Bayesian Estimation of Parametrized Efficiency

John T. Whelan
john.whelan@ligo.org

Center for Comp. Relativity & Gravitation, Rochester Institute of Technology
and
Max-Planck-Institut für Gravitationsphysik (Albert-Einstein-Institut), Hannover

Data Analysis Group Seminar
Max-Planck-Institut für Gravitationsphysik
(Albert-Einstein-Institut), Hannover
2015 July 29
LIGO-G1500977-v2

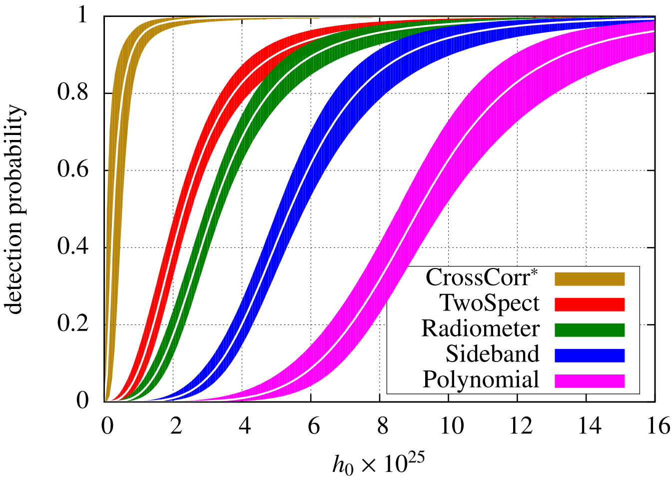
Outline

Outline

Background/Motivation

- In the Sco X-1 MDC [Messenger et al *PRD* **92**, 023006 (2015)] 5 pipelines each reported detection/non-detection for each of 50 signals w/different h_0 and optimal SNR
- Actual detection efficiency depended on numerous factors; for comparison, wanted simple estimated efficiency curves
- How to turn 50 yes/no answers into an estimated efficiency as a function of h_0 (or optimal SNR)?
- Added complication: one search found all 50 signals Can we make estimate more interesting/realistic than 100% efficiency?

Efficiency Curves from Sco X-1 MDC Paper



Bernoulli Experiments

- General formalism: N Bernoulli trials labelled by $i \in [1, N]$
- Results are $\{D_i\}$:
  $D_i = 1$ means success on ith trial, $D_i = 0$ means failure
- Properties (e.g., signal strength) of ith trial are $\mathbf{x}_i$
- "Efficiency" $\varepsilon(\mathbf{x}_i; \boldsymbol{\theta})$ depends on model parameters $\boldsymbol{\theta}$
- Likelihood function

$$P(\{D_i\}|\boldsymbol{\theta}, I) = \prod_i \left(\varepsilon(\mathbf{x}_i; \boldsymbol{\theta})\right)^{D_i} \left(1 - \varepsilon(\mathbf{x}_i; \boldsymbol{\theta})\right)^{1-D_i}$$

  ($I \equiv$ background information about expt, our knowledge, etc)
- Goal: use results $\{D_i\}$ to say something about $\boldsymbol{\theta}$
  and the expected efficiency $\varepsilon(\mathbf{x}_0; \boldsymbol{\theta})$ of a trial w/props $\mathbf{x}_0$
- Orthodox method: "bin" together trials w/similar props & use fraction of detections in bin as estimate of ε for that bin

Bayesian Method

- Likelihood function

$$P(\{D_i\}|\boldsymbol{\theta},I) = \prod_i (\varepsilon(\mathbf{x}_i;\boldsymbol{\theta}))^{D_i} (1-\varepsilon(\mathbf{x}_i;\boldsymbol{\theta}))^{1-D_i}$$

- Construct posterior pdf on parameters $\boldsymbol{\theta}$:

$$f(\boldsymbol{\theta}|\{D_i\},I) = \frac{f(\boldsymbol{\theta}|I)\,P(\{D_i\}|\boldsymbol{\theta},I)}{P(\{D_i\}|I)} \propto f(\boldsymbol{\theta}|I)\,P(\{D_i\}|\boldsymbol{\theta},I)$$

- Posterior pdf on efficiency ε_0 for a trial w/properties $\mathbf{x}_0$:

$$f(\varepsilon_0|\{D_i\},I) = \int d\boldsymbol{\theta}\,\delta\left(\varepsilon_0 - \varepsilon(\mathbf{x}_0;\boldsymbol{\theta})\right) f(\boldsymbol{\theta}|\{D_i\},I)$$

- Cumulative distribution function

$$P(\varepsilon_0 \le \varepsilon_0'|\{D_i\},I) = \int_{\varepsilon(\mathbf{x}_0;\boldsymbol{\theta})\le\varepsilon_0'} d\boldsymbol{\theta}\, f(\boldsymbol{\theta}|\{D_i\},I)$$

Examples

1. Binomial: all trials the same, so $\mathbf{x}_i$ irrelevant
   Parameter $\boldsymbol{\theta}$ is just ε itself, or e.g., $\lambda = \ln \frac{\varepsilon}{1-\varepsilon}$

Examples

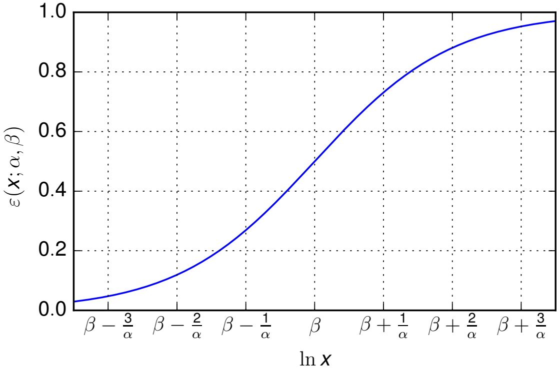
1. Binomial: all trials the same, so $\mathbf{x}_i$ irrelevant
   Parameter $\boldsymbol{\theta}$ is just ε itself, or e.g., $\lambda = \ln \frac{\varepsilon}{1-\varepsilon}$
2. Sigmoid: $\mathbf{x}_i$ is single signal strength $x_i > 0$
   Parameters $\boldsymbol{\theta}$ are α, β

$$\varepsilon(x_i; \alpha, \beta) = \frac{1}{1 + e^{-\alpha(\ln x_i - \beta)}}$$

Examples

2. Sigmoid: $\mathbf{x}_i$ is single signal strength $x_i > 0$
   Parameters $\boldsymbol{\theta}$ are α, β

$$\varepsilon(x_i; \alpha, \beta) = \frac{1}{1 + e^{-\alpha(\ln x_i - \beta)}}$$



Examples

1. Binomial: all trials the same, so $\mathbf{x}_i$ irrelevant
   Parameter $\boldsymbol{\theta}$ is just ε itself, or e.g., $\lambda = \ln \frac{\varepsilon}{1-\varepsilon}$
2. Sigmoid: $\mathbf{x}_i$ is single signal strength $x_i > 0$
   Parameters $\boldsymbol{\theta}$ are α, β

$$\varepsilon(x_i; \alpha, \beta) = \frac{1}{1 + e^{-\alpha(\ln x_i - \beta)}}$$

3. Bradley-Terry-Zermelo/paired comparisons:
   (Zermelo 1929, Bradley & Terry 1952)
   $\mathbf{x}_i \equiv$ choice of items A, B to compare
   $\boldsymbol{\theta} \equiv$ item strengths $\{\pi_A\}$ (up to overall constant factor)

$$\varepsilon(A, B|\{\pi_A\}) = \frac{\pi_A}{\pi_A + \pi_B}$$

Outline

Binomial Experiment: All Trials Identical

- Likelihood determined by $D = \sum_i D_i$,
  total # of successes in N trials ("sufficient statistic"):

$$P(D|\varepsilon, I) \propto \varepsilon^D (1-\varepsilon)^{N-D}$$

- Posterior pdf for efficiency $\varepsilon \in [0,1]$:

$$f(\varepsilon|D, I) \propto f(\varepsilon|I)\, \varepsilon^D (1-\varepsilon)^{N-D}$$

