Bayesian Inference for Climate Science: Examples using RSTAN

Kayla Montgomery, Ben Lee and Murali Haran
Department of Statistics, The Pennsylvania State University and the Network for Sustainable Climate Risk Management

Introduction

Goals:

- Describe basic ideas of Bayesian Inference
- Show how RSTAN can be used for Bayesian Inference
- Show how Bayesian Inference can be used for handling missing data

Bayesian Inference

- Bayesian Inference views parameters as random variables while frequentists ("classical statistics") view them as fixed
- Each parameter has a prior distribution which is then updated with the data
- Using the model for the data, information about the parameter is updated based on observations
- The updated distribution for the parameter is called the posterior distribution
- Outline of Bayesian Inference
  - Data: x, Parameter: θ
  - Probability model for data, $f(x|\theta)$
  - Prior for θ, $p(\theta)$
  - Posterior for θ, $\pi(\theta|x) \propto f(x|\theta)p(\theta)$
  - $f(x|\theta)$ with x observed is referred to as the likelihood function $L(x|\theta)$ or $L(\theta;x)$

Example: Linear Regression

- Example: Linear Regression
  - $Y = \beta_0 + \beta_1 X + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$
  - $\beta_0 \sim$ Normal(0, 100)
  - $\beta_1 \sim$ Normal(0, 100)
  - $\sigma^2 \sim$ Gamma(.001, .001)

Bayesian Computation

- Draw samples from the posterior distribution using the Metropolis-Hastings algorithm
- Markov Chain Monte Carlo
  - Can approximate population means of the distribution using sample means
- We can write our own code for the Metropolis-Hastings algorithm for a given posterior distribution
- RSTAN can take a model description and construct the Metropolis-Hastings code
  - Allows for more models to be fit more quickly/routinely

