A First Look at Bayesian Adaptive Methods

Fan Bu

October 12, 2021

Overview

1. Basics of Bayesian hypothesis testing
2. Bayesian sequential testing
3. Examples of Bayesian adaptive methods application
4. Methodological gaps BST can address

Basics of Bayesian hypothesis testing

Basic setup

- Observe data $X \sim p(x \mid \theta)$ (data model) w/. parameter θ
- Wish to test

$$H_0 : \theta \in \Theta_0 \qquad \text{v.s.} \qquad H_1 : \theta \in \Theta_1$$

Basic setup

- Observe data $X \sim p(x \mid \theta)$ (data model) w/. parameter θ
- Wish to test

$$H_0 : \theta \in \Theta_0 \qquad \text{v.s.} \qquad H_1 : \theta \in \Theta_1$$

- prior beliefs about the hypotheses: $p := P(H_0)$ (then $1 - p = P(H_1)$)
- priors for θ: π_i under H_i $(i = 0, 1)$

Evidence checking via posterior inference

- posterior distribution for θ under H_i $(i = 0, 1)$:

$$p(\theta \mid x, H_i) = \frac{p(x \mid \theta)\pi_i(\theta)}{m_i(x)}; \tag{1}$$

- $m_i(x) = \int p(x \mid \theta)\pi_i(\theta)d\theta$: marginal distribution of data X (or data evidence) under H_i.

Evidence checking via posterior inference

- posterior distribution for θ under H_i $(i = 0, 1)$:

$$p(\theta \mid x, H_i) = \frac{p(x \mid \theta)\pi_i(\theta)}{m_i(x)}; \tag{1}$$

- $m_i(x) = \int p(x \mid \theta)\pi_i(\theta)d\theta$: marginal distribution of data X (or data evidence) under H_i.
- posterior odds in favor of H_0:

$$\frac{P(H_0 \mid x)}{P(H_1 \mid x)} = \frac{m_0(x)p}{m_1(x)(1-p)} = \frac{m_0(x)}{m_1(x)}\frac{p}{1-p}; \tag{2}$$

Evidence checking via posterior inference

- posterior distribution for θ under H_i $(i = 0, 1)$:

$$p(\theta \mid x, H_i) = \frac{p(x \mid \theta)\pi_i(\theta)}{m_i(x)}; \tag{1}$$

- $m_i(x) = \int p(x \mid \theta)\pi_i(\theta)d\theta$: marginal distribution of data X (or data evidence) under H_i.
- posterior odds in favor of H_0:

$$\frac{P(H_0 \mid x)}{P(H_1 \mid x)} = \frac{m_0(x)p}{m_1(x)(1-p)} = \frac{m_0(x)}{m_1(x)}\frac{p}{1-p}; \tag{2}$$

- Bayes factor (posteriors odds when $p = 1/2$):

$$BF_{01} = \frac{m_0(x)}{m_1(x)}. \tag{3}$$

Evidence checking (Cont'd)

- For simple v.s. simple hypothesis testing, BF is same as likelihood ratio:

$$BF_{01} = \frac{p(x \mid \theta_0)}{p(x \mid \theta_1)}. \tag{4}$$

Evidence checking (Cont'd)

- For simple v.s. simple hypothesis testing, BF is same as likelihood ratio:

