Using Bayes Factor to Detect Spectral Line

- Problem Introduction
- Ways to compute Bayes Factor(BF)
- Simulation Study
- Discussion

Problem Introduction

- The most simple case. Let *i* be bin indicator.

$$Y_i \sim \mathrm{Pois}(\Lambda_i)$$

- Two models:

$$H_0: \Lambda_i = \alpha E_i^{-\beta}$$

$$H_a: \Lambda_i = \alpha E_i^{-\beta} + \lambda I_{i=\mu}$$

What is BF

- Given a model selection problem, suppose we observe data D, denote two different models as M1 and M2, the Bayes factor K is given by

$$\frac{\Pr(D|M_1)}{\Pr(D|M_2)} = \frac{\int \Pr(D|\theta_1, M_1)\,dF_{\theta_1}}{\int \Pr(D|\theta_2, M_2)\,dF_{\theta_2}}$$

- BF is known to be dependent on prior choice

Why Using Bayes Factor

- Classical Likelihood ratio test doesn’t work here.
- There are ways to sample the posterior distribution of all parameters(PCGS)
- Interested to see how BF performs compared to other tools like Posterior Predictive P-value and also how much does it depend on the prior

Ways to Compute BF

- It's a HARD numerical integration problem.

$$\int \Pr(D|\theta_1, M_1)\, dF_{\theta_1}$$

- Methods to compute it include:
  - Brute Force
  - Gaussian Approximation
  - Monte Carlo method

Gaussian Approximation

- Officially called **Laplace approximation:**

$$I = \int \Pr(D|\theta_1, M_1) dF_{\theta_1} = \int \Pr(D|\theta_1, M_1) \Pr(\theta_1) d\theta_1 \propto \int \Pr(\theta_1|D, M_1) d\theta_1$$

- Assume we have large sample size, posterior dist'n "would" be approximately Gaussian around its mode.

$$\hat{I} = (2\pi)^{d/2} |FOI^{-1}|^{1/2} \cdot Pr\left(y \middle| \widehat{\theta_1}, M_1\right) \cdot \Pr(\widehat{\theta_1})$$

- It works when the Gaussian Approximation assumption is valid.

Monte Carlo Method_1

- Recall we need calculate:

$$I = \int \Pr(y|\theta, M) \cdot \Pr(\theta)\, d\theta$$

- If we have a sample from the prior dist'n:

$$\hat{I} = \frac{1}{m} \cdot \sum_{i=1}^{m} p(y|\theta^i, M)$$

- It is simple/ easy to sample the prior

- If likelihood is peaked around the mode, the sum would be dominated by a few samples.

Monte Carlo Method_1

- If we have a sample from the posterior dist’n
- With little trick:

$$\hat{I} = \left(\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\mathrm{p}(\mathrm{y}|\theta^{\mathrm{i}},\mathrm{M})}\right)^{-1}$$

- It’s still simple/ we know how to sample posterior
- Likelihood on the denominator = disaster…

Ex: Simulation Study_1

- Assume powerlaw model for continuum

$$\lambda_i = \alpha \cdot E_i^{-\beta} \text{ vs } \lambda_i = \alpha \cdot E_i^{-\beta} + \omega \cdot I_{\{i=\mu\}}$$