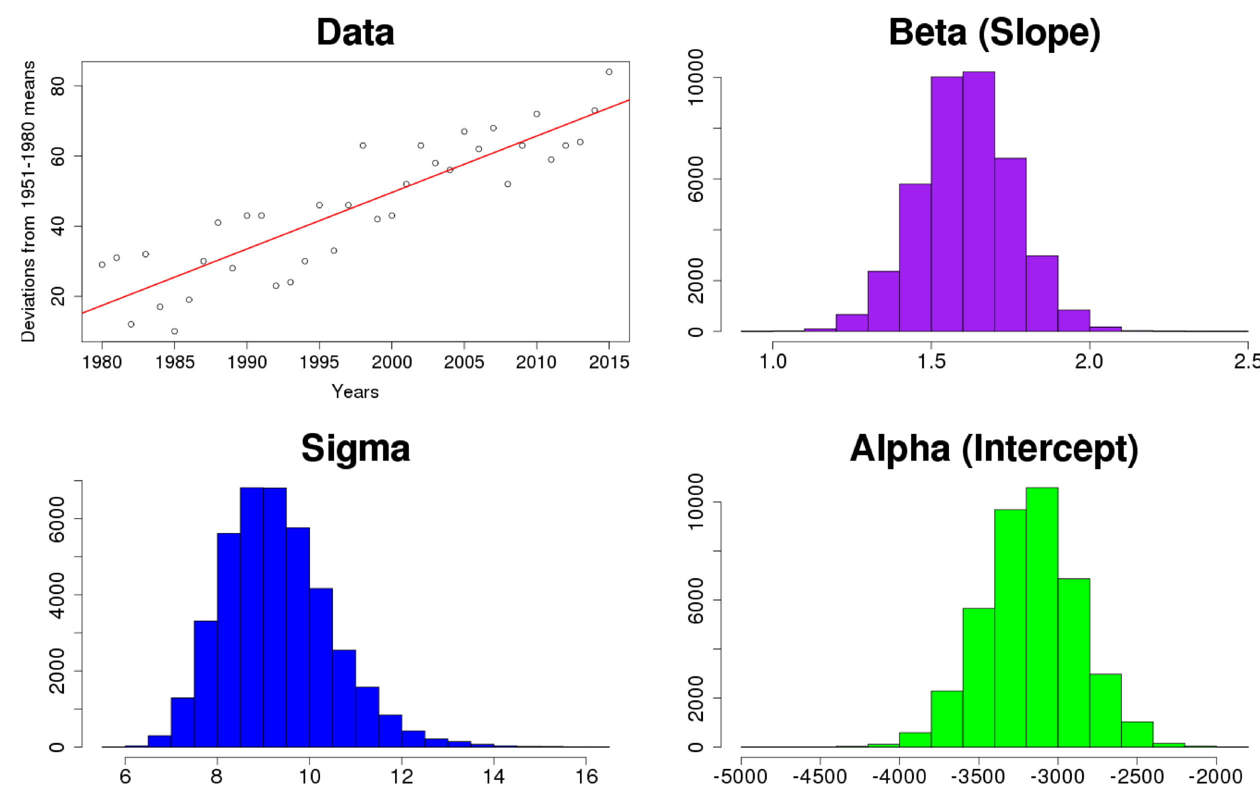
RSTAN

RSTAN provides a convenient way for users to specify models and provide data. It then generates MCMC based inference for the posterior distribution

Example:

- Y_i: global temperature average deviations at ith year from the 1950-1980 means
- X_i……X_n are years after 1980
- Used RSTAN with 20,000 iterations and 4 chains.
- Poster mean of Mean of
  - alpha (Intercept): 1.61
  - sigma: 9.30
  - beta (slope): -3171.59
- Data from NASA website
- Model: $Y_i = \alpha + \beta X_i + \varepsilon_i$, where $\varepsilon \sim N(0, \sigma^2)$

