Final project

Monte Carlo applications

problem 1 (individual): 10 %   problem 2 (team): 20%

due: December 13, 2019 2:00 pm

Team A: Russel, Seibert, Le

Team B: Thouvenin, Liang, Spisak

Team C: Correa, Haughton, Alavi

Team D: Phan, Roberts, Pham

Team E: Kruger, Levin

Problem 1 prepared and submitted individually

Consider the probability distribution p(**x**) which is the mixture of two multivariate Gaussian distributions in two variables with **x**=(x_1,x_2):

$$p(\boldsymbol{x}) = \frac{1}{2}\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_1, \Sigma_1) + \frac{1}{2}\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_2, \Sigma_2)$$

where $\boldsymbol{\mu}_1 = [0,0]^T, \boldsymbol{\mu}_2 = [5,5]^T, \Sigma_1 = \Sigma_2 = \text{diag}\{0.25, 2\}$

1(A) Plot the p(**x**) probability density function in the (x_1,x_2) variables

1(B) Calculate the mean of the vector **x**=(x_1,x_2) using Markov Chain Monte Carlo with Metropolis importance sampling. Compare the histogram with the **1(A)** plot.

1(C) Calculate the Monte Carlo error of the mean of the vector **x**=(x_1,x_2) using Markov Chain Monte Carlo with Metropolis importance sampling.

1(D) After consultation with the TA, estimate the autocorrelation time (separation of independent MC configurations) for correct error estimates.

1(E) Compare the MC results with the analytic expectations, including error estimates.

Problem 1 (phys 239 only)

prepared and submitted individually

1(F) Calculate the mean of $\sin^2(x_1)\cdot\sin^2(x_2)$ with your best error estimate for the distribution p(**x**), defined above as,

$$p(\boldsymbol{x}) = \frac{1}{2}\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \frac{1}{2}\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$$

where $\boldsymbol{\mu}_1 = [0,0]^T, \boldsymbol{\mu}_2 = [5,5]^T, \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathrm{diag}\{0.25, 2\}$

1(G) Compare the quality of Metropolis MC sampling with Gibbs sampling.

Problem 2 submitted by the team

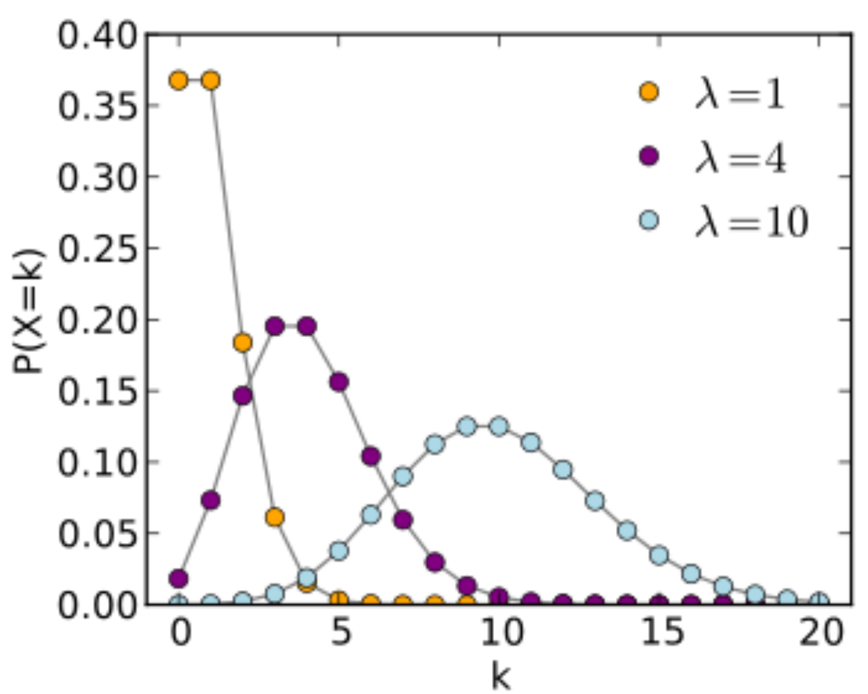
A CASE STUDY: CHANGE-POINT DETECTION

The task of change-point detection is of major importance in a number of scientific disciplines, ranging from engineering and sociology to economics and environmental studies.

The aim of the change-point identification task is to detect partitions in a sequence of observations, in order for the data in each block to be statistically "similar," in other words, to be distributed according to a common probability distribution.

