Bayesian Data Analysis on Climate Data

Bayesian Computation - Spring 2019

Jonas Wolter

June 19, 2019

Why Climate Data?

Hot Topic.

Why Climate Data?

Hot Topic.

Lots of Data available.

Why Climate Data?

Hot Topic.

Lots of Data available.

Bayesian Approach offers a lot of freedom.

Outline

1. Dataset
2. Models for analysis
3. Methods for approximation
4. Comparison of Models and Approximations
5. Conclusion

Outline

Outline

1. Dataset
2. Models for analysis
3. Methods for approximation
4. Comparison of Models and Approximations
5. Conclusion

Outline

1. Dataset
2. Models for analysis
3. Methods for approximation
4. Comparison of Models and Approximations
5. Conclusion

Outline

1. Dataset
2. Models for analysis
3. Methods for approximation
4. Comparison of Models and Approximations
5. Conclusion

Dataset - Earth Surface Temperature Data

Reference: `https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data`

1. Global temperatures since 1750
   - Land average temperature
   - Monthly data
   - Missing Data and Accuracy
2. Temperatures for all countries of the world
   - Year of first data varies

Dataset - Earth Surface Temperature Data

Reference: `https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data`

1. Global temperatures since 1750
   - Land average temperature
   - Monthly data $\Rightarrow$ convert to yearly
   - Missing Data and Accuracy
2. Temperatures for all countries of the world
   - Year of first data varies

Dataset - Earth Surface Temperature Data

Reference: `https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data`

1. Global temperatures since 1750
   - Land average temperature
   - Monthly data $\Rightarrow$ convert to yearly
   - Missing Data and Accuracy $\Rightarrow$ Start from year 1830
2. Temperatures for all countries of the world
   - Year of first data varies

Dataset - Earth Surface Temperature Data

Reference: `https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data`

1. Global temperatures since 1750
   - Land average temperature
   - Monthly data ⇒ convert to yearly
   - Missing Data and Accuracy ⇒ Start from year 1830
2. Temperatures for all countries of the world
   - Year of first data varies
3. Other Data available

Dataset - Earth Surface Temperature Data

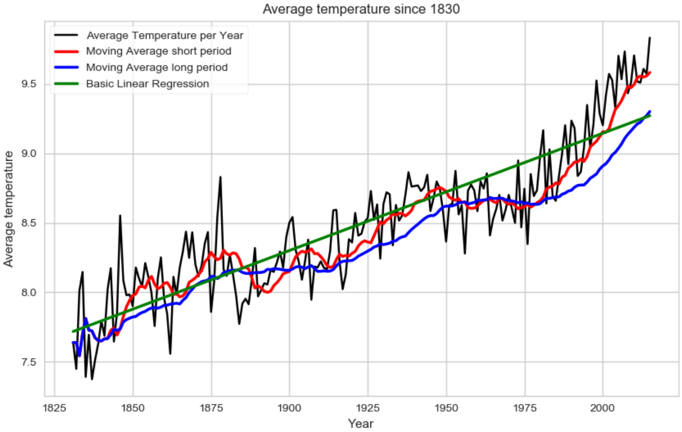


Figure 1: Yearly average temperature plotted together with moving averages and a basic linear regression line.

Models for analysis (I) - Linear Regression

- General setting:
  - $x = \text{Years} - \text{StartYear}$
  - $y = \text{Average temperature per year}$

Models for analysis (I) - Linear Regression

- General setting:
  - $x = \text{Years} - \text{StartYear}$
  - $y = \text{Average temperature per year}$
- Very basic Model:

$$y = b + ax + \epsilon, \quad (a, b) \sim \mathcal{N}(\mu, \Sigma) \quad \epsilon \sim \mathcal{N}(0, \sigma)$$

- Leads to a multivariate-Gamma distribution as a prior.
- Conjugate Model: can be solved analytically.

Models for analysis (II) - Polynomial Regression of degree 3

- Mathematical Model:

$$y = (a_0, a_1, a_2, a_3)\left(1, x, x^2, x^3\right)^T + \sigma S_d$$
$$(a_0, a_1, a_2, a_3) \sim \mathcal{N}(\mu, \Sigma), \quad \sigma \sim \mathrm{Exp}(\lambda), \quad d \sim \Gamma(k)$$

- Student Noise to account for outliers.
- More complex model but MLE solution known.
- σ and d need to be positive.

