Generalized Correspondence-LDA Models (GC-LDA) for Identifying Functional Regions in the Brain

Timothy N. Rubin
SurveyMonkey

Oluwasanmi Koyejo
Univ. of Illinois, Urbana-Champaign

Michael N. Jones
Indiana University

Tal Yarkoni
University of Texas at Austin

Abstract

This paper presents Generalized Correspondence-LDA (GC-LDA), a generalization of the Correspondence-LDA model that allows for variable spatial representations to be associated with topics, and increased flexibility in terms of the strength of the correspondence between data types induced by the model. We present three variants of GC-LDA, each of which associates topics with a different spatial representation, and apply them to a corpus of neuroimaging data. In the context of this dataset, each topic corresponds to a functional brain region, where the region's spatial extent is captured by a probability distribution over neural activity, and the region's cognitive function is captured by a probability distribution over linguistic terms. We illustrate the qualitative improvements offered by GC-LDA in terms of the types of topics extracted with alternative spatial representations, as well as the model's ability to incorporate a-priori knowledge from the neuroimaging literature. We furthermore demonstrate that the novel features of GC-LDA improve predictions for missing data.

1 Introduction

One primary goal of cognitive neuroscience is to find a mapping from neural activity onto cognitive processes–that is, to identify functional networks in the brain and the role they play in supporting macroscopic functions. A major milestone towards this goal would be the creation of a "functional-anatomical atlas" of human cognition, where, for each putative cognitive function, one could identify the regions and brain networks within the region that support the function.

Efforts to create such functional brain atlases are increasingly common in recent years. Most studies have proceeded by applying dimensionality reduction or source decomposition methods such as Independent Component Analysis (ICA) [4] and clustering analysis [9] to large fMRI datasets such as the Human Connectome Project [10] or the meta-analytic BrainMap database [8]. While such work has provided valuable insights, these approaches also have significant drawbacks. In particular, they typically do not jointly estimate regions along with their mapping onto cognitive processes. Instead, they first extract a set of neural regions (e.g., via ICA performed on resting-state data), and then in a separate stage—if at all—estimate a mapping onto cognitive functions. Such approaches do not allow information regarding cognitive function to constrain the spatial characterization of the regions. Moreover, many data-driven parcellation approaches involve a hard assignment of each brain voxel to a single parcel or cluster, an assumption that violates the many-to-many nature of functional brain networks. Ideally, a functional-anatomical atlas of human cognition should allow the spatial and functional correlates of each atom or unit to be *jointly* characterized, where the function of each region constrains its spatial boundaries, and vice-versa.

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

In the current work, we propose Generalized Correspondence LDA (GC-LDA) – a novel generalization of the Correspondence-LDA model [2] for modeling multiple data types, where one data type describes the other. While the proposed approach is general and can be applied to a variety of data, our work is motivated by its application to neuroimaging meta-analysis. To that end, we consider several GC-LDA models that we apply to the Neurosynth [12] corpus, consisting of the document text and neural activation data from a large body of neuroimaging publications. In this context, the models extract a set of neural "topics", where each topic corresponds to a functional brain region. For each topic, the model describes its spatial extent (captured via probability distributions over neural activation) and cognitive function (captured via probability distributions over linguistic terms). These models provide a novel approach for jointly identifying the spatial location and cognitive mapping of functional brain regions, that is consistent with the many-to-many nature of functional brain networks. Furthermore, to the best of our knowledge, one of the GC-LDA variants provides the first automated measure of the lateralization of cognitive functions based on large-scale imaging data.

The GC-LDA and Correspondence-LDA models are extensions of Latent Dirichlet Allocation (LDA) [3]. Several Bayesian methods with similarities (or equivalences) to LDA have been applied to different types of neuroimaging data. Poldrack et al. (2012) used standard LDA to derive topics from the text of the Neurosynth database and then projected the topics onto activation space based on document-topic loadings [7]. Yeo et al. (2014) used a variant of the Author-Topic model to model the BrainMap Database [13]. Manning et al. (2014) described a Bayesian method "Topographic Factor Analysis" to identify brain regions based on the raw fMRI images (but not text) extracted from a set of controlled experiments, which can later be mapped on functional categories [5].

Relative to the Correspondence-LDA model, the GC-LDA model incorporates: (i) the ability to associate different types of spatial distributions with each topic, (ii) flexibility in how strictly the model enforces a correspondence between the textual and spatial data within each document, and (iii) the ability to incorporate a-priori spatial structure, e.g., encouraging relatively homologous functional regions located in each brain hemisphere. As we show, these aspects of GC-LDA have a significant effect on the quality of the estimated topics, as well as on the models' ability to predict missing data.

2 Models

In this paper we propose a set of unsupervised generative models based on the Correspondence-LDA model [2] that we use to jointly model text and brain activations from the Neurosynth meta-analytic database [12]. Each of these models, as well as Correspondence-LDA, can be viewed as special cases of a broader model that we will refer to as Generalized Correspondence-LDA (GC-LDA). In the section below, we describe the GC-LDA model and its relationship to Correspondence-LDA. We then detail the specific instances of the model that we use throughout the remainder of the paper. A summary of the notation used throughout the paper is provided in Table 1.

2.1 Generalized Correspondence LDA (GC-LDA)

Each document d in the corpus is comprised of two types of data: a set of word tokens $\left\{w_1^{(d)}, w_2^{(d)}, ..., w_{N_w^{(d)}}^{(d)}\right\}$ consisting of unigrams and/or n-grams, and a set of peak activation tokens $\left\{x_1^{(d)}, x_2^{(d)}, ..., x_{N_x^{(d)}}^{(d)}\right\}$, where $N_w^{(d)}$ and $N_x^{(d)}$ are the number of word and activation tokens in document d, respectively. In the target application, each token x_i is a 3-dimensional vector corresponding to the peak activation coordinates of a value reported in fMRI publications. However, we note that this model can be directly applied to other types of data, such as segmented images, where each x_i corresponds to a vector of real-valued features extracted from each image segment (c.f. [2]).

GC-LDA is described by the following generative process (depicted in Figure 1.A):

1. For each topic $t \in \left\{1, ..., T\right\}$[1]:
   (a) Sample a Multinomial distribution over word types $\phi^{(t)} \sim$ Dirichlet(β)
2. For each document $d \in \{1, ..., D\}$:

[1]To make the model fully generative, one could additionally put a prior on the spatial distribution parameters $\Lambda^{(t)}$ and sample them. For the purposes of the present paper we do not specify a prior on these parameters, and therefore leave this out of the generative process.

Table 1: Table of notation used throughout the paper

Model specification	
Notation	Meaning
w_i, x_i	The ith word token and peak activation token in the corpus, respectively
$N_w^{(d)}$, $N_x^{(d)}$	The number of word tokens and peak activation tokens in document d, respectively
D	The number of documents in the corpus
T	The number of topics in the model
R	The number of components/subregions in each topic's spatial distribution (subregions model)
z_i	Indicator variable assigning word token w_i to a topic
y_i	Indicator variable assigning activation token x_i to a topic
$\mathbf{z}^{(d)}$, $\mathbf{y}^{(d)}$	The set of all indicator variables for word tokens and activation tokens in document d
N_{td}^{YD}	The number of activation tokens within document d that are assigned to topic t
c_i	Indicator variable assigning activation token y_i to a subregion (subregion models)
$\Lambda^{(t)}$	Placeholder for all spatial parameters for topic t
$\mu^{(t)}$, $\sigma^{(t)}$	Gaussian parameters for topic t
$\mu_r^{(t)}$, $\sigma_r^{(t)}$	Gaussian parameters for subregion r in topic t (subregion models)
$\phi^{(t)}$	Multinomial distribution over word types for topic t
$\phi_w^{(t)}$	Probability of word type w given topic t
$\theta^{(d)}$	Multinomial distribution over topics for document d
$\theta_t^{(d)}$	Probability of topic t given document d
$\pi^{(t)}$	Multinomial distribution over subregions for topic t (subregion models)
$\pi_r^{(t)}$	Probability of subregion r given topic t (subregion models)
β, α, γ	Model hyperparameters
δ	Model hyperparameter (subregion models)

