a single study only. It would be desirable to be able to extend the model to handle multiple domains and domain types with the capability to distinguish domain-specific and domain-type-specific factors vs factors common across several or all domains and/or domain types, and I will discuss ideas to do this in Section 5. Also, both models require Markov Chain Monte Carlo sampling, requiring a significant amount of computation time for data as complex as fMRI data.

4 Preliminary Work

Here I describe the work extending the Gaussian Naïve Bayes (GNB) classifier so that data from multiple subjects can be shared while accounting for the inter-subject variations that might exist. Section 4.1 describes the GNB classifier along with a couple of training procedures for the GNB classifier proposed previously. In Section 4.2, the hierarchical normal model, which is the basis for the proposed classifier, is described. I describe the proposed classifier—called the hierarchical Gaussian Naïve Bayes classifier—in

Section 4.3. Section 4.4 describes the experiments done using the GNB classifier and the variants of the hierarchical GNB classifier, including a description on the two datasets used for the experiments. Finally, in Section 4.7, I discuss the implications of the results of the experiments.

4.1 Gaussian Naïve Bayes Classifier

The Bayes classifier chooses the class c among K classes $(c_k)_{1 \leq k \leq K}$ which maximizes the posterior probability of the class given the data $\mathbf{y}$:

$$c = \arg\max_{c_k} P(C = c_k|\mathbf{y}) \propto \arg\max_{c_k} P(C = c_k)p(\mathbf{y}|C = c_k).$$

The data $\mathbf{y}$ is assumed to be a vector of length n composed of *features* $y_j, 1 \leq j \leq n$. The Naïve Bayes classifier makes the additional assumption that the class-conditional probability for each feature j is independent. In other words,

$$p(\mathbf{y}|C = c_k) = \prod_{j=1}^{n} p(y_j|C = c_k).$$

If, in addition to the above, we have the assumption that for each feature j,

$$p(y_j|C = c_k) = \mathcal{N}(y_j|\theta_j^{(k)}, (\sigma_j^{(k)})^2),$$

i.e. if we assume that the class-conditional density for each feature j with respect to class k is a Gaussian with mean $\theta_j^{(k)}$ and variance $(\sigma_j^{(k)})^2$, we have what is called the Gaussian Naïve Bayes (GNB) classifier (Mitchell et al. (2004)).

In a typical application of the GNB classifier, the classifier is trained by obtaining estimates $\hat{\theta}_j^{(k)}$ and $(\hat{\sigma}_j^{(k)})^2$ for each feature j from the training data. Now, I will describe two maximum-likelihood methods for learning estimates for the parameters of the GNB classifier that have been previously proposed in the context of classification of fMRI data.

Individual Method (Mitchell et al. (2004)) This method estimates parameters separately for each human subject. That is, for each class k, feature j, and subject s,

$$\hat{\theta}_{sj}^{(k)} = \frac{1}{n_s} \sum_{i=1}^{n_s} y_{sji}^{(k)}$$

$$(\hat{\sigma}_{sj}^{(k)})^2 = \frac{1}{n_s - 1} \sum_{i=1}^{n_s} (y_{sji}^{(k)} - \hat{\theta}_{sj}^{(k)})^2.$$

Note that there is no incorporation of information from the other subjects' data. When there is only one training example, in order to avoid degenerate situations, I use 0.001 as the variance estimate. The classifier learned using this method will be called *GNB-indiv*.

Pooled Method (Wang et al. (2004)) This method assumes that all the data comes from one subject, or equivalently, that there exists no variations across subjects after normalization of the data to a common template. That is, for each class k, feature j, and for all subjects,

$$\hat{\theta}_j^{(k)} = \frac{1}{n}\sum_{s=1}^{S}\sum_{i=1}^{n_s} y_{sji}^{(k)}$$
$$(\hat{\sigma}_j^{(k)})^2 = \frac{1}{n-1}\sum_{s=1}^{S}\sum_{i=1}^{n_s}(y_{sji}^{(k)} - \hat{\theta}_{sj}^{(k)})^2,$$

where $n = \sum_{s=1}^{S} n_s$. The estimates are the same for all subjects, and inherently, the method ignores possible variations across subjects. The classifier learned using this method will be called *GNB-pooled*.

4.2 Hierarchical Normal Model

The hierarchical normal model is a hierarchical Bayes model used to model interrelated Gaussian data. The general form of the model with two levels of hierarchy is as follows.

Let there be S groups in the data. In the context of fMRI, these groups will be equivalent to domains as defined previously, with a focus on subjects as the domains. Let y_{si} be the i-th instance of the data from group s. In fMRI context, y_{si} is the data from trial i for the sth subject. The generation of data from group s is modeled as

$$y_{si} \sim \mathcal{N}(\theta_s, \sigma^2),$$

assuming a common variance σ^2 across all the groups, and a group-specific mean θ_s. Furthermore, we assume that θ_s for each group is generated as

$$\theta_s \sim \mathcal{N}(\mu, \tau^2).$$

In other words, the distribution of the θ_s's is Gaussian with mean μ and variance τ^2.

Next, I consider two different assumptions about whether σ is known or unknown, while the rest of the parameters are unknown.

4.2.1 Known σ

The unknown parameters are θ, μ, and τ. With $\sigma_{s'}^2 = \frac{\sigma^2}{n_s}$, the joint posterior of the unknown parameters is

$$\begin{aligned} p(\theta, \mu, \tau|y) &\propto p(\mu, \tau)p(\theta|\mu, \tau)p(y|\theta) \\ &\propto p(\mu, \tau)\prod_{s=1}^{S}\mathcal{N}(\theta_s|\mu, \tau^2)\prod_{s=1}^{S}\mathcal{N}(\overline{y}_{s\cdot}|\theta_s, \sigma_{s'}^2). \end{aligned} \tag{1}$$

I use the uniform prior for (μ, τ), i.e. $p(\mu, \tau) \propto 1$.

4.2.2 Unknown σ

Now the unknown parameters include σ. Using uniform prior for $(\mu, \log\sigma, \tau)$, we have the joint posterior of the unknown parameter as

$$p(\theta, \mu, \sigma, \tau|y) \propto \sigma^{-2}\prod_{s=1}^{S}\mathcal{N}(\theta_s|\mu, \tau^2)\prod_{s=1}^{S}\prod_{i=1}^{n_s}\mathcal{N}(y_{si}|\theta_s, \sigma^2). \tag{2}$$

4.3 Hierarchical Gaussian Naïve Bayes Classifier

Now I am going to incorporate the formulation given in Section 4.2 in the estimation of parameters for the Gaussian Naïve Bayes classifier. I will call the modified classifier the *hierarchical Gaussian Naïve Bayes* classifier. Note that this is different from the hierarchical Na "ive Bayes classifier mentioned in McCallum et al. (1998), which deals with classes arranged in a hierarchy, but similar to the hierarchical Naïve Bayes classifier mentioned in Rosenstein et al. (2005), with the difference being that in Rosenstein et al. (2005), the classifier deals with multinomial data while the classifier described here deals with continuous Gaussian data. To avoid the notation from becoming cluttered, in this section, dependencies of the parameters and the hyperparameters on the class and feature are not stated explicitly but implied.