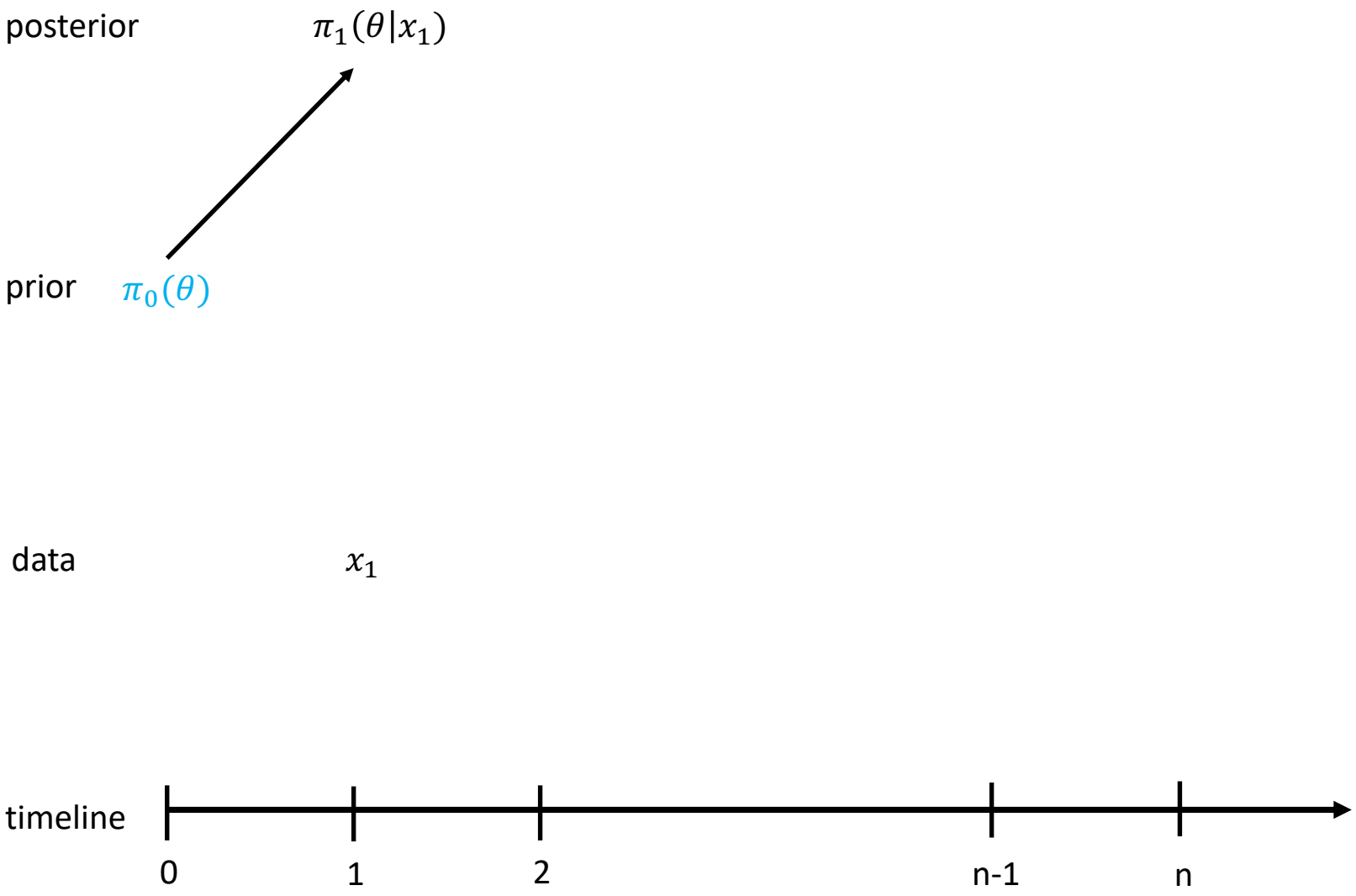
$$BF_{01} = \frac{p(x \mid \theta_0)}{p(x \mid \theta_1)}. \tag{4}$$

- With threshold A, would reject H_0 if $BF_{01} < A$;
- Or, equivalently, with threshold δ_L, reject H_0 if $P(H_0 \mid x) < \delta_L$.

Bayesian sequential testing

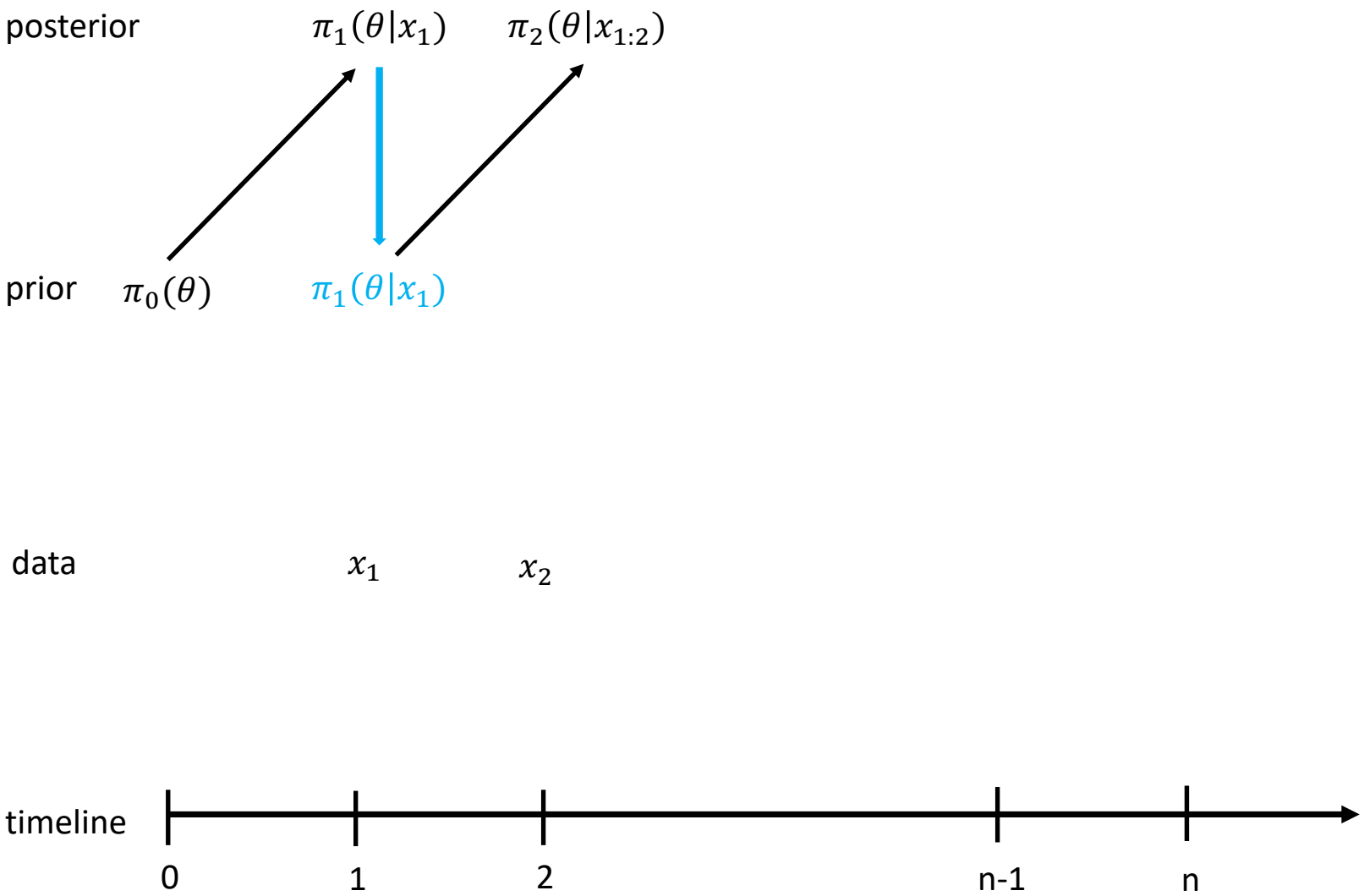
Bayesian analysis is intrinsically sequential