Models for analysis (III) - Linear Regression with Breaks

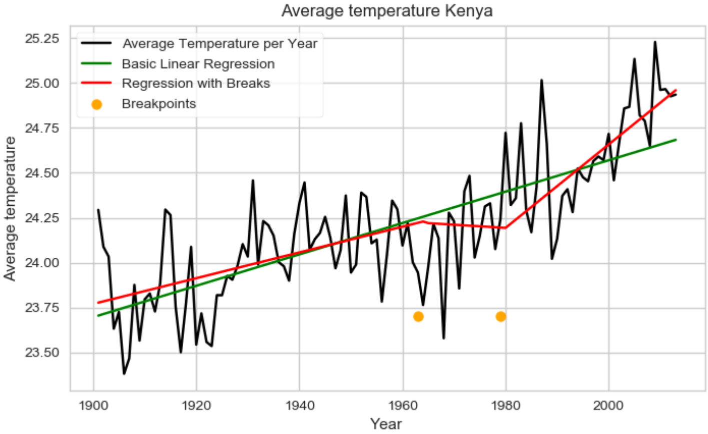


Figure 2: Example of Model III for the yearly average temperature in Kenya since 1900.

Models for analysis (III) - Linear Regression with Breaks

- Mathematical Model:

$$y = a_0 + \begin{cases} a_1 x & \text{if} \quad x < B_1 \\ a_1 B_1 + a_2 x & \text{if} \quad B_1 < x < B_2 \\ a_1 B_1 + a_2 B_2 + a_3 x & \text{if} \quad B_2 < x \end{cases} \quad + \sigma S_d$$

$$(a_0, a_1) \sim \mathcal{N}(\mu, \Sigma), \quad a_2, a_3 \sim \mathcal{N}(\mu, \tau) \quad \sigma \sim \mathrm{Exp}(\lambda), \quad d \sim \Gamma(k)$$

$$B_1 \sim \mathrm{Uniform}(0, \mathrm{Endpoint}), B_2 \sim \mathrm{Uniform}(B_1, \mathrm{Endpoint})$$

- Student Noise to account for outliers.
- σ and d need to be positive.
- B_1 and B_2 need to satisfy $0 \leq B_1 \leq B_2 \leq \mathrm{Endpoint}$.

Methods for approximation (I) - Laplace Approximation

- General Setting:
  - Unnormalised posterior $\tilde{f}(\theta|d)$
- Idea: $\tilde{f}(\theta|d) \sim C \exp\left(-\frac{1}{2}(\theta-\mu)^T \Sigma^{-1}(\theta-\mu)\right)$

Methods for approximation (I) - Laplace Approximation

- General Setting:
  - Unnormalised posterior $\tilde{f}(\theta|d)$
- Idea: $\tilde{f}(\theta|d) \sim C \exp\left(-\frac{1}{2}(\theta-\mu)^T \Sigma^{-1}(\theta-\mu)\right)$
- Approximate mean as mode of distribution.

$$\mu = \bar{\theta} = \mathrm{argmax}_\theta \left(\tilde{f}\left(\bar{\theta}|d\right)\right)$$

- Approximate Covariance as log curvature at the mode.

$$\Sigma = \left(-H_\theta\left(\log \tilde{f}\left(\bar{\theta}|d\right)\right)\right)^{-1}$$

- It follow $C = \tilde{f}\left(\bar{\theta}|d\right)$.

Methods for approximation (I) - Laplace Approximation

- General Setting:
  - Unnormalised posterior $\tilde{f}(\theta|d)$
- Idea: $\tilde{f}\left(\theta|d\right) \sim C \exp\left(-\frac{1}{2}(\theta-\mu)^T\Sigma^{-1}(\theta-\mu)\right)$
- Approximate mean as mode of distribution.

$$\mu = \bar{\theta} = \mathrm{argmax}_\theta \left(\tilde{f}\left(\bar{\theta}|d\right)\right)$$

  - How to find the maximum?
- Approximate Covariance as log curvature at the mode.

$$\Sigma = \left(-H_\theta\left(\log\ \tilde{f}\left(\bar{\theta}|d\right)\right)\right)^{-1}$$

  - How to find the Hessian?
- It follow $C = \tilde{f}\left(\bar{\theta}|d\right)$.