I will first give an overview of methods of parameter estimation in the context of hierarchical models. Then I will present variants of the hierarchical Gaussian Naïve Bayes classifier based on the combination of the assumption on the variance and the estimation method used. The algorithms will be summarized in Table 1 at the end of the section.

4.3.1 Maximum-A-Posteriori (MAP) Estimation

In the *maximum-a-posteriori* (MAP) estimation approach we obtain point estimates of the parameters that maximize the posterior distribution of the parameters. Below I will describe two approaches to account for the contribution of the hyperparameters Φ for the estimation of the useful parameters Θ in a general hierarchical model setting. In the context of the model described in Section 4.2, $\Theta = \{\theta, \sigma\}$ and $\Phi = \{\mu, \tau\}$.

Parametric Empirical Bayes (MAP-PEB) The quantity of interest in this discussion is the estimator of Θ that maximizes the marginal posterior $p(\Theta|y)$. In the MAP-PEB approach, we approximate the marginal posterior by the conditional posterior $p(\Theta|\Phi, y)$. We need to select the value of Φ to be used in the conditional posterior. In this paper, I choose Φ that maximizes the marginal likelihood $p(y|\Phi)$. This particular selection is called the *type II maximum likelihood* prior selection Berger (1985). It is equivalent to maximizing the marginal posterior $p(\Phi|y)$ when the hyperprior $p(\Phi)$ is uniform. Notice that Φ is estimated from the data, and so the approach in general is called the *empirical Bayes* approach. It is called the *parametric empirical Bayes* approach because the prior for $\Theta, p(\Theta|\Phi)$ has a parametric form.

Direct Maximization of the Marginal Posterior (MAP-direct) One drawback of the MAP-PEB approach is that it does not take into account the distribution of the hyperparameters Φ. This can especially be problematic if the marginal posterior of $\Phi, p(\Phi|y)$ is spread out, which means that Φ itself contains some uncertainty. We can avoid this problem by working with $p(\Theta|y)$ directly. So in contrast with the previous approach, we now would like to obtain Θ that maximizes $p(\Theta|y)$.

4.3.2 Known σ

In this subsection, I am assuming that σ is known, but in reality, it also needs to be estimated. The following estimator for σ^2 will be used:

$$\sigma^2 = \text{median}(S_s) \quad \text{where} \quad S_s = \frac{1}{n_s - 1} \sum_{i=1}^{n_s} (y_{si} - \overline{y}_{s\cdot})^2. \tag{3}$$

MAP-PEB Given the joint posterior in equation (1), we have that the conditional posterior for the θ_s's are independent and distributed as (see also Section 5.4 of Gelman et al. (2003))

$$\theta_s|\mu, \tau, y \sim \mathcal{N}(\hat{\theta_s}, V_s), \tag{4}$$

where

$$\hat{\theta_s} = \frac{\frac{1}{\sigma_{s'}^2}\overline{y}_{s\cdot} + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma_{s'}^2} + \frac{1}{\tau^2}} \quad \text{and} \quad V_s = \frac{1}{\frac{1}{\sigma_{s'}^2} + \frac{1}{\tau^2}}.$$

Also, the marginal likelihood $p(y|\mu, \tau)$ can be expressed in terms of the group means $\overline{y}_{s\cdot}$:

$$\overline{y}_{s\cdot}|\mu, \tau \sim \mathcal{N}(\mu, \sigma_{s'}^2 + \tau^2).$$

When maximizing the marginal likelihood with respect to μ, we obtain the following estimator for μ at the maximum:

$$\mu_{\text{MP}} = \frac{\sum_{s=1}^{S} \overline{y}_{s\cdot}/(\sigma_{s'}^2 + \tau^2)}{\sum_{s=1}^{S} 1/(\sigma_{s'}^2 + \tau^2)}.$$

On the other hand, when maximizing the marginal likelihood with respect to τ, we obtain the following relationship:

$$\sum_{s=1}^{S} \left(-\frac{\tau}{\sigma_{s'}^2 + \tau^2} + \frac{\tau(\overline{y}_{s\cdot} - \mu)^2}{(\sigma_{s'}^2 + \tau^2)^2} \right) = 0.$$

It is not clear whether it is possible to obtain a closed form expression for the marginal-likelihood-maximizing τ_{MP} in this case. However, if we make the assumption that $\sigma_{s'}^2$ is small, and if we choose estimators that are unbiased, we can make the following approximations for μ_{MP} and τ_{MP}^2:

$$\mu_{\text{MP}} = \frac{1}{S}\sum_{s=1}^{S} \overline{y}_{s\cdot} \tag{5}$$

$$\tau_{\text{MP}}^2 = \frac{1}{S-1}\sum_{s=1}^{S} (\overline{y}_{s\cdot} - \mu_{\text{MP}})^2. \tag{6}$$

In other words, we take μ_{MP} as the sample mean of the group means, and τ_{MP}^2 as the unbiased sample variance of the group means. Since the conditional posterior given in equation (4) is normal, the estimator for θ_s that maximizes the conditional posterior is given by

$$\theta_s = \frac{\frac{1}{\sigma_{s'}^2}\overline{y}_{s\cdot} + \frac{1}{\tau_{\text{MP}}^2}\mu_{\text{MP}}}{\frac{1}{\sigma_{s'}^2} + \frac{1}{\tau_{\text{MP}}^2}}. \tag{7}$$

MAP-direct In this approach, we maximize the marginal posterior $p(\theta|y)$ directly. It turns out that

$$\log p(\theta|y) = C - \frac{S-3}{2}\log\left(\frac{1}{S}\sum_{s=1}^{S}\theta_s^2 - (\frac{1}{S}\sum_{s=1}^{S}\theta_s)^2\right) - \frac{1}{2\sigma^2}\sum_{s=1}^{S}\sum_{i=1}^{n_s}(y_{si} - \theta_s)^2,$$

where C is a constant that does not depend on θ. Maximizing $\log p(\theta|y)$ with respect to θ_s, we obtain the following expression for θ_s at the maximum:

$$\theta_s = \frac{\frac{n_s}{\sigma^2}\overline{y}_{s\cdot} + \frac{(S-3)/S}{\frac{1}{S}\sum_{s=1}^{S}\theta_s^2 - (\frac{1}{S}\sum_{s=1}^{S}\theta_s)^2}\overline{\theta}_{\cdot}}{\frac{n_s}{\sigma^2} + \frac{(S-3)/S}{\frac{1}{S}\sum_{s=1}^{S}\theta_s^2 - (\frac{1}{S}\sum_{s=1}^{S}\theta_s)^2}}, \tag{8}$$

where $\overline{\theta}_{\cdot}$ is the sample mean of the θ_s's. There is a self-dependency on θ, so an iterative method has to be used to obtain the estimate.

