Research Statement

Dongwoo Kim

As a machine learning researcher, I study statistical models to uncover hidden structures of data. Today, a vast amount of information is digitized and stored into various forms. Despite the collaborative effort to organize the information into structuralized forms, a large body of the stored information still remains in unstructured forms such as text and images. I structuralize the unstructured data by inferring hidden structures which cannot be directly observed from the data. The hidden structures of data correspond to abstract concepts, like themes, behavioral or mental states, and are used to understand the data themselves as well as improve many other further tasks. In my research approach, I infer the hidden structure of data by 1) defining a **Bayesian latent variable model** based on a plausible generative process of the data, 2) inferring the hidden structure of the data through the **posterior inference** on the model, and 3) verifying the inferred structure on further tasks such as a classification or regression.

Specifically, my research aims to build probabilistic topic models based on the Bayesian nonparametric theory. The Bayesian topic models have emerged as an important tool to uncover the underlying semantic patterns, called topics, from a set of unstructured documents without any supervision. One important direction of the topic model involves the Bayesian nonparametric theory. With the model built upon the nonparametric method, the number of parameters of the model grows as more data is observed. As a result, this approach is well suited for the streaming and/or large-scale data. Despite the advantages of the nonparametrics, there has been not much work on the nonparametric topic models because constructing nonparametirc models typically entails defining a complex model construction and solving intractable posterior probability. During my Ph.D. studies, I tried to reduce the gap between the parametric and nonparametric topic models by developing nonparametric topic models in two different ways: directly extending the existing parametric topic models and building new model construction methods based on the nonparametric theory.

My next research goal is to model the hidden structure of the complex social process by using a corpus which contains various types of human activities. With the emergence of social media and life-logging platforms, we are facing a unique oppertunity to observe and quantify the complex social processes at a scale much larger than ever before. Analyzing and understanding this data will change our understanding of the complex social processes that underlie society. As a next step, I am planning to use what I have learned through my past research in topic modeling to develop new models to analyze the human behavior patterns beneath the social process.

1 Nonparametric Topic Models with Metadata

Bayesian nonparametric topic models, rooted in the hierarchical Dirichlet processes, generalize the parametric models by placing an infinite dimensional prior over the topic space. The introduction of Bayesian nonparametrics greatly increases the flexibility of the modeling assumptions. For example, we developed a hierarchical topic model [4]. The model infers an appropriate depth and width of topic hierarchy that cannot be modeled in a parametric way. Despite its potential capabilities, the nonparametric method has not been not widely applied to topic modeling because the model construction is complex, and posterior inference is intractable for many interesting models.

On the parametric side, incorporating metadata of documents into the existing topic models has been widely studied to capture the different aspect of topics. For example, with the time index of documents, one may infer how topics can change over time; with the citation network among documents, one may infer how the topic effects on the citation behavior; and with the authors of documents, one may infer the topic usage of each author. I addressed the problem of constructing a nonparametric topic models with various types of metadata, and proposed several models to infer the latent topics which reflect the underlying structure correlated with the metadata. A selection of my research is described below.

1.1 Dirichlet Process with Mixed Random Measure (DP-MRM)

A topic is a multinomial distribution over the vocabulary, and typically, it is represented by a set of high probability words. How can we interpret these probability distributions? One possible solution is to tag a

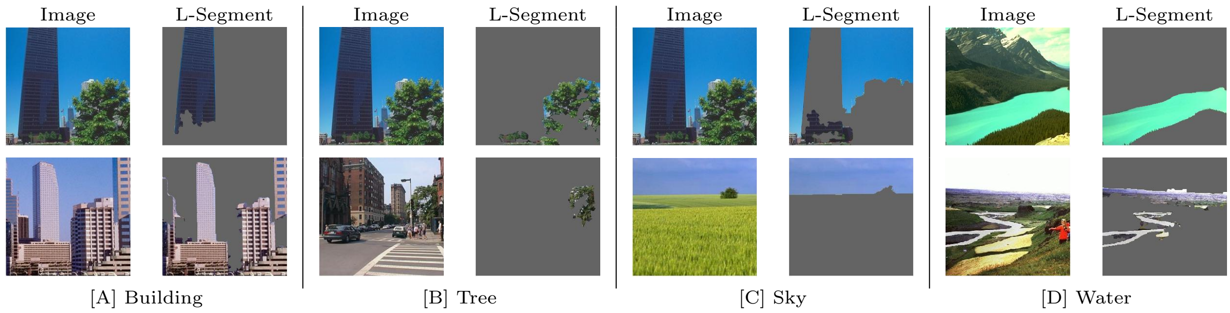

Figure 1: Labeled Segments (L-Segment) from posterior samples of DP-MRM. From left to right, the segments are extracted from *building*, *water*, and *tree* labels. With the list of objects without their location information, DP-MRM segments the images and links the segmented parts to the corresponding objects, simultaneously.

topic with metadata such as categories and tages. Ramage et al. proposed labeled-LDA as a tagging process of latent topics. In the labeled-LDA, each word of a document is assigned to one of the document's labels. As a result, each inferred topic is tagged by one of the labels. This mapping improves the interpretability of topics. At the same time, the one-to-one relationship between topics and labels overly restricts the flexibility of the model because the topics of a document are restricted by the given labels of the document.