- Change variables to $\lambda = \ln \frac{\varepsilon}{1-\varepsilon} \in (-\infty, \infty)$ (log odds ratio)
  Densities transform as $f(\lambda)\, d\lambda = f(\varepsilon)\, d\varepsilon$, so

$$f(\varepsilon) = \frac{f(\lambda)}{\varepsilon(1-\varepsilon)}$$

Choosing a Prior for Binomial Efficiency

- Bayes/Laplace (Laplace 1814); uniform in efficiency

$$f(\varepsilon|I_{\mathrm{BL}}) = 1, \qquad 0 \leq \varepsilon \leq 1$$

- Haldane (Haldane 1932); uniform in log-odds, $f(\lambda|I_H) = \mathrm{const}$

$$f(\varepsilon|I_{\mathrm{H}}) \propto \frac{1}{\varepsilon(1-\varepsilon)}, \qquad 0 < \varepsilon < 1$$

  Advocated by Jaynes (2003) as "complete ignorance"; not normalizable

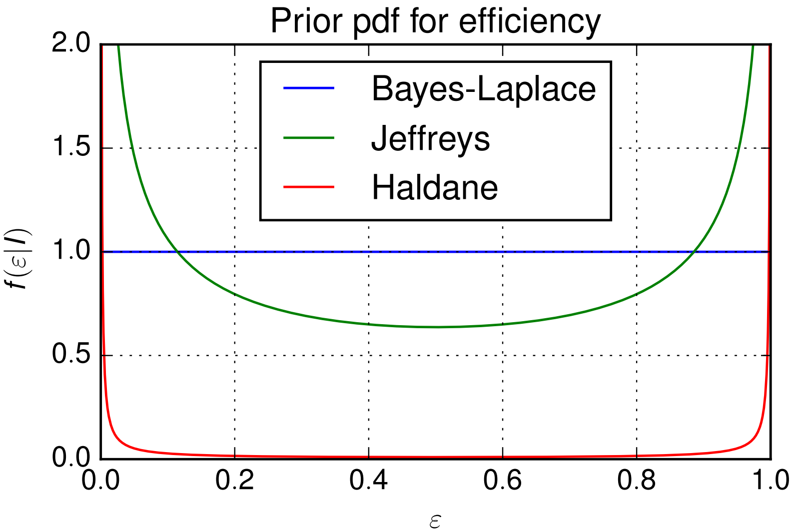
- Jeffreys (Jeffreys 1946); "in-between"

$$f(\varepsilon|I_{\mathrm{J}}) = \frac{1}{\pi\sqrt{\varepsilon(1-\varepsilon)}}, \qquad 0 < \varepsilon < 1$$

All special cases of Beta dist w/params $\nu > \delta > 0$:

$$f(\varepsilon;\nu,\delta) = \frac{\Gamma(\nu)}{\Gamma(\delta)\Gamma(\nu-\delta)}\varepsilon^{\delta-1}(1-\varepsilon)^{\nu-\delta-1}, \qquad 0 < \varepsilon < 1$$

Comparison of prior pdfs



Note: Definition of Jeffreys prior

- "Jeffreys prior" sometimes (mis)-used to mean uniform-in-log
- Actual definition is $f(\boldsymbol{\theta}|I_{\mathrm{J}}) \propto \sqrt{\mathcal{I}(\boldsymbol{\theta})}$ where

$$\mathcal{I}(\boldsymbol{\theta}) = \det\left\{-E\left[\frac{\partial^2 P(\{D_i\}|\boldsymbol{\theta}, I)}{\partial\theta^\alpha\,\partial\theta^\beta}\right]\right\}$$

  is the determinant of the Fisher information matrix
- Depends on the likelihood
- Defined so equivalent pdf derived after change of variables
- For binomial, usually derived in terms of ε;
  derive in terms of $\lambda = \ln\frac{\varepsilon}{1-\varepsilon}$ for illustration

Jeffreys prior for binomial experiment

- Likelihood in terms of log odds ratio λ:

$$P(D|\lambda, I) \propto \frac{e^{\lambda D}}{(1+e^{\lambda})^N}$$

- Log-likelihood

$$\ell(\lambda) = \ln P(D|\lambda, I) = \lambda D - N \ln(1+e^{\lambda}) + \text{constant}$$

- After some differentiation ...

$$\ell''(\lambda) = -N\frac{e^{-\lambda}}{(1+e^{\lambda})^2} \quad \Longrightarrow \quad f(\lambda|I_{\text{J}}) \propto \frac{e^{-\lambda/2}}{1+e^{\lambda}} = \varepsilon^{1/2}(1-\varepsilon)^{1/2}$$

- Since $f(\varepsilon) = \frac{f(\lambda)}{\varepsilon(1-\varepsilon)}$,

$$f(\varepsilon|I_{\text{J}}) \propto \varepsilon^{-1/2}(1-\varepsilon)^{-1/2}$$

Aside: Conjugate Prior Family

- Beta distribution priors convenient here; conjugate prior family for Binomial
- I.e., if the prior has that form

$$f(\varepsilon|I) \propto \varepsilon^{\delta-1}(1-\varepsilon)^{\nu-\delta-1}$$

then the posterior does too, w/$\delta \longrightarrow \delta + D$ & $\nu \longrightarrow \nu + N$:

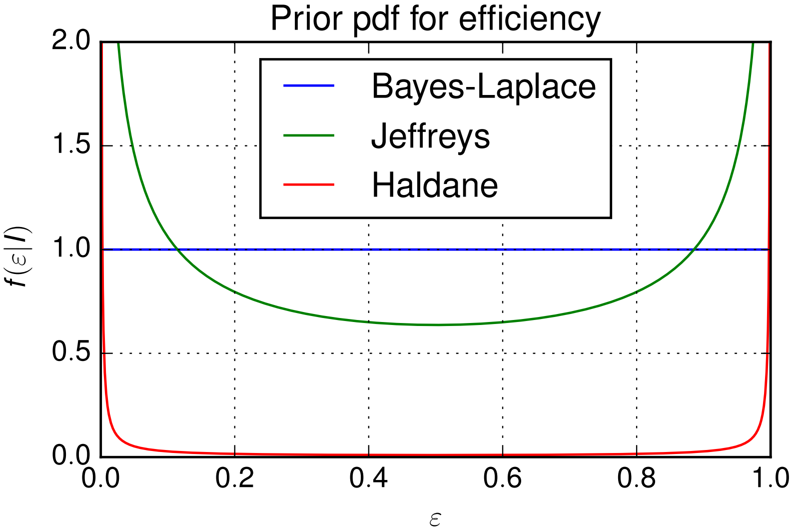
$$f(\varepsilon|D,I) \propto \varepsilon^{(\delta+D)-1}(1-\varepsilon)^{(\nu+N)-(\delta+D)-1}$$