(a) Sample a Multinomial distribution over topics $\theta^{(d)} \sim$ Dirichlet(α)

(b) For each peak activation token $x_i,\ i \in \{1, ..., N_x^{(d)}\}$:

  i. Sample indicator variable y_i from Multinomial($\theta^{(d)}$)

  ii. Sample a peak activation token x_i from the spatial distribution: $x_i \sim f(\Lambda^{(y_i)})$

(c) For each word token $w_i,\ i \in \{1, ..., N_w^{(d)}\}$:

  i. Sample indicator variable z_i from Multinomial$\left(\frac{N_{1d}^{YD}+\gamma}{N_x^{(d)}+\gamma*T}, \frac{N_{2d}^{YD}+\gamma}{N_x^{(d)}+\gamma*T}, ..., \frac{N_{Td}^{YD}+\gamma}{N_x^{(d)}+\gamma*T}\right)$, where N_{td}^{YD} is the number of activation tokens y in document d that are assigned to topic t, $N_x^{(d)}$ is the total number of activation tokens in d, and γ is a hyperparameter

  ii. Sample a word token w_i from Multinomial($\phi^{(z_i)}$)

Intuitively, in the present application of GC-LDA, each topic corresponds to a functional region of the brain, where the linguistic features for the topic describe the cognitive processes associated with the spatial distribution of the topic. The resulting joint distribution of all observed peak activation tokens, word tokens, and latent parameters for each individual document in the GC-LDA model is as follows:

$$p(\mathbf{x}, \mathbf{w}, \mathbf{z}, \mathbf{y}, \theta) = p(\theta|\alpha) \cdot \left(\prod_{i=1}^{N_x^{(d)}} p(y_i|\theta^{(d)}) p(x_i|\Lambda^{(y_i)}) \right) \cdot \left(\prod_{j=1}^{N_w^{(d)}} p(z_j|\mathbf{y}^{(d)}, \gamma) p(w_j|\phi^{(z_j)}) \right) \quad (1)$$

Note that when $\gamma = 0$, and the spatial distribution for each topic is specified as a single multivariate Gaussian distribution, the model becomes equivalent to a smoothed version of the Correspondence LDA model described by Blei & Jordan (2003) [2].[2]

[2]We note that [2] uses a different generative description for how the z_i variables are sampled conditional on the $y_i^{(d)}$ indicator variables; in [2], z_i is sampled uniformly from $(1, ..., N_y^{(d)})$, and then w_i is sampled from the multinomial distribution of the topic $y_i^{(d)}$ that z_i points to. This ends up being functionally equivalent to the generative description for z_i given here when $\gamma = 0$. Additionally, in [2], no prior is put on $\phi^{(t)}$, unlike in GC-LDA. Therefore, when using GC-LDA with a single multivariate Gaussian and $\gamma = 0$, it is equivalent to a *smoothed* version of Correspondence-LDA. Dirichlet priors have been demonstrated to be beneficial to model performance [1], so including a prior on $\phi^{(t)}$ in GC-LDA should have a positive impact.

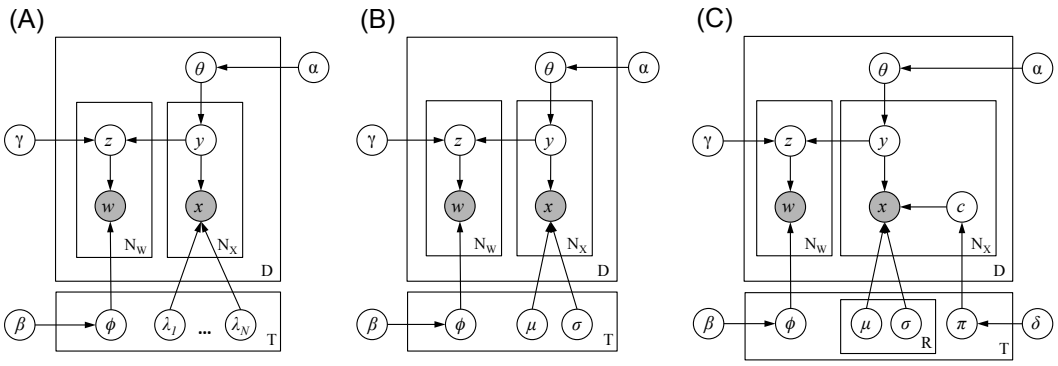

Figure 1: **(A)** Graphical model for the Generalized Correspondence-LDA model, GC-LDA. **(B)** Graphical model for GC-LDA with spatial distributions modeled as a single multivariate Gaussian (equivalent to a smoothed version of Correspondence-LDA if $\gamma = 0$)[2]. **(C)** Graphical model for GC-LDA with subregions, with spatial distributions modeled as a mixture of multivariate Gaussians

A key aspect of this model is that it induces a correspondence between the number of activation tokens and the number of word tokens within a document that will be assigned to the same topic. The hyperparameter γ controls the strength of this correspondence. If $\gamma = 0$, then there is zero probability that a word for document d will be sampled from topic t if no peak activations in d were sampled from t. As γ becomes larger, this constraint is relaxed. Although intuitively one might want γ to be zero in order to maximize the correspondence between the spatial and linguistic information, we have found that setting $\gamma > 0$ leads to significantly better model performance. We conjecture that using a non-zero γ allows the parameter space to be more efficiently explored during inference, and that it improves the model's ability to handle data sparsity and noise in high dimensional spaces, similar to the role that the α and β hyperparameters serve in standard LDA [1].

2.2 Versions of GC-LDA Employed in Current Paper

There are multiple reasonable choices for the spatial distribution $p(x_i|\Lambda^{(y_i)})$ in GC-LDA, depending upon the application and the goals of the modeler. For the purposes of the current paper, we considered three variants that are motivated by the target application. The **first model** shown in Figure 1.B employs a single multivariate Gaussian distribution for each topic's spatial distribution – and is therefore equivalent to a smoothed version of Correspondence-LDA if setting $\gamma = 0$. The generative process for this model is the same as specified above, with generative step (b.ii) modified as follows: Sample peak activation token x_i from from a Gaussian distribution with parameters $\mu^{(y_i)}$ and $\sigma^{(y_i)}$. We refer to this model as the "no-subregions" model.

The **second model** and **third model** both employ Gaussian mixtures with $R = 2$ components for each topic's spatial distribution, and are shown in Figure 1.C. Employing a Gaussian mixture gives the model more flexibility in terms of the types of spatial distributions that can be associated with a topic. This is notably useful in modeling spatial distributions associated with neural activity, as it allows the model to learn topics where a single cognitive function (captured by the linguistic distribution) is associated with spatially discontiguous patterns of activations. In the second GC-LDA model we present—which we refer to as the "unconstrained subregions" model—the Gaussian mixture components are unconstrained. In the third version of GC-LDA—which we refer to as the "constrained subregions" model—the Gaussian components are constrained to have symmetric means with respect to their distance from the origin along the horizontal spatial axis (a plane corresponding to the longitudinal fissure in the brain). This constraint is consistent with results from meta-analyses of the fMRI literature, where most studied functions display a high degree of bilateral symmetry [6, 12].

The use of mixture models for representing the spatial distribution in GC-LDA requires the additional parameters c, π, and hyperparameter δ, as well as additional modifications to the description of the generative process. Each topic's spatial distribution in these models is now associated with a multinomial probability distribution $\pi^{(t)}$ giving the probability of sampling each component r from each topic t, where $\pi_r^{(t)}$ is the probability of sampling the rth component (which we will refer to as a

subregion) from the tth topic. Variable c_i is an indicator variable that assigns each activation token x_i to a subregion r of the topic to which it is assigned via y_i. A full description of the generative process for these models is provided in Section 1 of the supplementary materials[3].