4.3.3 Unknown σ

Now instead of assuming σ known and using the estimator in equation (3), I consider the case when σ is assumed unknown and its estimation is incorporated into the model.

MAP-PEB Based on the joint posterior given in equation (2), we have that the conditional posterior for the θ_s's is given by

$$\theta_s|\mu,\sigma,\tau,y \sim \mathcal{N}(\hat{\theta}_s, V_{\theta_s}),$$

where

$$\hat{\theta}_s = \frac{\frac{1}{\tau^2}\mu + \frac{n_s}{\sigma^2}\overline{y}_{s\cdot}}{\frac{1}{\tau^2}+\frac{n_s}{\sigma^2}} \quad \text{and} \quad V_{\theta_s} = \frac{1}{\frac{1}{\tau^2}+\frac{n_s}{\sigma^2}}.$$

The conditional posterior for σ is given by

$$\sigma^2|\theta,\mu,\tau,y \sim \text{Inv-}\chi^2(n,\hat{\sigma}^2),$$

where $\hat{\sigma}^2 = \frac{1}{n}\sum_{s=1}^{S}\sum_{i=1}^{n_s}(y_{si}-\theta_s)^2$.

I use the estimators for μ_{MP} and τ_{MP} given in equations (5) and (6). Using these estimators and getting the modes of the conditional posterior distributions for θ and σ, I obtain the following estimators for θ and σ, with $n = \sum_{s=1}^{S} n_s$:

$$\theta_s = \frac{\frac{1}{\tau^2_{\text{MP}}}\mu_{\text{MP}} + \frac{n_s}{\sigma^2}\overline{y}_{s\cdot}}{\frac{1}{\tau^2_{\text{MP}}}+\frac{n_s}{\sigma^2}} \tag{9}$$

$$\sigma^2 = \frac{1}{n+2}\sum_{s=1}^{S}\sum_{i=1}^{n_s}(y_{si}-\theta_s)^2. \tag{10}$$

Note that there are mutual dependencies between the expressions for θ_s and σ^2, so I resort to a coordinate ascent method.

MAP-direct In this case, we maximize the marginal posterior $p(\theta,\sigma^2|y)$ directly. It turns out that

$$\begin{aligned}\log p(\theta,\sigma^2|y) = C + &-\frac{n+2}{2}\log\sigma^2 \\ &-\frac{S-3}{2}\log\left(\frac{1}{S}\sum_{s=1}^{S}\theta_s^2 - (\frac{1}{S}\sum_{s=1}^{S}\theta_s)^2\right) \\ &-\frac{1}{2\sigma^2}\sum_{s=1}^{S}\sum_{i=1}^{n_s}(y_{si}-\theta_s)^2.\end{aligned}$$

Maximizing with respect to θ_s, we obtain at the maximum

$$\theta_s = \frac{\frac{n_s}{\sigma^2}\overline{y}_{s\cdot} + \frac{(S-3)/S}{\frac{1}{S}\sum_{s=1}^{S}\theta_s^2 - (\frac{1}{S}\sum_{s=1}^{S}\theta_s)^2}\overline{\theta}_{\cdot}}{\frac{n_s}{\sigma^2}+\frac{(S-3)/S}{\frac{1}{S}\sum_{s=1}^{S}\theta_s^2 - (\frac{1}{S}\sum_{s=1}^{S}\theta_s)^2}}, \tag{11}$$

and maximizing with respect to σ^2, we obtain at the maximum

$$\sigma^2 = \frac{1}{n+2} \sum_{s=1}^{S} \sum_{i=1}^{n_s} (y_{si} - \theta_s)^2. \tag{12}$$

Again, there are mutual dependencies between θ_s and σ^2, so I again resort to a coordinate ascent method.

1. For MAP-PEB, calculate μ_{MP} and τ_{MP} using equations (5) and (6) respectively.
2. When assuming known variance, calculate σ^2 using equation (3).
3. Calculate θ_s (and σ^2 when assuming unknown variance) using the respective estimator(s):
   - Known-variance MAP-PEB: use equation (7).
   - Known-variance MAP-direct: use equation (8) iteratively.
   - Unknown-variance MAP-PEB: use equations (9) and (10) iteratively.
   - Unknown-variance MAP-direct: use equations (11) and (12) iteratively.
4. Use θ_s and σ^2 as the parameters of the corresponding class and feature of the GNB classifier.

Table 1: Algorithm for learning a hierarchical GNB classifier

4.4 Experiments

For the experiments, I use two real-world fMRI datasets. The datasets are described below.

4.5 Datasets

4.5.1 Starplus Dataset (Reichle et al. (2000))

In the starplus study, in each trial each subject looked at a pairing of a sentence and a picture and had to decide whether the sentence described the picture. Data from 13 subjects are used for this paper. There were 40 trials for each subject. In half of the trials, the picture was presented before the sentence, and in the other half, the sentence was presented before the picture. Only a subset of the brain, selected based on the expected activated areas, was captured. The data is divided into 24 anatomically defined regions of interest (ROIs). Standard preprocessing for fMRI data were applied to the data. The data was normalized to the ROI space by averaging the voxels for each ROI. The data was normalized to have mean 0 and variance 1 both across all the time points during each trial and across the voxels at a given time point.

For the classification task, there is an example from each trial, each example containing 16 images corresponding to the first stimulus from the ROI representing the calcarine sulcus. Hence, the number of features selected is 16, each feature corresponding to the averaged activation at the calcarine sulcus ROI at a specific time point. The task is to classify each example into whether a sentence or picture was presented during the acquisition of that example. For each subject, the data from each class is randomly divided into two folds with equal number of trials (20 for each fold, with 10 trials for each class in each fold), and cross-validation is performed on the folds with different numbers of training examples from the training fold. For methods that incorporate data from the other subjects, all the data from the other subjects are used along with the training fold. 10 repetitions are performed, each with a random choice of fold assignments. The averages and standard deviations of the cross-validation accuracies across all subjects are reported in Section 4.6.