Let x_n be a discrete random variable that corresponds to the count of an event, for example, the number of requests for telephone calls within an interval of time, requests for individual documents on a web server, particle emissions in radioactive materials, number of accidents in a working environment, and so on. We adopt the Poisson process to model the distribution of x_n, that is,

$$P(x;\lambda) = \frac{(\lambda\tau)^x}{x!} e^{-\lambda\tau} \quad x=0,1,2,\ldots$$

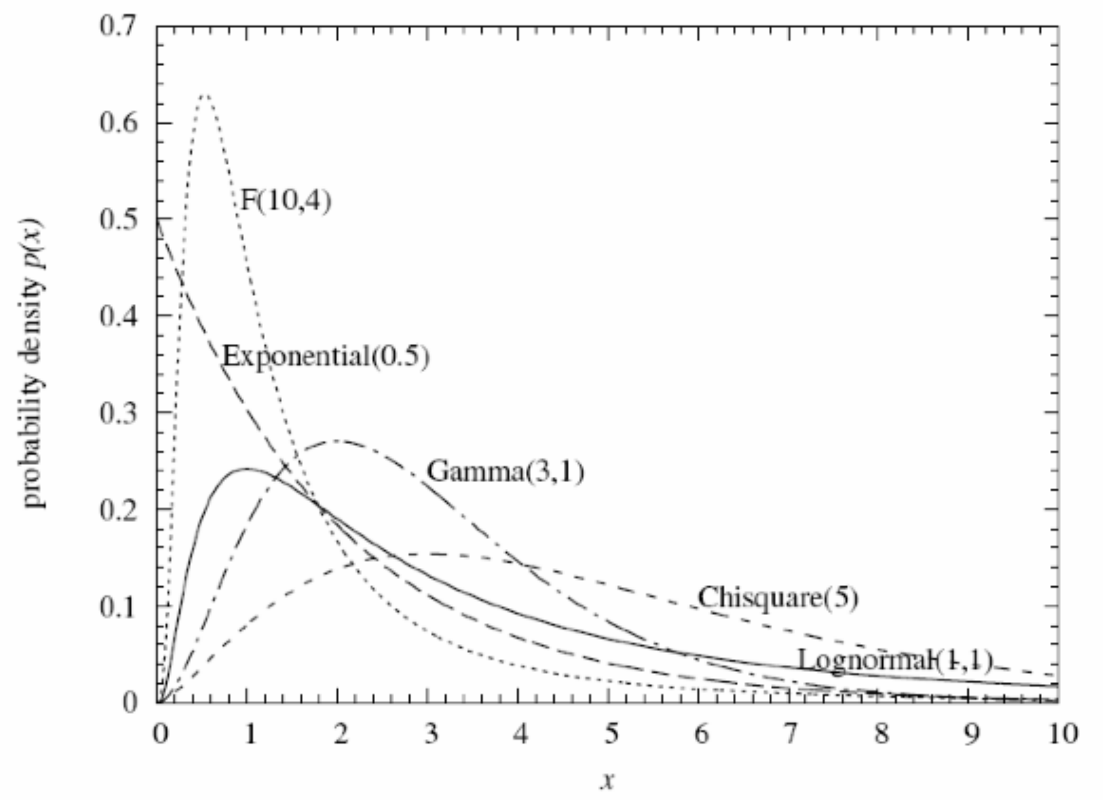


Poisson processes have been widely used to model the number of events that take place in a time interval, τ. For our example, we have chosen $\tau = 1$. The parameter λ is known as the *intensity* of the process

Gamma distribution:

$$x \sim \text{Gamma}(\alpha, \beta), \qquad \alpha > 0,\ \beta > 0$$

$$p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \qquad x > 0$$



$$\text{Mean}\{\text{Gamma}(\alpha, \beta)\} = \alpha/\beta$$

$$\text{Var}\{\text{Gamma}(\alpha, \beta)\} = \alpha/\beta^2$$

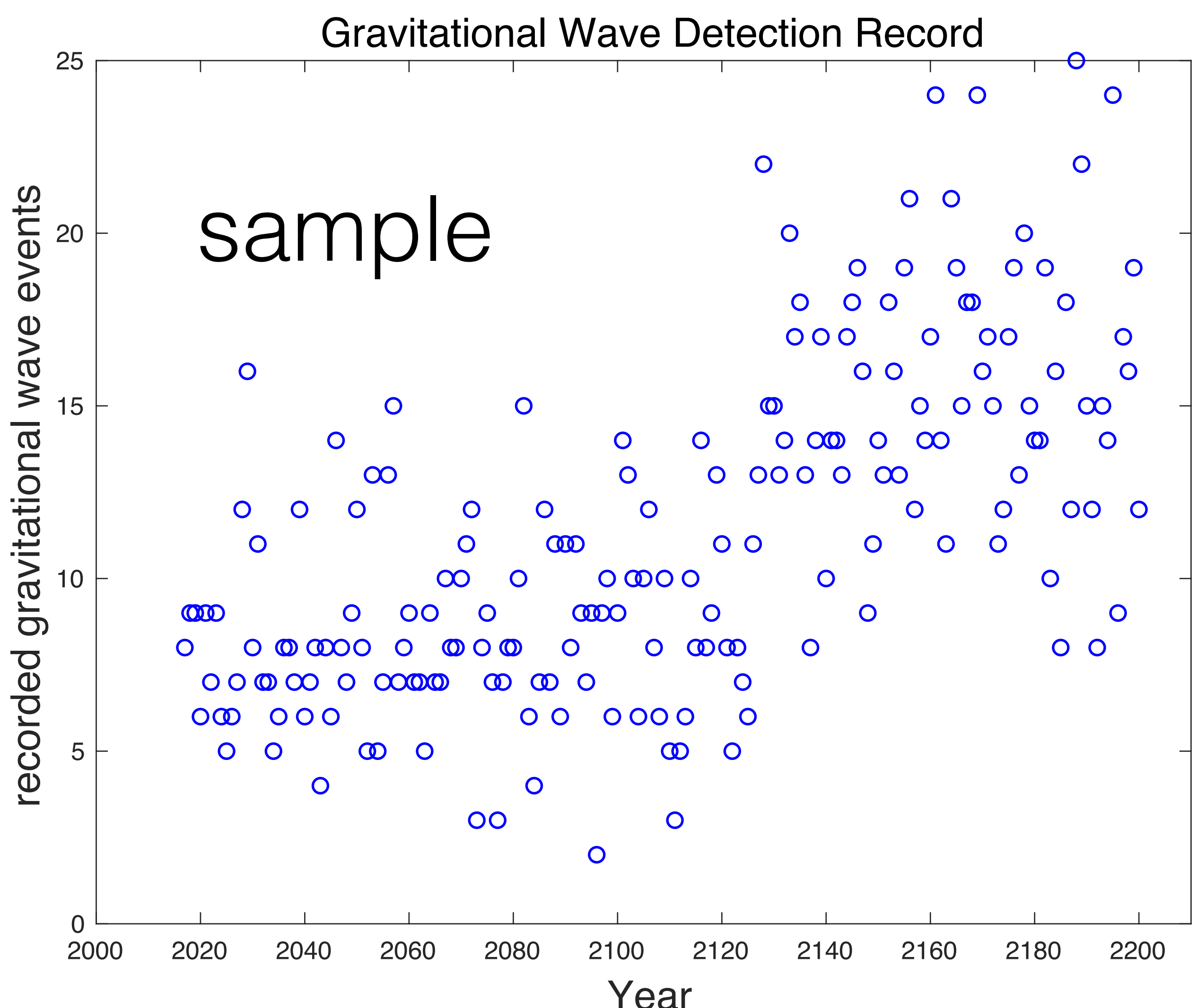
When $\alpha \geq 1$ there is a single mode at $x = (\alpha - 1)/\beta$ (peak)

Problem 2 submitted by the team

plot your figure of gravitational wave detection events between 2017-2200 from data files custom made for each team (linked to Final web page)