Methods for approximation (II) - Gaussian Variational Approximation

- Idea: $f(\theta|d) \sim \mathcal{N}(\mu, \Sigma)$
  - Minimize $\mathrm{KL}(g(\theta), f(\theta|d))$
  - Maximise

$$ELBO(\mu, \Sigma) = -\mathbb{E}\left(\phi(\mu + \exp(L)\tau)\right) + \frac{d}{2}\log(2\pi e) + Tr(L),$$

$$L = \frac{\log(\Sigma)}{2}, \quad \theta = \mu + S_\Sigma \tau, \quad S_\Sigma S_{Sigma}^T = \Sigma$$

Methods for approximation (II) - Gaussian Variational Approximation

- Idea: $f(\theta|d) \sim \mathcal{N}(\mu, \Sigma)$
  - Minimize $\mathrm{KL}(g(\theta), f(\theta|d))$
  - Maximise

$$ELBO(\mu, \Sigma) = -\mathbb{E}(\phi(\mu + \exp(L)\tau)) + \frac{d}{2}\log(2\pi e) + Tr(L),$$

$$L = \frac{\log(\Sigma)}{2}, \quad \theta = \mu + S_\Sigma \tau, \quad S_\Sigma S_{Sigma}^T = \Sigma$$

- This requires $\nabla_\theta ELBO$ and $\nabla_L ELBO$.

Methods for approximation (II) - Gaussian Variational Approximation

- Idea: $f(\theta|d) \sim \mathcal{N}(\mu, \Sigma)$
  - Minimize $\mathrm{KL}(g(\theta), f(\theta|d))$
  - Maximise

$$ELBO(\mu, \Sigma) = -\mathbb{E}\left(\phi(\mu + \exp(L)\tau)\right) + \frac{d}{2}\log(2\pi e) + Tr(L),$$

$$L = \frac{\log(\Sigma)}{2}, \quad \theta = \mu + S_\Sigma \tau, \quad S_\Sigma S_{Sigma}^T = \Sigma$$

- This requires $\nabla_\theta ELBO$ and $\nabla_L ELBO$.
- Use stochastic gradient descent to find optimal μ, L.

Methods for approximation (II) - Gaussian Variational Approximation

- Idea: $f(\theta|d) \sim \mathcal{N}(\mu, \Sigma)$
  - Minimize $\mathrm{KL}(g(\theta), f(\theta|d))$
  - Maximise

$$ELBO(\mu, \Sigma) = -\mathbb{E}\left(\phi(\mu + \exp(L)\tau)\right) + \frac{d}{2}\log(2\pi e) + Tr(L),$$
$$L = \frac{\log(\Sigma)}{2}, \quad \theta = \mu + S_\Sigma \tau, \quad S_\Sigma S_{Sigma}^T = \Sigma$$

- This requires $\nabla_\theta ELBO$ and $\nabla_L ELBO$.
- Use stochastic gradient descent to find optimal μ, L.
- Two possible methods have been implemented.

Methods for approximation (III) - Metropolis Hastings

- Idea: Building a Markov Chain which has $f(\theta|d)$ as stationary distribution.
    - Start with any initial chain and correct *flow*.

Methods for approximation (III) - Metropolis Hastings

- Idea: Building a Markov Chain which has $f(\theta|d)$ as stationary distribution.
  - Start with any initial chain and correct *flow*.
- Algorithm
  1. Start with initial point P_0
  2. Generate proposal $P_{prop} = P_n + \lambda\eta_n$, where η_n need to be symmetric.
     1. Accept or reject proposal if probability rises.
     2. Accept a fraction of steps when probability falls.
  3. Other proposals possible.
  4. Possibly delete *burn-in* period.