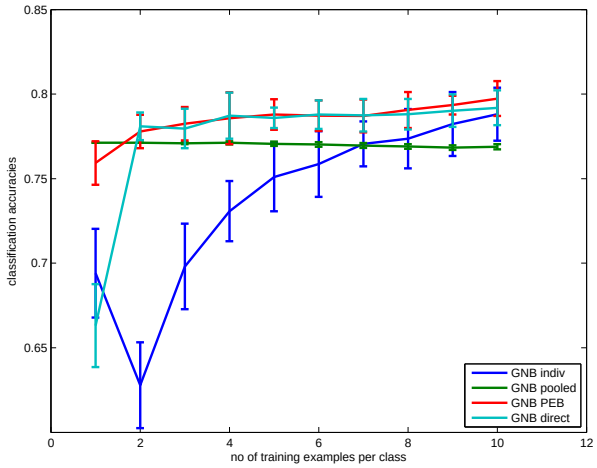

Figure 1: Starplus dataset, classification accuracies vs number of training examples for the variants of the known-variance hierarchical GNB and the two reference methods (GNB-indiv and GNB-pooled).

4.5.2 Twocategories Dataset (Pereira et al. (2006))

In this study, words from categories "tools" and "dwellings" were presented to each subject. The subjects had to think about the properties of each word and decide which category each word belongs to. Data from 6 subjects are used for this paper. 14 words were used in the study, 7 for each category. A word was presented during a trial, and each word was presented 6 times, so there are 42 examples for each category. The whole brain was captured. Standard preprocessing for fMRI was applied to the data. The data from each subject was then normalized to the MNI space Evans et al. (1993). Furthermore, for each trial and each voxel, only the average of time points 5 to 8 is used. The data was normalized to have mean 0 and variance 1 across the voxels for each trial.

The classification task is to classify the category of the word ("tools" or "dwellings") that the subject viewed in each trial. 300 voxels are chosen based on the Fisher's median separability criterion Saito and Coifman (1995); hence, the number of features in the data is 300 (the number 300 is chosen arbitrarily). For each subject, the data for each class is randomly split into two folds with equal number of trials (21 for each fold and each class), and cross-validation is performed on the folds with different numbers of training examples from the training fold. For methods that incorporate data from the other subjects, all the data from the other subjects is used along with the training group. 5 repetitions are performed, each with a random choice of fold/group members. The averages and standard deviations of the cross-validation accuracies across all subjects are reported in Section 4.6.

4.6 Results

4.6.1 Known σ

Figures 1 and 2 show the classification accuracies of the GNB classifiers trained with the methods described in Section 4.1, along with those of the hierarchical GNB classifiers trained with the MAP-PEB and MAP-direct methods assuming known σ, for the two datasets described in Section 4.5. For GNB-indiv, the accuracy is low when the number of training examples is small, and it increases as the number of training examples increases. This is expected since with a small number of training examples from the test subject and not being able to leverage data from the other subjects, the classifier cannot make a good prediction of the test data. GNB-pooled, on the other hand, is able to leverage data from the other subjects when the number of training examples is small, giving a better classification accuracy than GNB-indiv. However, as

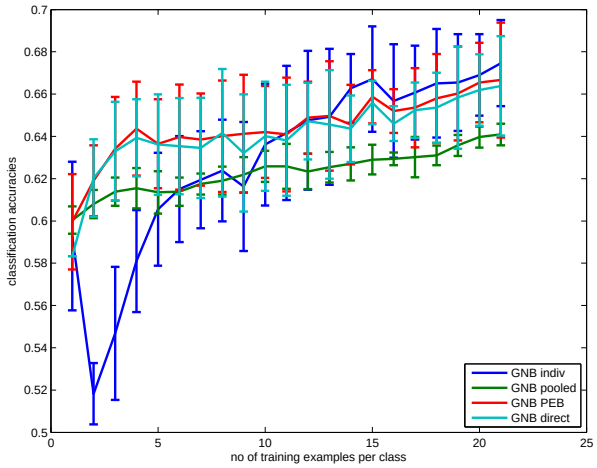

Figure 2: Twocategories dataset, classification accuracies and their standard errors vs number of training examples for the variants of the known-variance hierarchical GNB and the two reference methods.

the number of training examples for the test subject grows, the effect of the other subjects' data still dominates in such a way that the improvement in classification is diminished; in particular, for the starplus dataset, we see that GNB-pooled does not show any improvements as the number of training examples from the test subject grows because the data is dominated by that of the other twelve subjects.

On the other hand, the two variants of the hierarchical GNB classifier are able to leverage the other subjects' data when the number of training examples from the test subject is small, yielding better classification accuracies than GNB-indiv, and also reduce the contribution of the other subjects' data when that number increases, yielding better accuracies than GNB-pooled. In particular, note that when the number of training examples is large, the accuracies of the hierarchical GNB classifier match those of GNB-indiv. We also see that the accuracies of the two variants are not significantly different. There is an exception in the case of the starplus dataset when the number of training examples of the test subject is one, where the performance of the classifier using the MAP-direct method is lower than the one using the MAP-PEB method, and is also lower compared to the performance of GNB-pooled. However, this phenomenon is not observed for the twocategories dataset, so it is not a general trend of the methods.

4.6.2 Unknown σ

In figures 3 and 4, I include the results for the variants of the hierarchical GNB classifier when the variance is assumed to be unknown. The results for the classifier trained using the MAP-PEB method when the variance is assumed to be known are also included for comparison. In particular, we see that making the assumption that the variance is unknown does not result in significant differences in the performance of the variants of the hierarchical GNB classifier. Also, note that for the variant with the unknown variance assumption trained with the MAP-direct method, we see again a lower performance in the starplus dataset compared to the MAP-PEB variant when the number of training examples of the test subject is one, a trend similar to that for the MAP-direct variant with the assumption of known variance.

