Bayesian Models for Combining Data Across Subjects and Studies in Predictive fMRI Data Analysis

Thesis Proposal

Indrayana Rustandi

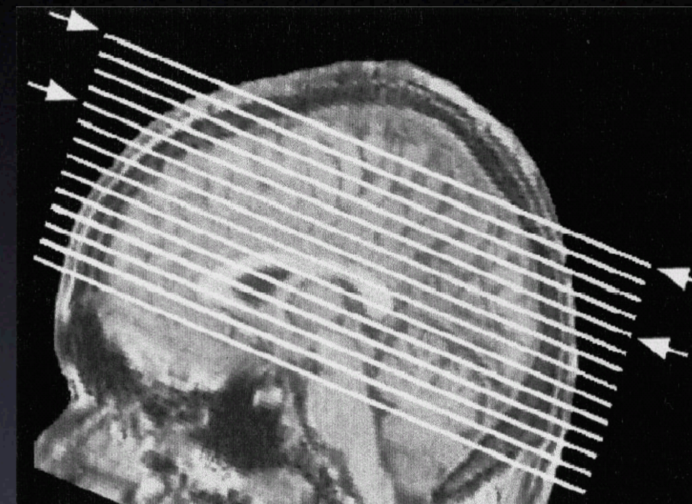
April 3, 2007

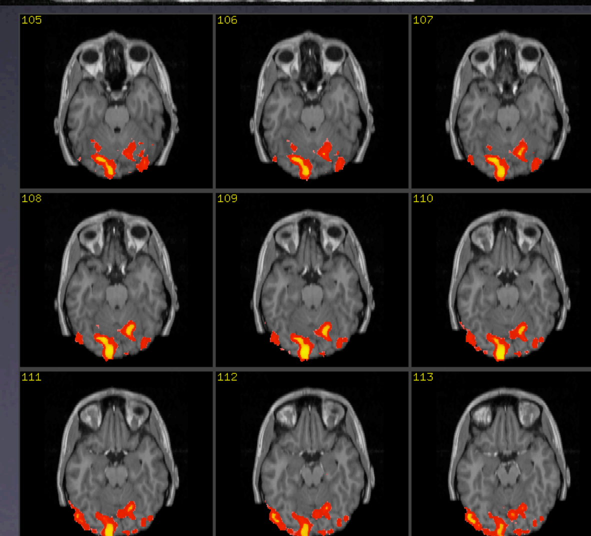
Outline

- Motivation and Thesis
- Preliminary results: Hierarchical Gaussian Naive Bayes
- Proposed work, including schedule

fMRI

- 3D images of hemodynamic activations in the brain
    - assumed to be correlated with local neural activations
- ~10,000 spatial features (voxels, analogous to pixels)
- Temporal component
- ~10-100 trials





