9:00–15:00 / January 5–7, 2015 / Øster Farimagsgade 5, Room 4.2.26

Applied Introduction to Bayesian Data Analysis

Copenhagen Graduate School of Social Sciences

Johannes Karreth

Assistant Professor, Department of Political Science, Rockefeller College of Public Affairs & Policy
University at Albany, State University of New York, USA
Email: jkarreth@albany.edu

Course website: http://www.jkarreth.net/bayes-cph.html (with links to workshop materials)
Course registration: http://phdcourses.dk/Course/35395#.VH5Bd74_6TD

Workshop description and goals

This workshop provides an applied overview of Bayesian tools for data analysis and inference. The Bayesian approach to data analysis and inference offers social scientists a number of advantages including, but not limited to: a high degree of flexibility in estimating parameters at nested levels, dealing with incomplete data, estimating and incorporating uncertainty in measurement models, and using prior information to refine model estimates and predictions.

The overall goal of the workshop is to enable participants to use Bayesian tools in their own research and to give participants useful tools to communicate Bayesian results to other scholars. For that reason, participants are encouraged to bring their own projects to the workshop and (as far as feasible) use them for assignments. Upon conclusion of this workshop, participants will:

- have learned the origins of and logic behind Bayesian inference,
- have learned the differences and similarities between Bayesian and frequentist inference,
- be able to use Bayesian techniques for analyzing social science data, and
- be prepared to take more in-depth and advanced courses in Bayesian methods or study more advanced materials independently.

Over three days, the workshops will explore the following topics (with modifications depending on time and participant interest):

- Why use Bayesian inference?
- Philosophical and theoretical foundations for Bayesian inference
- Mechanics: Markov Chain Monte Carlo tools and sampling
- Building and estimating Bayesian linear and generalized linear models
- Bayesian approaches to measurement
- Bayesian approaches to structured and multilevel data
- Model presentation, workflow, and reproducibility

Data management and analysis rely on R and R packages or other software designed for Bayesian estimation such as MCMCpack, JAGS, or Stan (all accessed through R). Most applications shown in the workshop will use R and JAGS, but MCMCpack and Stan will be introduced as well. Prior knowledge of R is useful, but not required. All computer code demonstrated in the workshop will be explained in detail and made available to participants.

Literature

Participants should have access to one of the following books for background reading. All are great resources. Each of them covers a majority of the workshop's contents and will be useful as a reference for participants beyond this workshop. Also, each book has helpful code and the data used in the books available on dedicated websites. The first four use R and JAGS (or the nearly-identical BUGS language) for model fitting.

DBDA Kruschke, J. (2014). *Doing Bayesian Data Analysis, Second Edition: A Tutorial with R, JAGS, and STAN*. Academic Press / Elsevier, Oxford. [Most accessible]

ARM Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York, NY. [Good coverage of multilevel models]

BASS Jackman, S. (2009). *Bayesian Analysis for the Social Sciences*. Wiley, Chichester. [Very detailed]

BM Gill, J. (2014). *Bayesian Methods: A Social and Behavioral Sciences Approach, Third Edition*. Chapman and Hall/CRC, Boca Raton, FL. [Detailed, but broad coverage]

BDA Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis, Third Edition*. Chapman and Hall/CRC, Boca Raton, FL. [Most advanced, uses R and Stan]

As a general primer for R, I recommend:

- Kabacoff, R. Quick-R. Available at `statmethods.net`. [Introductory]
- Fox, J. and Weisberg, S. (2011). *An R Companion to Applied Regression, Second Edition*. Sage, Thousand Oaks. [Comprehensive]

Software and Preparation

Participants are encouraged to bring their laptop computers to the workshop. Before the start of the workshop, participants should try to install the following programs on their laptops:

1. R is an open-source software package and available for download at `http://www.r-project.org`.
2. RStudio is a convenient integrated development environment for R and available for free at `http://www.rstudio.com`.
3. JAGS is a command-line program for Bayesian estimation. We will use it through R. Participants should install it from `http://sourceforge.net/projects/mcmc-jags/files/`.

We will go over how to use these programs on the first day of the workshop, using a detailed tutorial with step-by-step instructions. We will also have time to catch up on installation problems on the first day.

Workshop outline

The following time slots and topics may be modified as the workshop proceeds. The most current version of this document can be found at http://www.jkarreth.net/files/bayes-cph.pdf.