4.7 Discussion

We see that for the two datasets presented here, the hierarchical GNB classifier is able to leverage the other subjects' data when the number of training examples is small, and at the same time reduce the contribution of the other subjects' data when the number of training examples increases. Of course, there are many

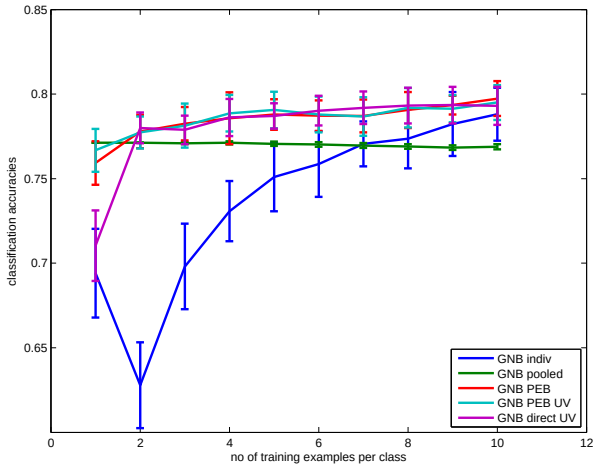

Figure 3: Starplus dataset, classification accuracies and their standard errors vs number of training examples for the known-variance MAP-PEB hierarchical GNB, the unknown-variance variants of hierarchical GNB, and the two reference methods.

more fMRI datasets available, and many more will be collected in the future. Will the results generalize to other datasets? The key factor is whether the model assumptions hold for the datasets in question. In a lot of cases, these assumptions hold because we know that there are similar patterns of brain activations for the same feature across subjects when the subjects are doing the same cognitive task. However, it might also be the case that for the same feature, the patterns of brain activations might be different for different subjects, even when the cognitive task is the same. An example is given by the study presented in Kamitani and Tong (2005), in which it is found that orientation columns in V1 are encoded by the signal captured by fMRI; however, it is not clear whether after normalization, the same voxel, i.e. feature, across subjects will encode the same orientation column. So applying the hierarchical GNB classifier in that case might not give a significant benefit when the number of training examples for the test subject is small. Nevertheless, when the number of training examples is large, the hierarchical GNB classifier should be less adversely affected because of its ability to reduce contribution from the other subjects' data.

Besides the validity of the model assumptions, another factor that can affect the performance of the hierarchical GNB classifier is the number of subjects. Intuitively, the variance of the estimates of the group parameters (i.e. the hyperparameters) will be high when the number of subjects is low. However, we can see from the estimators for the MAP-direct method (in particular, in equations (8) and (11)), that when the number of subjects is three, there will be no contribution from the other subjects' data. This can be used as a guideline that preferably there will be more than three subjects available when applying the hierarchical GNB classifier. In the results, I have shown that the classifier can work well when there are as few as six subjects available.

Given the four variants of the hierarchical GNB classifier, which one should be used? Even though the results presented here do not give a conclusive evidence that one variant is better than the others, there might be other fMRI datasets where the variants will perform differently. Let me highlight differences in the variants that can affect the choice of one variant over the others in a future application. First, in terms of computational complexity, the other variants besides the known-variance MAP-PEB method involve iterative methods, i.e. the coordinate ascent method. Therefore, the known-variance MAP-PEB method is the fastest to compute. Another difference is caused by the two methods of estimation. In the MAP-PEB variants, the hyperparameters are estimated while in the MAP-direct variants, they are marginalized. The difference between these two approaches will be small when the hyperparameters are concentrated on particular values, but when they are not, then the MAP-PEB variants do not account for the uncertainty

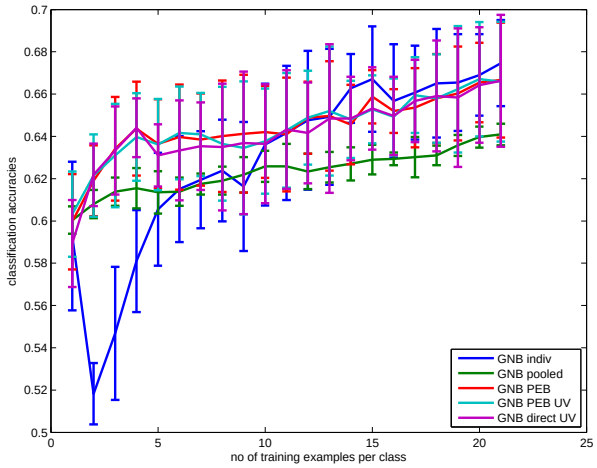

Figure 4: Twocategories dataset, classification accuracies vs number of training examples for the known-variance MAP-PEB hierarchical GNB, the unknown-variance variants of hierarchical GNB, and the two reference methods.

of the hyperparameters unlike the MAP-direct variants. And there is also the difference in whether the variance is assumed known or unknown. When the variance is assumed unknown, the estimation takes into account the revised estimates for the subjects' means, while when the variance is assumed known, the estimation uses the individual subjects' means, so the two assumptions might result in significantly different accuracies when the variations of the data around the different kinds of means are significantly different.

Regardless of the variants, the results obtained are promising, in the sense that there is virtue in sharing data across instances in a domain, in this case, subjects. The results suggest that it is also possible to reduce the number of trials for some subjects and yet maintain almost the same level of accuracies as when the number of trials is not reduced. This suggests that we can have a classifier that is initially trained on some subjects and then trained on other subjects using fewer number of trials. Further work might make this more flexible, for instance, having classifiers trained initially on some studies and then trained on other studies. In the next section, I present my ideas about how to make this work, among others.

5 Proposed Work

There are opportunities for sharing data beyond sharing data across subjects. In this section, I discuss the issues involved in sharing data across multiple domains. I plan to address some of these issues in my thesis.

5.1 Sharing in A Single Domain Type

Let me first consider the problem of sharing data in a single domain. The previous section discusses an approach that can be used for sharing in a particular domain, namely sharing across subjects. Now I will consider sharing within other domains relevant to the fMRI domain.

5.1.1 Sharing Across Voxels

Let us consider the voxel domain. One question might be whether the methods described in the previous section are applicable for sharing across voxels. But before addressing this question, let us step back and

consider the nature of the voxel domain. I will consider only the classification task. Note that in classification of fMRI data, the voxels are *features*, which are direct inputs to the classifier, in contrast with subjects, which virtually are classification tasks in the sense of multitask learning. Note that typically a classifier takes the features describing the data, but the classifier is trained and tested only for a specific task. So the problem of sharing across voxels is the problem of sharing across features. The GNB classifier, used as baseline in the previous section, makes the assumption that each feature is conditionally independent given the class. So sharing across features can relax this assumption. In the setting of a Gaussian Bayes classifier, this means that we need to remove the naïve conditional independence assumption in the class-conditional probability/density $p(\mathbf{y}|C = c_k)$ and model the class-conditional covariance structure of $\mathbf{y}$.