fMRI Data Analysis

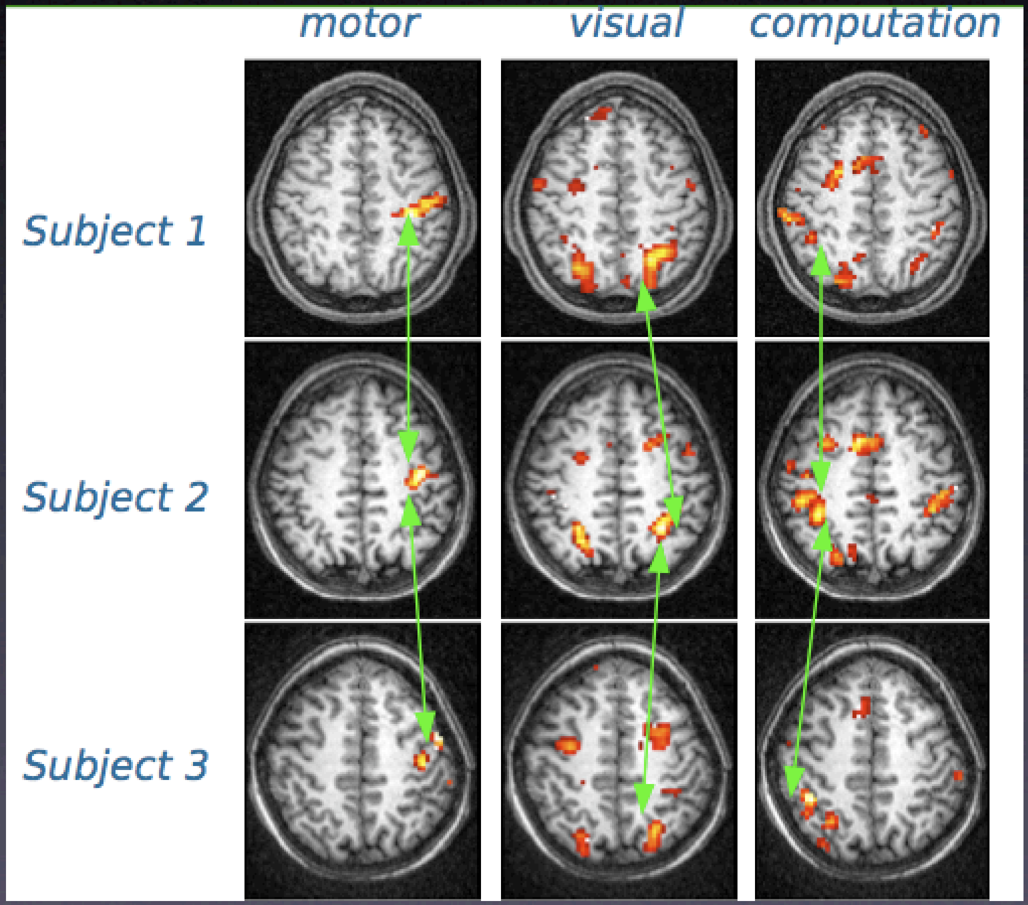
- Descriptive
  - Locations of activations correlated with a cognitive phenomenon
  - Most common paradigm used
- Predictive
  - Prediction of the cognitive phenomenon underlying brain activations
  - Classification of cognitive tasks, prediction of levels of stimulus presence (EBC competition)

Motivation: Subject-Level

- For predictive analysis, analysis is done separately for individual subjects
  - Problem: lack of training examples, can potentially improve performance by incorporating data from other subjects
- Simple solution: pool the data for all the subjects together
  - Problem: for some subjects, might not be reasonable to pool data (e.g. subjects with different conditions)
  - Problem: inter-subject variability is ignored

Inter-Subject Variability

- Human brains have similar functional structures, but there are differences in shapes and volumes (different feature spaces for different human subjects)
- Normalization to a common space is possible, but can result in the distortion of the data
- Even after normalization, the activations are also governed by personal experience, and affected by environment



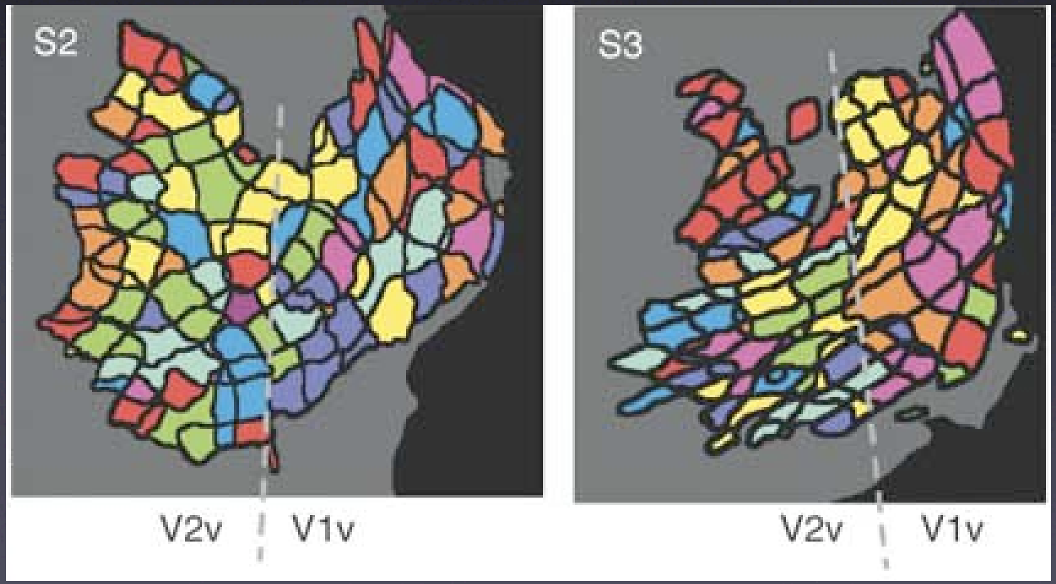
Thirion et al. (2006)