Methods for approximation (III) - Metropolis Hastings

- Idea: Building a Markov Chain which has $f(\theta|d)$ as stationary distribution.
  - Start with any initial chain and correct *flow*.
- Algorithm
  1. Start with initial point P_0
  2. Generate proposal $P_{prop} = P_n + \lambda\eta_n$, where η_n need to be symmetric.
     1. Accept or reject proposal if probability rises.
     2. Accept a fraction of steps when probability falls.
  3. Other proposals possible.
  4. Possibly delete *burn-in* period.
- Stepsize is critical!

Comparison of Models - Basic Linear Regression

- All methods yield good results in less than a minute runtime. Laplace is the fastest.
- GVA and Laplace almost identical.

Comparison of Models - Basic Linear Regression

- All methods yield good results in less than a minute runtime. Laplace is the fastest.
- GVA and Laplace almost identical.
- Marginal Variance for intercept and slope are small.
- Error estimate σ is around 0.3. Actual standard deviation is 0.27.

Comparison of Models - Basic Linear Regression

- All methods yield good results in less than a minute runtime. Laplace is the fastest.
- GVA and Laplace almost identical.
- Marginal Variance for intercept and slope are small.
- Error estimate σ is around 0.3. Actual standard deviation is 0.27.
- Not very good for forecasting because it is sluggish.

Comparison of Models - Basic Linear Regression

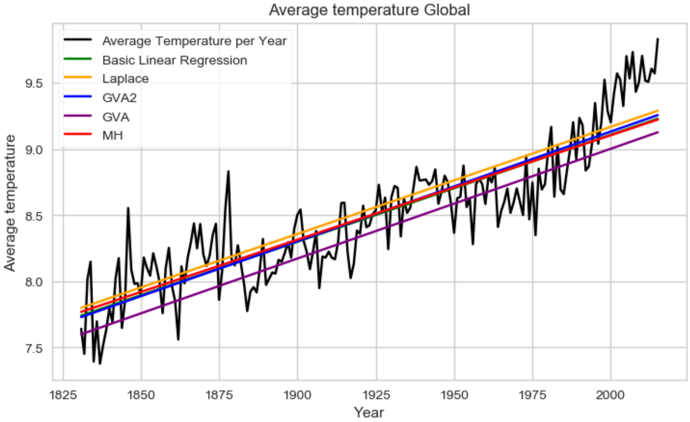
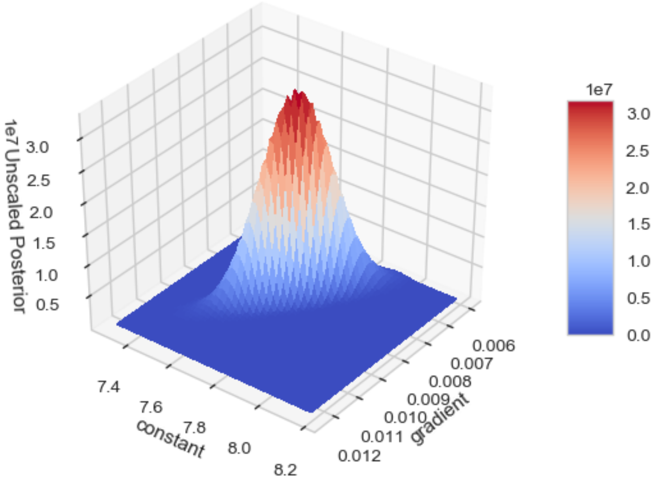


Figure 3: Basic Bayesian linear regression with different approximation methods.

Comparison of Models - Basic Linear Regression



Comparison of Models - Polynomial Regression

- MH yield very good results in a decent amount of time.
- Laplace and GVA take long to arrive at acceptable results.
- Variance for coefficients is very small.
  - Problems
    1. σ and d might not be normally distributed.
    2. Calamity of multimodality.
    3. Posterior is steep.
    4. Optimisation issues.

Comparison of Models - Polynomial Regression

- MH yield very good results in a decent amount of time.
- Laplace and GVA take long to arrive at acceptable results.
- Variance for coefficients is very small.
  - Problems
    1. σ and d might not be normally distributed.
    2. Calamity of multimodality.
    3. Posterior is steep.
    4. Optimisation issues.
- Not very good for forecasting because of overfitting.