A candidate method for modeling the covariance structure is the method outlined in the previous section, namely a hierarchical modeling of the voxel means with a common normal prior. Here we need to take into account the spatial nature of the voxel domain. There have been findings that spatially localized groups of voxels tend to exhibit similar activations. So one possibility is to have separate hierarchical normal models for spatial clusters of voxels. One problem is how we can determine which voxels belong together in a cluster. To do this, heuristics such as those employed in Niculescu et al. (2006) can be employed to determine spatially neighboring voxels that share similar activation patterns so that they can be modeled together. However, the clustering assumption can be relaxed through through a prior distribution that encodes the prior knowledge for spatial dependence of nearby voxels, such as the Gaussian Markov Random Field (Rue and Held (2005)). With this prior distribution, instead of having separate models for different clusters, we can still have one model for the whole-brain activations.

5.1.2 Sharing Across Studies

In some cases, one subject participates in more than one fMRI study. In these cases, there are no cross-subject variations, but there are cross-study variations. Can we model these cross-study variations using a model similar to that presented for sharing data across subjects? First, let us consider the implications. Using a hierarchical normal model for modeling cross-study variations means that we assume that for that particular subject, a voxel's mean activation in different studies comes from a normal distribution with a common-study mean. This assumption might be reasonable when the studies are relatively similar, in that they exercise similar cognitive functions. When the studies exercise completely different cognitive functions, the activations across studies will be different for voxels discriminately activated by these studies, so intuitively, it is less reasonable that these different activations will come from a normal distribution. So the applicability of sharing across studies can be affected by the similarity of the studies. Of course, it is also reasonable to consider some subjects more similar than others. I will discuss this topic in greater detail in section 5.3, when I consider determining groups or clusters within domains.

In reality, it is not very common that a subject participates in multiple similar fMRI studies. Instead, usually different studies involve different subjects. So in a lot of cases, we need to consider sharing across studies and subjects together. This is an instance of sharing across multiple domains, which is covered next.

5.2 Sharing Across Multiple Domain Types

I have described potential for sharing in different kinds of domain types. The next question is, is it possible to share across multiple domain types simultaneously? I think the answer is yes, and in this thesis I will develop methods for doing so. Below, I will first discuss possible methods for sharing across two domain types, and then I will cover sharing across an arbitrary number of domain types.

5.2.1 Sharing Across Two Domain Types

I first consider sharing across two domain types. Some examples are sharing across both subjects and studies, and sharing across both subjects and voxels. How can we augment the model to simultaneously handle sharing across two domain types? Here I describe two possible approaches: adding another level to the hierarchy, or combining two separate submodels, one for each domain.

Multiple Levels of the Hierarchy

$$
\begin{aligned}
y_{s^{(m)}i} &\sim \mathcal{N}(\theta_{s^{(m)}}, \sigma^2) \\
\theta_{s^{(m)}} &\sim \mathcal{N}(\mu^{(m)}, (\tau^{(m)})^2) \\
\mu^{(m)} &\sim \mathcal{N}(\alpha, \beta^2),
\end{aligned}
$$

where m represents a specific study. In other words, we add another level to model the variations (assumed normally distributed) of the study-specific means. I use the normal distribution for convenience, but it is not clear whether it is the appropriate distribution to use. If there is sufficient data, it might be possible to model the variations using a distribution with a nonparametric density function.

Independent Submodels for Each Domain An alternative possible model is the following:

$$
\begin{aligned}
y_{s^{(m)}i} &\sim \mathcal{N}(f(\theta_s, \xi_m), \sigma^2) \\
\theta_s &\sim \mathcal{N}(\mu, \tau^2) \\
\xi_m &\sim \mathcal{N}(\alpha, \beta^2).
\end{aligned}
$$

Here, we model the contributions from the subject group and the study group as independent, and $f()$ can be any arbitrary function. For simplicity, $f()$ can be assumed to be linear, i.e. $f(x, y) = ax + by$ for some constants a and b.

5.2.2 Sharing Across n Domain Types

Going beyond two domain types, we can consider a hybrid approach, using multiple levels of hierarchy for a subset of the domain types and independent submodels for another subset of the domain types. A key question is how to determine which option to use for which subsets of the domain types. Furthermore, can such a decision be automated; in other words, can the model be made to be flexible enough to decide between these options based on the data?

5.3 Determining Groups to Share

So far, I have assumed that when sharing across a domain type, we can share across all instances of that domain type. However, as mentioned in section 5.1.2, it is more reasonable to share across similar domains than across different domains. This information might be given in advance, for instance, it is usually the case that we have to treat subjects with some clinical conditions differently from normal subjects. Nonetheless, is it possible to determine which domains to share and not to share automatically from the data, such that the predictive performance is maximized?

I believe the answer to the preceding question is yes, and a way to do so is by using a mixture model. I will consider this in the context of sharing across subjects, but the principle should be also applicable to sharing across the other domain types.

Here is a model for sharing across subjects augmented with a mixture model:

$$
\begin{aligned}
y_{si} &\sim \mathcal{N}(\theta_s, \sigma^2) \\
\theta_s &\sim \mathcal{N}(\mu^{(k)}, (\tau^{(k)})^2) \\
k &\sim \text{Multinomial}(\pi_1, \cdots, \pi_K)
\end{aligned}
$$

where k is the group indicator, and K is the total number of groups. We can describe the model as having the groups represented as Gaussian mixture models with K mixtures. One major question is how

one can choose the correct value of K. We can do this using cross-validation or reversible-jump Markov Chain Monte Carlo (RJMCMC). Alternatively, we can reformulate the model as

$$\begin{aligned} y_{si} &\sim \mathcal{N}(\theta_s, \sigma^2) \\ \theta_s &\sim \mathcal{N}(\mu^{(k)}, (\tau^{(k)})^2) \\ k &\sim \mathrm{DP}(\alpha_0, G_0) \end{aligned}$$

where $\mathrm{DP}()$is the Dirichlet process (Ferguson (1973)) with parameter α_0 and base measure G_0, which we can take from the conjugate family (i.e. normal for a hierarchical normal model). Inference can be performed using Markov Chain Monte Carlo (Neal (2000)) or variational methods (Blei and Jordan (2006)).

5.4 Factor Analysis

Up to now, I have discussed doing classification using voxels as the features. An alternative is provided by assuming that in the data there exist unobserved factors, some of which might be discriminative of the classes in the classification task. We can try to extract these factors using factor analysis methods. The information extracted about the factors, including the information about which voxels belong to which factors, can potentially yield novel insights about the workings of the brain.