2.3 Inference for GC-LDA

Exact probabilistic inference for the GC-LDA model is intractable. We employed collapsed Gibbs sampling for posterior inference – collapsing out $\theta^{(d)}$, $\phi^{(t)}$, and $\pi^{(t)}$ while sampling the indicator variables y_i, z_i and c_i. Spatial distribution parameters $\Lambda^{(t)}$ are estimated via maximum likelihood. The per-iteration computational complexity of inference is $O(T(N_W + N_X R))$, where T is the number of topics, R is the number of subregions, and N_W and N_X are the total number of word tokens and activation tokens in the corpus, respectively. Details of the inference methods and sampling equations are provided in Section 2 of the supplement.

3 Experimental Evaluation

We refer to the three versions of GC-LDA described in Section 2 as (1) the "no subregions" model, for the model in which each topic's spatial distribution is a single multivariate Gaussian distribution, (2) the "unconstrained subregions" model, for the model in which each topic's spatial distribution is a mixture of $R = 2$ unconstrained Gaussian distributions, and (3) the "constrained subregions" model, for the model in which each topic's spatial distribution is a mixture of $R = 2$ Gaussian distributions whose means are constrained to be symmetric along the horizontal spatial dimension with respect to their distance from the origin.

Our empirical evaluations of the GC-LDA model are based on the application of these models to the Neurosynth meta-analytic database [12]. We first illustrate and contrast the qualitative properties of topics that are extracted by the three versions of GC-LDA[4]. We then provide a quantitative model comparison, in which the models are evaluated in terms of their ability to predict held out data. These results highlight the promise of GC-LDA and this type of modeling for jointly extracting the spatial extent and cognitive functions of neuroanatomical brain regions.

Neurosynth Database: Neurosynth [12] is a publicly available database consisting of data automatically extracted from a large collection of functional magnetic resonance imaging (fMRI) publications[5]. For each publication, the database contains the abstract text and all reported 3-dimensional peak activation coordinates (in MNI space) in the study. The text was pre-processed to remove common stop-words. For the version of the Neurosynth database employed in the current paper, there were 11,362 total publications, which had on average 35 peak activation tokens and 46 word tokens after preprocessing (corresponding to approximately 400k activation and 520k word tokens in total).

3.1 Visualizing GC-LDA Topics

In Figure 2 we present several illustrative examples of topics for all three GC-LDA variants that we considered. For each topic, we illustrate the topic's distribution over word types via a word cloud, where the sizes of words are proportional to their probabilities $\phi_w^{(t)}$ in the model. Each topic's spatial distribution over neural activations is illustrated via a kernel-smoothed representation of all activation tokens that were assigned to the topic, overlaid on an image of the brain. For the models that represent spatial distributions using Gaussian mixtures (the unconstrained and constrained subregions models), activations are color-coded based on which subregion they are assigned to, and the mixture weights for the subregions $\pi_r^{(t)}$ are depicted above the activation image on the left. In the constrained subregions model (where the means of the two Gaussians were constrained to be symmetric along the horizontal axis) the two subregions correspond to a 'left' and 'right' hemisphere subregion. The following parameter settings were used for generating the images in Figure 2: $T = 200$, $\alpha = .1$, $\beta = .01$, $\gamma = .01$, and for the models with subregions, $\delta = 1.0$.

[3] Note that these models are still instances of GC-LDA as presented in Figure 1.1; they can be equivalently formulated by marginalizing out the c_i variables, such that the probability $f(x_i|\Lambda^{(t)})$ depends directly on the parameters of each component, and the component probabilities given by $\pi^{(t)}$.

[4] A brief discussion of the stability of topics extracted by GC-LDA is provided in Section 3 of the supplement

[5] Additional details and Neurosynth data can be found at `http://neurosynth.org/`

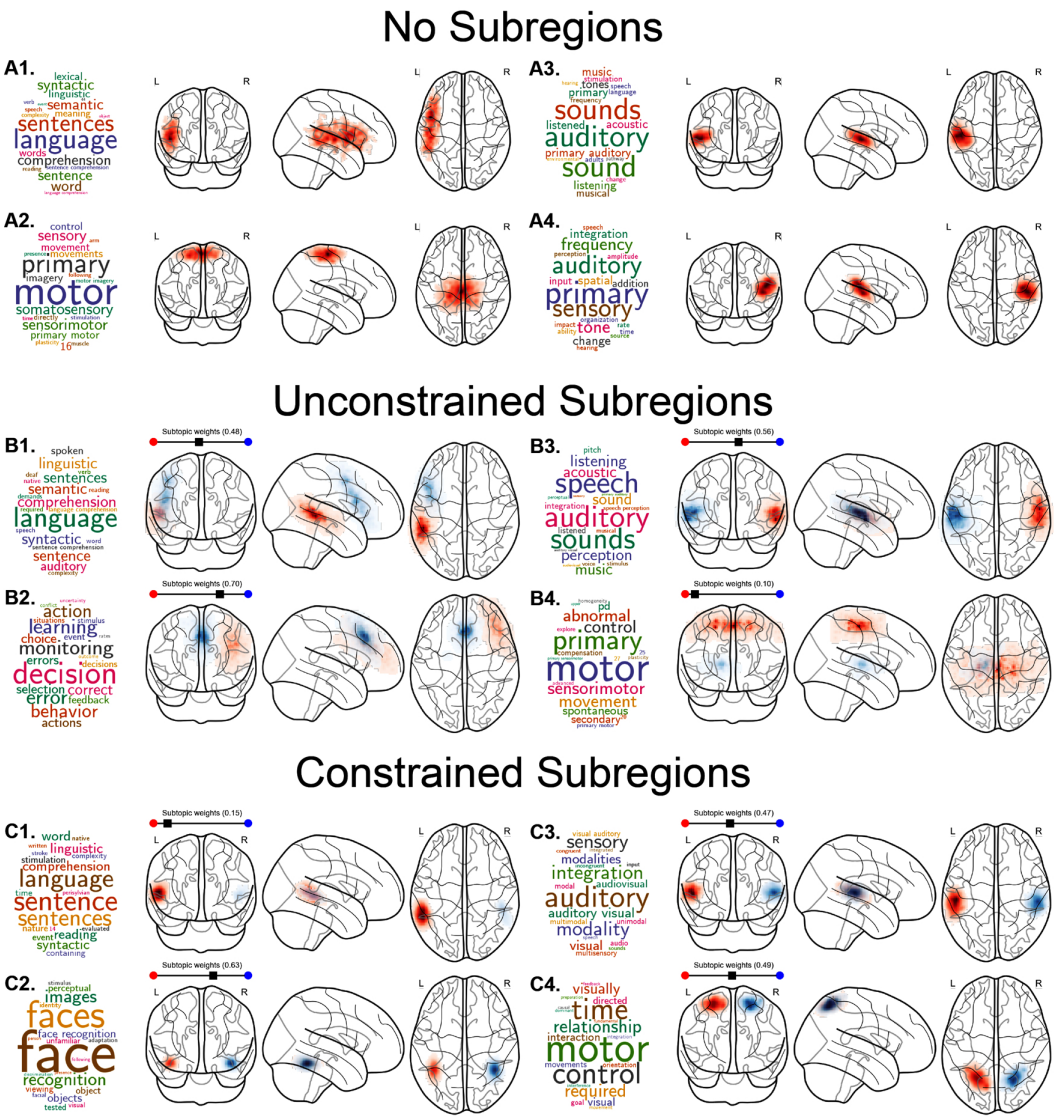

Figure 2: Illustrative examples of topics extracted for the three GC-LDA variants. Probability distributions over word types $\phi^{(t)}$ are represented via word clouds, where word sizes are proportional to $\phi_w^{(t)}$. Spatial distributions are illustrated using kernel-smoothed representations of all activation tokens assigned to each topic. For the models with subregions, each activation token's color (blue or red) corresponds to the subregion r that the token is assigned to.