- In particular

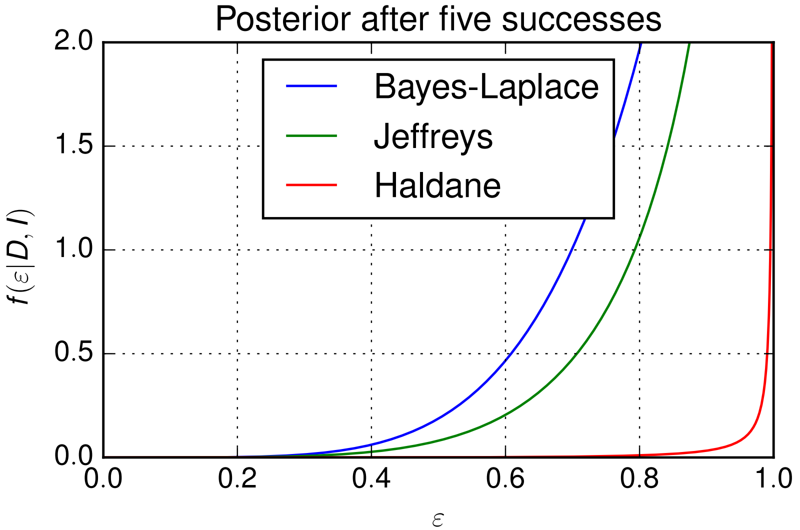
$$E[\varepsilon|D,I] = \int_0^1 d\varepsilon\, \varepsilon\, f(\varepsilon|D;I) = \frac{\delta+D}{\nu+N}$$

- Note if $D=N$ (all trials succeeded): $E[\varepsilon|D=N,I_{\mathrm{H}}]=1$; $E[\varepsilon|D=N,I_{\mathrm{BL}}]=\frac{N}{N+1}$; $E[\varepsilon|D=N,I_{\mathrm{J}}]=\frac{2N}{2N+1}$

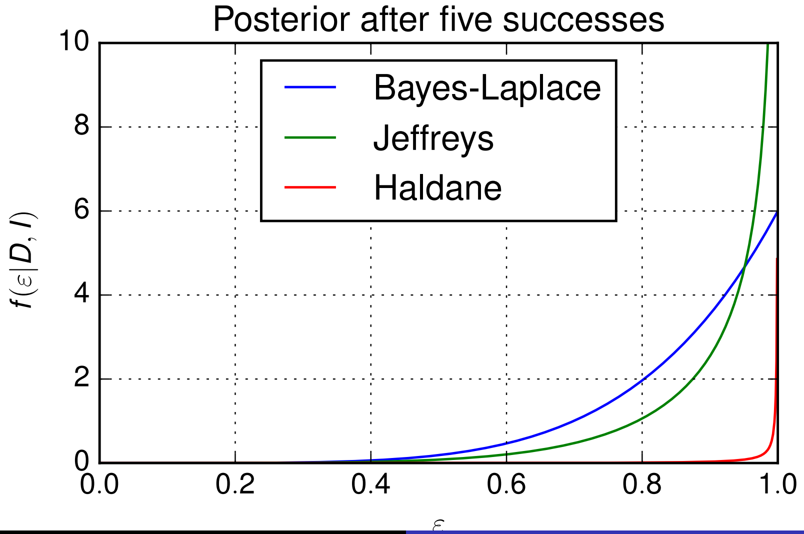
Comparison of prior pdfs



Comparison of posterior pdfs



Comparison of posterior pdfs



Outline

1. General Problem
   - Background
   - Bayesian Method
   - Examples
2. Example: Binomial Trials
   - Basics
   - Choice of Prior
   - Posterior distributions
3. Example: Sigmoid Fitting
   - Basics
   - Ignorance Priors
   - Jeffreys Prior

Overview of sigmoid efficiency estimate

- Properties $\mathbf{x}_i$ become signal strength $x_i > 0$ for each trial
- Sigmoid $\varepsilon(x_i;\alpha,\beta) = \frac{1}{1+e^{-\alpha(\ln x_i - \beta)}}$ gives likelihood

$$P(\{D_i\}|\alpha,\beta,I) \propto \prod_i \frac{e^{D_i\alpha(\ln x_i - \beta)}}{1+e^{\alpha(\ln x_i - \beta)}}$$

- Posterior for sigmoid parameters

$$f(\alpha,\beta|\{D_i\},I) \propto f(\alpha,\beta|I)\,P(\{D_i\}|\alpha,\beta,I)$$

- Cume dist fcn of efficiency ε_0 for trial at signal strength x_0:

$$P(\varepsilon_0 \le \varepsilon_0'|\{D_i\},I) = \int_{-\infty}^{\infty} d\alpha \int_{-\infty}^{\infty} d\beta \Big|_{\varepsilon(x_0;\alpha,\beta)\le\varepsilon_0'} f(\alpha,\beta|\{D_i\},I)$$

- Everything is doable; “just” need a prior $f(\alpha,\beta|I)$

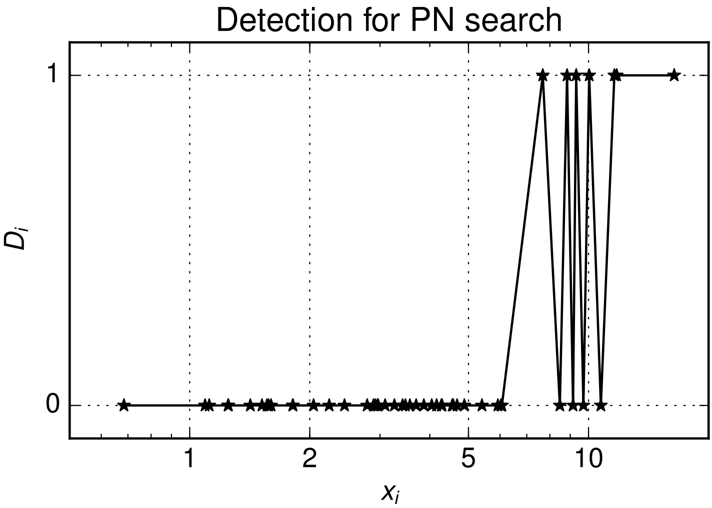
Choice of prior for sigmoid fitting

- Can try to choose Haldane-like “ignorance” prior
- One method is transformation groups (Jaynes 2003, Ch.12)
- Prior ought to be invariant under $x \to bx$ & $x \to x^a$
- β is a location parameter for $\ln x$; α is a scale parameter
- Depending on whether you try to invoke $x \to bx^a$ or $x \to (bx)^a$ you seem to get either $f(\alpha, \beta | I_{H1}) \propto \frac{1}{\alpha}$ or $f(\alpha, \beta | I_{H2}) = \text{constant}$
- For concreteness, try I_{H1}, so prior is uniform density in $\ln \alpha$ & β

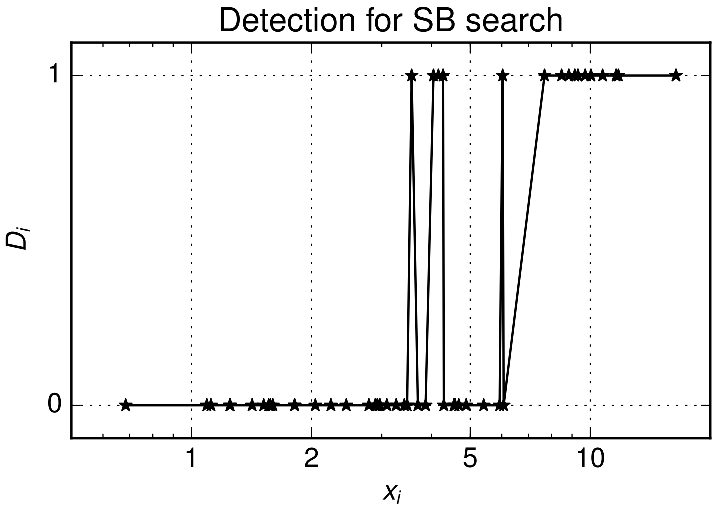
Sigmoid posterior examples w/ignorance prior

- Examples from Sco X-1 MDC:
  50 signals w/$x_i = 10^{25} h_0$ drawn from (sort of) log-normal dist
- Posterior is proportional to likelihood with this prior
- Look at specific cases ...

