A gentle introduction to Bayesian statistics and modeling - Part I

Brian Groenke

July 2024

Table of Contents

Probability according to Cicero

"That is probable which for the most part usually comes to pass, or which is a part of the ordinary beliefs of mankind, or which contains in itself some resemblance to these qualities, whether such resemblance be true or false."

– Cicero, ca. 80 BCE

The Frequentist vs. The Bayesian

The Frequentist argues that probability is strictly defined by the *frequency* of an event over an *infinite* number of samples or experiments.

The Frequentist vs. The Bayesian

The Frequentist argues that probability is strictly defined by the *frequency* of an event over an *infinite* number of samples or experiments.

The Bayesian argues that probability quantifies our **uncertainty** about a prediction, parameter, or hypothesis.

The Frequentist vs. The Bayesian

The Frequentist argues that probability is strictly defined by the *frequency* of an event over an *infinite* number of samples or experiments.

The Bayesian argues that probability quantifies our **uncertainty** about a prediction, parameter, or hypothesis.

For the Frequentist, the "true" (population) parameters are fixed while the observed data (samples) are random.

The Frequentist vs. The Bayesian

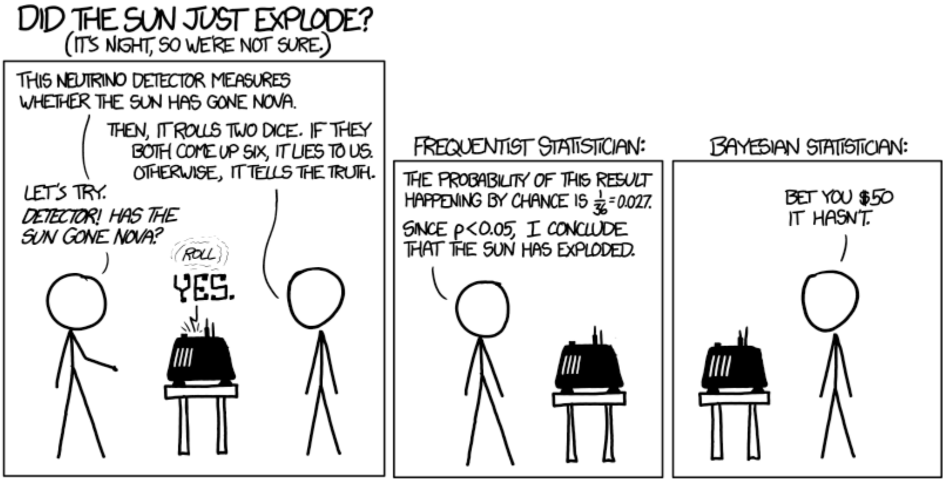
The Frequentist argues that probability is strictly defined by the *frequency* of an event over an *infinite* number of samples or experiments.

The Bayesian argues that probability quantifies our **uncertainty** about a prediction, parameter, or hypothesis.

For the Frequentist, the "true" (population) parameters are fixed while the observed data (samples) are random.

For the Bayesian, all variables are random variables, including the parameters!

The Frequentist vs. The Bayesian



Frequentist inference

Under the frequentist interpretation of probability, the parameters of the population are fixed quantities to be estimated.

Frequentist inference

Under the frequentist interpretation of probability, the parameters of the population are fixed quantities to be estimated.

The frequentist is therefore prohibited from making *probabilistic* statements about parameters and hypotheses.

Frequentist inference

Under the frequentist interpretation of probability, the parameters of the population are fixed quantities to be estimated.

The frequentist is therefore prohibited from making *probabilistic* statements about parameters and hypotheses.

Probabilities and uncertainties can only be expressed in terms of *repeated sampling* from some target population.

Null hypothesis significance testing

Common practice "null hypothesis significance test" (NHST) inference:

Null hypothesis significance testing

Common practice "null hypothesis significance test" (NHST) inference:

1. Define your experiment, along with the relevant **null** (H_0) and **alternative** (H_A) hypotheses, and collect your data.