Motivation: Study-Level

- fMRI studies are expensive; it is desirable to incorporate data from existing similar studies
- Problem: problems from the subject-level
- Problem: variability due to different experimental conditions (e.g. the use of different stimuli, different magnetic field strength)
- Problem: which studies are similar

Motivation: Generalization

- How much commonality exists across different individuals with respect to a particular cognitive task
- Influence how much can be shared across different individuals (or groups)
- Example: sharing for classification of picture vs sentence might be easy, but sharing for classification of orientation of visual stimuli using V1/V2 voxels might be hard



Kamitani and Tong
Nature Neuroscience, 2005

Thesis

Machine learning and statistical techniques to

- Combine data from multiple subjects and studies
- Improve predictive performance (compared to separate analyses for individual subjects and studies)
- Distinguish common patterns of activations versus subject-specific or study-specific patterns of activations

Framework of choice is Bayesian statistics, in particular hierarchical Bayesian modeling

- Offer a principled way to account for uncertainties and the different levels of data generation involved

Related Work in fMRI

- Classification
  - Pooled data from multiple subjects (Wang et al. (2004), Davatzikos et al. (2005), Mourao-Miranda et al. (2006))
- Group analysis: multiple subjects in a specific study
  - Focus: descriptive, increase in sensitivity for detection of activations
  - Mixed-effects model (Woods (1996), Holmes and Friston (1998), Beckmann et al. (2003))
  - Hierarchical Bayes model (Friston et al. (2002))

Related Work in ML/ Statistics

- Multitask learning/inductive transfer
  - Caruana (1997)
  - Generative setting: Rosenstein et al. (2005), Roy and Kaelbling (2007)

Preliminary Work

- Combining data from multiple subjects in a given study
- Extension of the Gaussian Naive Bayes classifier
- The use of hierarchical Bayes modeling
- Designed for data after feature space normalization
  - Simplify the problem, even though not ideal

Gaussian Naive Bayes (GNB)

- Bayesian classifier: pick the class with maximum class posterior probability (proportional to product of class prior and class-conditional probability of the data)

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

- Naive Bayes: independence of features conditional on the class

$$P(\mathbf{y}|C) = \prod_{j=1}^{J} P(y_j|C)$$

- Gaussian Naive Bayes: for each feature j, the class-conditional distribution is Gaussian

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

GNB, Learning

Use maximum likelihood (sample mean and sample variance)