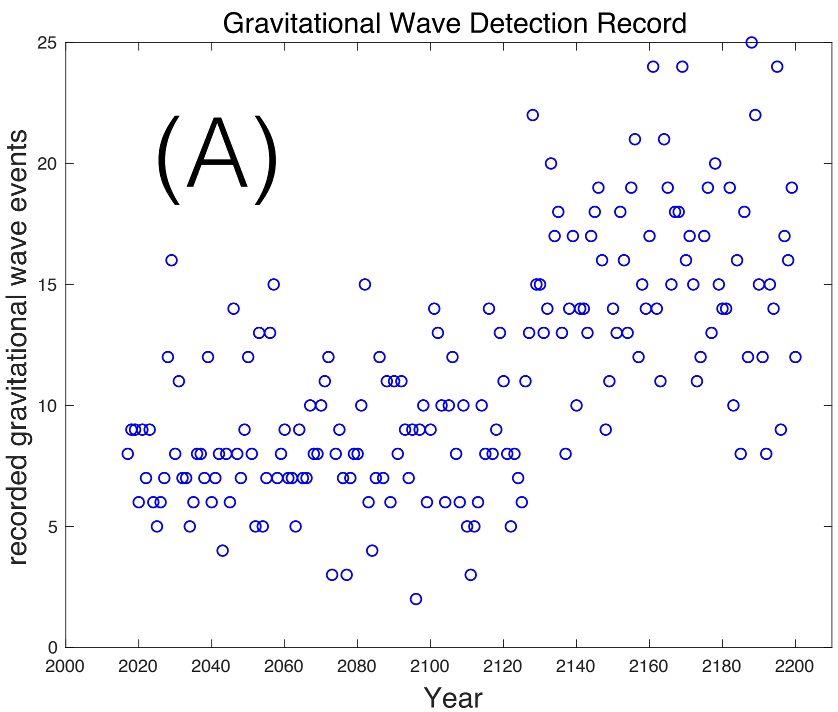
estimate with prior based MCMC the year when a space based gravitational wave detector Lisa (space based laser interferometer) came online and changed the average annual rate.

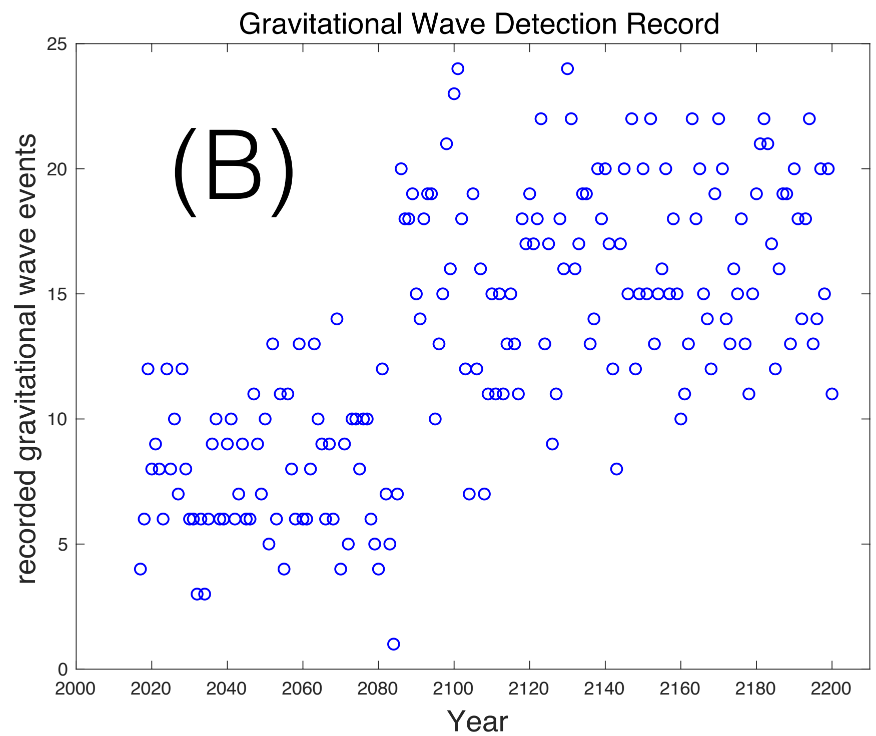
Bayesian prior is parametrized with the Gamma distribution where a=8 and b=1 are reasonable choices although results are not sensitive to the choices