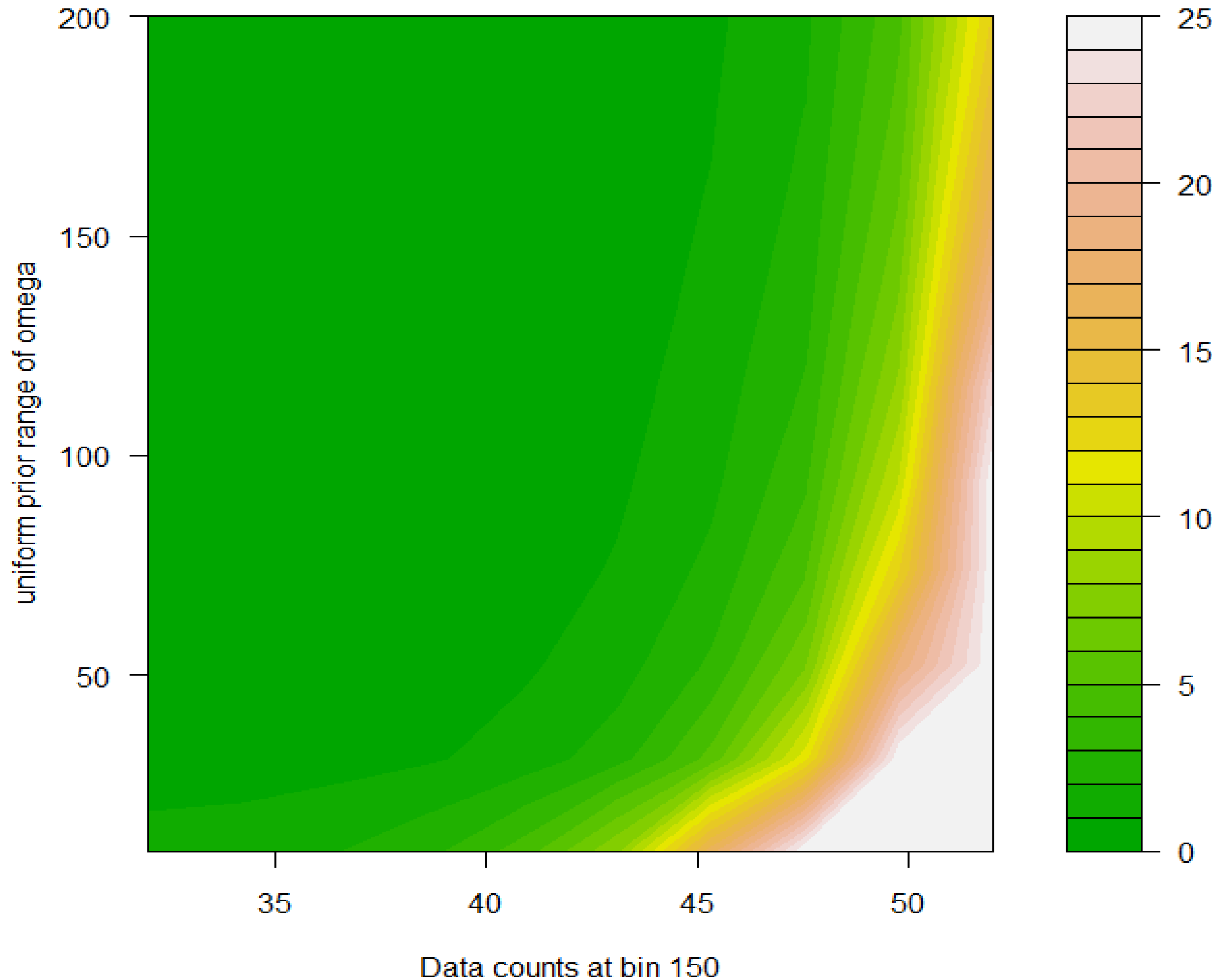
- Line_location is assumed to be known in this study. Assign uniform prior to all other parameters

- True line_location is @bin[150], where continuum intensity is equal to 32.