A current "posterior" can become the "prior" for future inference.



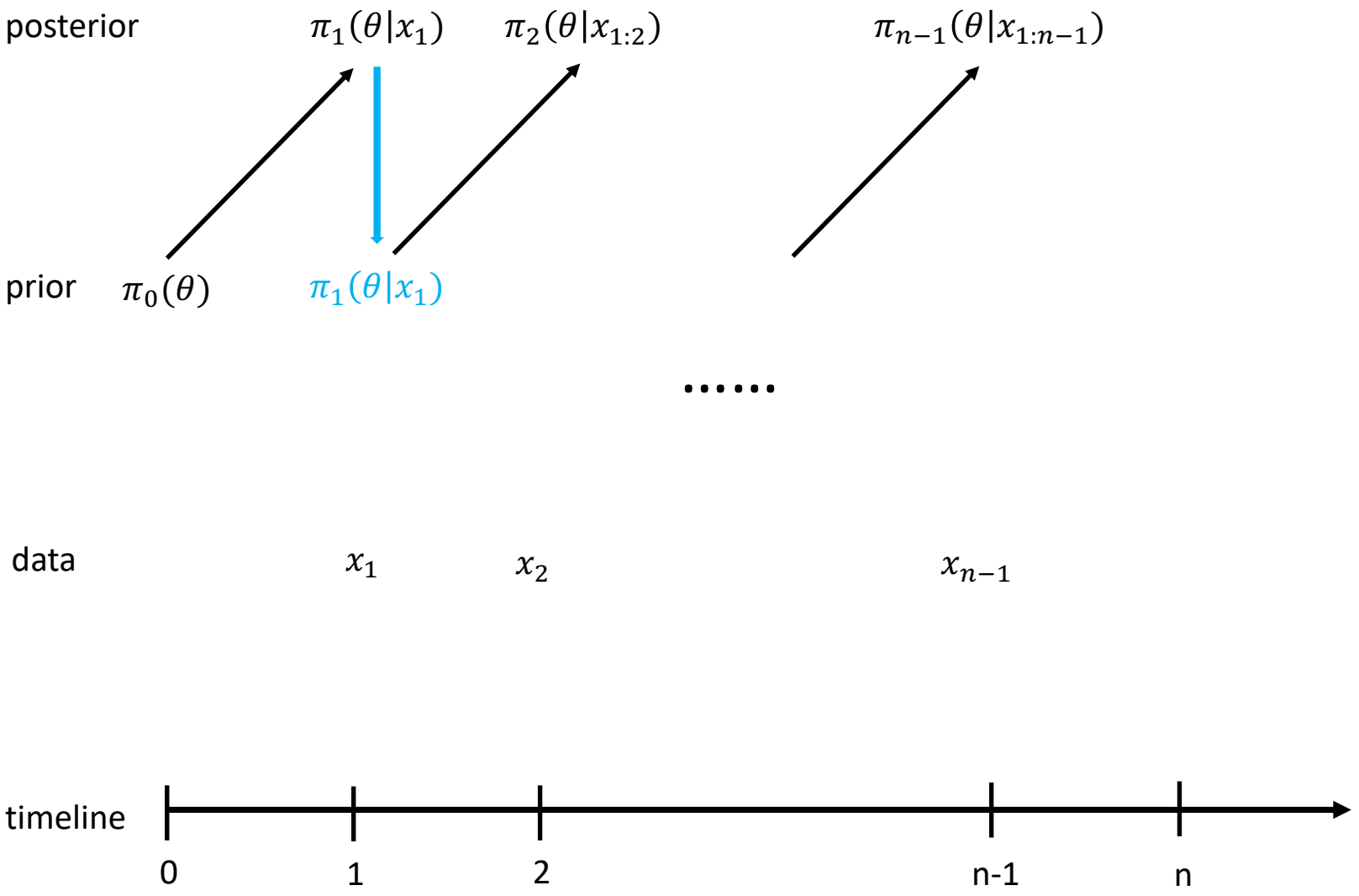
Bayesian analysis is intrinsically sequential