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

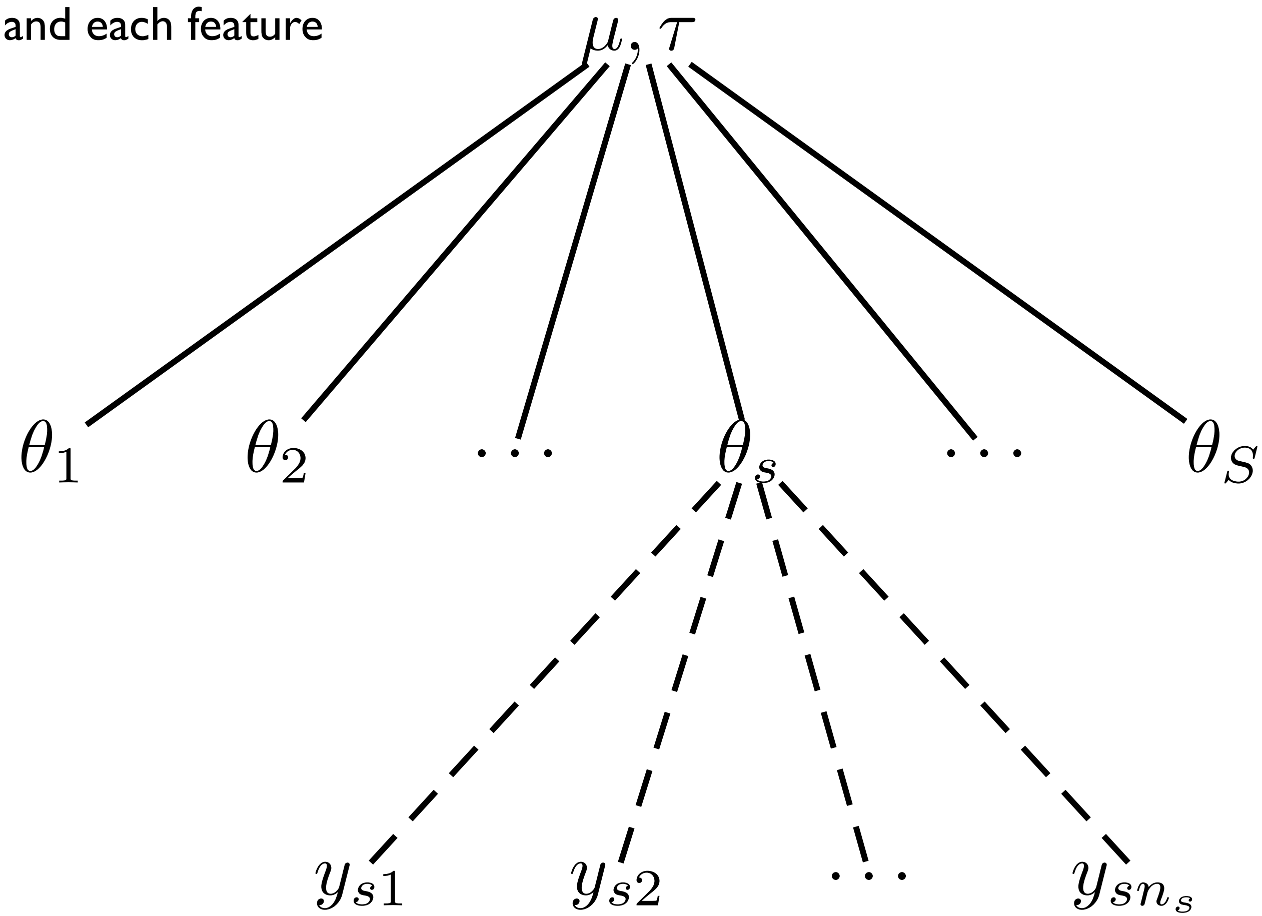
$$(\hat{\sigma}_{sj}^{(k)})^2 = \frac{1}{n_s - 1} \sum_{i=1}^{n_s} (y_{sji}^{(k)} - \hat{\theta}_{sj}^{(k)})^2$$

s: subject
j: feature
i: instance
k: class

For pooled data, aggregate the data over all the subjects (estimates will be the same for all subjects)

Hierarchical Normal Model

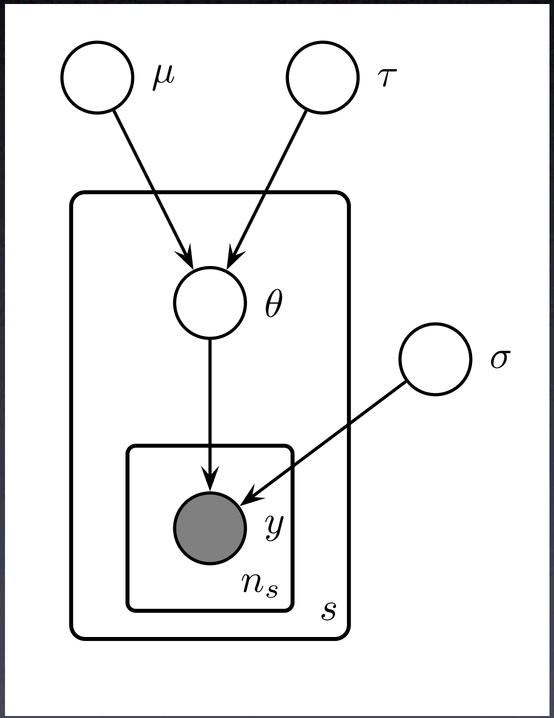
For each class and each feature



Hierarchical Normal Model

- The tool to extend the Gaussian Naive Bayes classifier to handle multiple subjects
- Gelman et al. (2005), also used in Friston et al. (2002) for group analysis (aim: hypothesis testing)
- Modeling Gaussian data for different but related groups; the means for each group has a common Gaussian distribution
- Generative model:

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



s: group (subject)
i: instance

Hierarchical GNB (HGNB)

- Use the hierarchical normal model as a class-conditional generative model for each feature, as a way to integrate data from multiple subjects
- Assume data has been normalized to a common space
- Same variance for all subjects
- Estimate variance separately, taking the median of sample variances for all the subjects

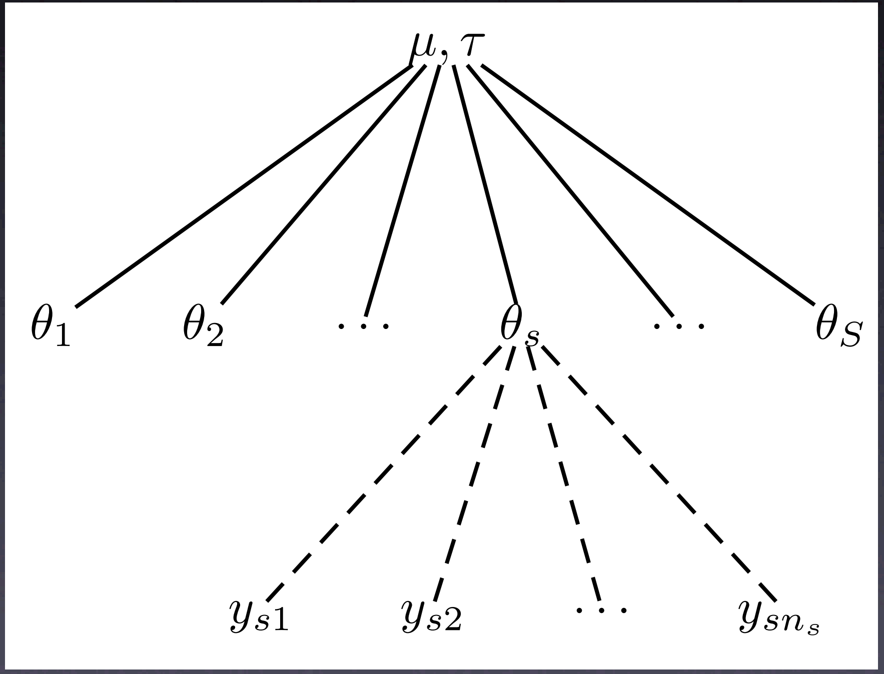
MAP, Empirical Bayes

estimates that (approximately) maximize the marginal likelihood (the probability of data given hyperparameters)

$$\mu_{\mathrm{MP}} = \frac{1}{S}\sum_{s=1}^{S} \bar{y}_{s\cdot}$$

$$\tau^2_{\mathrm{MP}} = \frac{1}{S-1}\sum_{s=1}^{S} (\bar{y}_{s\cdot} - \mu_{\mathrm{MP}})^2$$

MP: point estimate
s: subject



maximum of the posterior of θ_s conditional on the data and the hyperparameters

$$\theta_s = \frac{\frac{n_s}{\sigma^2}\bar{y}_{s\cdot} + \frac{1}{\tau^2_{\mathrm{MP}}}\mu_{\mathrm{MP}}}{\frac{n_s}{\sigma^2} + \frac{1}{\tau^2_{\mathrm{MP}}}}$$

When the number of examples is small, HGNB behaves like GNB on pooled data
When the number of examples is large, HGNB behaves like GNB on the individual subject's data

+

-

*

It is not true that the plus is above the star.

Datasets

Starplus

- Classification of the types of first stimuli (picture or sentence) given a window of fMRI data
- Spatial normalization: use average of voxels in each region of interest (ROI)
- Feature selection: use ROI for visual cortex
- 16 features (each time point is a feature)
- 20 trials per class per subject
- 13 subjects