For nearly all of the topics shown in Figure 2, the spatial and linguistic distributions closely correspond to functional regions that are extensively described in the literature (e.g., motor function in primary motor cortex; face processing in the fusiform gyrus, etc.). We note that a key feature of all versions of the GC-LDA model, relative to the majority of existing methods in the literature, is that the model is able to capture the one-to-many mapping from neural regions onto cognitive functions. For example, in all model variants, we observe topics corresponding to auditory processing and language processing (e.g., the topics shown in panels B1 and B3 for the subregions model). While these cognitive processes are distinct, they have partial overlap with respect to the brain networks they recruit – specifically, the superior temporal sulcus in the left hemisphere.

For functional regions that are relatively medial, the no-subregions model is able to capture bilateral homologues by consolidating them into a single distribution (e.g., the topic shown in A2, which spans the medial primary somatomotor cortex in both hemispheres). However, for functional regions that are more laterally localized, the model cannot capture bilateral homologues using a single topic. For cognitive processes that are highly lateralized (such as language processing, shown in A1, B1

and C1) this poses no concern. However, for functional regions that are laterally distant and do have spatial symmetry, the model ends up distributing the functional region across multiple topics–see, e.g., the topics shown in A3 and A4 in the no-subregions model, which correspond to the auditory cortex in the left and right hemisphere respectively. Given that these two topics (and many other pairs of topics that are not shown) correspond to a single cognitive function, it would be preferable if they were represented using a single topic. This can potentially be achieved by increasing the flexibility of the spatial representations associated with each topic, such that the model can capture functional regions with distant lateral symmetry or other discontiguous spatial features using a single topic. This motivates the unconstrained and constrained subregions models, in which topic's spatial distributions are represented by Gaussian mixtures.

In Figure 2, the topics in panels B3 and C3 illustrate how the subregions models are able to handle symmetric functional regions that are located on the lateral surface of the brain. The lexical distribution for each of these individual topics in the subregions models is similar to that of both the topics shown in A3 and A4 of the no-subregions model. However, the spatial distributions in B3 and C3 each capture a summation of the two topics from the no subregions model. In the case of the constrained subregion model, the symmetry between the means of the spatial distributions for the subregions is enforced, while for the unconstrained model the symmetry is data-driven and falls out of the model.

We note that while the unconstrained subregions model picks up spatial symmetry in a significant subset of topics, it does not always do so. In the case of language processing (panel A1), the lack of spatial symmetry is consistent with a large fMRI literature demonstrating that language processing is highly left-lateralized [11]. And in fact, the two subregions in this topic correspond approximately to Wernicke's and Broca's areas, which are integral to language comprehension and production, respectively. In other cases, (e.g., the topics in panels B2 and B4), the unconstrained subregions model partially captures spatial symmetry with a highly-weighted subregion near the horizontal midpoint, but also has an additional low-weighted region that is lateralized. While this result is not necessarily wrong per se, it is somewhat inelegant from a neurobiological standpoint. Moreover, there are theoretical reasons to prefer a model in which subregions are always laterally-symmetrical. Specifically, in instances where the subregions are symmetric (the topic in panel B3 for the unconstrained subregions model and all topics for the constrained subregions model), the subregion weights provide a measure of the relative lateralization of function. For example, the language topic in panel C1 of the constrained subregions model illustrates that while there is neural activation corresponding to linguistic processing in the right hemisphere of the brain, the function is strongly left-lateralized (and vice-versa for face processing, illustrated in panel C2). By enforcing the lateral symmetry in the constrained subregions model, the subregion weights $\pi_r^{(t)}$ (illustrated above the left activation images) for each topic inherently correspond to an automated measure of the lateralization of the topic's function. Thus, the constrained model produces what is, to our knowledge, the first data-driven estimation of region-level functional hemispheric asymmetry across the whole brain.

3.2 Predicting Held Out Data

This section describes quantitative comparisons between three GC-LDA models in terms of their ability to predict held-out data. We split the Neurosynth dataset into a training and test set, where approximately 20% of all data in the corpus was put into the test set. For each document, we randomly removed $\left\lfloor .2N_x^{(d)} \right\rfloor$ peak activation tokens and $\left\lfloor .2N_w^{(d)} \right\rfloor$ word tokens from each document. We trained the models on the remaining data, and then for each model we computed the log-likelihood of the test data, both for the word tokens and peak tokens.

The space of possible hyperparameters to explore in GC-LDA is vast, so we restrict our comparison to the aspects of the model which are novel relative to the original Correspondence-LDA model. Specifically, for all three model variants, we compared the log-likelihood of the test data across different values of γ, where $\gamma \in \{0, 0.001, 0.01, 0.1, 1\}$. We note again here that the no-subregions model with $\gamma = 0$ is equivalent to a smoothed version of Correspondence-LDA [2] (see footnote 2 for additional clarification). The remainder of the parameters were fixed as follows (chosen based on a combination of precedent from the topic modeling literature and preliminary model exploration): $T = 100$, $\alpha = .1$, and $\beta = .01$ for all models, and $\delta = 1.0$ for the models with subregions. All models were trained for 1000 iterations.

Figure 3 presents the held out log-likelihoods for all models across different settings of γ, in terms of (i) the total log-likelihood for both activation tokens and word tokens (left) (ii) log-likelihood for activation tokens only (middle), and (iii) log likelihood for word tokens only (right). For both activation tokens and word tokens, for all three versions of GC-LDA, using a non-zero γ leads to significant improvement in performance. In terms of predicting activation tokens alone, there is a monotonic relationship between the size of γ and log-likelihood. This is unsurprising, since increasing γ reduces the extent that word tokens constrain the spatial fit of the model. In terms of predicting word tokens (and overall log-likelihood), the effect of γ shows an inverted-U function, with the best performance in the range of .01 to .1. These patterns were consistent across all three variants of GC-LDA. Taken together, our results suggest that using a non-zero γ results in a significant improvement over the Correspondence-LDA model.

In terms of comparisons across model variants, we found that both subregions models were significant improvements over the no-subregions models in terms of total log-likelihood, although the no-subregions model performed slightly better than the constrained subregions model at predicting word tokens. In terms of the two subregions models, performance is overall fairly similar. Generally, the constrained subregions model performs slightly better than the unconstrained model in terms of predicting peak tokens, but slightly worse in terms of predicting word tokens. The differences between the two subregions models in terms of total log-likelihood were negligible. These results do not provide a strong statistical case for choosing one subregions model over the other; instead, they suggest that the modeler ought to choose between models based on their respective theoretical or qualitative properties (e.g., biological plausibility, as discussed in Section 3.1).

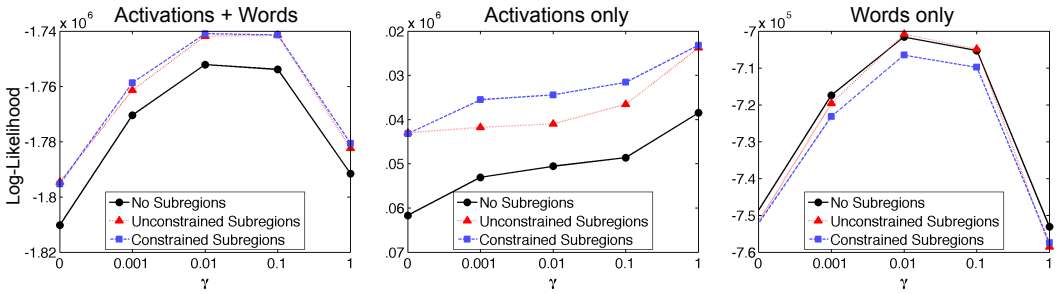

Figure 3: Log Likelihoods of held out data for the three GC-LDA models as a function of model parameter γ. Left: total log-likelihood (activation tokens + word tokens). Middle: log-likelihood of activation tokens only. Right: log-likelihood of word tokens only.