Most factor analysis methods, such as those given in West (2003) and Griffiths and Ghahramani (2006), are formulated in the context of data from a single domain. In order to extract the different kinds of factors, a naïve approach would be to combine data from the multiple domains and domain types and look at the contribution to (in factor analysis terminology, the loading of) each factor. This might reveal factors whose loadings come from data of a specific domain or domain type versus factors whose loadings come from data across domains and/or domain types. The implication is that the latter factors are shared across domains and domain types. In the factor regression context, it is also useful to look at the contribution of each factor to the regression prediction to see how much these various factors contribute to the prediction. Doing so can reveal how much can be shared versus how much is domain-specific and/or domain-type-specific for the phenomenon to predict. An alternative is to explicitly encode the different classes of factors (shared vs domain-specific and/or domain-type-specific) in the model or as prior distributions. It would be interesting to find out how similar or different the results will be compared to those of the naïve approach.

Computation is a major challenge when doing factor regression given the kind of data that I consider here. One question that I also plan to address is whether approximate inference techniques such as variational (Jordan et al. (1999)) or expectation propagation (Minka (2001)) methods can be applied to get a computationally efficient but accurate approximation for inference.

5.5 Methods for Unnormalized Data

In the previous discussions, I have assumed that the fMRI data have been normalized to a common space. However, it is likely that the normalization procedure causes some loss of information. The causes include the uncertainty of the mapping of voxels from the original subject's space to the normalized space, and the interpolations and extrapolations involved in the process. It might be desirable to incorporate the normalization procedure into the model presented—instead of done as a preprocessing step as is common currently—so that we can account for the uncertainty introduced by the normalization procedure. There are models that have been proposed for using Bayesian inference to perform registration of images (e.g. Marroquin et al. (2002) and Pohl et al. (2006)). As a first step, these Bayesian registration methods can be used as a baseline for augmenting the model.

5.6 Incorporating Temporal Information

I have assumed that voxel values at a given time point can be features in the classification task, conditionally independent given the class in the case of the GNB classifier (the starplus dataset in the preliminary work).

The time element can also be removed by averaging the voxels over time (the twocategories dataset in the preliminary work). In reality, especially in fMRI data analysis, the temporal information available might be important, for instance, how the hemodynamic responses at particular locations evolve for different kinds of cognitive tasks might indicate different demands for these tasks. However, inherent in the consideration of temporal information is the notion of forward progress. In most cases, there is a dependence on values at previous time points. Hence, the notion of sharing for time is different from the notion of sharing for the other domain types I have discussed so far.

A way to incorporate temporal information is through the use of state space models (Durbin and Koopman (2001)), or, from the Bayesian perspective, the dynamic linear models (West and Harrison (1997)). When using state space models, we need to specify the state transition and observation matrices. The observation matrix can be formulated by taking into account the hemodynamic response exhibited in fMRI data. However, formulating the state transition matrix is more challenging. This is a problem of *system identification* (Ljung (1998)), and it is possible that existing system identification methods might be applicable.

5.7 Sharing Across Modalities

In this context, fMRI is a modality used to gather information about the workings of the brain. There are also other modalities that are used in the cognitive neuroscience field, such as EEG. There have been instances on simultaneous analysis of EEG and fMRI data (e.g. Christmann et al. (2007)). However, the analysis is done independently on each subject and is usually done for a specific study. Hence, some of the issues faced in the course of the thesis are also relevant for integrating data across modalities.

5.8 Levels of Granularity and Process Decomposition

Cognitive processes usually exist at several simultaneous levels of granularity or can be broken down into subprocesses. When considering the process of reading a word, for instance, an example of the former is considering the process of reading a noun as a specialization of the read-word process, while an example of the latter is breaking the read-word process into processes for viewing the letters, recognizing the letters, grouping the letters together, and associating the grouped letters with a word. Implicit in classification is choosing a level of granularity or a specific level of processes in the above sense. The problem is that, when focusing on a more specific level for either case, the analysis becomes more sensitive to the idiosyncrasies of each subject. Of interest also is how lower levels in either case influence higher levels. I believe that it is possible to incorporate both granularity levels and process levels into the models for sharing across domain types into the model, potentially allowing us to make inference on a specific level of granularity or a specific subprocess component of some cognitive process. In particular, hierarchical classification methods (e.g., Shahbaba and Neal (2007)) might be applicable to the problem of dealing with different levels of granularities.

6 Research Plan

Here is a plan of the research that I will undertake for the thesis work:

- Remainder of spring 2007: undertake research on models for combining data for multiple subjects and studies, including extending the sparse Bayesian factor regression model.
- Fall 2007: undertake research on incorporating spatial normalization across subjects in the model.
- Spring 2008: investigate how to extend the model so that temporal information can be incorporated.

References

Bakker, B. and Heskes, T. (2003). Task Clustering and Gating for Bayesian Multitask Learning. *Journal of Machine Learning Research*, 4:83–99.

Beckmann, C. F., Jenkinson, M., and Smith, S. M. (2003). General multilevel linear modeling for group analysis in fMRI. *NeuroImage*, 20:1052–1063.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer, second edition.

Blei, D. M. and Jordan, M. I. (2006). Variational Inference for Dirichlet Process Mixtures. *Bayesian Analysis*, 1(1):121–144.

Caruana, R. (1997). Multitask Learning. *Machine Learning*, 28(1):41–70.

Carvalho, C., Chang, J., Lucas, J., Nevins, J. R., Wang, Q., and West, M. (2005). High-dimensional sparse factor modelling. ISDS Discussion Paper 2005-15, Duke University.

Casey, B., Cohen, J. D., O'Craven, K., Davidson, R. J., Irwin, W., Nelson, C. A., Noll, D. C., Hu, X., Lowe, M. J., Rosen, B. R., Truwitt, C. L., and Turski, P. A. (1998). Reproducibility of fMRI Results across Four Institutions Using a Spatial Working Memory Task. *NeuroImage*, 8:249–261.

Christmann, C., Koeppe, C., Braus, D. F., Ruf, M., and Flor, H. (2007). A simultaneous EEG-fMRI study of painful electric stimulation. *NeuroImage*, 34:1428–1437.

Davatzikos, C., Ruparel, K., Fan, Y., Shen, D., Acharyya, M., Loughead, J., Gur, R., and Langleben, D. (2005). Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection. *NeuroImage*, 28:663–668.

Durbin, J. and Koopman, S. (2001). *Time Series Analysis by State Space Methods*. Oxford University Press.

Evans, A. C., Collins, D. L., Mills, S. R., Brown, E. D., Kelly, R. L., and Peters, T. M. (1993). 3D statistical neuroanatomical models from 305 MRI volumes. In *Proc. IEEE-Nuclear Science Symposium and Medical Imaging Conference*.

Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209–230.

Friston, K. J., Glaser, D. E., Henson, R. N. A., Kiebel, S., Phillips, C., and Ashburner, J. (2002a). Classical and Bayesian inference in neuroimaging: Applications. *NeuroImage*, 16:484–512.