hammer

palace

Datasets

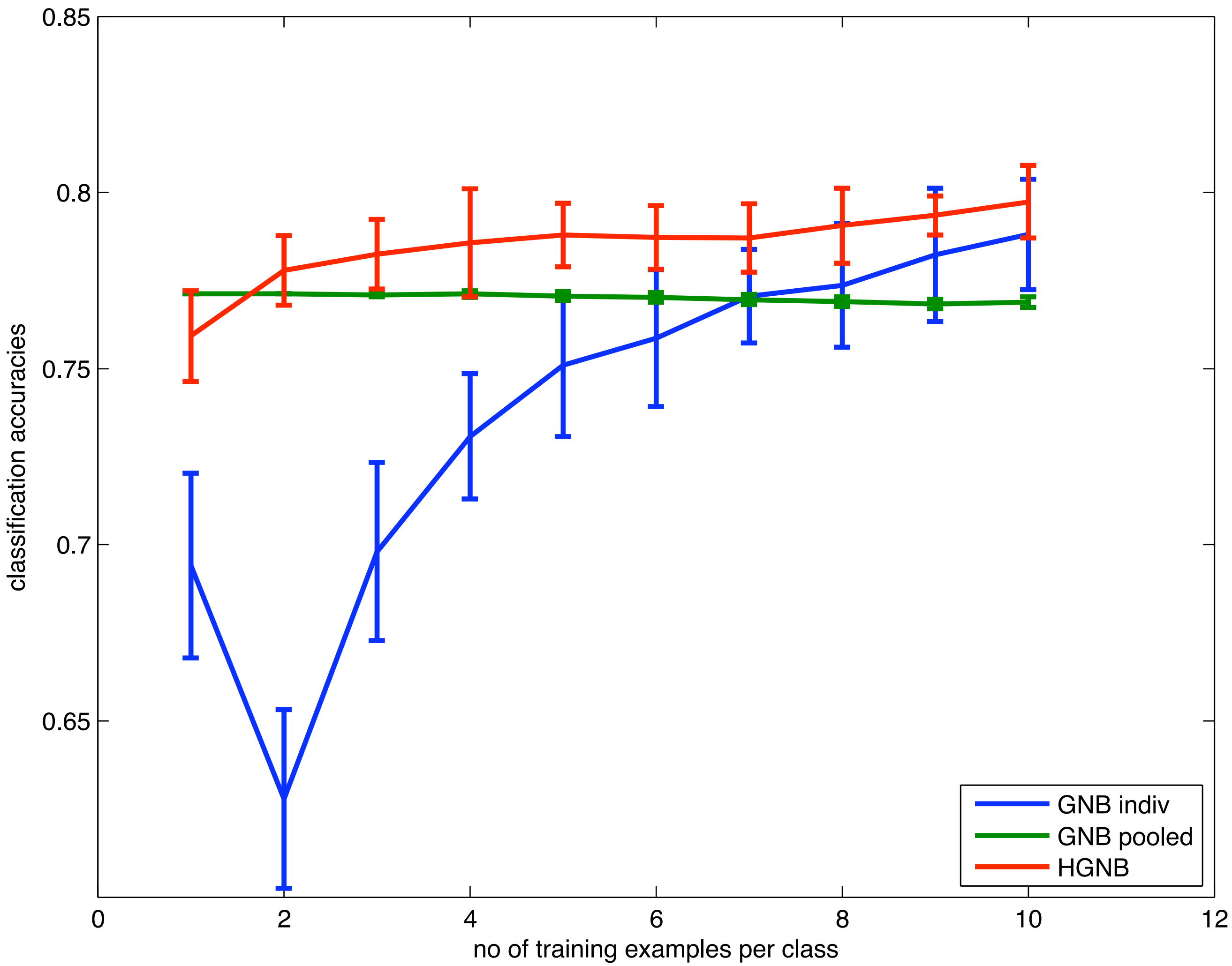
Twocategories

- Classification of the category of word (tools or dwellings) given a window of fMRI data
- Spatial normalization: use transformation to a common brain template (MNI template)
- Feature selection: 300 voxels ranked using Fisher's LDA
- 300 features (averaged over time)
- 42 trials per class per subject
- 6 subjects

Experiment

- Iterate over the subjects, designating the current one as the test subject
- 2-fold cross-validation, varying the number of training examples used from the test subject for each class; fold randomly chosen (repeated several times)
- GNB indiv: GNB learned using data from the test subject only
- GNB pooled: GNB learned using data from the test subject and the other subjects (assuming no inter-subject variability)
- HGNB using data from the test subject and the other subjects