4 Summary

We have presented generalized correspondence LDA (GC-LDA) – a generalization of the Correspondence-LDA model, with a focus on three variants that capture spatial properties motivated by neuroimaging applications. We illustrated how this model can be applied to a novel type of metadata—namely, the spatial peak activation coordinates reported in fMRI publications—and how it can be used to generate a relatively comprehensive atlas of functional brain regions. Our quantitative comparisons demonstrate that the GC-LDA model outperforms the original Correspondence-LDA model at predicting both missing word tokens and missing activation peak tokens. This improvement was demonstrated in terms of both the introduction of the γ parameter, and with respect to alternative parameterizations of topics' spatial distributions.

Beyond these quantitative results, our qualitative analysis demonstrates that the model can recover interpretable topics corresponding closely to known functional regions of the brain. We also showed that one variant of the model can recover known features regarding the hemispheric lateralization of certain cognitive functions. These models show promise for the field of cognitive neuroscience, both for summarizing existing results and for generating novel hypotheses. We also expect that novel features of GC-LDA can be carried over to other extensions of Correspondence-LDA in the literature.

In future work, we plan to explore other spatial variants of these models that may better capture the morphological features of distinct brain regions – e.g., using hierarchical priors that can capture the hierarchical organization of brain systems. We also hope to improve the model by incorporating features such as the correlation between topics. Applications and extensions of our approach for more standard image processing applications may also be a fruitful area of research.

References

[1] Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. On smoothing and inference for topic models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 27–34. AUAI Press, 2009.

[2] David M Blei and Michael I Jordan. Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 127–134. ACM, 2003.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[4] Vince D Calhoun, Jingyu Liu, and Tülay Adalı. A review of group ica for fmri data and ica for joint inference of imaging, genetic, and erp data. *Neuroimage*, 45(1):S163–S172, 2009.

[5] Jeremy R Manning, Rajesh Ranganath, Kenneth A Norman, and David M Blei. Topographic factor analysis: a bayesian model for inferring brain networks from neural data. *PloS one*, 9(5):e94914, 2014.

[6] Adrian M Owen, Kathryn M McMillan, Angela R Laird, and Ed Bullmore. N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human brain mapping*, 25(1):46–59, 2005.

[7] Russell A Poldrack, Jeanette A Mumford, Tom Schonberg, Donald Kalar, Bishal Barman, and Tal Yarkoni. Discovering relations between mind, brain, and mental disorders using topic mapping. *PLoS Comput Biol*, 8(10):e1002707, 2012.

[8] Stephen M Smith, Peter T Fox, Karla L Miller, David C Glahn, P Mickle Fox, Clare E Mackay, Nicola Filippini, Kate E Watkins, Roberto Toro, Angela R Laird, et al. Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences*, 106(31):13040–13045, 2009.

[9] Bertrand Thirion, Gaël Varoquaux, Elvis Dohmatob, and Jean-Baptiste Poline. Which fmri clustering gives good brain parcellations? *Frontiers in neuroscience*, 8(167):13, 2014.

[10] David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, WU-Minn HCP Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.

[11] Mathieu Vigneau, Virginie Beaucousin, Pierre-Yves Herve, Hugues Duffau, Fabrice Crivello, Olivier Houde, Bernard Mazoyer, and Nathalie Tzourio-Mazoyer. Meta-analyzing left hemisphere language areas: phonology, semantics, and sentence processing. *Neuroimage*, 30(4):1414–1432, 2006.

[12] Tal Yarkoni, Russell A Poldrack, Thomas E Nichols, David C Van Essen, and Tor D Wager. Large-scale automated synthesis of human functional neuroimaging data. *Nature methods*, 8(8):665–670, 2011.

[13] BT Thomas Yeo, Fenna M Krienen, Simon B Eickhoff, Siti N Yaakub, Peter T Fox, Randy L Buckner, Christopher L Asplund, and Michael WL Chee. Functional specialization and flexibility in human association cortex. *Cerebral Cortex*, page bhu217, 2014.

Supplement: Generalized Correspondence-LDA Models (GC-LDA) for Identifying Functional Regions in the Brain

Section 1 of this supplement presents the generative process for the GC-LDA variants that use Gaussian mixtures to model each topic's spatial component. Section 2 provides inference details for all versions of GC-LDA considered in this paper[1]. Section 3 provides an analysis of the stability of the topics extracted by GC-LDA.

The notation used for both model specification and inference throughout the supplement is summarized in Table 1.

1 Generative Process and Joint Distribution for GC-LDA with Gaussian Mixtures

For completeness, we present here a modified version of the generative process for the GC-LDA models in which the spatial distributions are modeled as mixtures of multivariate Gaussians with R components. We only present the updated process for generating topics t and activation tokens x_i, as the generative process for sampling word tokens w_i does not depend on the parameterization of the spatial distributions:

1. For each topic $t \in \{1, ..., T\}$:
   (a) Sample a Multinomial distribution over word types $\phi^{(t)} \sim$ Dirichlet(β)
   (b) Sample a Multinomial distribution over subregions $\pi^{(t)} \sim$ Dirichlet(δ)
2. For each document $d \in \{1, ..., D\}$:
   (a) For each peak activation token $x_i \in \{1, ..., N_x^{(d)}\}$:
      i. Sample indicator variable y_i from Multinomial($\theta^{(d)}$)
      ii. Sample indicator variable c_i from Multinomial($\pi^{(y_i)}$)
      iii. Sample a peak activation token x_i from the spatial distribution for subregion $r_{c_i}^{(y_i)}$: $x_i \sim$ Gaussian($\mu_{c_i}^{(y_i)}, \sigma_{c_i}^{(y_i)}$)

The joint distribution of all observed peak activation tokens, word tokens, and latent parameters for each individual document in the GC-LDA model with a mixture of Gaussian spatial distributions is as follows:

$$p(\mathbf{x}, \mathbf{w}, \mathbf{z}, \mathbf{y}, \mathbf{c}, \theta) = p(\theta|\alpha) \cdot \left(\prod_{i=1}^{N_x^{(d)}} p(y_i|\theta^{(d)}) p(c_i|\pi^{(y_i)}) p(x_i|\mu_{c_i}^{(y_i)}, \sigma_{c_i}^{(y_i)}) \right) \cdot \left(\prod_{j=1}^{N_w^{(d)}} p(z_j|\mathbf{y}^{(d)}, \gamma) p(w_j|\phi^{(z_j)}) \right) \tag{1}$$

[1] An implementation of GC-LDA is available at `http://github.com/timothyrubin/python_gclda`

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

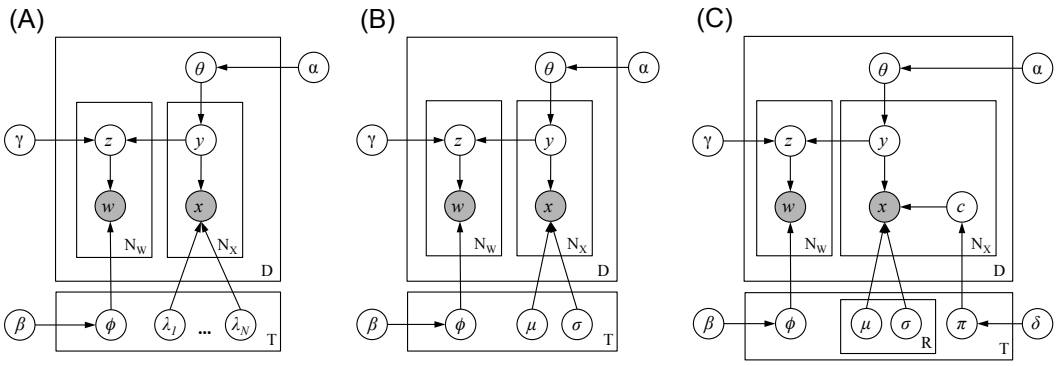

Figure 1: **(A)** Plate notation for the Generalized Correspondence-LDA model, GC-LDA. **(B)** Plate notation for GC-LDA with spatial distributions modeled as a single multivariate Gaussian (Equivalent to a smoothed version of Correspondence-LDA if $\gamma = 0$). **(C)** Plate notation for GC-LDA with subregions, with spatial distributions modeled as a mixture of multivariate Gaussians

2 Inference for GC-LDA

During inference, we seek to estimate the posterior distribution across all unobserved model parameters. As is typical with topic models, exact probabilistic inference for the GC-LDA model is intractable. Inference for the original Correspondence LDA model [2] used Variational Bayesian methods. Here, we employ a mixture of MCMC techniques based on Gibbs Sampling [4], since Gibbs sampling approaches have often outperformed variational methods for inference in LDA [1, 5]. The per-iteration computational complexity is $O(T(N_W + N_X R))$, where T is the number of topics, R the number of subregions, and N_W and N_X are the total number of word tokens and activation tokens in the corpus, respectively.