Posterior Inference from RSTAN



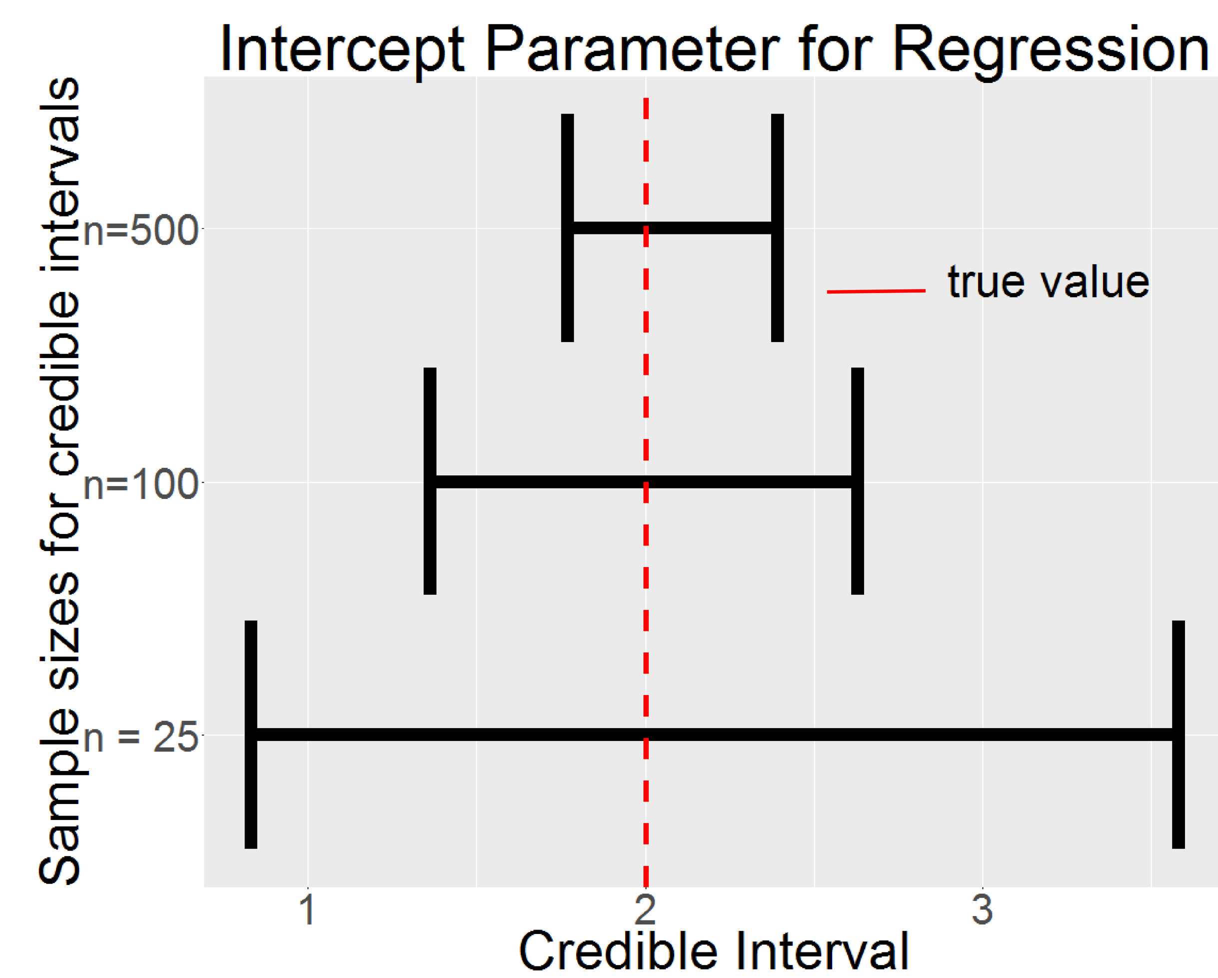
RSTAN Example Code

Model string for RSTAN for the global means data

```
model_string <- "
data {
 int N;
 vector[N] x;
 vector[N] y;
}
parameters {
 real alpha;
 real beta;
 real sigma;
}
model {
 y ~ normal(alpha + beta * x, sigma);
}"
```

Credible Interval

- Bayesian approach uses Credible Intervals to summarize information about parameters
  - Roughly analogous to a classical Confidence Interval
  - 95% Credible Interval is the shortest interval that contains 95% of the data in a posterior distribution
- Length of the 95% credible interval shrinks with more data



Credible v Confidence Intervals

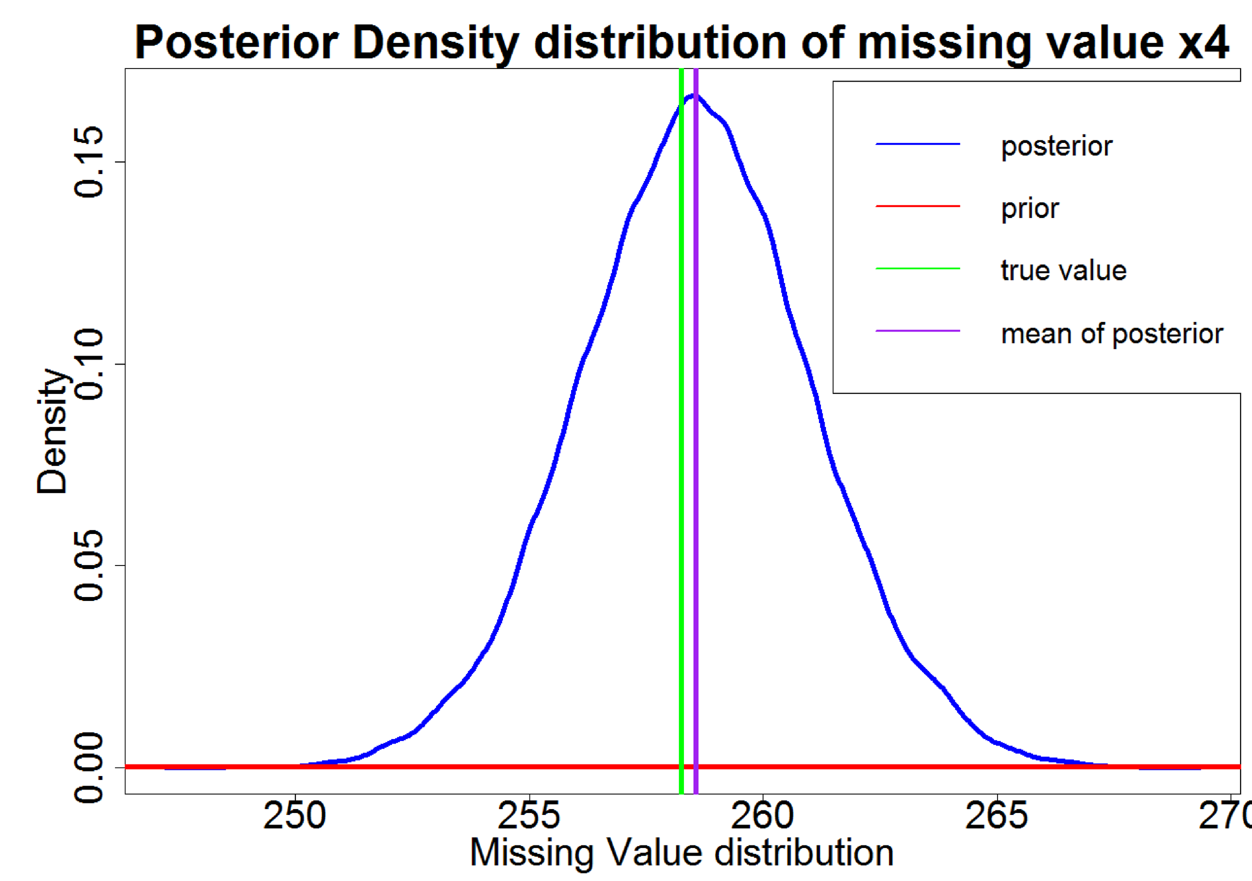
- Compared average length and coverage percentage for credible (Bayesian) and confidence intervals (Frequentist)
- Model:
  - $Y_i = \alpha + \beta X_i + \varepsilon_i$
  - $\varepsilon_1, \varepsilon_2, \varepsilon_3$ iid $N(0, \sigma^2)$
  - Parameters: α, β, σ^2
  - Data: (X_1, Y_1), ….., (X_n, Y_n)
  - Posterior: $\pi(\alpha, \beta, \sigma^2 | Y, X)$
- Priors:
  - $\alpha \sim N(0, 10)$, $p(\alpha)$
  - $\beta \sim N(0, 10)$, $p(\beta)$
  - $\sigma^2 \sim$ Gamma(.001, .001), $p(\sigma^2)$
- Ran 100 simulations of 25 data points to calculate average length of credible/confidence intervals and coverage
- Coverage is the proportion of intervals that contain the true parameter value
- α (slope): similar coverage and similar length
- β (intercept): similar coverage and similar length
- σ: Bayesian approach had better coverage but credible interval was twice as long as the Confidence Interval