I directly extended the labeled-LDA to DP-MRM where the number of topics per label is automatically inferred by the nonparametric method [5]. As a result, the model infers an appropriate number of topics per label. For example, the model infers more topics for the label 'sports' than the more specific and narrow label 'soccer'. I further enhance the model by relaxing exchangeability assumption of words to model the local dependencies between words. The final model is applied to multi-labeled images for image segmentation and object labeling problems by modeling both **local dependencies** between pixels (words) and the **global dependencies** between labels across the different images (documents). Given a set of images and a list of objects for each image, DP-MRM automatically segments the images and links the segmented parts to the corresponding objects without pixel level supervision. Figure 1 shows how the model segments and labels images.

1.2 Hierarchical Dirichlet Scaling Process (HDSP)

DP-MRM successfully applied to labeled documents and images, but sometimes, the model is not appropriate for the different types of side information. For example, if authors of a document is used as labels, then the model allocates a unique set of topics per author, which can not model the shared interest across the authors. I relaxed the assumption behind DP-MRM and developed HDSP[7] which **allocates the latent correlation between topics and labels** and then infers the correlation from a corpus.

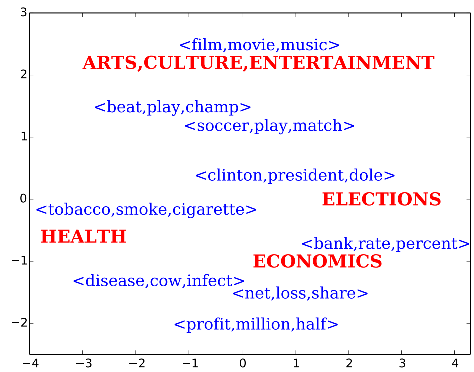

Figure 2: Relative locations of categories (capital letters in red) and latent topics (small letters in blue) inferred by HDSP from the Reuters corpus.

To incorporate the correlation into the topic model framework, I proposed novel construction method for HDSP. Within HDSP, the first level random measure is constructed through the widely used method, stick breaking process, and the second level random measures are constructed through the normalized gamma process. This construction escapes from the widely used construction method, two levels of stick breaking processes, and yields a highly controllable and flexible model.

HDSP models topics correlated with numerical labels as well as categorical labels. The model was successfully applied to the academic corpus, where the authors are used as labels, and the product review corpus, where the numerical ratings and category of a product are used as labels. Given a product category of a review, the model shows an improved performance on classifying the rating of review. The relationship between inferred topics and labels of a news corpus are visualized in Figure 2. HDSP and DP-MRM have been published at ICML 2014 and 2012.

1.3 Topics over time

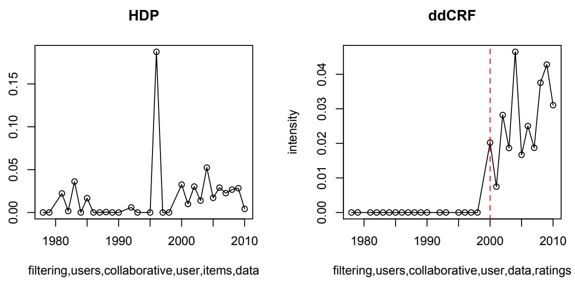

Figure 3: The emergences of 'collaborative filtering' topic inferred from academic corpora by HDP and ddCRF.

A corpus over a long term period may expose the topical characteristics correlated with the written time of each document. With the nonparametric approach, I model the **emergence** and **dissapearance** of topics over time which cannot be modeled by the parametric approach because, in the parametric models, the number of topics should be fixed a priori. With the preliminary study to capture the emergency and disappearance of topics [3], I develop the distance-dependent Chinese restaurant franchise (ddCRF) model where topics emerge and decay over time [2]. The ddCRF relaxes the exchangeability between documents and allows the topics in nearby documents are more likely to be similar. With a corpus collected over several years, the model captures the emergence of topics over time.

1.4 Conditional generative process of a corpus

The models above, HDSP, DP-MRM, and ddCRF, generate the topic proportions of a document as a dependent variable of observable meta information. This modeling approach differs from the traditional definition of a generative process where the every observable variables are generated from a latent variable or parameter. For example, the supervised-LDA and max-margin LDA propose generative processes where the observable labels are generated from a topic proportion of a document. However, I believe that a more natural model of the **human writing process** is to decide what to write about (e.g., categories) before writing the content of a document. I proposed nonparametric models based on this perspective on the generative process and showed improved performances on real world datasets.