In describing the inference procedure, we will provide update equations for the three variants of GC-LDA that were used in our experiments, depicted in Figures 1.B and 1.C. Specifically, we describe the updates for the GC-LDA model where each topic's spatial component is represented by a single Gaussian distribution (Figure 1.B), and for the two GC-LDA models where each topic's spatial distribution is represented by a mixture of Gaussian distributions (Figure 1.C). As a reminder, the difference between the two versions of the model that use Gaussian mixtures (referred to as the "unconstrained subregions" and "constrained subregions" models), is that we constrain the mean of the two Gaussian components to be symmetric with respect to their distance from the origin along the horizontal spatial axis in the "constrained subregions" model. In places where the updates for the versions of the models are different, we will first describe the update for the model with single a Gaussian distribution, and then describe how it is modified for the models that use Gaussian mixtures.

After model initialization, our Gibbs Sampling method involves sequentially updating the spatial distribution parameters $\Lambda^{(t)}$ for all topics, the assignments z_i of word tokens to topics, and the assignments y_i of peak activation tokens to topics (and additionally the assignments c_i of activation tokens to subregions when using a Gaussian mixture model for each topic's spatial distribution). We first provide an overview of the sampling algorithm sequence, and then describe in detail the update equations used at each step. We also note here that the update equations presented here will generalize to any variant of the GC-LDA model using a single parametric or mixture of parametric spatial distributions, provided the updates for the spatial parameter estimates are modified appropriately.

2.1 Overview of Inference Procedure

Configuring and running the model consists of two phases: (1) Model initialization, and (2) Inference. We first describe model initialization, and give an overview of the sequence in which model parameters are updated. We will then provide the exact update equations for each of the steps used during inference.

Table 1: Table of notation used throughout the appendix

Model specification	
Notation	Meaning
w_i, x_i	The ith word token and peak activation token in the corpus
$N_x^{(d)}$, $N_w^{(d)}$	The number of word tokens and peak activation tokens in document d, respectively
D	The number of documents in the corpus
T	The number of topics in the model
R	The number of components/subregions in each topic's spatial distribution (subregions model)
z_i	Indicator variable assigning word token w_i to a topic
y_i	Indicator variable assigning activation token x_i to a topic
$\mathbf{z}^{(d)}$, $\mathbf{y}^{(d)}$	The set of all indicator variables for word tokens and activation tokens in document d
N_{td}^{YD}	The number of activation tokens within document d that are assigned to topic t
c_i	Indicator variable assigning activation token y_i to a subregion (subregion models)
$\Lambda^{(t)}$	Placeholder for all spatial parameters for topic t
$\mu^{(t)}$, $\sigma^{(t)}$	Gaussian parameters for topic t
$\mu_r^{(t)}$, $\sigma_r^{(t)}$	Gaussian parameters for subregion r in topic t (subregion models)
$\phi^{(t)}$	Multinomial distribution over word types for topic t
$\phi_w^{(t)}$	Probability of word type w given topic t
$\theta^{(d)}$	Multinomial distribution over topics for document d
$\theta_t^{(d)}$	Probability of topic t given document d
$\pi^{(t)}$	Multinomial distribution over subregions for topic t (subregion models)
$\pi_r^{(t)}$	Probability of subregion r given topic t (subregion models)
β, α, γ	Model hyperparameters
δ	Model hyperparameter (subregion models)
Count matrices used during model inference	
Notation	Meaning
$N_{t\cdot}^{YT}$	The number of activation tokens that are assigned via y_i to topic t
$N_{td,-i}^{YD}$	The number of activation tokens in document d that are assigned via y_i to topic t, excluding the ith token
$N_{z_j d}^{YD*}$	The number of activation tokens in document d that would be assigned to the topic indicated by z_j, given the proposed update of y_i
$N_{rt,-i}^{CT}$	The number of activation tokens that are assigned via c_i to subregion r in topic t, excluding the ith token (subregion models)
$N_{wt,-i}^{ZT}$	The number of times word type w is assigned via z_i to topic t, excluding the ith token
N_{td}^{ZD}	The number of word tokens in document d that are assigned via z_i to topic t

2.1.1 Model Initialization

To initialize the model, we first randomly assign all y_i indicator variables to one of the topics $y_i \sim \text{uniform}(1, ..., T)$. The z_i indicator variables are randomly sampled from the multinomial distribution conditioned on $y_i^{(d)}$ as defined in the generative model: $z_i \sim \text{Multinomial}\big(\frac{N_{1d}^{YD}+\gamma}{N_x^{(d)}+\gamma*T}, \frac{N_{2d}^{YD}+\gamma}{N_x^{(d)}+\gamma*T}, ..., \frac{N_{Td}^{YD}+\gamma}{N_x^{(d)}+\gamma*T}\big)$. In the model that uses an unconstrained mixture of Gaussians with $R = 2$, the initial c_i are randomly assigned: $c_i \sim \text{uniform}(1, ...R)$. In "constrained subregions" model we used a deterministic initial assignment, where we set $c_i = 1$ if the x-coordinate of the activation token was less than or equal to zero (i.e., if the activation peak fell within the left hemisphere of the brain), and $c = 2$ otherwise.

2.1.2 Parameter Update Sequence

After initialization, the model inference procedure entails repeating the following three parameter update steps until the algorithm has converged:

1. For each topic t, update the estimate of the spatial distribution parameters $\Lambda^{(t)}$ conditioned on the subset of peaks x_i with indicator variables $y_i = t$. When using a model with subregions for the topics' spatial components, update the estimate of the spatial distribution

parameters $\Lambda_r^{(t)}$ conditioned on the subsets of peaks x_i with indicator variables $y_i = t$ and $c_i = r$.

2. For each activation token x_i in each document d, update the corresponding indicator variable y_i assigning the token to a topic, conditioned on the current estimates of all spatial distribution parameters $\Lambda^{(\cdot)}$, the current assignments of $\mathbf{z}^{(d)}$ of all word tokens to topics in document d, and the current estimate of the document's multinomial distribution over topics $\theta^{(d)}$. When using a model with subregions, instead jointly update the indicator variables y_i of the token to a topic and c_i of the token to a subregion within topic y_i. This update is additionally conditioned on the current estimate of all topic's multinomial distributions over subregions $\pi^{(\cdot)}$.
3. For each word token w_i in each document d, update the corresponding indicator variable z_i assigning the token to a topic, conditioned on the current estimates of all topics' multinomial distributions over words $\phi^{(\cdot)}$, and the current assignments $\mathbf{y}^{(d)}$ of all peaks to topics in document d.

Note that we do not need to directly update the $\theta^{(d)}$, $\phi^{(t)}$ or $\pi^{(t)}$ parameters during inference, because these distributions are "collapsed out" [5] and are estimated directly from the current state of indicator variables **y**, **z**, and **c**, respectively. Convergence of this algorithm is evaluated by computing the log-likelihood of the observed data after every iteration of the sampler; when the log-likelihood is no increasing over multiple iterations, we halt the algorithm and compute a final estimate of all parameters.

We now provide the update equations for each of these steps.