Null hypothesis significance testing

Common practice "null hypothesis significance test" (NHST) inference:

1. Define your experiment, along with the relevant **null** (H_0) and **alternative** (H_A) hypotheses, and collect your data.

2. Compute the relevant *test statistic* $\tilde{z}$ for your null hypothesis.

Null hypothesis significance testing

Common practice "null hypothesis significance test" (NHST) inference:

1. Define your experiment, along with the relevant **null** (H_0) and **alternative** (H_A) hypotheses, and collect your data.

2. Compute the relevant *test statistic* $\tilde{z}$ for your null hypothesis.

3. Compute the p-value $p = P(\tilde{z}|H_0)$, i.e. the probability of the observed test statistic under the null distribution.

Null hypothesis significance testing

Common practice "null hypothesis significance test" (NHST) inference:

1. Define your experiment, along with the relevant **null** (H_0) and **alternative** (H_A) hypotheses, and collect your data.

2. Compute the relevant *test statistic* $\tilde{z}$ for your null hypothesis.

3. Compute the p-value $p = P(\tilde{z}|H_0)$, i.e. the probability of the observed test statistic under the null distribution.

4. Reject H_0 if $p < \alpha$ where α is an arbitrary significance threshold.

Problems with NHST

There are several problems with this procedure.

Problems with NHST

There are several problems with this procedure.

1. The p-value, $p = P(\tilde{z}|H_0)$, is misleading.

Problems with NHST

There are several problems with this procedure.

1. The p-value, $p = P(\tilde{z}|H_0)$, is misleading.

- It is **not** the probability of the null hypothesis being "true"; this would be $P(H_0|\mathcal{D})$

Problems with NHST

There are several problems with this procedure.

1. The p-value, $p = P(\tilde{z}|H_0)$, is misleading.

- It is **not** the probability of the null hypothesis being "true"; this would be $P(H_0|\mathcal{D})$
- It is **not** the probability that your data was produced by "random chance".

Problems with NHST

There are several problems with this procedure.

1. The p-value, $p = P(\tilde{z}|H_0)$, is misleading.

- It is **not** the probability of the null hypothesis being "true"; this would be $P(H_0|\mathcal{D})$
- It is **not** the probability that your data was produced by "random chance".
- It does **not** imply anything about the size of the effect!

Problems with NHST

There are several problems with this procedure.

1. The p-value, $p = P(\tilde{z}|H_0)$, is misleading.

- It is **not** the probability of the null hypothesis being "true"; this would be $P(H_0|\mathcal{D})$
- It is **not** the probability that your data was produced by "random chance".
- It does **not** imply anything about the size of the effect!
- It says **absolutely nothing** about the alternative hypothesis, H_A.

Problems with NHST

2. The null hypothesis is almost never true.

Problems with NHST

2. The null hypothesis is almost never true.

- When is the "true" linear relationship between two values ever *exactly* zero?

Problems with NHST

2. The null hypothesis is almost never true.

- When is the "true" linear relationship between two values ever *exactly* zero?
- When is the "true" difference between means every *exactly* zero?

Problems with NHST

2. The null hypothesis is almost never true.

- When is the "true" linear relationship between two values ever *exactly* zero?
- When is the "true" difference between means every *exactly* zero?
- What is the point of falsifying something that we know *a priori* must be false?

Problems with NHST

3. Inference based on a test statistic $\tilde{z}$ is inference based on **hypothetical data** that was never observed.

- The null distribution $\mathrm{P}(\tilde{z}|\mathrm{H}_0)$ represents an **assumed distribution over data**, not over the hypothesis itself.

Problems with NHST

3. Inference based on a test statistic $\tilde{z}$ is inference based on **hypothetical data** that was never observed.

- The null distribution $P(\tilde{z}|H_0)$ represents an **assumed distribution over data**, not over the hypothesis itself.

- The p-value is the probability of observing your data or something "more extreme" under this distribution.

Problems with NHST

3. Inference based on a test statistic $\tilde{z}$ is inference based on **hypothetical data** that was never observed.

- The null distribution $P(\tilde{z}|H_0)$ represents an **assumed distribution over data**, not over the hypothesis itself.

- The p-value is the probability of observing your data or something "more extreme" under this distribution.