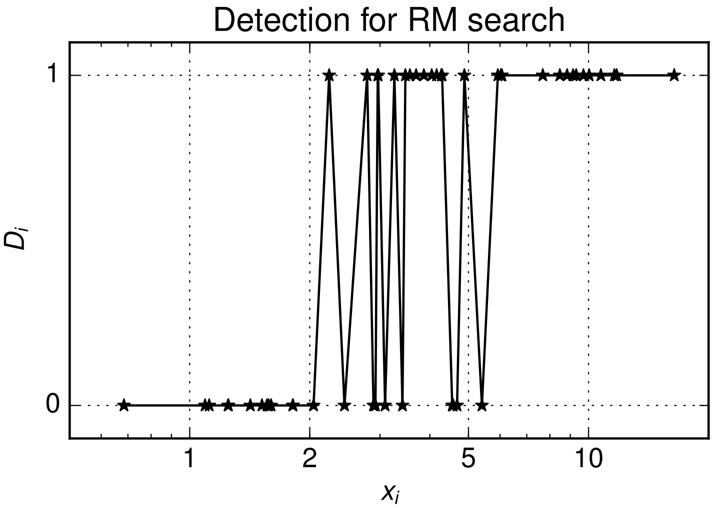
Example of Bernoulli data



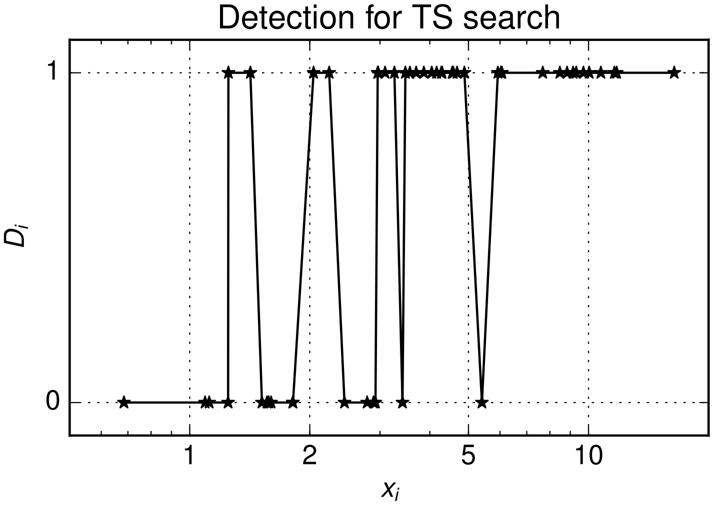
Example of Bernoulli data



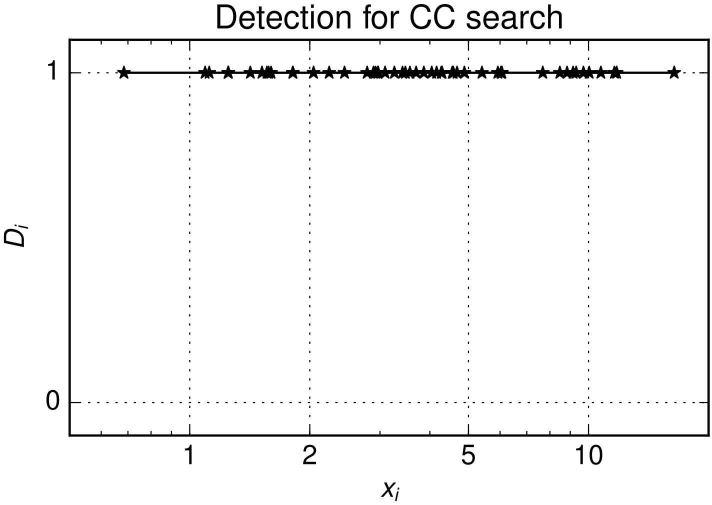
Example of Bernoulli data



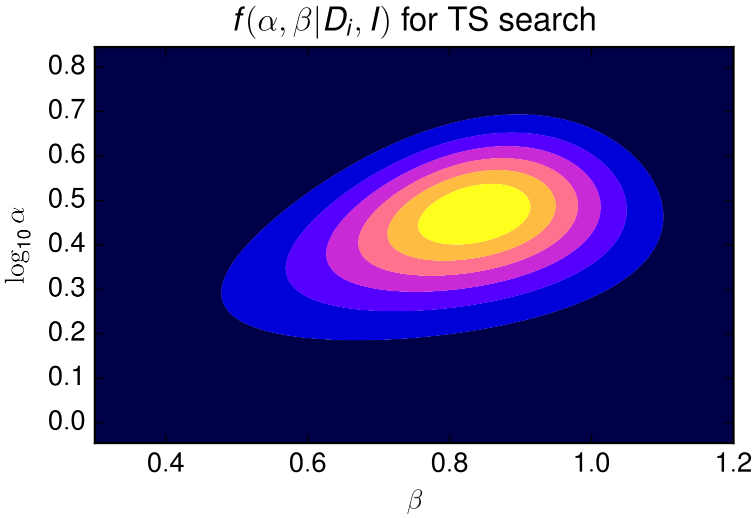
Example of Bernoulli data



Example of Bernoulli data



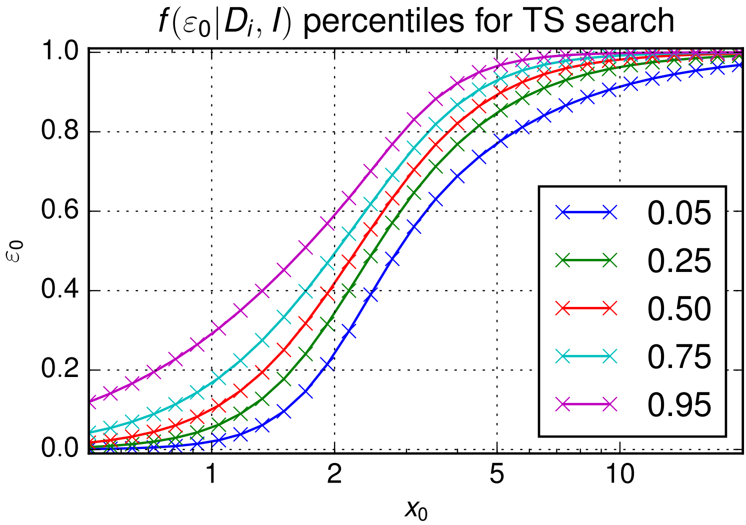
Posterior for sigmoid parameters



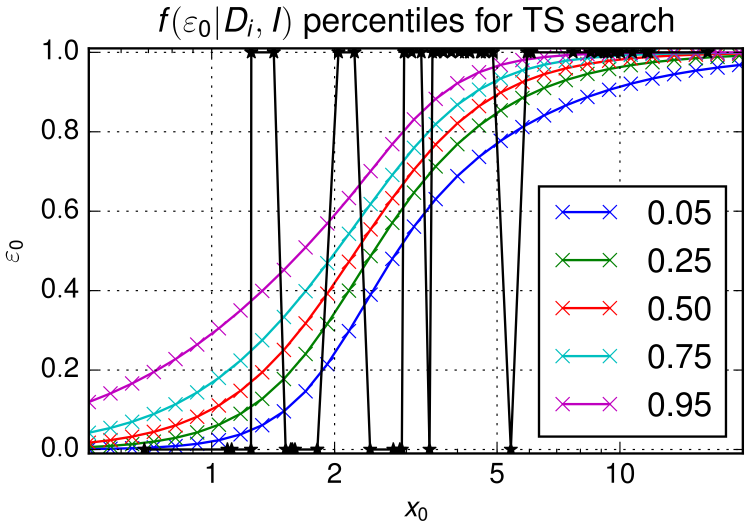
Algorithm for finding percentiles of efficiency posterior

- Lay out a grid in $\ln\alpha$ & β (could try to find mean & variance of posterior, but in practice easier to use trial & error)
- Calculate $f(\ln\alpha,\beta|\{D_i\},I)$ at each point in grid
- For each x_0 of interest, calculate $\varepsilon(x_0;\alpha,\beta)$ & sort $(\ln\alpha,\beta)$ pairs by this
- Find cumulative sum of $f(\ln\alpha,\beta|\{D_i\},I)\,\Delta\ln\alpha\,\Delta\beta$ This is cdf for ε_0 at each $\varepsilon(x_0;\alpha,\beta)$ value
- Note where cdf reaches 0.05, 0.25, 0.50, 0.75, 0.95