2.2 Updating Spatial Distribution Estimates: $\Lambda^{(t)}$

To estimate the spatial distributions, we compute the maximum likelihood estimates of the spatial distribution for each topic t, conditioned on the subset of peak activation tokens that are assigned to t. When each topic is associated with a single multivariate Gaussian distribution:

$$\hat{\mu}^{(t)} = \frac{\sum_{i,y_i=t} x_i}{N_{t\cdot}^{YT}} \tag{2}$$

$$\hat{\sigma}^{(t)} = \frac{\sum_{i,y_i=t} (x_i - \hat{\mu}^{(t)})^2}{N_{t\cdot}^{YT}} \tag{3}$$

where $N_{t\cdot}^{YT}$ is the total number of peak activation tokens x_i that are assigned (via y_i) to t. When using a mixture of Gaussians for the spatial distributions, the same estimates are used to estimate the means and covariances for each subregion, $\hat{\mu}_r^{(t)}$ and $\hat{\sigma}_r^{(t)}$, except that the sums are computed over the subset of peak activation tokens for which $y_i = t$ and $c_i = r$. Similarly, for any arbitrary choice of spatial distribution not specifically considered in this paper (e.g., a kernel density estimator), one can use the standard maximum likelihood estimator.

In the "constrained subregions" model, where the Gaussian component means are constrained to be symmetric about the horizontal spatial axis (with respect to the distance from the origin), we must further modify the estimation procedure. We estimate a single mean for the two subregions, with respect to it's location along the horizontal axis in terms of distance from the origin (corresponding to the longitudinal fissure of the brain), by computing the average coordinates of all x_i tokens that are assigned to t after taking the absolute value of the tokens' distance from the origin. This estimate is then used as the mean of the 2nd subregion along the horizontal axis, and the mean of the 1st subregion is set equal to the same mean, reflected about the horizontal axis (so that along this coordinate, $\hat{\mu}_1^{(t)} = -\hat{\mu}_2^{(t)}$). The covariance matrices of the two subregions are estimated independently using equation 3. We note that these updates correspond to maximum likelihood estimates, subject to the constraint that the mean is symmetric along the horizontal axis.

2.3 Updating Assignments y_i of Activation Tokens x_i to Topics

This update step, in which peak activation tokens x_i to are assigned to topics via the indicator variables y_i, is dependent upon the choice of the spatial distribution. Specifically, when using a

model with topic subregions (e.g., where each topic is associated with a Gaussian mixture), this step involves additionally updating the c_i assignments of tokens to subregions. We first provide the update equations for the model that uses a Gaussian distribution for each topic, and then describe the modification to this update needed when using a subregions model.

2.3.1 Updating y_i Assignments for GC-LDA Models Using Single Multivariate Gaussian Spatial Distributions

Here, we wish to update the indicator variable $y_i^{(d)}$, which is the assignment of the ith peak activation token x_i of document d to a topic. This update is conditioned on the current estimates of all spatial distribution parameters Λ, the current vector $\mathbf{z}^{(d)}$ of assignments of words to topics in document d, and the current estimate of the document's multinomial distribution over topics $\theta^{(d)}$.

We employ a Gibbs Sampling step to update each indicator variable using a proposal distribution. The proposal distribution is used to compute the relative probabilities that x_i should be assigned to a specific topic $t = 1, ...T$. Once the relative probabilities are computed across all topics, we randomly sample a topic-assignment y_i from the proposal distribution, normalized such that the probability of assigning the word to a topic sums to 1 across all topics. The update equation is as follows:

$$\begin{aligned} p(y_i = t|x_i, \mathbf{z}^{(d)}, \mathbf{y}_{-i}^{(d)}, \Lambda^{(t)}, \gamma, \alpha) &\sim p(x_i|\Lambda^{(t)}) \cdot p(t|\theta^{(d)}) \cdot p(\mathbf{z}^{(d)}|\mathbf{y}^{(d)*}, \gamma) \\ &\sim p(x_i|\Lambda^{(t)}) \cdot (N_{td,-i}^{YD} + \alpha) \cdot \prod_{j=1}^{N_w^{(d)}} \frac{N_{z_j d}^{YD*} + \gamma}{N_x^{(d)} + \gamma * T} \\ &\sim p(x_i|\Lambda^{(t)}) \cdot (N_{td,-i}^{YD} + \alpha) \cdot \left(\frac{N_{td,-i}^{YD} + \gamma + 1}{N_{td,-i}^{YD} + \gamma} \right)^{N_{td}^{ZD}} \end{aligned} \tag{4}$$

To understand this equation and the notation, we consider the three main terms in the equation in detail.

The first term, $p(x_i|\Lambda^{(t)})$, is the probability that peak activation x_i was generated from the spatial distribution associated with topic t. For example, if each topic is associated with a single multivariate Gaussian distribution, this term corresponds to the multivariate Gaussian probability density function with parameters $\mu^{(t)}$ and $\sigma^{(t)}$ evaluated at location x_i.

The second term, $(N_{td,-i}^{YD} + \alpha)$ is an estimate of the probability of sampling topic t from $\theta^{(d)}$, using an estimate of $\theta^{(d)}$ that is computed from the set of all indicator variables $\mathbf{y}_{-i}^{(d)}$ in document d excluding the indicator variable for the token i that is currently being sampled. In the notation above, $N_{td,-i}^{YD}$ is equal to the number of activation tokens in document d that are currently assigned via y to topic t, where $-i$ indicates that the current token that we are sampling is removed from these counts.

The third term, $\prod_{j=1}^{N_w^{(d)}} \frac{N_{z_j d}^{YD*} + \gamma}{N_x^{(d)} + \gamma * T}$ is the multinomial probability of sampling all of the current indicator variables $\mathbf{z}^{(d)}$ for words in document d, given the count matrix $N_{\cdot d}^{YD*}$ that results from the proposed update of the indicator variables for the peak assignment y_i. In this notation, $\frac{N_{z_j d}^{YD*} + \gamma}{N_x^{(d)} + \gamma * T}$ is the multinomial probability of sampling the indicator variable z_j from the proposed vector of peak-topic assignments $\mathbf{y}^{(d)*}$, where $N_{z_j d}^{YD*}$ is the number of y indicator variables that would be assigned to the same topic as indicator variable z_j given the proposed update of y_i. In the context of Gibbs sampling, the third term can be simplified as shown in the final form of the equation, in which N_{td}^{ZD} corresponds to the number of word tokens in document d that are currently assigned via z to topic t.

2.3.2 Updating y_i and c_i Assignments for GC-LDA models Using Mixtures of Multivariate Gaussian Spatial Distributions

In the GC-LDA model in which each topic's spatial distribution is a mixture of multivariate Gaussian distributions, we use a modified Gibbs sampling procedure in which we jointly sample both the y_i assignment of the peak activation token to a topic, and the c_i assignment of the peak activation token to a subregion, according to the following update update equation:

$$p(y_i = t, c_i = r | x_i, \mathbf{z}^{(d)}, \mathbf{y}^{(d)}_{-i}, \Lambda^{(t)}_r, \pi^{(t)}, \delta, \gamma, \alpha)$$

$$\begin{aligned}
&\sim p(x_i|\Lambda^{(t)}_r) \cdot p(t|\theta^{(d)}) \cdot p(r|\pi^{(t)}) \cdot p(\mathbf{z}^{(d)}|\mathbf{y}^{(d)*}, \gamma) \\
&\sim p(x_i|\Lambda^{(t)}_r) \cdot (N^{YD}_{td,-i} + \alpha) \cdot \frac{N^{CT}_{rt,-i} + \delta}{\sum_{r'=1}^{R}(N^{CT}_{r't,-i} + \delta)} \cdot \prod_{j=1}^{N^{(d)}_w} \frac{N^{YD*}_{z_j d} + \gamma}{N^{(d)}_x + \gamma * T} \\
&\sim p(x_i|\Lambda^{(t)}_r) \cdot (N^{YD}_{td,-i} + \alpha) \cdot \frac{N^{CT}_{rt,-i} + \delta}{\sum_{r'=1}^{R}(N^{CT}_{r't,-i} + \delta)} \cdot \left(\frac{N^{YD}_{td,-i} + \gamma + 1}{N^{YD}_{td,-i} + \gamma} \right)^{N^{ZD}_{td}}
\end{aligned} \tag{5}$$

This update equation is the same as the update equation for the model with a single multivariate Gaussian distribution per topic, with the exception of the first and third terms. The first term $p(x_i|\Lambda^{(t)}_r)$ now corresponds the probability that peak activation x_i was generated from the spatial distribution associated with subregion r of topic t. The third term is the probability $\pi^{(t)}_r$ of sampling subregion r from topic t. The notation $N^{CT}_{rt,-i}$ corresponds to the total number of subregion indicator variables c_i that are currently assigned to subregion r within topic t, excluding the count of the token that is currently being sampled.