- Inference based on this probability implicitly conditions on hypothetical data that was never observed!

Problems with NHST

4. Pre-defined hypothesis test procedures often make a wide range of assumptions about the data.

Problems with NHST

4. Pre-defined hypothesis test procedures often make a wide range of assumptions about the data.

Student's t-test assumptions

- The sample mean is normally distributed.
- The sample variance is χ^2 distributed (or the samples are normally distributed).
- The sample mean and sample variance are independent.
- The variances of both populations are equal (relaxed for Welch's t-test).

Problems with NHST

5. Uncertainties expressed in the limit of infinite samples are often not helpful.

Problems with NHST

5. Uncertainties expressed in the limit of infinite samples are often not helpful.

- We never have infinite data.

Problems with NHST

5. Uncertainties expressed in the limit of infinite samples are often not helpful.

- We never have infinite data.

- In many areas of study such as geoscience, we have very little control over the data generating process.

Problems with NHST

5. Uncertainties expressed in the limit of infinite samples are often not helpful.

- We never have infinite data.

- In many areas of study such as geoscience, we have very little control over the data generating process.

- Uncertainty should be quantified in terms of how well our hypotheses fit the data we have and not against hypothetical data that was never observed (see The Likelihood Principle).

Problems with NHST

These criticisms are not new...

- Berkson J. **Some difficulties of interpretation encountered in the application of the chi-square test.** Journal of the American Statistical Association. 1938.
- Rozeboom WW. **The fallacy of the null-hypothesis significance test.** Psychological bulletin. 1960.
- Berger JO, Sellke T. **Testing a point null hypothesis: The irreconcilability of p values and evidence.** Journal of the American statistical Association. 1987.
- Johnson DH. **The insignificance of statistical significance testing.** The journal of wildlife management. 1999.

Problems with NHST

. . . and have continued over time:

- Wasserstein RL, Schirm AL, Lazar NA. **Moving to a world beyond "p< 0.05"**. 2019.
- Amrhein V, Greenland S, McShane B. **Scientists rise up against statistical significance.** Nature. 2019 Mar;567(7748):305-7.
- Gelman A, Stern H. **The difference between "significant" and "not significant" is not itself statistically significant.** The American Statistician. 2006 Nov 1;60(4):328-31.
- Vasishth S, Mertzen D, Jäger LA, Gelman A. **The statistical significance filter leads to overoptimistic expectations of replicability.** Journal of Memory and Language. 2018 Dec 1;103:151-75.

What about confidence intervals?

Confidence intervals are generally preferable to p-values.

What about confidence intervals?

Confidence intervals are generally preferable to p-values.

However, they are still easy to misinterpret (Morey et al. 2015):

1. A particular 95% confidence interval does not have a 95% probability of containing the true parameter value.

What about confidence intervals?

Confidence intervals are generally preferable to p-values.

However, they are still easy to misinterpret (Morey et al. 2015):

1. A particular 95% confidence interval does not have a 95% probability of containing the true parameter value.
2. Smaller confidence intervals do not imply "more precise" results.

What about confidence intervals?

Confidence intervals are generally preferable to p-values.

However, they are still easy to misinterpret (Morey et al. 2015):

1. A particular 95% confidence interval does not have a 95% probability of containing the true parameter value.
2. Smaller confidence intervals do not imply "more precise" results.
3. Confidence intervals do not correspond to plausible or likely parameter values.

What about confidence intervals?

Confidence intervals are generally preferable to p-values.

However, they are still easy to misinterpret (Morey et al. 2015):

1. A particular 95% confidence interval does not have a 95% probability of containing the true parameter value.
2. Smaller confidence intervals do not imply "more precise" results.
3. Confidence intervals do not correspond to plausible or likely parameter values.

Just like p-values, CIs represent **pre-observational probabilities** over the sampling procedure itself.

This is confusing...

- Interpreting p-values and confidence intervals is highly counter-intuitive.

This is confusing. . .

- Interpreting p-values and confidence intervals is highly counter-intuitive.

- "Significance" testing forces researchers to make arbitrary "decisions" about their data that are not necessary.