Day	Time	Unit
Monday	9:00-9:15	Welcome & Introductions
Monday	9:15-10:00	1
Monday	10:00-12:00	2
Monday	13:00-14:00	3
Monday	14:00-15:00	Lab: First steps in R and JAGS
Tuesday	9:00-10:30	4
Tuesday	10:45-12:00	4
Tuesday	13:00-14:00	5
Tuesday	14:00-15:00	6
Tuesday	18:00	Dinner at Madklubben Steak
Wednesday	9:00-10:30	7
Wednesday	10:45-12:00	8
Wednesday	13:00-14:00	8
Wednesday	14:30-15:00	9

The "background readings" below are suggestions for additional self-study after the respective workshop units. **You are not expected to have read all assigned readings before the workshop**, although it will be helpful to have access to some of them if you want to read up on questions you might have right away. The textbook readings provide substantial details for the topics that we explore in a more introductory fashion. Alternative chapters from each book are listed here to give you some direction for further reading after the workshop. Some of the journal articles are more in-depth discussions of specific problems, challenges, or techniques; others are straightforward applications of the techniques encountered in the workshop.

Unit 1: Why be Bayesian?

Background reading:

- Siegfried, T. (2010). Odds are, it's wrong: Science Fails to Face the Shortcomings of Statistics. *Science News*, 177(7):26–29
- Senn, S. (2003). Bayesian, Likelihood, and Frequentist Approaches to Statistics. *Applied Clinical Trials*, 12(8):35–38

Unit 2: Foundations for Bayesian inference

Probability, prior distributions, Bayes' rule, posterior distributions
Background reading:

- Western, B. and Jackman, S. (1994). Bayesian Inference for Comparative Research. *American Political Science Review*, 88(2):412–423
- DBDA: chs. 2, 4, 5
- BASS: ch. 1, 2
- BM: chs. 1, 2, 5
- BDA: ch. 1.

Unit 3: Mechanics for Bayesian inference

Markov Chain Monte Carlo tools, sampling methods, convergence diagnostics
Background reading:

- DBDA: chs. 7, 8
- ARM: chs. 18, 19
- BASS: chs. 3–6
- BM: chs. 9, 10, 14
- BDA: chs. 10–12
- Robert, C. and Casella, G. (2010). *Introducing Monte Carlo Methods with R*. Springer, New York, NY, ch. 8
- Cowles, M. K. and Carlin, B. P. (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association*, 91(434):883–904
- Geyer, C. (2011). Introduction to Markov Chain Monte Carlo. In Brooks, S., Gelman, A., Jones, G., and Meng, X.-L., editors, *Handbook of Markov Chain Monte Carlo*, pages 3–48. Chapman and Hall/CRC
- Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News*, 6(1):7–11
- Plummer, M. (2013). JAGS Version 3.4.0 User Manual
- Spiegelhalter, D. J., Thomas, A., Best, N. G., and Lunn, D. (2003). WinBUGS Version 1.4 User Manual (although we're not using WinBUGS in this workshop, this manual offers some useful background information on Gibbs sampling.)
- Martin, A. D., Quinn, K. M., and Park, J. H. (2011). MCMCpack: Markov Chain Monte Carlo in R. *Journal of Statistical Software*, 42(9):22
- Tsai, T.-h. and Gill, J. (2012). superdiag: A Comprehensive Test Suite for Markov Chain Non-Convergence. *The Political Methodologist*, 19(2):12–18

Unit 4: Building and estimating Bayesian linear and generalized linear models

Regression modeling of continuous, binary, ordinal, and categorical outcomes
Background reading:

- DBDA: chs. 15–24
- ARM: ch. 16, 18, 19, appendix A
- BASS: section 2.5, ch. 8
- BM: ch. 4, 11
- BDA: chs. 14, 16

Unit 5: Model fit & model checking

Posterior predictive checks and Bayes factors
Background reading:

- ARM: ch. 24
- BM: ch. 6
- BDA: chs. 6, 7
- Gelman, A., Goegebeur, Y., Tuerlinckx, F., and Mechelen, I. V. (2000). Diagnostic Checks for Discrete Data Regression Models Using Posterior Predictive Simulations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 49(2):247–268

Unit 6: Processing and presenting useful quantities of interest from Bayesian models

Numerical and graphical summaries of posterior distributions
Background reading:

- ARM: ch. 21
- Karreth, J. Lab exercises for presenting Bayesian model output (available to participants)

Unit 7: Measurement models

Bayesian factor analysis and item response modeling
Background reading:

- BASS: ch. 9
- Quinn, K. M. (2004). Bayesian Factor Analysis for Mixed Ordinal and Continuous Responses. *Political Analysis*, 12(4):338–353
- Treier, S. and Jackman, S. (2008). Democracy as a Latent Variable. *American Journal of Political Science*, 52(1):201–217
- Pemstein, D., Meserve, S. A., and Melton, J. (2010). Democratic Compromise: A Latent Variable Analysis of Ten Measures of Regime Type. *Political Analysis*, 18(4):426–449
- Gray, J. and Slapin, J. B. (2012). How Effective are Preferential Trade Agreements? Ask the Experts. *Review of International Organizations*, 7(3):309–333
- Bakker, R. (2009). Re-measuring Left–Right: A Comparison of SEM and Bayesian Measurement Models for Extracting Left–Right Party Placements. *Electoral Studies*, 28(3):413–421
- Fariss, C. J. (2014). Respect for Human Rights has Improved Over Time: Modeling the Changing Standard of Accountability . *American Political Science Review*, 108(2):297–318
- Linzer, D. A. and Staton, J. K. (ND). A Measurement Model for Synthesizing Multiple Comparative Indicators: The Case of Judicial Independence. *Working paper*
- Grimmer, J. (2010). A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases. *Political Analysis*, 18(1):1–35
- Bafumi, J., Gelman, A., Park, D. K., and Kaplan, N. (2005). Practical Issues in Implementing and Understanding Bayesian Ideal Point Estimation. *Political Analysis*, 13(2):171–187
- Barberá, P. (Forthcoming). Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data. *Political Analysis*

Unit 8: Analyzing structured data: multilevel modeling

Bayesian hierarchical/multilevel models for nested and time-series cross-sectional data
Background reading:

- ARM: chs. 11–15, 17
- BASS: ch. 7
- BM: ch. 12
- BDA: ch. 15
- On multilevel modeling:
  - Steenbergen, M. R. and Jones, B. S. (2002). Modeling Multilevel Data Structures. *American Journal of Political Science*, 46(1):218–237
  - Bell, A. and Jones, K. (2015). Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data. *Political Science Research and Methods*, 3(1):133–153

  - Shor, B., Bafumi, J., Keele, L., and Park, D. (2007). A Bayesian Multilevel Modeling Approach to Time-Series Cross-Sectional Data. *Political Analysis*, 15(2):165–181
  - Pang, X. (2010). Modeling Heterogeneity and Serial Correlation in Binary Time-Series Cross-sectional Data: A Bayesian Multilevel Model with AR(p) Errors. *Political Analysis*, 18:470–498

- Application examples:

  - Ward, M. D., Siverson, R. M., and Cao, X. (2007). Disputes, Democracies, and Dependencies: A Reexamination of the Kantian Peace. *American Journal of Political Science*, 51(3):583–601
  - Blaydes, L. and Linzer, D. A. (2012). Elite Competition, Religiosity and Anti-Americanism in the Islamic World. *American Political Science Review*, 106(2):225–243
  - Cobb, R. V., Greiner, D. J., and Quinn, K. M. (2012). Can Voter ID Laws Be Administered in a Race-Neutral Manner? Evidence from the City of Boston in 2008. *Quarterly Journal of Political Science*, 7(1):1–33
  - Stegmueller, D. (2013). How Many Countries for Multilevel Modeling? A Comparison of Frequentist and Bayesian Approaches. *American Journal of Political Science*, 57(3):748–761
  - Chaudoin, S., Milner, H. V., and Pang, X. (2014). International Systems and Domestic Politics: Linking Complex Theories with Empirical Models in International Relations. *International Organization*, forthcoming

- On multilevel regression with post-stratification (MRP):

  - Park, D. K., Gelman, A., and Bafumi, J. (2004). Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls. *Political Analysis*, 12(4):375–385
  - Lax, J. R. and Phillips, J. H. (2009). How Should We Estimate Public Opinion in The States? *American Journal of Political Science*, 53(1):107–121
  - Lock, K. and Gelman, A. (2010). Bayesian Combination of State Polls and Election Forecasts. *Political Analysis*, 18(3):337–348

Unit 9: Conclusion; Workflow optimization & reproducibility

Background reading:

- Gelman, A. (2008). Objections to Bayesian Statistics. *Bayesian Analysis*, 3(3):445–450
- Broman, K. (N.d.). Initial Steps Toward Reproducible Research. Online at http://kbroman.org/steps2rr/
- Gandrud, C. (2013). *Reproducible Research with R and RStudio*. Chapman and Hall/CRC, Boca Raton, FL
- Dafoe, A. (2014). Science Deserves Better: The Imperative to Share Complete Replication Files. *PS: Political Science & Politics*, 47(1):60–66