Classification Accuracies, Starplus

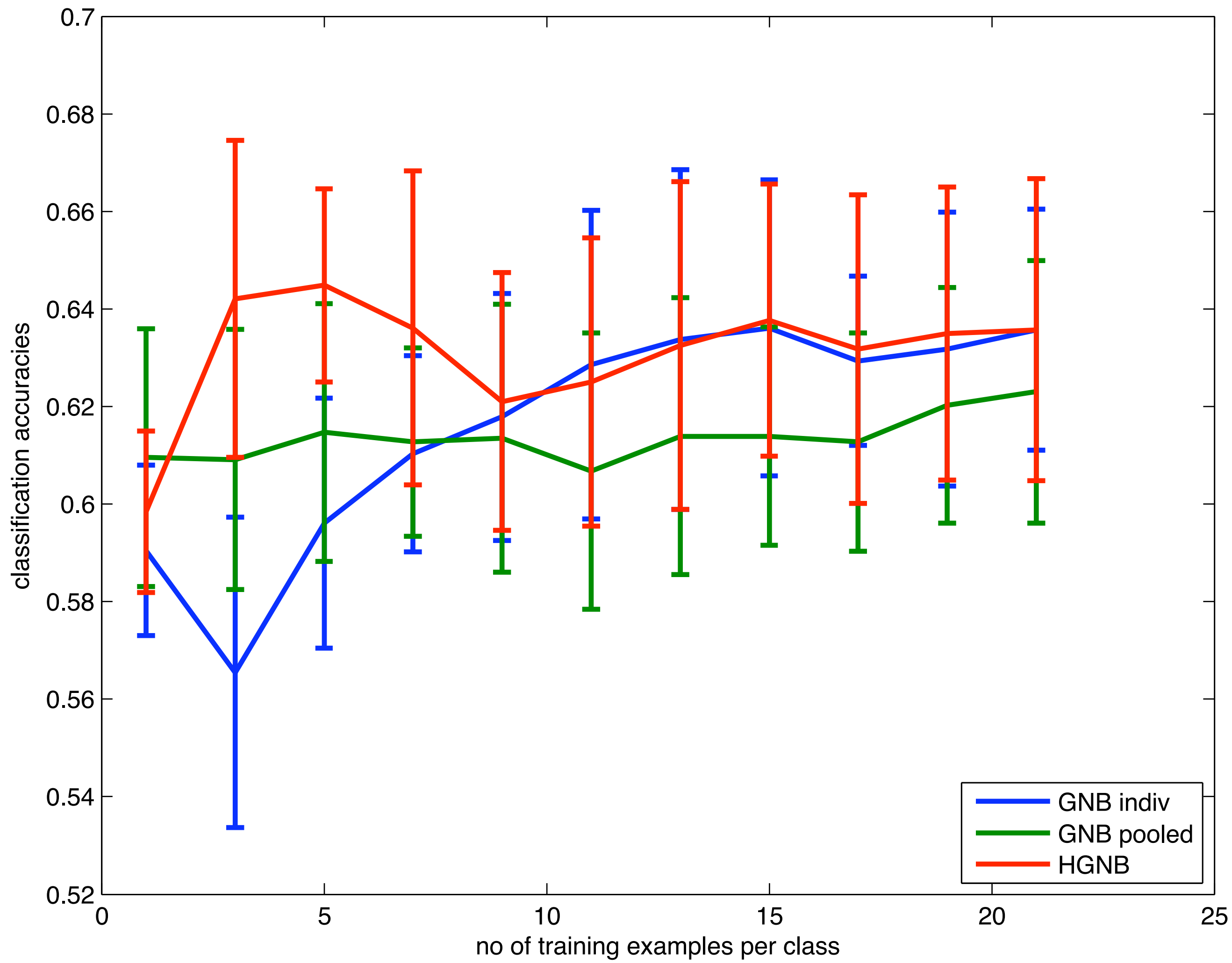


classification accuracies

no of training examples per class

Classification Accuracies, Twocategories

classification accuracies



no of training examples per class

HGNB Recap

- Classifier to combine data across multiple subjects in a study
  - Improvement in predictive performance over separate analyses and pooling data
- Assume that each cognitive task to predict generates similar brain activations on all the subjects
- Show that hierarchical Bayes modeling can model inter-subject variability

Proposed Work

- Goals that have not been addressed by HGNB:
  1. sharing across studies, or both subjects and studies
  2. determining groups to share
  3. determining cross-subject/study commonality of particular cognitive tasks (related to generalisability)
  4. dealing with the distortion caused by normalization
- Work proposed to address the above goals:
  - Variations on HGNB
  - Latent structure in data
  - Accounting for normalization