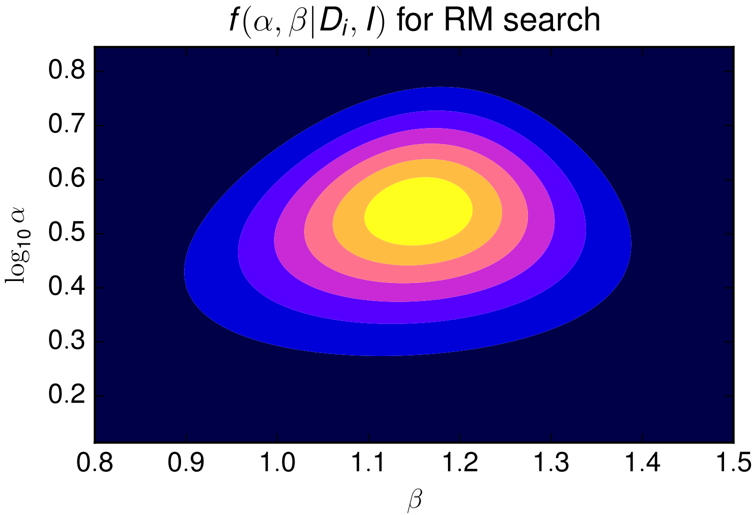
Percentiles for efficiency posterior



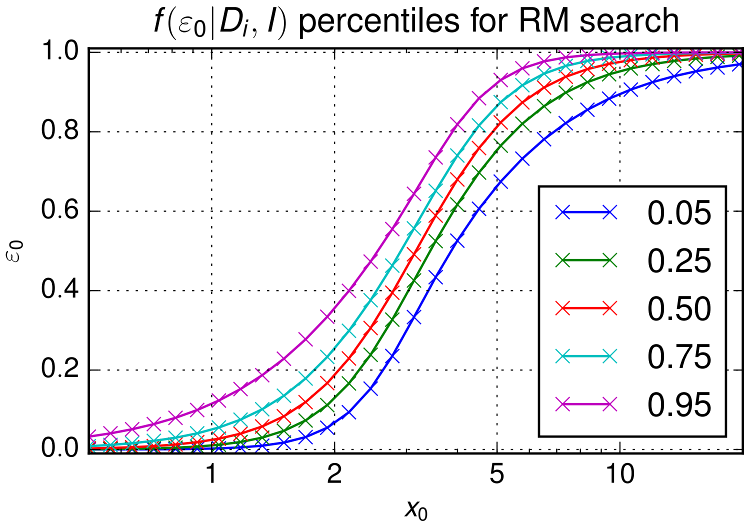
Percentiles for efficiency posterior



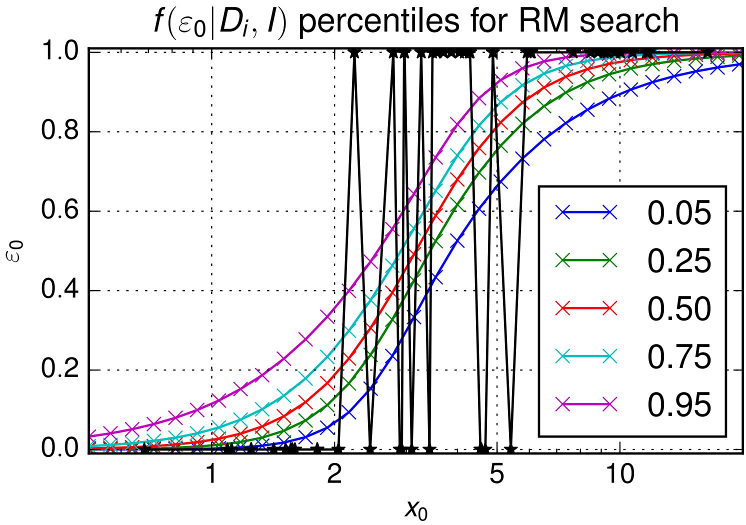
Posterior for sigmoid parameters



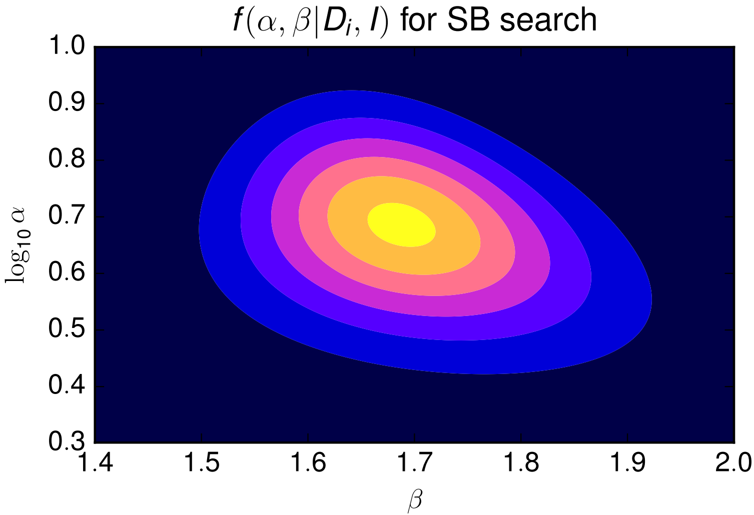
Percentiles for efficiency posterior



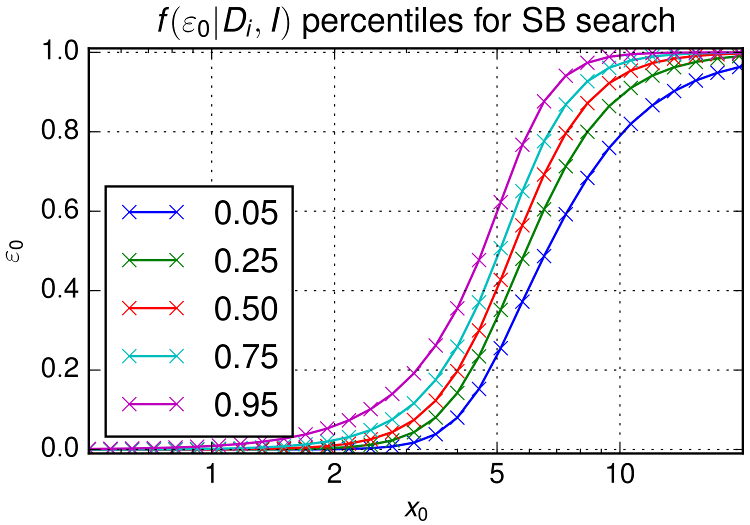
Percentiles for efficiency posterior



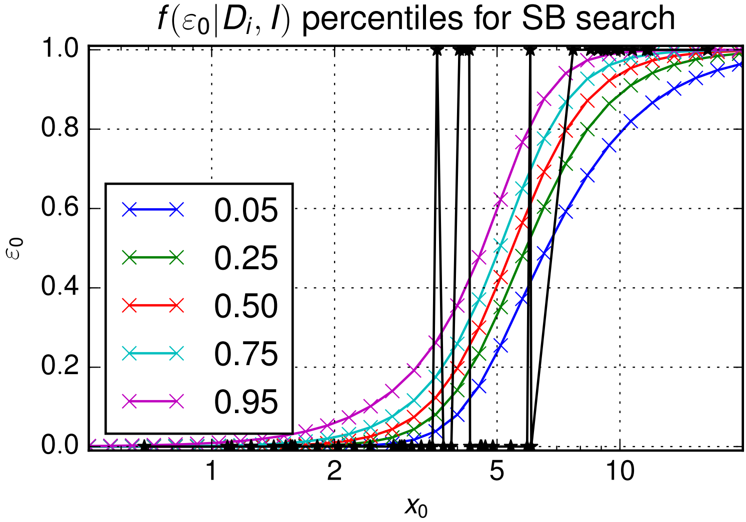
Posterior for sigmoid parameters



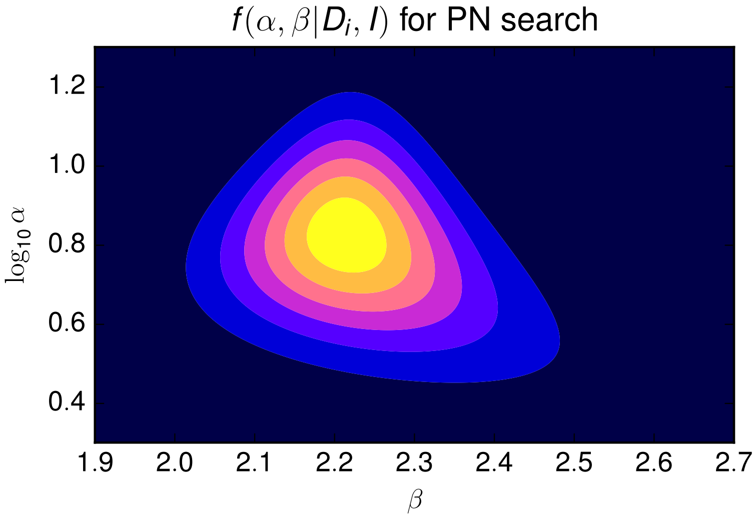
Percentiles for efficiency posterior



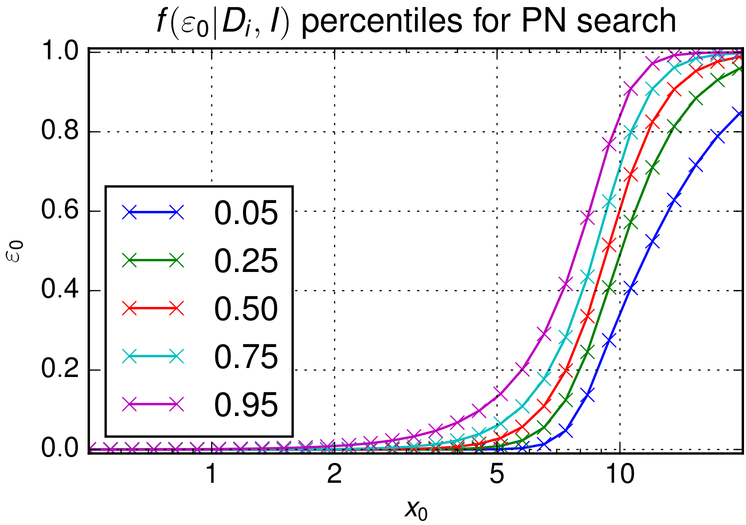
Percentiles for efficiency posterior