2 Modeling Human Behavior Patterns from Corpus

My next research aims to infer the **hidden structure of complex social behavior** by modeling a corpus which contains various types of social processes and activities. My main research theme has been focused on finding the hidden structure which corresponds to latent themes of a corpus. As the corpus is written by many authors, we can uncover various types of hidden structure which reflects the latent characteristics and behavioral patterns of these authors. Moreover, the emergence of social media and online communities give researchers a unique opportunity to observe and quantify complex user behavior which has not been observed so far.

My early research addressed a range of problems relating to the analysis of textual information used in complex social processes. I developed a new approach to extract interests of Twitter users [1] and also examined how diversity seeking users (e.g. the users who always crave new stories) can benefit the online news outlet which provides the social feed to users [6]. These new approaches try to overcome the limitations of traditional social science research; small scale experiments using controlled study or questionnaire-based methods.

A two-stage approach was used to analyze the user behavior; 1) the latent topics of corpus are inferred by relatively simple models such as LDA, and then 2) the inferred topics are combined with additional information, such as a social network structure, for the further analysis on user behavior. Advantages of using the Bayesian latent variable models are the model built upon explicit assumptions and prior beliefs on data, and thus, yielding mathematically rigorous and easily interpretable models. The two-stage approach excludes side information from the first stage, therefore the inferred topics or latent variables may not reflect the latent semantics that we would like to know about. One possible direction is to devise a holistic model with various types of information together in order to infer the latent semantics that correspond to the generative process defined by appropriate assumptions about user behavior.

For example, in our ongoing study, we examine the conversation patterns of bilingual and monolingual in multilingual society through HDSP. We assume that the differences in linguistic fluency may result in different conversation patterns. We answer this question by looking at Switzerland, a highly multilingual society, with a large corpus of geotagged Twitter data. Inferred topics from HDSP, as shown in Figure **??**, show the topical differences among different linguistic groups.

This research direction is inherently interdisciplinary. Social scientists identify the most vital research questions, while machine learning researchers contribute for developing novel computational tools. This approach has the potential to change our understanding of the complex social processes that underlie society. I am open to collaborate with people from various field, and I will contribute to both machine learning and Bayesian statistics, as well as computational social science.

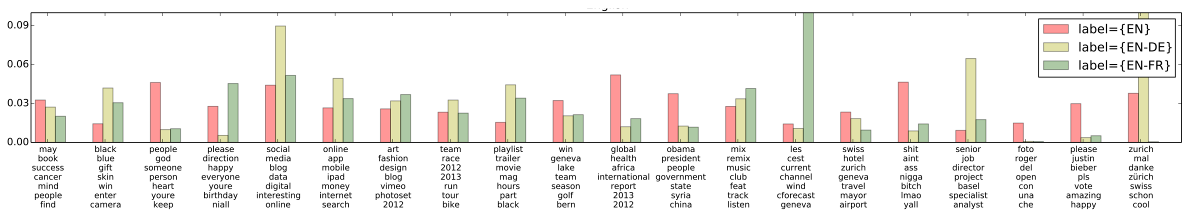

Figure 4: Conversation themes with respect to the monolingual and different types of bilinguals in Switzerland. The results are analyzed by HDSP on the Tweets of the Swiss.

References

[1] **Dongwoo Kim**, Yohan Jo, Il-Chul Moon, and Alice Oh. Analysis of twitter lists as a potential source for discovering latent characteristics of users. In *Workshop on Microblogging at the ACM Conference on Human Factors in Computer Systems*, 2010.

[2] **Dongwoo Kim** and Alice Oh. Accounting for data dependencies within a hierarchical dirichlet process mixture model. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management (CIKM)*, 2011.

[3] **Dongwoo Kim** and Alice Oh. Topic chains for understanding a news corpus. In *Proceedings of the 12th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING)*, 2011.

[4] Joon Hee Kim, **Dongwoo Kim**, Suin Kim, and Alice Oh. Modeling topic hierarchies with the recursive chinese restaurant process. In *Proceedings of the 21th ACM Conference on Information and Knowledge Management (CIKM)*, 2012.

[5] **Dongwoo Kim**, Suin Kim, and Alice Oh. Dirichlet process with mixed random measures: a nonparametric topic model for labeled data. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012.

[6] Jooyeon Kim, Joon Hee Kim, **Dongwoo Kim**, and Alice Oh. Diversity-seeking users and their influence on social news sites. In *Workshop on Data Science for News Publishing (NewsKDD) at KDD 2014*, 2014.

[7] **Dongwoo Kim** and Alice Oh. Hierarchical dirichlet scaling process. In *Proceedings of the 31th International Conference on Machine Learning (ICML)*, 2014.