Variations on HGNB

- Goals (1st and 2nd)
  - sharing across studies, or both subjects and studies
  - determining groups to share
- Variation/extension of the HGNB classifier

Sharing

- Across studies: use the hierarchical normal model to model cross-study variations
- Across subjects and studies:
  - Add another level of the hierarchy (study -> subject -> data or subject -> study -> data)
  - Independent models for subjects and studies

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

Determining Groups to Share

- More reasonable to share across some subjects than others (e.g. subjects with similar clinical conditions)
- Also across some studies than others (not as useful to share data from a study on the visual system and data from a language study)
- Automatically determine grouping
- Clustering, mixture model

$$y_{si} \sim \mathcal{N}(\theta_s, \sigma^2)$$
$$\theta_s \sim \mathcal{N}(\mu^{(k)}, (\tau^{(k)})^2)$$
$$k \sim \mathrm{Multinomial}(\pi_1, \cdots, \pi_K)$$

s: subject
i: instance
k: class

- Dirichlet process mixture model

$$y_{si} \sim \mathcal{N}(\theta_s, \sigma^2)$$
$$\theta_s \sim \mathcal{N}(\mu_s, \tau^2)$$
$$\mu_s \sim G$$
$$G \sim \mathrm{DP}(\alpha, G_0)$$

Latent structure in data

- Goal (3rd): determining cross-subject/study commonality of particular cognitive tasks (related to generalisability)
- Assume there are latent factors underlying the data, with a lot fewer factors than voxels
- Determine commonality by looking at the shared latent factors
  - If the information for a certain cognitive task is shareable among a certain group of subjects and/or studies, there will be common factors for the elements of the group
- Dimensionality reduction, sparsity

Sparse Factor Regression

- West (2003)
- Similar to (probabilistic) factor analysis or PCA, with a regression component

$$\mathbf{x}_i = \mathbf{B}\lambda_i + \nu_i$$
$$y_i = \theta'\lambda_i + \varepsilon_i$$

x_i: i-th instance of data (px1)
y_i: i-th response (scalar)
λ_i: factor for i-th instance (kx1)
B: data factor loading (pxk)
θ: response factor loading (1xk)
ν_i: data noise for i-th instance
ϵ_i: response noise for i-th instance

- k factors, (k << p), k determined in advance
- Sparsity assumption on the factor loading matrix B
- For testing, assume the corresponding y to be missing data

Sparse Factor Regression for fMRI

- The images share a common factor loading matrix B (even for different subjects and studies)
- θ indicates which factors are relevant for prediction (can add sparsity prior for θ)
- Allow θ to be different for different subjects and different studies
- Shareability is determined by how many non-zero elements of θ are shared
- How many factors to use? May use the Indian buffet process (Griffiths and Ghahramani, (2006)) as a prior, which can also facilitate sparsity of factors

Topics

- Can think of latent factors in terms of topics in a topic model (e.g. Latent Dirichlet Allocation (LDA), Blei et al. (2003))
- LDA:
  - A document is a mixture of topics
  - A topic specifies a distribution over words
- LDA for fMRI data:
  - A brain activation image is a mixture of latent factors
  - A latent factor specifies a distribution over voxel activations

LDA for fMRI Data

- Sparsity: each latent factor determines the distribution for only a subset of the voxels
- Because each image is a mixture of latent factors, shareability is determined by the number of predictive latent factors shared
- Details need to be worked out

Accounting for Normalization

- Goal (4th): dealing with the distortion caused by normalization
- Incorporate the uncertainties introduced by normalization in the prediction or analysis
- Approach:
  - probabilistic voxel correspondence

Probabilistic voxel correspondence

- Probabilistic model for normalization
- Model the correspondence among voxels across different brains
- Use a probabilistic atlas as a prior
    - Available from the International Consortium for Brain Mapping (ICBM)
- Incorporate information about the brain structure (available from structural images)
- A lot still needs to be investigated

Schedule

- December 2007: variations on HGNB and latent structure in fMRI data
  - variations on HGNB
  - sparse factor regression
  - formulate topic model for fMRI
- December 2008: accounting for normalization
  - probabilistic voxel correspondence