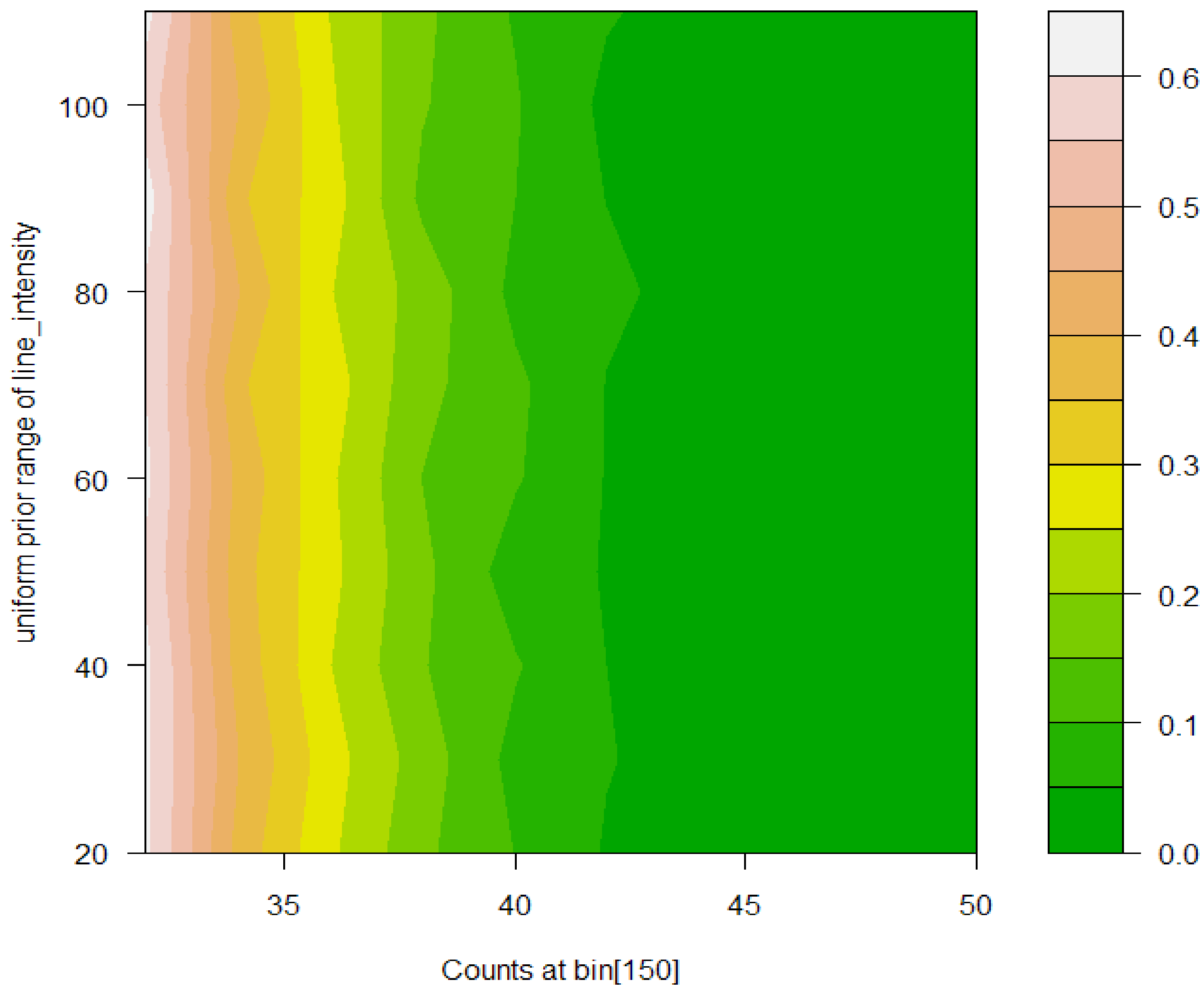
- Posterior distribution does look Gaussian

- Only method that works is Laplace Approximation.

Heatplot of BF for Simulation_1



Heatmap of PPP for Simulation_1



Ex: Simulation study_2

- Still assume powerlaw for continuum

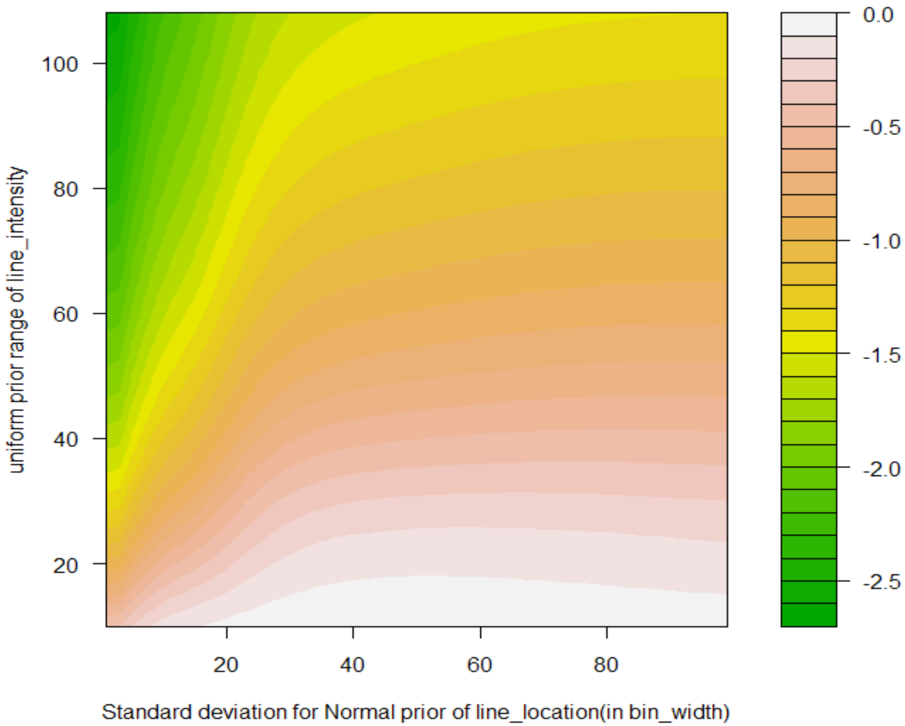
$$\lambda_i = \alpha \cdot E_i^{-\beta} \text{ vs } \lambda_i = \alpha \cdot E_i^{-\beta} + \omega \cdot I_{\{i=\mu\}}$$