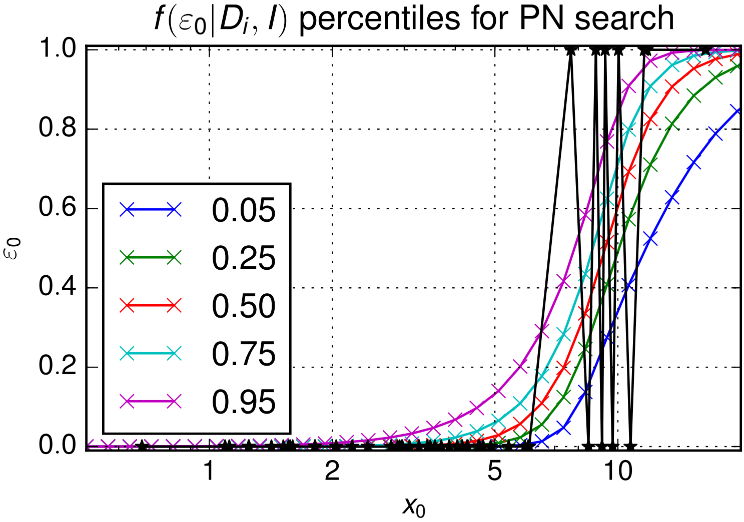
Posterior for sigmoid parameters



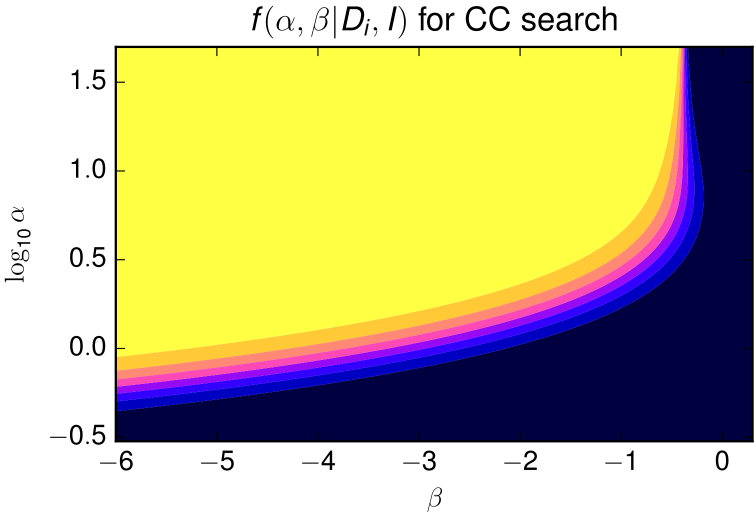
Percentiles for efficiency posterior



Percentiles for efficiency posterior



Posterior for sigmoid parameters



Jeffreys prior for sigmoid fit

- With ignorance prior, perfect results lead to the estimate $\varepsilon = 1$ at all signal strengths
- We don't literally believe this; use some regularizing prior
- One choice is Jeffreys prior $f(\boldsymbol{\theta}|I_{\mathrm{J}}) \propto \sqrt{\mathcal{I}(\boldsymbol{\theta})}$
- Easier to calculate if we change variables to α & $\gamma = \alpha\beta$, so likelihood is