Comparison of Models - Polynomial Regression

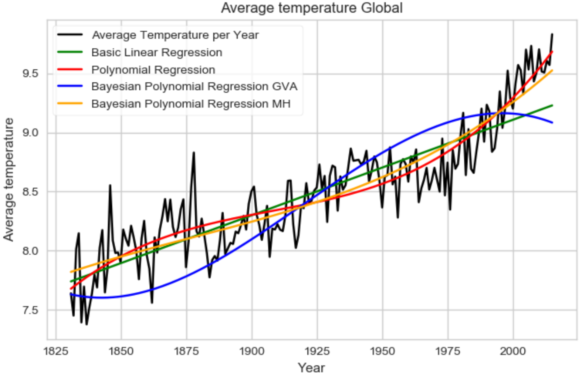


Figure 5: Result of polynomial regression for different approximation methods.

Comparison of Models - Polynomial Regression

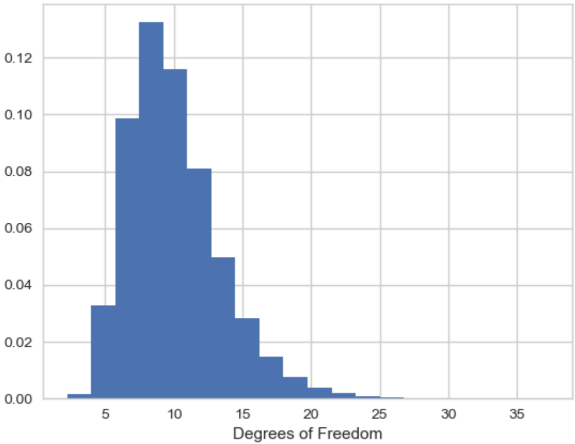


Figure 6: Histogram for parameter d obtained using the MH-algorithm.

Comparison of Models - Polynomial Regression

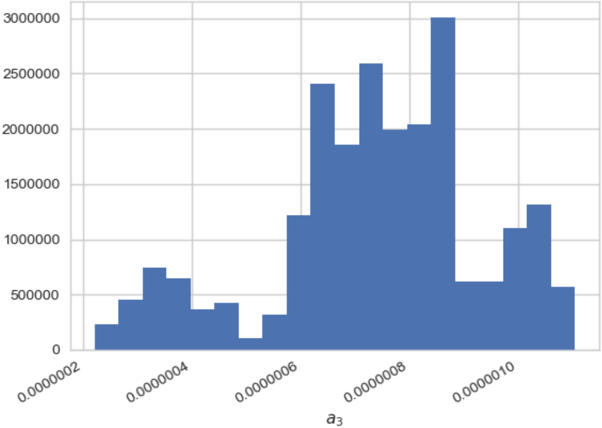


Figure 7: Histogram for coefficient of cubic term obatained using MH-algorithm with 1,000,000 samples.

Comparison of Models - Linear Regression with Breaks

- All methods yield acceptable results. MH has best performance.
- GVA without full Hessian does not work.
- GVA is slow because of the computation of the Hessian.
- Variance for coefficients is very small.
- Similar problems as for polynomial Regression but not as severe.

Comparison of Models - Linear Regression with Breaks

- All methods yield acceptable results. MH has best performance.
- GVA without full Hessian does not work.
- GVA is slow because of the computation of the Hessian.
- Variance for coefficients is very small.
- Similar problems as for polynomial Regression but not as severe.
- Sum of Squares if less than for polynomial regression.
- Model is easier to compute than polynomial regression.
- Best model for forecasting.