A current "posterior" can become the "prior" for future inference.



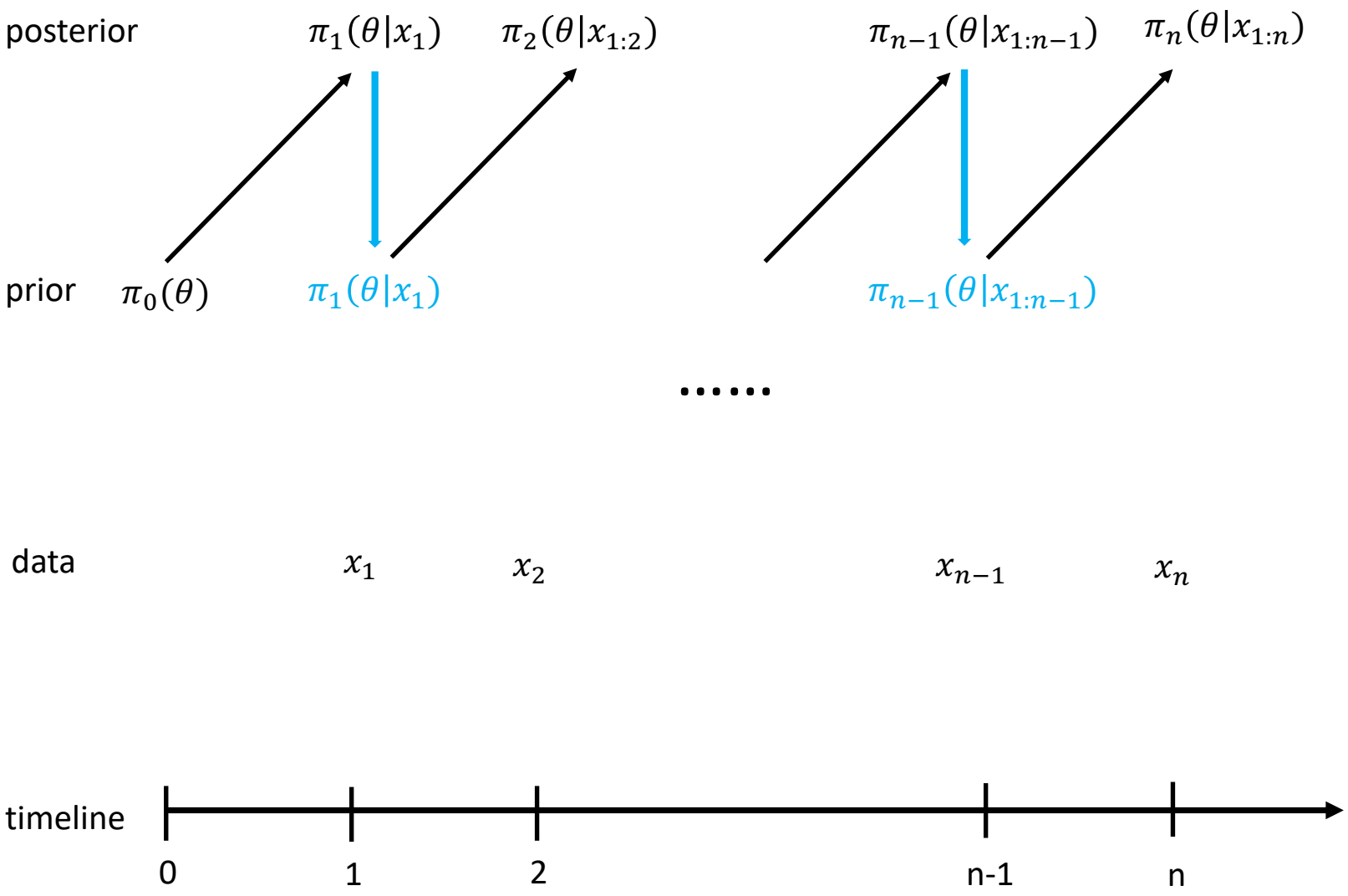
Bayesian analysis is intrinsically sequential

A current "posterior" can become the "prior" for future inference.



Bayesian analysis is intrinsically sequential

A current "posterior" can become the "prior" for future inference.