$$P(\{D_i\}|\alpha,\gamma,I) \propto \prod_i \frac{e^{D_i(\alpha \ln x_i - \gamma)}}{1 + e^{\alpha \ln x_i - \gamma}} = e^{\ell(\alpha,\gamma)}$$

- Fisher information is determinant of 2×2 matrix:

$$\mathcal{I}(\alpha,\gamma) = \frac{\partial^2 \ell}{\partial \alpha^2} \frac{\partial^2 \ell}{\partial \gamma^2} - \left(\frac{\partial^2 \ell}{\partial \alpha \partial \gamma}\right)$$

Form of Jeffreys prior for sigmoid

- $f(\alpha, \gamma | I_J) \propto \sqrt{\mathcal{I}(\alpha, \gamma)}$ with Fisher information

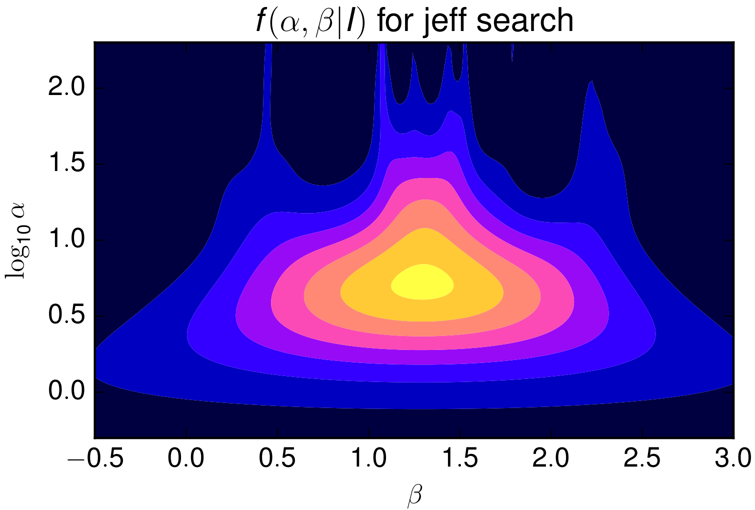
$$\mathcal{I}(\alpha, \gamma) = \frac{1}{2} \sum_i \sum_j W_i(\alpha, \gamma) W_j(\alpha, \gamma) (\ln x_i - \ln x_j)^2$$

  where

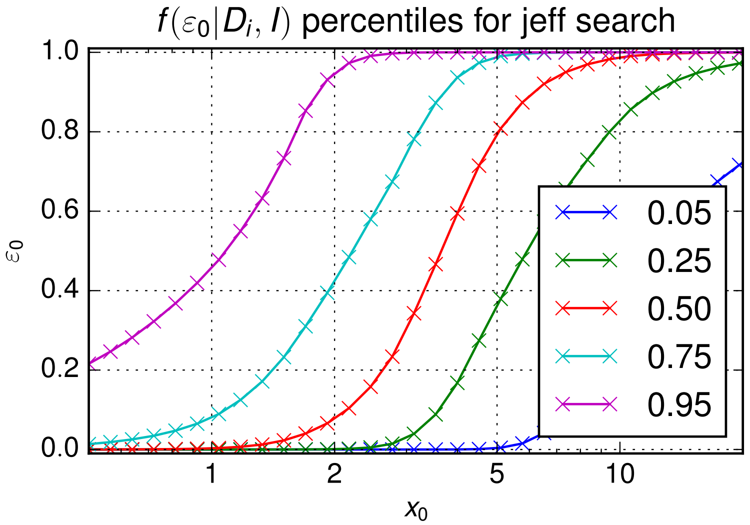
$$W_i(\alpha, \gamma) = \varepsilon(x_i; \alpha, \beta) \left(1 - \varepsilon(x_i; \alpha, \beta)\right)$$

- Depends on strengths of trials. (A little unsatisfying, but we chose those strengths because we think that's where the efficiency turns over.)