Comparison of Models - Forecasting

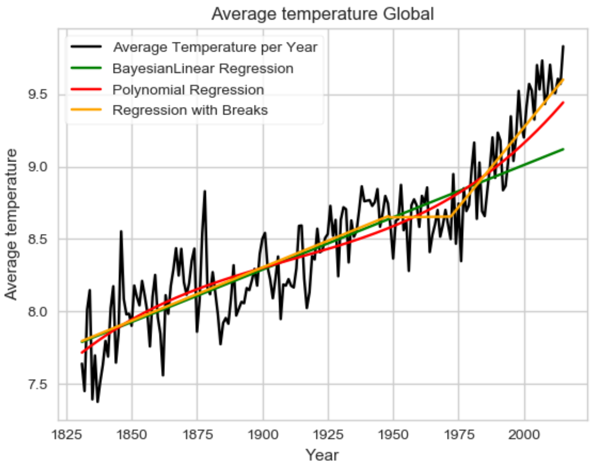


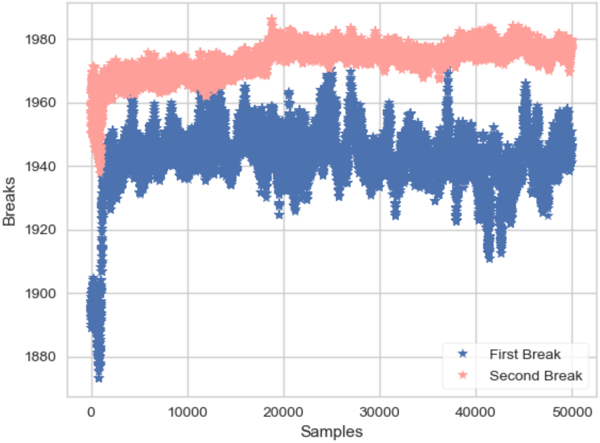
Figure 8: Different Models plotted when climate data until 1990 is considered.

Conclusions

1. Laplace approximation is fast but generally MH works best.
    - MH yields very accurate results.
    - Burn-in period is not too long.

Conclusions

1. Laplace approximation is fast but generally MH works best.
   - MH yields very accurate results.
   - Burn-in period is not too long.



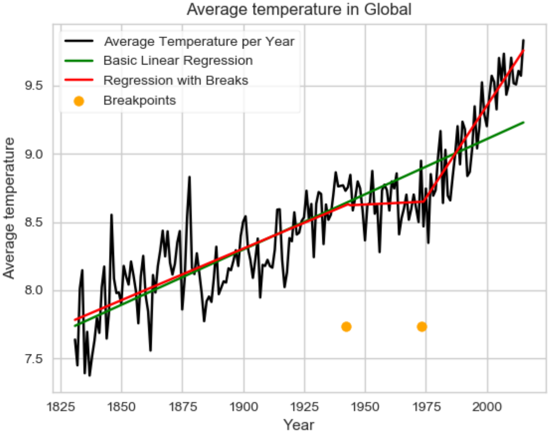
Conclusions

1. Laplace approximation is fast but generally MH works best.
   - MH yields very accurate results.
   - Burn-in period is not too long.
2. GVA is not very suitable for this problem.

Conclusions

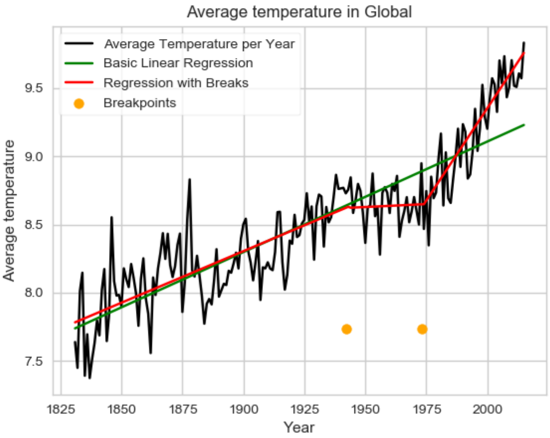
1. Laplace approximation is fast but generally MH works best.
   - MH yields very accurate results.
   - Burn-in period is not too long.
2. GVA is not very suitable for this problem.
3. Model III - Linear Regression with breaks works best for the given data.

Conclusion



Average temperature in Global
Average Temperature per Year
Basic Linear Regression
Regression with Breaks
Breakpoints
Average temperature
9.5
9.0
8.5
8.0
7.5
1825
1850
1875
1900
1925
1950
1975
2000
Year

Conclusion



It allows to see the man-made climate change.