How it usually works

Given thresholds $B \geq 1 \geq A$, with Bayes factor $BF_{01}^{(n)}$ acquired at step n [1]:

- if $BF_{01}^{(n)} > B$, stop the study and accept H_0;
- if $A < BF_{01}^{(n)} < B$, continue the study;
- if $BF_{01}^{(n)} < A$, stop the study and reject H_0.

[1] See, for example, [Bar46, Wet61, BBW94, BW88, BBW97, BBW99].

How it usually works

Given thresholds $B \geq 1 \geq A$, with Bayes factor $BF_{01}^{(n)}$ acquired at step n [1]:

- if $BF_{01}^{(n)} > B$, stop the study and accept H_0;
- if $A < BF_{01}^{(n)} < B$, continue the study;
- if $BF_{01}^{(n)} < A$, stop the study and reject H_0.

Of course, we can make decisions based on the posterior probability $P(H_0 \mid x)$ or $P(H_1 \mid x)$ with specified thresholds (e.g., [Cor66]).

[1] See, for example, [Bar46, Wet61, BBW94, BW88, BBW97, BBW99].

Examples of Bayesian adaptive methods application

The setup

- Compare effect θ (e.g., log relative risk) with baseline θ_0:

$$H_0 : \theta \leq \theta_0 \qquad \text{v.s.} \qquad H_1 : \theta > \theta_0.$$

The setup

- Compare effect θ (e.g., log relative risk) with baseline θ_0:

$$H_0 : \theta \leq \theta_0 \qquad \text{v.s.} \qquad H_1 : \theta > \theta_0.$$

- Or, consider a "minimal practical increase" δ:

$$H_0 : \theta \leq \theta_0 + \delta \qquad \text{v.s.} \qquad H_1 : \theta > \theta_0 + \delta.$$

Binary decisions

- Specify lower and upper probability bounds δ_L and δ_U;
- **Early stopping for signal** if $P(H_1 \mid x) > \delta_U$;
- **Early stopping for futility** if $P(H_1 \mid x) < \delta_L$.
- Common choice: $\delta_L = 0.05, 0.1$, $\delta_U = 0.8, 0.95$, OR calibrated through simulations. [2]

[2] See, for example, [TSE95, SJM+06, ZLK+08, BCLM10, LSR20], etc.

Non-binary decisions

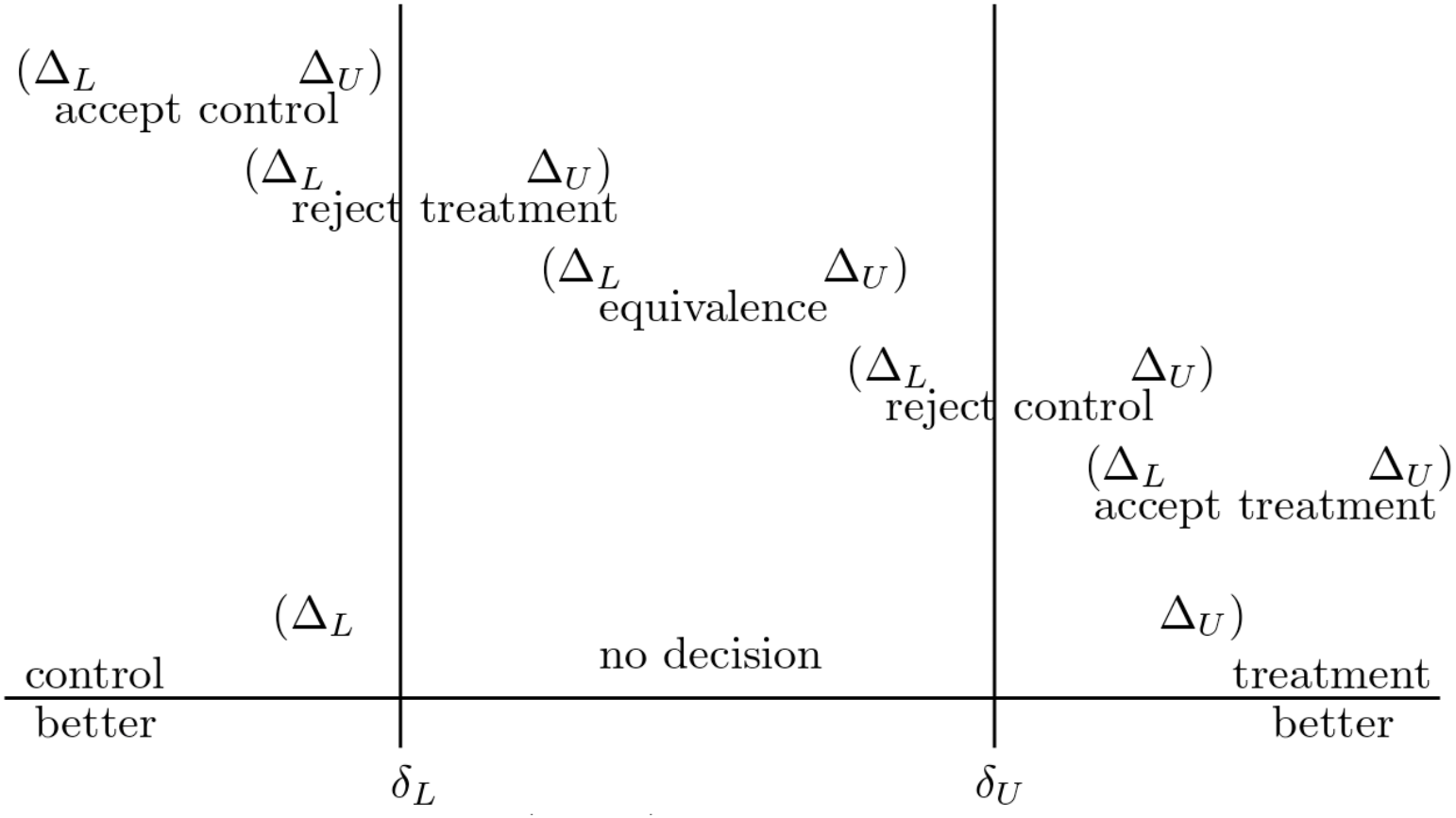
We can draw different conclusions about the strength of evidence in support of either hypothesis.