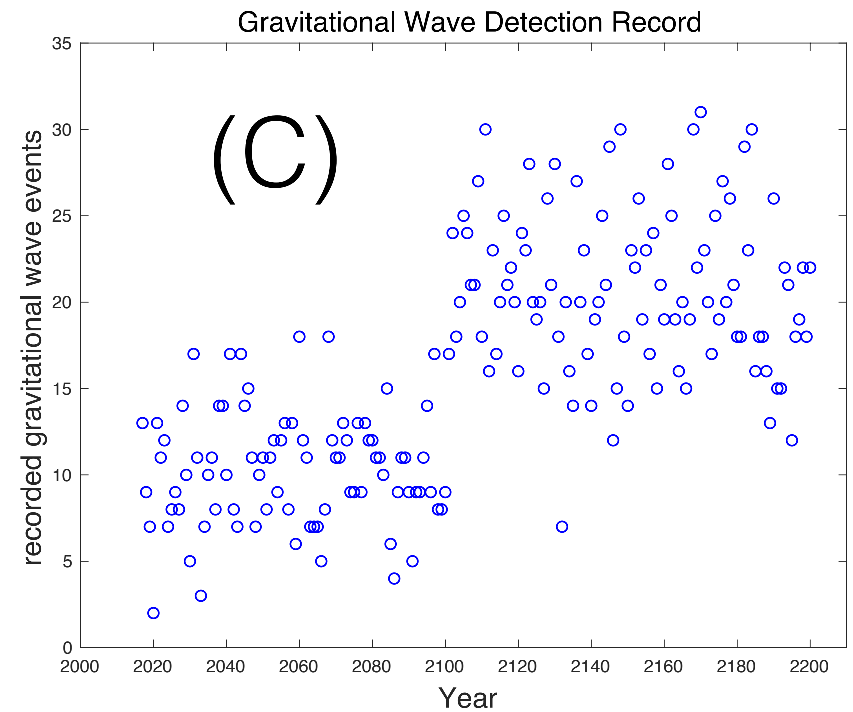


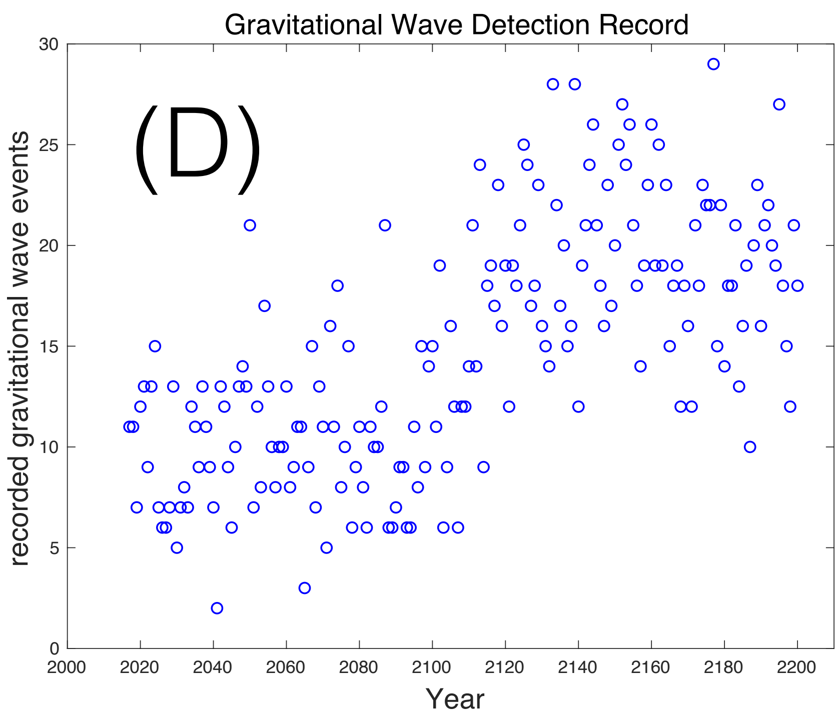
Problem 2 submitted by the team

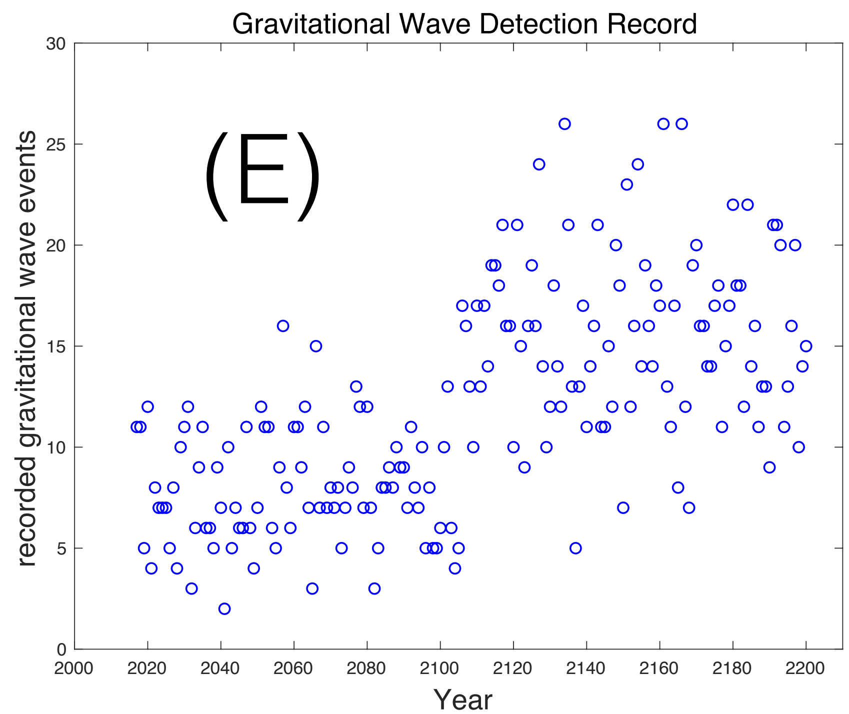
five sets for teams:











Problem 2 submitted by the team

We assume that our observations, x_n, $n = 1, 2, \ldots, N$, have been generated by two different Poisson processes, $P(x; \lambda_1)$ and $P(x; \lambda_2)$. Also, the change of the model has taken place suddenly at an unknown time instant, n_0. Our goal is to estimate the posterior,

$$P(n_0 | \lambda_1, \lambda_2, \boldsymbol{x}_{1:N}).$$

Moreover, the exact values of λ_1 and λ_2 are not known. The only available information is that the Poisson process intensities, λ_i, $i = 1, 2$, are distributed according to a (prior) gamma distribution, that is,

$$p(\lambda) = \text{Gamma}(\lambda | a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda),$$

for some known positive values a, b. We will finally assume that we have no prior information on when the time of change occurred; thus, the prior is chosen to be the uniform distribution, $P(n_0) = \frac{1}{N}$. Based on the previous assumptions, the corresponding joint distribution is given by,

$$p(n_0, \lambda_1, \lambda_2, \boldsymbol{x}_{1:N}) = p(\boldsymbol{x}_{1:N} | \lambda_1, \lambda_2, n_0) p(\lambda_1) p(\lambda_2) P(n_0)$$ ← the Bayesian

or

$$p(n_0, \lambda_1, \lambda_2, \boldsymbol{x}_{1:N}) = \prod_{n=1}^{n_0} P(x_n | \lambda_1) \prod_{n=n_0+1}^{N} P(x_n | \lambda_2) p(\lambda_1) p(\lambda_2) P(n_0).$$

Taking the logarithm in order to get rid of the products, and integrating out respective variables, the following conditionals needed in Gibbs sampling are obtained **(to prove!)**

Problem 2 submitted by the team

2(A) prove the conditional probabilities to prepare for Gibbs sampling:

Taking the logarithm in order to get rid of the products, and integrating out respective variables, the following conditionals needed in Gibbs sampling are obtained

$$p(\lambda_1|n_0,\lambda_2,\boldsymbol{x}_{1:N}) = \mathrm{Gamma}(\lambda_1|a_1,b_1),$$

with

$$a_1 = a + \sum_{n=1}^{n_0} x_n, \quad b_1 = b + n_0,$$

$$p(\lambda_2|n_0,\lambda_1,\boldsymbol{x}_{1:N}) = \mathrm{Gamma}(\lambda_2|a_2,b_2),$$

$$a_2 = a + \sum_{n=n_0+1}^{N} x_n, \quad b_2 = b + (N - n_0),$$

and

$$\ln \tilde{P}(n_0|\lambda_1,\lambda_2,\boldsymbol{x}_{1:N}) = \ln\lambda_1 \sum_{n=1}^{n_0} x_n - n_0\lambda_1 + \ln\lambda_2 \sum_{n=n_0+1}^{N} x_n$$
$$- (N - n_0)\lambda_2, \quad n_0 = 1, 2, \ldots, N.$$

the two definitions of the Γ- function on p.5 and p.8 are equivalent with the a→α and b→β substitutions.

The Γ- function on p.5 is mapped to the Poisson distribution of p.4 with the substitution in the Γ- function α→x (x now in the Poisson dist.) β→λ (λ now in the Poisson dist.) and setting x=1 in the Γ- function.

So the combination of the Poisson dist. and the Γ- function prior becomes the product of two Γ- functions which facilitates the integration over the prior

The last line for ln(P) just gives the log of the products of independent Poisson probabilities once the Poisson intensities $\lambda_{1,2}$ are determined from the Gamma distributions for a particular n_0 . λ_1 up to year n_0 and λ_2 from year n_0+1 to year N. ~ indicates the normalization factor which has to be taken into account. See next page for how to draw Gibbs sampling from discrete probabilities.