Jeffreys prior for sigmoid parameters

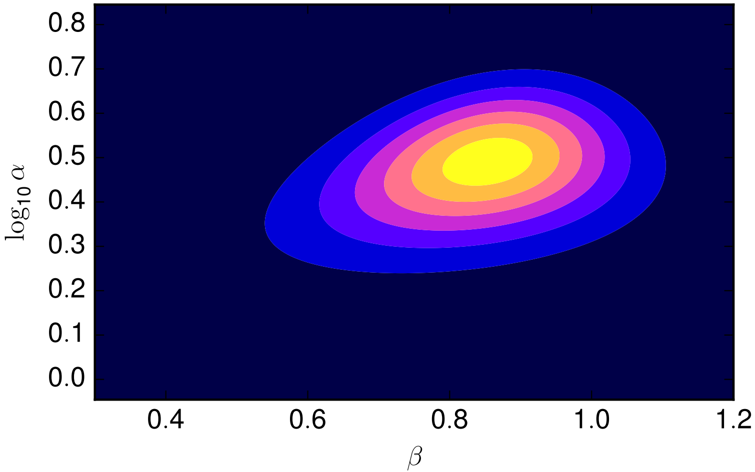


Percentiles for efficiency prior

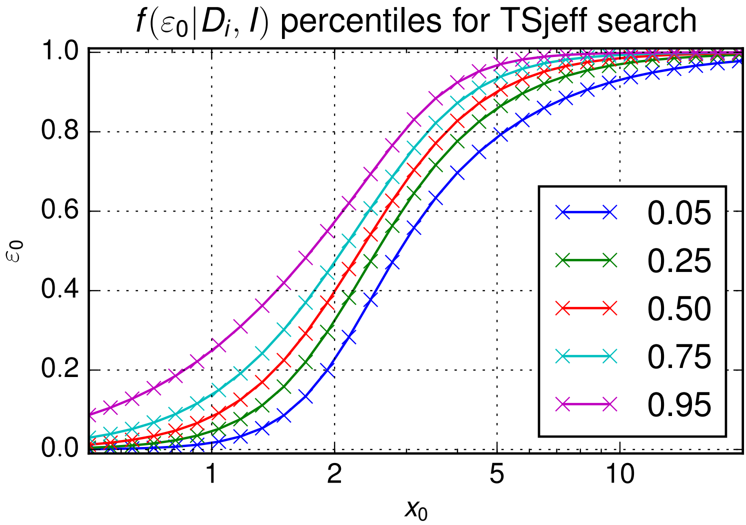


Posterior for sigmoid parameters

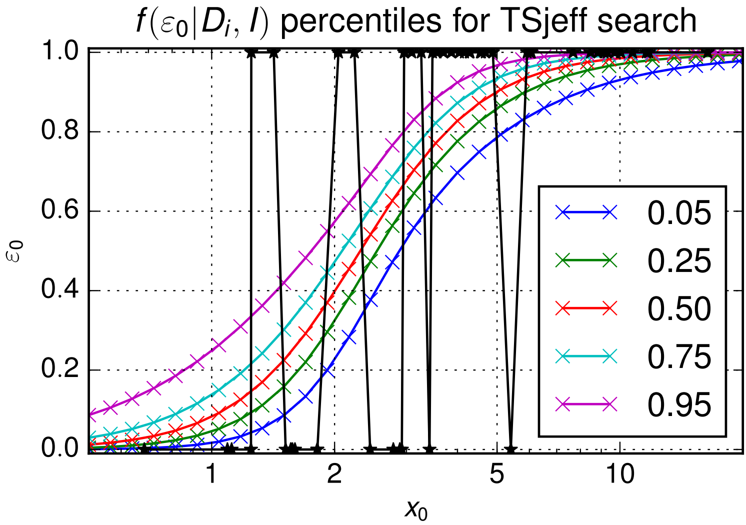
$f(\alpha, \beta | D_i, I)$ for TSjeff search



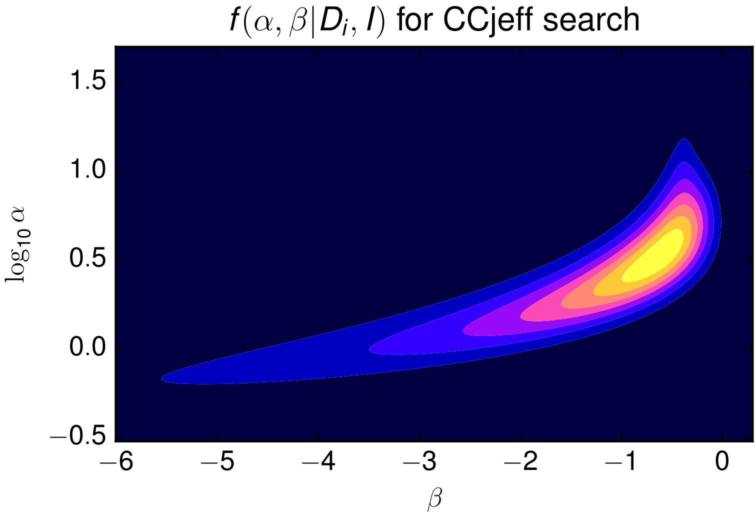
Percentiles for efficiency posterior



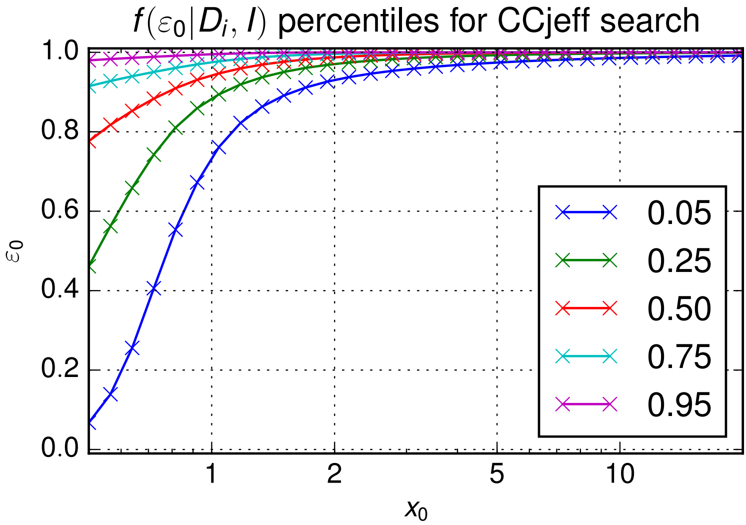
Percentiles for efficiency posterior



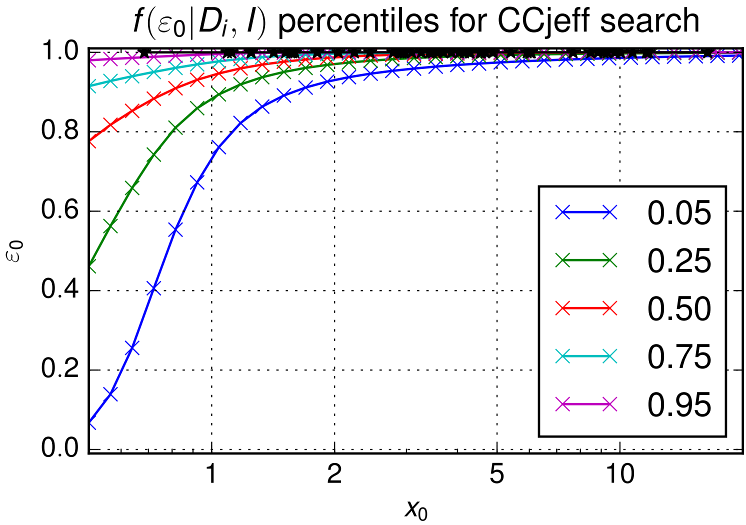
Posterior for sigmoid parameters



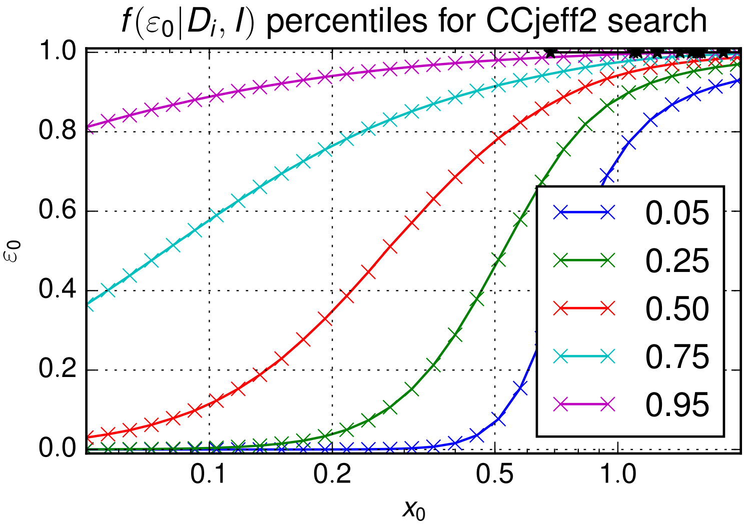
Percentiles for efficiency posterior



Percentiles for efficiency posterior



Percentiles for efficiency posterior



Summary

- Bayesian methods allow construction of posterior probabilities for efficiencies given a set of Bernoulli trials
- No binning necessary, but need a parametrized model and prior probabilities for the parameters
- “Ignorance priors” often lead one to predict 100% efficiency given perfect results
- Other priors, e.g. Jeffreys, give more conservative results even if all trials successful

References

- Bradley and Terry 1952 *Biometrika* **39**, 324
- Haldane 1932 *Math. Proc. Cambridge Phil. Soc.* **28**, 1469
- Jaynes 2003 Probability Theory: The Logic of Science
- Jeffreys 1946 *Proc. Royal Soc. London A* **186**, 1007
- Laplace 1814 Essai philosophique sur les probabilités
- Messenger et al 2015 *Physical Review D* **92**, 023006
- Zermelo 1929 *Mathematische Zeitschrift* **29**, 436