Decisions based on BF_{10} (Bayes factor in favor of H_1) [Jef98, KR95, SWZP17]:

- $1 < BF_{10} < 3$: anecdotal evidence
- $3 < BF_{10} < 10$: moderate evidence
- $10 < BF_{10} < 30$: strong evidence
- $BF_{10} > 30$ very strong evidence.

Non-binary decisions (Cont'd) - zone method

Or, given an indifference zone $[\delta_L, \delta_U]$, make decisions based on credible interval for $\Delta = \theta - \theta_0$ [BCLM10].



Advantages and methodological gaps

Advantages

- Easy to incorporate historical data through priors:
  - e.g., historical adverse event incidence rate 1%, then can use $\text{Beta}(1, 99)$ prior for incidence rate θ
  - can "discount" prior to address temporal changes [WH06]

Advantages

- Easy to incorporate historical data through priors:
  - e.g., historical adverse event incidence rate 1%, then can use Beta$(1, 99)$ prior for incidence rate θ
  - can "discount" prior to address temporal changes [WH06]
- Can test multiple hypotheses (e.g., detect multiple safety signals) simultaneously
  [GB98, BH99, SB06, LT07, GH10, KHM12, KM13, Kac14, Ber13]

Advantages

- Can incorporate model selection/averaging within the Bayesian framework (e.g., [Raf95, Was00, CGM+01, SPD+09, WT08, SDM20])
- Non-violation of **the likelihood principle**:
  - "Given a statistical model, all information relavent to inferences about model parameters θ is contained in the likelihood function $L(\theta; x)$";
  - "If two likelihood functions $L_1(\theta; x)$ and $L_2(\theta; x)$ are proportional then they contain same information about θ".
  - Conclusions about a study shouldn't depend on how we look at the data or how we decide when to stop - frequentist sequential analysis clearly violates this
  - Bayesian analysis respects the likelihood principle

Potential challenges

- Calibration of the thresholds
  - No universal rule (like $\alpha = 0.05$)
  - Theoretical results for matching frequentist error rate bounds available for simple models (e.g.,[BBW94, Cor66])

Potential challenges

- Calibration of the thresholds
  - No universal rule (like $\alpha = 0.05$)
  - Theoretical results for matching frequentist error rate bounds available for simple models (e.g.,[BBW94, Cor66])
  - Can use simulation-based or data-driven calibration (negative/positive controls utilizing OHDSI network) to satisfy frequentist operating characteristics requirements

Potential challenges

- Model-free or likelihood-free cases
  - When a data model cannot be be easily specified or a likelihood function doesn't exist;
  - Traditional Bayesian approach can't be applied, but generalized Bayesian methods for likelihood-free or loss-function-based inference might be employed (e,g., [TS14, LHW19, BHW16]).

Thank you!

Likelihood Principle Examples

Example 1: Two scientists are collaborating on an experiment. They have a joint graduate student who is conducting the experiment, and they both are watching her work.

- The observations $X_1, X_2, \ldots$ are i.i.d. $Bernoulli(\theta)$ random variables.
- The scientists agree they are testing $H_0 : \theta = 0.5$ versus $H_1 : \theta > 0.5$.
- After the ninth observation, they simultaneously say "That's enough data," and they tell the student to stop experimenting.
- The data consisted of 9 successes and 3 failures.
- Each scientist analyzes the data; when getting back together, they are shocked that they disagree.
  - Scientist 1 claims that there is not significant evidence at the 0.05 level.
  - Scientist 2 says there is significant evidence at the 0.05 level.
- How did this disagreement happen?