2.4 Updating z_i Assignments of Word Tokens w_i to topics

Here we wish to update the indicator variables $z^{(d)}_i$, giving the assignment of the ith word token w_i in document d to a topic. This update is conditioned on the current vector $\mathbf{y}^{(d)}$ of assignments of peaks to topics in d, and an estimate of each topic's multinomial distribution over word types $\phi^{(t)}$

This update involves a collapsed Gibbs sampling step similar in form to the one employed for inference in standard LDA [5]. The update equation is as follows:

$$\begin{aligned}
p(z_i = t | w_i, \mathbf{z}_{-i}, \mathbf{y}^{(d)}, \gamma, \beta) &\sim p(t|\mathbf{y}^{(d)}, \gamma) \cdot p(w_i|\phi^{(t)}) \\
&\sim (N^{YD}_{td,-i} + \gamma) \cdot \frac{N^{ZT}_{wt,-i} + \beta}{\sum_{w'=1}^{T}(N^{ZT}_{w't,-i} + \beta)}
\end{aligned} \tag{6}$$

The first term in this equation gives the probability of sampling topic t from document d, which is proportional to $N^{YD}_{td,-i}$—the count of the number of activation tokens in document d that are currently assigned to topic t—plus the smoothing parameter γ, as defined in the generative model. The second term in this equation is the probability of sampling word w_i from topic t, given the current estimates of the topic-word multinomial distributions. As with the estimate of $\theta^{(d)}$ computed during the $\mathbf{y}_i$ update steps, $\phi^{(t)}$ is computed from the counts of word token assignments, where $N^{ZT}_{wt,-i}$ is the number of times word type w is assigned to topic t across the vector of indicator variables $\mathbf{z}_{-i}$, ignoring the token that is currently being sampled.

2.5 Computing Final Parameter Estimates

We compute final estimates (as well as estimates to be used for log-likelihood computations during inference) of the model parameters as follows:

$$\hat{\theta}^{(d)}_t = \frac{N^{YD}_{td} + \alpha}{\sum_{t'=1}^{T}(N^{YD}_{t'd} + \alpha)} \tag{7}$$

$$\hat{\pi}^{(t)}_r = \frac{N^{CT}_{rt} + \delta}{\sum_{r'=1}^{R}(N^{CT}_{r't} + \delta)} \tag{8}$$

$$\hat{\phi}^{(t)}_w = \frac{N^{ZT}_{wt} + \beta}{\sum_{t'=1}^{T}(N^{ZT}_{w't} + \beta)} \tag{9}$$

The final estimates for the parameters of the spatial distributions are equivalent to estimates used during inference, described previously.

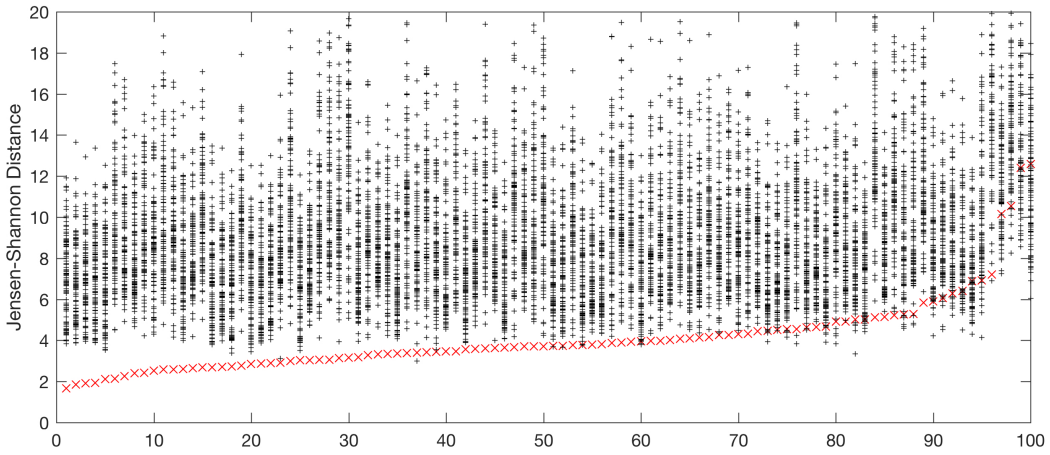

Figure 2: Jensen-Shannon (JS) distances between pairs of topics learned using distinct subsets of training documents. For each topic $t = 1...100$ learned using training Subset 1, we show the JS-distances between the topic t and all topics learned from training Subset 2. The JS-distance between topic t and the topic it was aligned with using a matching algorithm is indicated using a red '$\times$'. Topics from training Subset 1 are sorted in terms of the JS-distance between the topic and its aligned topic from Subset 2.

3 Topic Stability Analysis

Given that one goal of our models are to work towards a "functional neuro-anatomical atlas", it is important to consider how stable the topic solutions provided by the model are. That is, if the model is identifying functional regions that are consistent with true underlying neuroanatomical patterns, we expect that these regions should be consistently identified regardless of the specific training data used by the model. To investigate the stability of our topic solutions, we randomly partitioned the Neurosynth database into two equal halves—training Subset 1 and Subset 2—where each of these subsets contained 5,681 complete documents. For each of the two subsets, we trained a "constrained subregions" GC-LDA model using $\gamma = .01$ and all other hyper-parameters equal to those described in Section 3.2 of the main paper.

To evaluate the similarity between the topic solutions identified from the training subsets, we followed a procedure similar to the one described for alignment of standard LDA topics in [6] (although note that in [6] the authors used the same training data but different random initializations to produce two separate topic solutions). Specifically, we computed a T-by-T "dissimilarity" matrix, where element i, j of the matrix corresponded to the dissimilarity between the ith topic in training Subset 1 and the jth topic in Subset 2. We defined the dissimilarity between two topics as the sum of the Jensen-Shannon (JS) distances [3] between the probability distributions over words and the spatial probability distributions for the two topics. Given these dissimilarity matrices, we aligned each topic $t = 1...T$ learned from training Subset 1 with a single topic from training Subset 2, using a greedy algorithm which iterated T times over the following steps: (1) find the lowest remaining dissimilarity value in the dissimilarity matrix, and store the row and column indices as a mapping from the topics in Subset 1 to Subset 2, then (2) remove the corresponding rows and topics from the matrix.

Given the aligned topic sets, we qualitatively evaluated the similarities between the aligned topic pairs in terms of both their spatial and linguistic distributions. Additionally, for each topic t from training Subset 1, we visualized the distribution of JS-distances between its "aligned" topic and all non-aligned topics, as illustrated in Figure 2. From Figure 2, it is clear that for many of the best-aligned topics, the JS-distance between the aligned topics lies outside of the distribution of distances for the non-aligned topics. Based on these analyses, we estimate that approximately 50% of topics identified by the GC-LDA model are stable, and will be consistently extracted, independent of the specific training documents. We note that these analyses are only heuristic in nature, and in future work we hope to formalize a concrete procedure for assessing topic stability.

References

[1] Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. On smoothing and inference for topic models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 27–34. AUAI Press, 2009.

[2] David M Blei and Michael I Jordan. Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 127–134. ACM, 2003.

[3] Dominik Maria Endres and Johannes E Schindelin. A new metric for probability distributions. *IEEE Transactions on Information theory*, 2003.

[4] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*, volume 2. Taylor & Francis, 2014.

[5] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004.

[6] Mark Steyvers and Tom Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440, 2007.