Frequestist	Slope	Sigma	Intercept
Average Length of Confidence Interval	0.046	**0.878**	2.626
Percentage that have true values	95%	**65%**	95%
Bayesian	**Slope**	**Sigma**	**Intercept**
Average Length of Credible Interval	0.046	**1.954**	2.670
Percentage that have true values	94%	**97%**	95%

Missing Data

- Can use Bayesian approach to handle missing data
  - Missing data points are treated as parameters with a prior and posterior distribution
  - Inference for other parameters now includes uncertainty about missing data

Data was generated using the following model:
$Y \sim 6 + 3X + \varepsilon$, where $\varepsilon \sim N(0, 3)$
$X \sim 5 + 5W + \delta$, where $\delta \sim N(0, 6)$
W is a known covariate
10 values of x were removed and treated as missing



	4 of the 10 Missing Data		
	Credible Interval	Point Estimate	True Value
X1	(-247.14, -237.40)	-242.21	-241.19
X2	(-145.83, -136.14)	-140.91	-139.62
X3	(407.31, 416.94)	412.26	413.24
X4	(253.74, 263.62)	258.56	258.25

Conclusion

- Bayesian Inference is useful for estimating parameters
  - Particularly in handling missing data
- RSTAN is fast and convenient for Bayesian inference

References

1. Wood, S. N. (n.d.). Core statistics
2. Stan Development Team. (n.d.). Stan Modeling Language User's Guide and Reference Manual (2.9.0 ed.)
3. GLOBAL Land-Ocean Temperature Index in 0.01 degrees Celsius. (n.d.). Retrieved July 21, 2016, from http://data.giss.nasa.gov/gistemp/tabledata_v3/GLB.Ts dSST.txt

Contact info: Kayla Montgomery | knm5291@psu.edu | Department of Statistics, The Pennsylvania State University and the Network for Sustainable Climate Risk Management

Acknowledgements: This work was supported by the National Science Foundation through the Network for Sustainable Climate Risk Management (SCRiM) under NSF cooperative agreement GEO-1240507. Patrick Applegate provided the Global Mean Data. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.