Problem 2 submitted by the team

For discrete probabilities P_i , with u ~ U(0,1) uniform random number in the (0,1) interval:

- Define $a_k = \sum_{i=1}^{k-1} P_i$, $b_k = \sum_{i=1}^{k} P_i$, $k = 1, 2, \ldots, K$, $a_1 = 0$.
- **For** $i = 1, 2, \ldots,$ **Do**
  - $u \sim \mathcal{U}(0, 1)$
  - Select

$$x_k \text{ if } u \in [a_k, b_k), \quad k = 1, 2, \ldots, K$$

- **End For**

Problem 2 submitted by the team

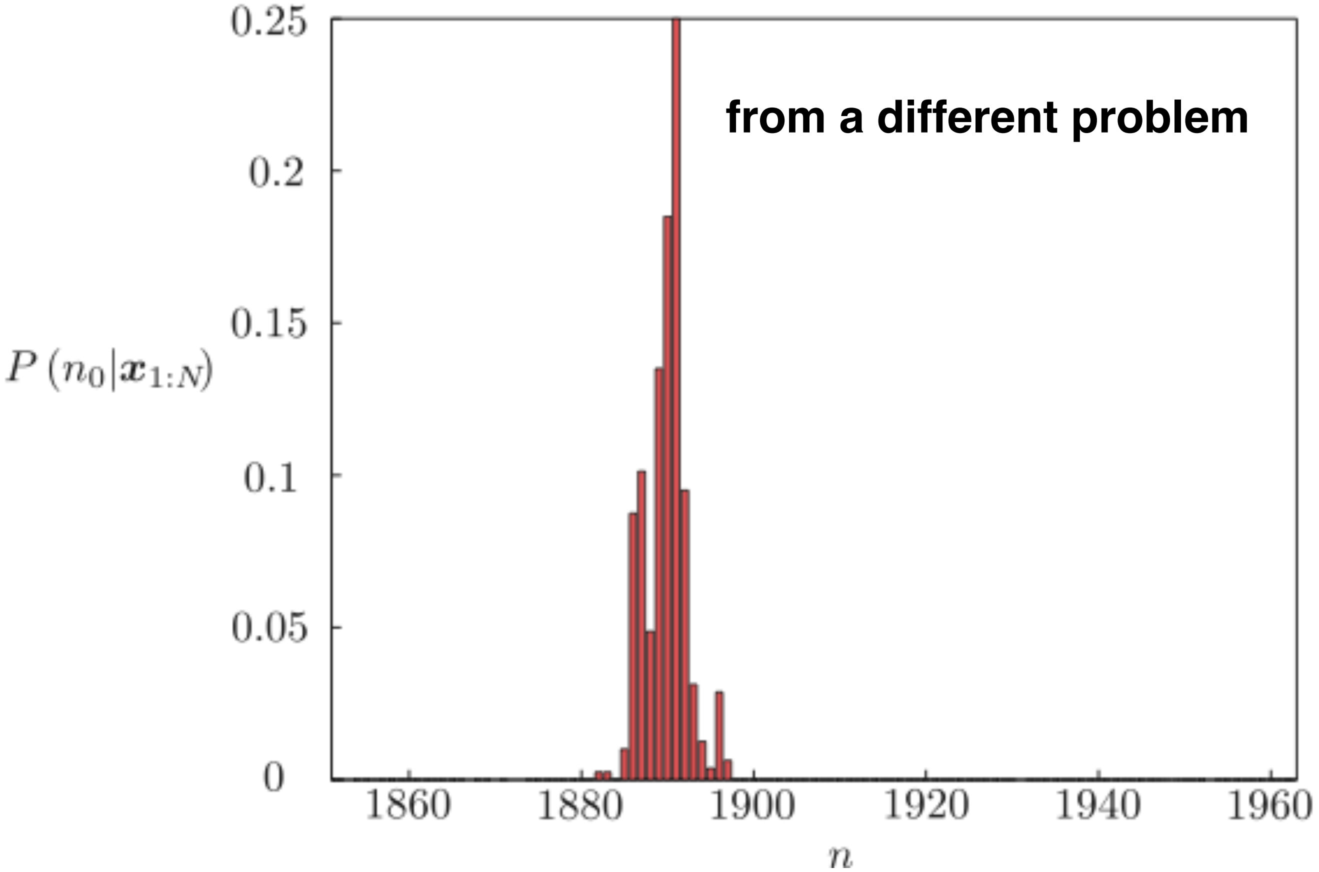
2(B) Implement the Gibbs sampling of the Markov Chain Monte Carlo:

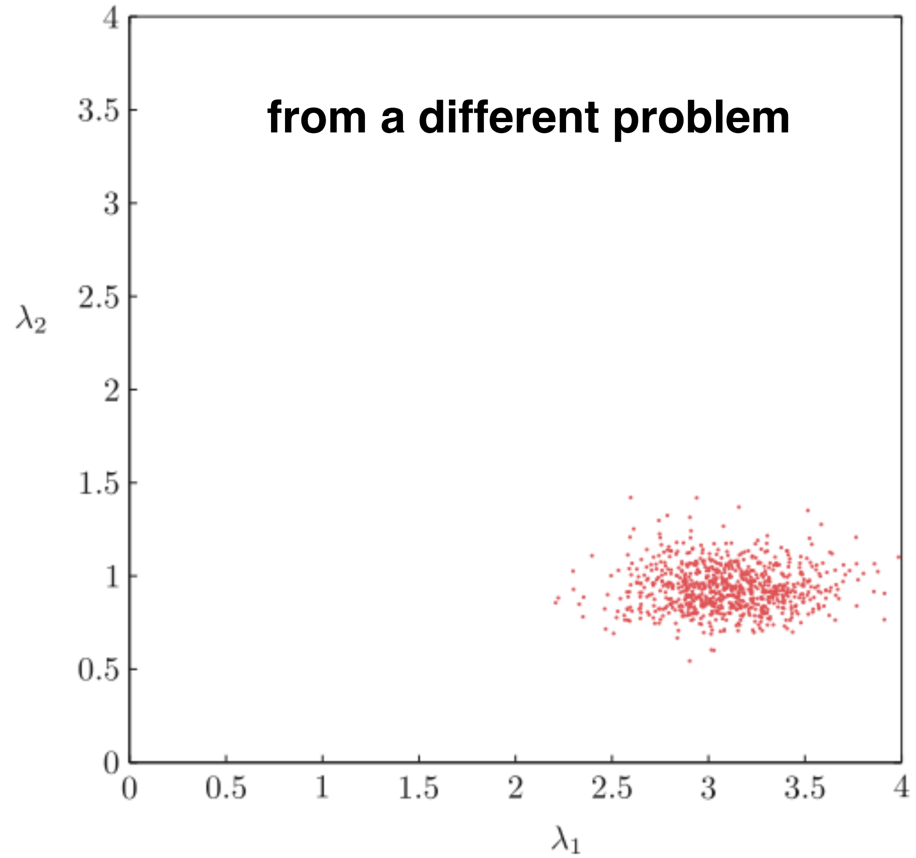
Gibbs sampling for change-point detection

- Having obtained $\boldsymbol{x}_{1:N} := \{x_1, \ldots, x_N\}$, select a and b.
- Initialize $n_0^{(0)}$
- **For** $i = 1, 2, \ldots,$ **Do**
  - $\lambda_1^{(i)} \sim \text{Gamma}\left(\lambda | a + \sum_{n=1}^{n_0^{(i-1)}} x_n,\ b + n_0^{(i-1)}\right)$
  - $\lambda_2^{(i)} \sim \text{Gamma}\left(\lambda | a + \sum_{n=n_0^{(i-1)}+1}^{N} x_n,\ b + (N - n_0^{(i-1)})\right)$
  - $n_0^{(i)} \sim P(n_0 | \lambda_1^{(i)}, \lambda_2^{(i)}, x_{1:N})$
- **End For**

Problem 2 submitted by the team

2(C) Plot the n_0 probably distribution and $\boldsymbol{\lambda}_1$, $\boldsymbol{\lambda}_2$ from Gibbs sampling:





2(D) What are the means of n_0, $\boldsymbol{\lambda}_1$, $\boldsymbol{\lambda}_2$?

2(E) Estimate the MC errors on n_0, $\boldsymbol{\lambda}_1$, $\boldsymbol{\lambda}_2$ from the independent MC configurations of the simulations

Problem 2 (phys 239 only) submitted by the team

2(F) Compare your Gibbs sampling based simulation with Metropolis Monte Carlo