NON-PARAMETRIC BAYESIAN ANALYSIS OF HETEROGENEOUS DATA

David Blei
Princeton University
March 2013
Final Report

DISTRIBUTION A: Approved for public release.

AIR FORCE RESEARCH LABORATORY
AF OFFICE OF SCIENTIFIC RESEARCH (AFOSR)
ARLINGTON, VIRGINIA 22203
AIR FORCE MATERIEL COMMAND
**Title and Subtitle**

Non-parametric Bayesian analysis of heterogeneous data

**Author(s)**

David Blei

**Performing Organization Name(s) and Address(es)**

Princeton University
4 New South Building
Princeton, NJ 08544

**Sponsoring/Monitoring Agency Name(s) and Address(es)**

AFOSR
875 N Randolph St
Arlington, VA 22203
Dr. Tristan Nguyen/RSL

**Distribution/Availability Statement**

Distribution A: Approved for Public Release

**Abstract**

Under this grant, my research focused on fusing heterogenous sources of data with Bayesian nonparametric models. We published many papers in the service of this goal. I would like to highlight the following papers about furthering Bayesian nonparametrics and examining the fusion of heterogenous data types in a diversity of settings. This is an extension of last year’s report. It is my final report.

**Subject Terms**

Non-parametric, Bayesian, heterogenous data, fusion, models, research, papers, furthering, nonparametrics, settings, report.
Under this grant, my research focused on fusing heterogenous sources of data with Bayesian nonparametric models. We published many papers in the service of this goal. I would like to highlight the following papers about furthering Bayesian nonparametrics and examining the fusion of heterogenous data types in a diversity of settings. This is an extension of last year’s report. It is my final report.

1. With Sam Gershman, we wrote a tutorial about Bayesian nonparametrics (Gershman and Blei, 2012).

2. With Peter Frazier and colleagues, we have worked on distance dependent Bayesian nonparametric models (Blei and Frazier, 2011; Gershman et al., 2011; Ghosh et al., 2011). These allow external data sources to influence the latent clustering (and latent feature representation) of a variety of data. We have applied these models to text, images, EEG, and stock prices.

3. With Lauren Hannah, we developed Dirichlet process mixtures of generalized linear models (Hannah et al., 2010, 2011). These allow covariates to affect the clustering of a response and exert a relationship on it.

4. With Chong Wang, we modeled collaborative filtering data—user preferences and content about the items (Wang and Blei, 2011). This work won the Best Student Paper Award at KDD 2011.

5. With Sean Gerrish, we built a model of legislative roll call data (i.e., votes on bills) and bill texts (Gerrish and Blei, 2011). This work won a Distinguished Application Award at ICML 2011. We recently furthered this work to model issue-adjusted ideal points (Gerrish and Blei, 2012).

6. John Paisley, Chong Wang, and I developed the Discrete Infinite Logistic Normal (DILN), which is a new kind of Bayesian nonparametric model (Paisley et al., 2011, 2012). DILN allows the atoms of an underlying random measure to exert correlation.
7. To perform inference with massive data sets, Matt Hoffman, Francis Bach, and I developed stochastic variational inference for Latent Dirichlet allocation (Hoffman et al., 2010a). Chong Wang, John Paisley, and I extended this algorithm to the hierarchical Dirichlet process, enabling us to fit Bayesian nonparametric models to massive data (Wang et al., 2011). Recently, Chong Wang and I developed a truncation-free variant of stochastic variational inference for this important class of models (Wang and Blei, 2012).

8. Jonathan Chang and I published the relational topic model, a model of documents and links (Chang and Blei, 2010). Unlike traditional network models, this model incorporates node content—it can predict content from links and links from content. Prem Gopalan and I developed stochastic inference for analyzing massive social networks (Gopalan et al., 2012).

9. Matt Hoffman and I wrote several papers about Bayesian nonparametric analysis of recorded music (Hoffman et al., 2009b,a,c, 2010b).


11. Chong Wang and I relaxed some of the assumptions made by the hierarchical Dirichlet process, coupling sparsity and smoothness (Wang and Blei, 2009a). With Sinead Williamson and Katherine Heller, we further extended this work to matrix factorization (Williamson et al., 2010).

References