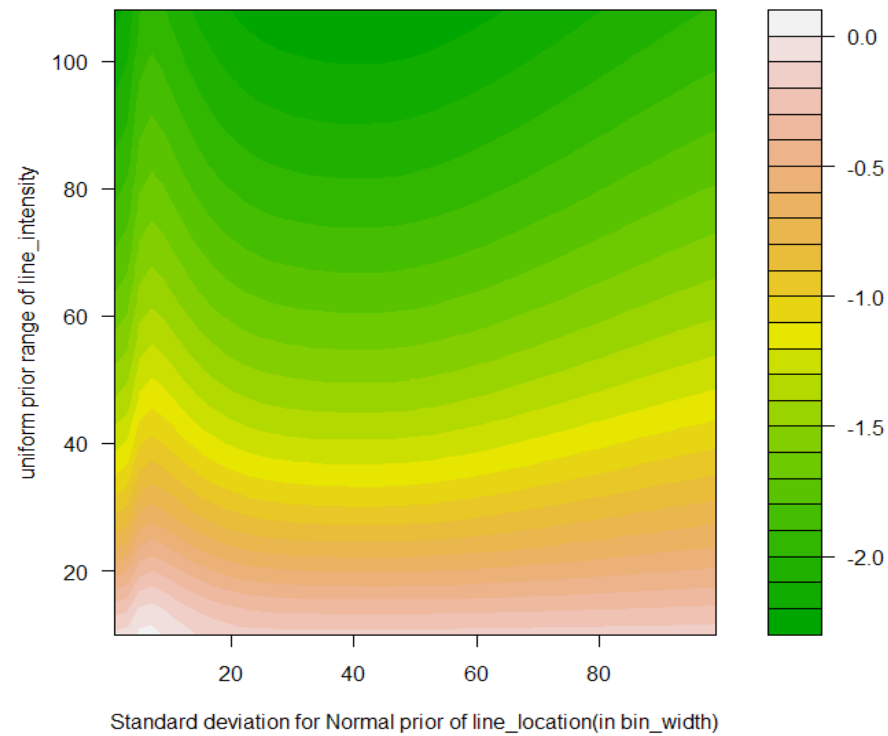
- Both line_intensity and line_location are unknown; Assume uniform prior for line_intensity; Gaussian derived discrete prior for line_location.

- True line_location is @bin[150], where continuum intensity is equal to 32.