Credit: James Berger

Scientist 1's analysis: He had planned from the beginning to take just 12 observations (but had not communicated this). Thus the number of successes, X, is $Binomial(12, \theta)$, and the p-value for the observed $x = 9$ is

$$p = Pr(X \geq 9 \mid \theta = 0.5) = \sum_{x=9}^{12} \binom{12}{x} 0.5^x (1 - 0.5)^{(12-x)} = .0730\,.$$

Scientist 2's analysis: She had planned to continue taking observations until observing 3 failures. Thus, for her, X has a $Negative - binomial(3, \theta)$ distribution, and the p-value is

$$p = Pr(X \geq 9 \mid \theta = 0.5) = \sum_{x=9}^{\infty} \binom{x+2}{x} 0.5^x (1 - 0.5)^3 = .0338\,.$$

- The two scientists had different *stopping rules*.
- But note that these were just thoughts in their heads; these thoughts had no effect on the actual experiment that was conducted or the results.
- The stopping rule principle says such thoughts should not matter; the stopping rule should not affect the analysis.

Credit: James Berger

For Scientist 1 the observed likelihood function was

$$\mathcal{L}_1(\theta) = \binom{12}{9}\theta^9(1-\theta)^3\,;$$

For Scientist 2 it was

$$\mathcal{L}_2(\theta) = \binom{11}{9}\theta^9(1-\theta)^3\,.$$

Since $\mathcal{L}_1(\theta) \propto \mathcal{L}_2(\theta)$, the Likelihood Principle also says that the evidence about θ from either viewpoint is the same.

Credit: James Berger

Example 2: A scientist enters the statistician's office with $n = 100$ observations, assumed to be independent and from a $N(\theta, 1)$ distribution, with the desire to test $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. She reports that $\bar{x}_n = 0.2$, so the standardized test statistic is $z_{100} = \sqrt{n}\ |\bar{x}_{100} - 0| = 2$.

- A careless classical statistician might simply conclude that there is significant evidence against H_0 at the 0.05 level.
- A careful classical statistician will ask the scientist "Why did you cease experimentation after 100 observations?"
  - If the scientist replies, "I just decided to take a batch of 100 observations," there would seem to be no problem.
  - But there is another important question that should be asked (from the classical perspective), namely: "What would you have done had the first 100 observations not yielded significance?"

Credit: James Berger

To see the reasons for this question, suppose the scientist replies: "I would then have taken another batch of 100 observations." This reply does not completely specify a stopping rule, but the scientist might agree that she was implicitly considering a stopping rule of the form

- take the first 100 observations;
  - if $z_{100} > k$ for some critical value k, then stop and reject H_0,
  - if $z_{100} < k$ then take another 100 observations and reject if $z_{200} > k$.
- For this procedure to have level $\alpha = 0.05$, k must be chosen to be 2.18 (Pocock, 1977). Since the actual data had $z_{100} = 2 < 2.18$, the scientist could not actually conclude significance, and hence would have to take the next 100 observations.

This strikes many people as peculiar. The interpretation of the results of an experiment depends not only on the data obtained and the way it was obtained, but also upon thoughts of the experimenter concerning plans for the future.

Credit: James Berger

The puzzled scientist scientist leaves and gets the next 100 observations.

- She reports that $z_{200} = 2.2 > 2.18$; has significance now been obtained?
- No, again the statistician asks what the scientist would have done had the results not been significant.
  - The scientist says, "If my grant renewal were to be approved, I would then take another 100 observations;
  - If the grant renewal were rejected, I would have no more funds and would have to stop the experiment at 200 observations."
  - The classical statistician *must* then advise her that, if the grant is rejected, significance has been obtained, but otherwise another 100 observations must be taken.

Note: This is not at all fanciful; the standard practice in psychology experiments is to do precisely the above: keep taking observations in batches until $p < 0.05$, ignoring the issue of the stopping rule.

Credit: James Berger

References I

- George A Barnard, *Sequential tests in industrial statistics*, Supplement to the Journal of the Royal Statistical Society **8** (1946), no. 1, 1–21.
- James O Berger, Lawrence D Brown, and Robert L Wolpert, *A unified conditional frequentist and bayesian test for fixed and sequential simple hypothesis testing*, The Annals of Statistics (1994), 1787–1807.
- James O Berger, Ben Boukai, and Yinping Wang, *Unified frequentist and bayesian testing of a precise hypothesis*, Statistical Science **12** (1997), no. 3, 133–160.
- James O Berger, Benzion Boukai, and Yinping Wang, *Simultaneous bayesian-frequentist sequential testing of nested hypotheses*, Biometrika **86** (1999), no. 1, 79–92.
- Scott M Berry, Bradley P Carlin, J Jack Lee, and Peter Muller, *Bayesian adaptive methods for clinical trials*, CRC press, 2010.

