Research Strategy

Significance

Big Models for Big Data

Need for more complex models for EHR

subsubsection on why

Difficulty to fit complex models

Stan is flexible and fast

Modeling should be accessible

Innovation

H MC Algorithm is faster

Approach

Multidimensional statistical and computational methods for analyzing, inspecting, displaying, representing, parsing, and searching high-dimensional data

Preliminary work

Funded through several mechanisms including the National Science Foundation and the National Institute of Health, the team submitting this proposal already developed and implemented many related algorithms and software packages. The proposed work is a direct continuation of the below described preliminary work of developing and implementing novel ground breaking algorithms for hierarchical modeling of complex data.

A novel algorithm for Bayesian inference: Hamiltonian Monte Carlo

Dr. Gelman, Betancourt and collaborators developed Hamiltonian Monte Carlo (HMC) methods, a novel approach to computationally implement complex hierarchical Bayesian inference through Monte Carlo simulation.

Open source computational implementation of HMC for diverse interfaces: Stan.

Drs. Gelman, Betancourt and Goodrich developed Stan, an open source multipurpose probabilistic programming language to build complex Bayesian models in several open source and commerical software including so far Stata, Mathlab, Python, Julia and R/Rstudio.

Implementation in the open source software environment R/Rstudio: rstan

Drs. Goodrich, Gelman, Betancourt and collaborators developed Rstan, a software package to use Hamiltonian Monte Carlo algorithms the open source statistical software enviroment R/Rstudio.

Prototype software development: shinyStan and rstanArm

Drs. Goodrich, Andreae, Gelman and Betancourt and collaborators developed two additional prototype software package for the open source statistical software environment R/Rstudio, called (1) rstanarm and (2) shinystan.

(1) rstanarm makes building advanced hierarchical Bayesian models accessible to data scientist without the need for an understanding of the complexities underlying Hamiltonian Monte Carlo. This software package uses the same notation for model description as other widely accepted software packages for mixed modeling in R/Rstudio like lme4.

(2) shinystan is an interactive tool to explore and diagnose the output of Monte Carlo simulations for Bayesian inference. Also a package for R/Rstudio, shinystan provides multidimensional statistical, graphical and computational tools for any analyzing, inspecting, displaying, representing, parsing, and searching high-dimensional MCMC output, but is optimized for HMC.

	Year 1	Year 2	Year 3	Year 4	Year 5
Stan GLM	meta-analysis	3-4 level models	semi- and parametric survival		change point
Shinystan	Binary posterior predictive check	dependence plot	tree depth plot		
New Methods	Dependence plot	Multilevel model priors			

Table 1: Timeline: The above timeline outlines our targets and milestones over the five year grant period.

Team experience in complex hierarchical modeling in medicine and software development

Our team has extensive experience in hierarchical and complex modeling in medicine and electronic medical records. Dr. Andreae published several systematic reviews and meta-analyses[1;2;3]. Dr. Goodrich Dr. Hall is nationally recognized for the development and application of change point models in epidemiology and surveillance. Dr. Gong is leading an NIH funded trial to predict and improve respiratory outcomes after intubation based on real time electronic medical records. Drs. Andreae, Goodrich and collaborators used the software prototype rstanarm and shinystan to build a multilevel hierarchical model to investigate health care disparities and quality of anesthesia delivery in the large National Clinical Outcomes Registry maintained by the American Society of Anesthesiology. Dr. Gelman is internationally recognized as a leader in hierarchical modeling with past and present funding and publication in pharmacodynamic modeling, ...

Our team has a solid background in the development of complex statistical software and in the visualization of model parameters, co-variance matrices and statistical data. Drs. Goodrich and Gelman developed several widely cited and used software packages including the probabilistic programming language Stan[4], the basis for the proposed work. Dr. Goodrich and Andreae work together on the preliminary software packages shinystan and rstanarm.

Best practices and proven methods for software design, construction, and implementation

Mechanisms for incorporating user reported corrections into the software.

Products of this project

The goal of this project is to encourage wider adoption of complex hierarchical modeling for Big Data by clinical data scientists and to support the creation and reuse of open-source extensions and applications.

Software sharing plan

The software developed for this grant will all be incorporated into Comprehensive R Archive Network of the open-source R Project for Statistical Computing. To facilitate its unimpeded utilization, our source code and documentation will be distributed under the least restrictive open-source licensing terms possible: R's licensing is under the GNU General Public License. The benefits are that our software products and subsequent developments

- will be freely available to researchers and the general public, including for commercial use;
- may be freely extended, customized, and incorporated into the other tools;
- can be maintained in the event of the original developers not being willing or able to;
- can be enhanced based on user-provided feedback for bug-fixes, examples, and enhancements; using GitHub pull requests with integration testing

Any documentation is released under the same license as Wikipedia, the Creative Commons Attribution/ShareAlike 4.0 license (CC BY-SA 4.0)[5]

Potential problems

Timeline

test

References

[1] M. H. Andreae and D. A. Andreae. Regional anaesthesia to prevent chronic pain after surgery: a Cochrane systematic review and meta-analysis. *Br J Anaesth*, 111(5):711–720, Nov 2013. PMCID: PMC3793661.

[2] M. H. Andreae, G. M. Carter, N. Shaparin, K. Suslov, R. J. Ellis, M. A. Ware, D. I. Abrams, H. Prasad, B. Wilsey, D. Indyk, M. Johnson, and H. S. Sacks. Inhaled Cannabis for Chronic Neuropathic Pain: A Meta-analysis of Individual Patient Data. *J Pain*, Sep 2015. PMID: 26362106.

[3] G. M. Carter, D. Indyk, M. Johnson, M. Andreae, K. Suslov, S. Busani, A. Esmaeili, and H. S. Sacks. Micronutrients in HIV: a Bayesian meta-analysis. *PLoS One*, 10(4):e0120113, 2015. PMID: 25830916.

[4] Stan Development Team. Stan: A C++ Library for Probability and Sampling, Version 2.5.0, 2014.

[5] Creativecommons.org. Creative Commons - Attribution-ShareAlike 4.0 International - CC BY-SA 4.0, 2015.