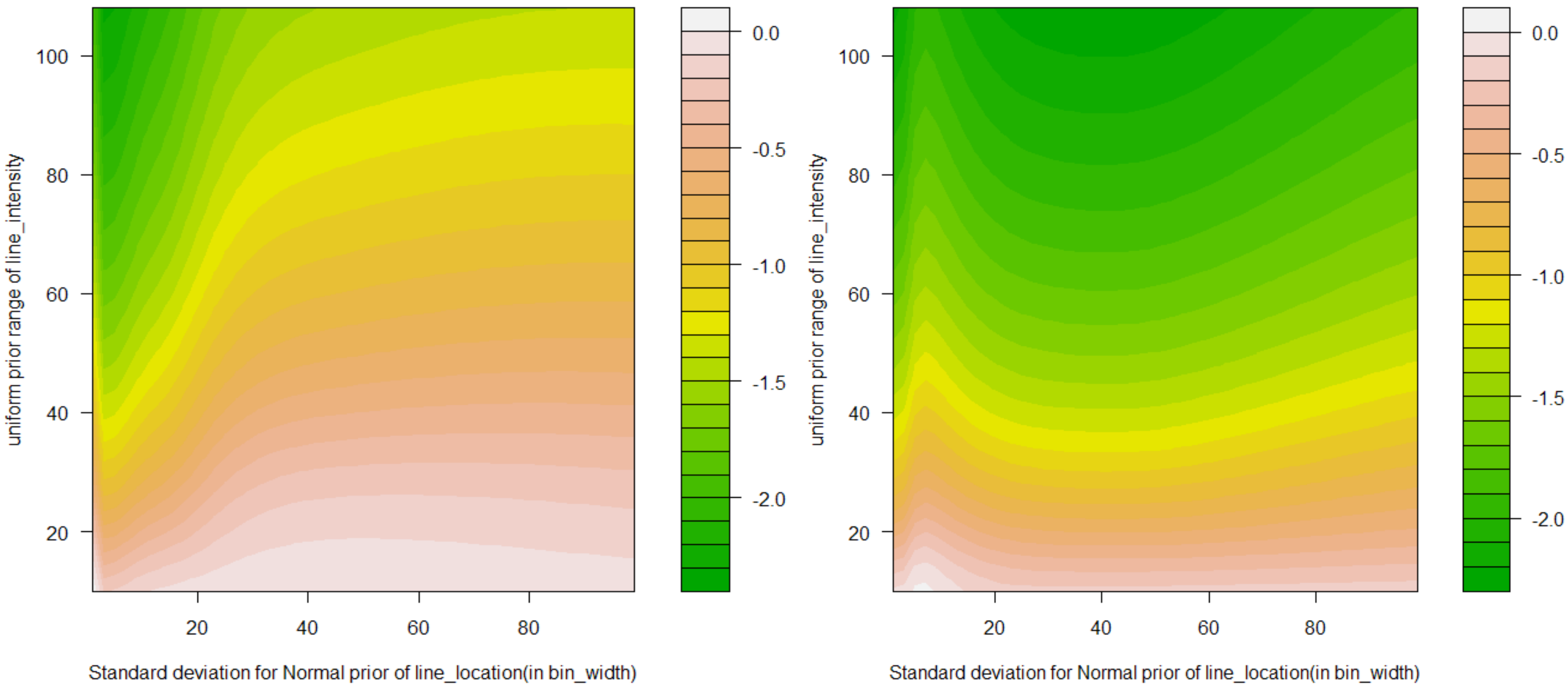
- Posterior distribution no longer looks Gaussian. Only method works is to use brute force.