References II

- James O Berger, *Statistical decision theory and bayesian analysis*, Springer Science & Business Media, 2013.
- Donald A Berry and Yosef Hochberg, *Bayesian perspectives on multiple comparisons*, Journal of Statistical Planning and Inference **82** (1999), no. 1-2, 215–227.
- Pier Giovanni Bissiri, Chris C Holmes, and Stephen G Walker, *A general framework for updating belief distributions*, Journal of the Royal Statistical Society. Series B, Statistical methodology **78** (2016), no. 5, 1103.
- James O Berger and Robert L Wolpert, *The likelihood principle*, IMS, 1988.

References III

- Hugh Chipman, Edward I George, Robert E McCulloch, Merlise Clyde, Dean P Foster, and Robert A Stine, *The practical implementation of bayesian model selection*, Lecture Notes-Monograph Series (2001), 65–134.
- Jerome Cornfield, *A bayesian test of some classical hypotheses—with applications to sequential clinical trials*, Journal of the American Statistical Association **61** (1966), no. 315, 577–594.
- Ramanan Gopalan and Donald A Berry, *Bayesian multiple comparisons using dirichlet process priors*, Journal of the American Statistical Association **93** (1998), no. 443, 1130–1139.
- Mengye Guo and Daniel F Heitjan, *Multiplicity-calibrated bayesian hypothesis tests*, Biostatistics **11** (2010), no. 3, 473–483.
- Harold Jeffreys, *The theory of probability*, OUP Oxford, 1998.

References IV

- KJ Kachiashvili, *The methods of sequential analysis of bayesian type for the multiple testing problem*, Sequential Analysis **33** (2014), no. 1, 23–38.
- KJ Kachiashvili, MA Hashmi, and A Mueed, *Sensitivity analysis of classical and conditional bayesian problems of many hypotheses testing*, Communications in Statistics-Theory and Methods **41** (2012), no. 4, 591–605.
- KJ Kachiashvili and A Mueed, *Conditional bayesian task of testing many hypotheses*, Statistics **47** (2013), no. 2, 274–293.
- Robert E Kass and Adrian E Raftery, *Bayes factors*, Journal of the american statistical association **90** (1995), no. 430, 773–795.

References V

- SP Lyddon, CC Holmes, and SG Walker, *General bayesian updating and the loss-likelihood bootstrap*, Biometrika **106** (2019), no. 2, 465–478.
- Rongxia Li, Brock Stewart, and Charles Rose, *A bayesian approach to sequential analysis in post-licensure vaccine safety surveillance*, Pharmaceutical statistics **19** (2020), no. 3, 291–302.
- Aurélie Labbe and Mary E Thompson, *Multiple testing using the posterior probabilities of directional alternatives, with application to genomic studies*, Canadian Journal of Statistics **35** (2007), no. 1, 51–68.
- Adrian E Raftery, *Bayesian model selection in social research*, Sociological methodology (1995), 111–163.

References VI

- James G Scott and James O Berger, *An exploration of aspects of bayesian multiple testing*, Journal of statistical planning and inference **136** (2006), no. 7, 2144–2162.
- SGJ Senarathne, Christopher C Drovandi, and James M McGree, *A laplace-based algorithm for bayesian adaptive design*, Statistics and Computing **30** (2020), no. 5, 1183–1208.
- Michael K Smith, Ieuan Jones, Mark F Morris, Andrew P Grieve, and Keith Tan, *Implementation of a bayesian adaptive design in a proof of concept study*, Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry **5** (2006), no. 1, 39–50.
- Klaas Enno Stephan, Will D Penny, Jean Daunizeau, Rosalyn J Moran, and Karl J Friston, *Bayesian model selection for group studies*, Neuroimage **46** (2009), no. 4, 1004–1017.

References VII

- Felix D Schönbrodt, Eric-Jan Wagenmakers, Michael Zehetleitner, and Marco Perugini, *Sequential hypothesis testing with bayes factors: Efficiently testing mean differences.*, Psychological methods **22** (2017), no. 2, 322.
- Brandon M Turner and Per B Sederberg, *A generalized, likelihood-free method for posterior estimation*, Psychonomic bulletin & review **21** (2014), no. 2, 227–250.
- Peter F Thall, Richard M Simon, and Elihu H Estey, *Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes*, Statistics in medicine **14** (1995), no. 4, 357–379.
- Larry Wasserman, *Bayesian model selection and model averaging*, Journal of mathematical psychology **44** (2000), no. 1, 92–107.

References VIII

- GB Wetherill, *Bayesian sequential analysis*, Biometrika **48** (1961), no. 3/4, 281–292.
- Mike West and Jeff Harrison, *Bayesian forecasting and dynamic models*, Springer Science & Business Media, 2006.
- J Kyle Wathen and Peter F Thall, *Bayesian adaptive model selection for optimizing group sequential clinical trials*, Statistics in medicine **27** (2008), no. 27, 5586–5604.
- Xian Zhou, Suyu Liu, Edward S Kim, Roy S Herbst, and J Jack Lee, *Bayesian adaptive design for targeted therapy development in lung cancer—a step toward personalized medicine*, Clinical Trials **5** (2008), no. 3, 181–193.