Friston, K. J., Penny, W., Phillips, C., Kiebel, S., Hinton, G., and Ashburner, J. (2002b). Classical and Bayesian inference in neuroimaging: Theory. *NeuroImage*, 16:465–483.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*. Chapman & Hall/CRC, second edition.

Griffiths, T. and Ghahramani, Z. (2006). Infinite latent feature models and the Indian buffet process. In Weiss, Y., Schölkopf, B., and Platt, J., editors, *Advances in Neural Information Processing Systems 18*, pages 475–482. MIT Press, Cambridge, MA.

Harville, D. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72:320–338.

Haynes, J.-D. and Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7:523–534.

Holmes, A. P. and Friston, K. J. (1998). Generalisability, random effects and population inference. *NeuroImage*, 7:S754.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. Wiley-Interscience.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37:183–233.

Kamitani, Y. and Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, 8:679–685.

Lazar, N. A., Luna, B., Sweeney, J. A., and Eddy, W. F. (2002). Combining brains: A survey of methods for statistical pooling of information. *NeuroImage*, 16:538–550.

Ljung, L. (1998). *System Identification: Theory for the User*. Prentice Hall, 2nd edition.

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., and Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature*, 412:150–157.

Lucas, J., Carvalho, C., Wang, Q., Bild, A., Nevins, J., and West, M. (2006). Sparse statistical modelling in gene expression genomics. In Do, K., Müller, P., and Vannucci, M., editors, *Bayesian Inference for Gene Expression and Proteomics*, pages 155–176. Cambridge University Press.

Marroquin, J., Vemuri, B., Botello, S., Calderon, F., and Fernandez-Bouzas, A. (2002). An Accurate and Efficient Bayesian Method for Automatic Segmentation of Brain MRI. *IEEE Transactions on Medical Imaging*, 21(8):934–945.

Marx, Z., Rosenstein, M. T., Kaelbling, L. P., and Dietterich, T. G. (2005). Transfer Learning with an Ensemble of Background Tasks. In *Inductive Transfer: 10 Years Later, NIPS Workshop*.

McCallum, A., Rosenfeld, R., Mitchell, T., and Ng, A. Y. (1998). Improving Text Classification by Shrinkage in a Hierarchy of Classes. In *Proceedings of the 15th International Conference on Machine Learning*.

Minka, T. P. (2001). Expectation Propagation for approximate Bayesian inference. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, pages 362–369.

Mitchell, T. M., Hutchinson, R., Niculescu, R. S., Pereira, F., Wang, X., Just, M., and Newman, S. (2004). Learning to decode cognitive states from brain images. *Machine Learning*, 57:145–175.

Mourão-Miranda, J., Reynaud, E., McGlone, F., Calvert, G., and Brammer, M. (2006). The impact of temporal compression and space selection on SVM analysis of single-subject and multi-subject fMRI data. *NeuroImage*, 33:1055–1065.

Neal, R. M. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, 9(2):249–265.

Niculescu, R. S., Mitchell, T. M., and Rao, R. B. (2006). Bayesian Network Learning with Parameter constraints. *Journal of Machine Learning Research*, 7:1357–1383.

Norman, K. A., Polyn, S. M., Detre, G. J., and Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, 10(9):424–430.

Penny, W., Holmes, A., and Friston, K. (2003). Random effects analysis. In Frackowiak, R., Friston, K., Frith, C., Dolan, R., Friston, K., Price, C., Zeki, S., Ashburner, J., and Penny, W., editors, *Human Brain Function*. Academic Press, 2nd edition.

Pereira, F., Mason, R., Mitchell, T., Just, M., and Kriegeskorte, N. (2006). Decoding of semantic category information from single trial fMRI activation in response to word stimuli, using searchlight voxel selection. In *12th Conference on Human Brain Mapping*.

Pohl, K. M., Fisher, J., Grimson, W. E. L., Kikinis, R., and Wells, W. M. (2006). A Bayesian model for joint segmentation and registration. *NeuroImage*, 31:228–239.

Reichle, E. D., Carpenter, P. A., and Just, M. A. (2000). The neural bases of strategy and skill in sentence-picture verification. *Cognitive Psychology*, 40:261–295.

Rosenstein, M. T., Marx, Z., Kaelbling, L. P., and Dietterich, T. G. (2005). To Transfer or Not To Transfer. In *Inductive Transfer: 10 Years Later, NIPS Workshop*.

Roy, D. M. and Kaelbling, L. P. (2007). Efficient Bayesian Task-Level Transfer Learning. In *Proceedings of the 20th Joint Conference on Artificial Intelligence*.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall/CRC.

Saito, N. and Coifman, R. R. (1995). Local discriminant bases and their applications. *Journal of Mathematical Imaging and Vision*, 5:337–358.

Shahbaba, B. and Neal, R. M. (2007). Improving Classification When a Class Hierarchy is Available Using a Hierarchy-Based Prior. *Bayesian Analysis*.

Wang, X., Hutchinson, R., and Mitchell, T. M. (2004). Training fMRI classifiers to discriminate cognitive states across multiple subjects. In *NIPS*.

Wei, X., Yoo, S.-S., Dickey, C. C., Zou, K. H., Guttmann, C. R., and Panych, L. P. (2004). Functional MRI of auditory verbal working memory: long-term reproducibility analysis. *NeuroImage*, 21:1000–1008.

West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. In Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M., editors, *Bayesian Statistics 7*, pages 733–742. Oxford University Press.

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*. Springer.

Woods, R. P. (1996). Modeling for intergroup comparisons of imaging data. *NeuroImage*, 4:S84–S94.

Worsley, K. J., Liao, C. H., Aston, J., Petre, V., Duncan, G. H., Morales, F., and Evans, A. C. (2002). A general statistical analysis for fMRI data. *NeuroImage*, 15:1–15.

Xue, Y., Liao, X., Carin, L., and Krishnapuram, B. (2007). Multi-Task Learning for Classification with Dirichlet Process Priors. *Journal of Machine Learning Research*, 8:35–63.

Yu, K., Tresp, V., and Schwaighofer, A. (2005). Learning Gaussian Processes from Multiple Tasks. In *Proceedings of the 22nd International Conference on Machine Learning*.

Zhang, J., Ghahramani, Z., and Yang, Y. (2006). Learning Multiple Related Tasks using Latent Independent Component Analysis. In Weiss, Y., Schölkopf, B., and Platt, J., editors, *Advances in Neural Information Processing Systems 18*, pages 1585–1592. MIT Press, Cambridge, MA.