- Data @Bin[150] = 32
- PPP = 0.36

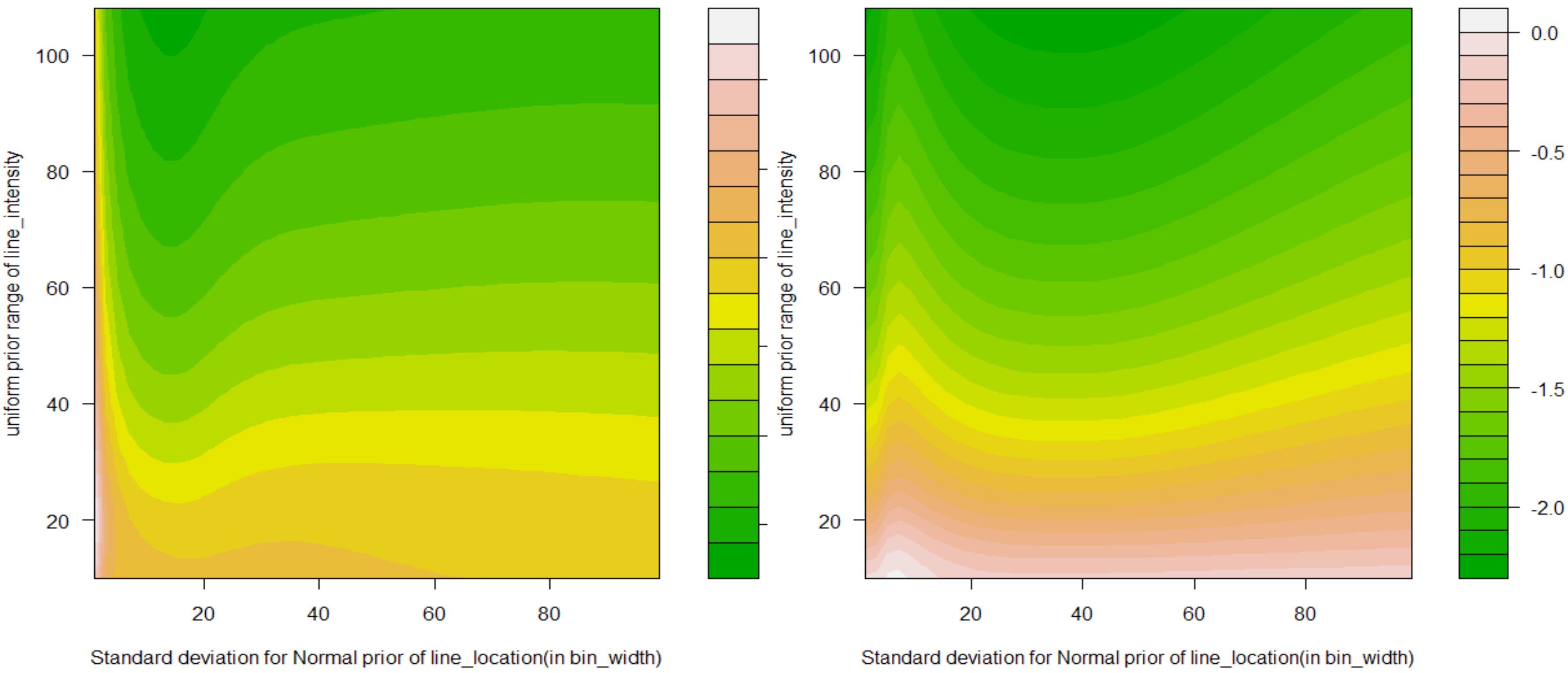




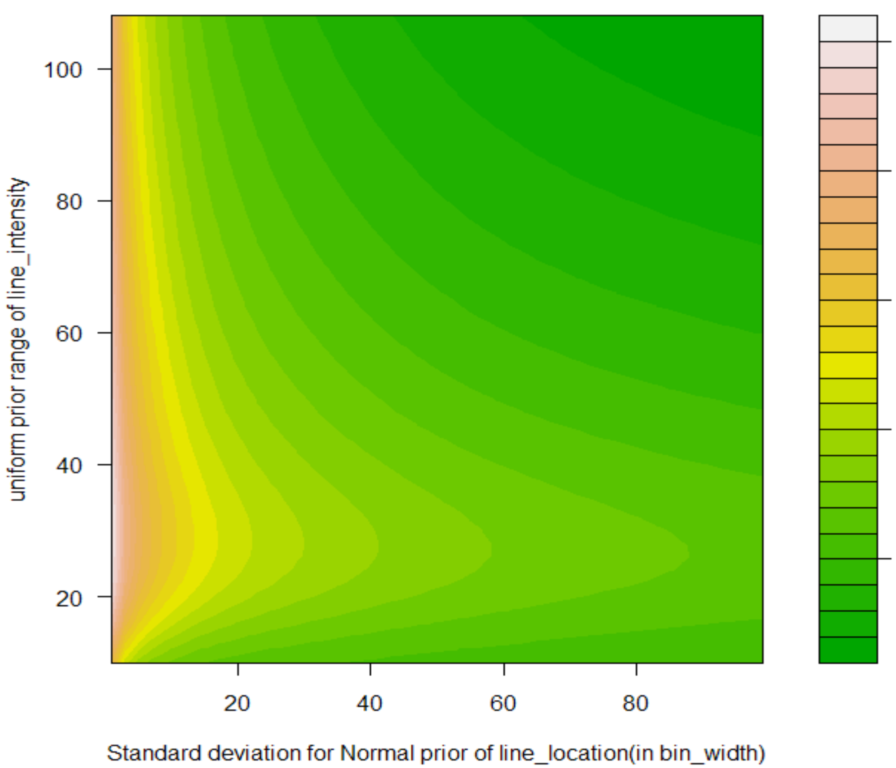
- Data @Bin[150] = 32+7
- PPP = 0.36

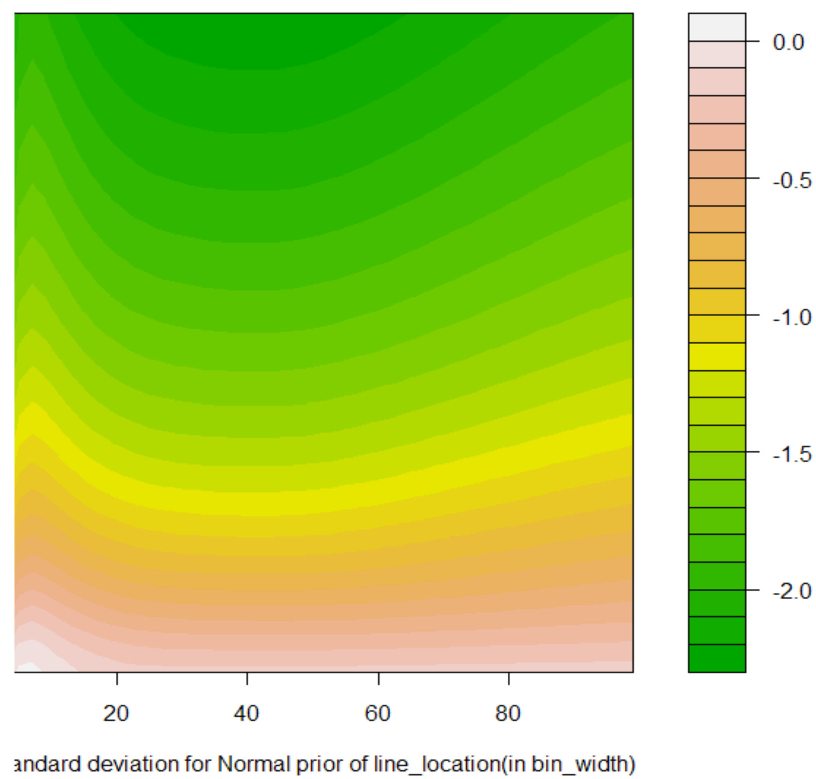


- Data @Bin[150] = 32 + 14
- PPP = 0.36



- Data @Bin[150] = 32 + 21
- PPP = 0.36





Cont’d: Simulation_2

- When prior mode for the line_location is at the “true” line_location, the power of BF is strongly dependent on the prior standand deviation of its prior.

- When prior mode for the line_location is away from the “true” line_location, the power of BF decreases a lot.

- What is the usual choice of the priors here?

Simulation_3 and Discussion

- Same problem settings except that all parameters are unknown.
- Posterior dist’n doesn’t look Gaussian.
- No method works.
- Two more methods are tried:
  - Bridge Sampling
  - Savage density ratio

Bridge Sampling

- Designed to calculate the ratio of two normalizing constant.
- BF is the ratio of normalizing constant for two posterior dist'n.
$$I = \int \Pr(D|\theta_1, M_1) dF_{\theta_1} = \int \Pr(D|\theta_1, M_1) \Pr(\theta_1) d\theta_1 \propto \int \Pr(\theta_1|D, M_1) d\theta_1$$
- Let $p_i = \frac{q_i}{c_i}$, then $\frac{c_1}{c_2} = \frac{E_2[q_1(\omega)\alpha(\omega)]}{E_1[q_2(\omega)\alpha(\omega)]}$
- However, bridge sampling requires both models have common support of parameter space.
- Can we re-write null model into: $\lambda_i = \alpha E_i^{-\beta} + \omega I_{i=\mu}$
But with constraints like: $\omega \sim \mathrm{unif}(0,1)$

Savage Density Ratio

- For nested models:

$$B_{10} = \frac{Pr(\psi = \psi_0 | M_1)}{Pr(\psi = \psi_0 | Y, M_1)}$$

- Work for us if line_location is known. Equivalent to test line_intensity is equal to zero.

- However, bridge sampling and savage density ratio also doesn’t perform as good as Laplace Approximation.

- Next step?