This is confusing. . .

- Interpreting p-values and confidence intervals is highly counter-intuitive.
- "Significance" testing forces researchers to make arbitrary "decisions" about their data that are not necessary.
- The basic theory of random sampling and relative frequency are deceptively intuitive. . .

This is confusing. . .

- Interpreting p-values and confidence intervals is highly counter-intuitive.
- "Significance" testing forces researchers to make arbitrary "decisions" about their data that are not necessary.
- The basic theory of random sampling and relative frequency are deceptively intuitive. . .
- . . . but using the corresponding procedures to answer common research questions is actually not!

When to still be frequentist

- Equivalent to the Bayesian solution (e.g. basic linear regression)

When to still be frequentist

- Equivalent to the Bayesian solution (e.g. basic linear regression)

- No prior information is available (rarely the case)

When to still be frequentist

- Equivalent to the Bayesian solution (e.g. basic linear regression)
- No prior information is available (rarely the case)
- Repeated (random) sampling is built into the problem

When to still be frequentist

- Equivalent to the Bayesian solution (e.g. basic linear regression)
- No prior information is available (rarely the case)
- Repeated (random) sampling is built into the problem
- Sampling distribution and experimental design are well controlled

When to still be frequentist

- Equivalent to the Bayesian solution (e.g. basic linear regression)
- No prior information is available (rarely the case)
- Repeated (random) sampling is built into the problem
- Sampling distribution and experimental design are well controlled
- Nonparametrics (e.g. bootstrapping) for large, high-dimensional datasets

Recommendations

- **Do not** equate "statistical significance" with *scientific significance*.

Recommendations

- **Do not** equate "statistical significance" with *scientific significance*.
- **Do not** selectively show results based on p-values or confidence intervals.

Recommendations

- **Do not** equate "statistical significance" with *scientific significance*.
- **Do not** selectively show results based on p-values or confidence intervals.
- **Do not** highlight certain results as "statistically significant" based on their p-values or confidence intervals.

Recommendations

- **Do not** equate "statistical significance" with *scientific significance*.
- **Do not** selectively show results based on p-values or confidence intervals.
- **Do not** highlight certain results as "statistically significant" based on their p-values or confidence intervals.
- **Avoid** NHST in favor of either the more rigorous Neyman-Pearson theory or a Bayesian approach.

Recommendations

- **Do** be careful and conservative with your interpretation of p-values and confidence intervals.

Recommendations

- **Do** be careful and conservative with your interpretation of p-values and confidence intervals.
- **Do** be as transparent as possible when reporting statistical results.

Recommendations

- **Do** be careful and conservative with your interpretation of p-values and confidence intervals.
- **Do** be as transparent as possible when reporting statistical results.
- **Do** prefer problem-specific error metrics and goodness of fit measures like R^2 to p-values.

Recommendations

- **Do** be careful and conservative with your interpretation of p-values and confidence intervals.
- **Do** be as transparent as possible when reporting statistical results.
- **Do** prefer problem-specific error metrics and goodness of fit measures like R^2 to p-values.
- **Do** use resampling or out-of-sample testing to assess robustness of your results.

When to be Bayesian

- Small, sparse, or partial data

When to be Bayesian

- Small, sparse, or partial data
- Valuable prior information is available

When to be Bayesian

- Small, sparse, or partial data
- Valuable prior information is available
- Ill-posed inverse problems

When to be Bayesian

- Small, sparse, or partial data
- Valuable prior information is available
- Ill-posed inverse problems
- Uncertainty quantification in non-standard models

When to be Bayesian

- Small, sparse, or partial data

- Valuable prior information is available

- Ill-posed inverse problems

- Uncertainty quantification in non-standard models

In general, it's good to be Bayesian by default :)

Bayes rule

We can prove Bayes rule in just three lines!

Bayes rule

We can prove Bayes rule in just three lines!

$$p(A, B) = p(B, A)$$

Bayes rule

We can prove Bayes rule in just three lines!

$$p(A, B) = p(B, A)$$

$$p(A|B)P(B) = p(B|A)p(A)$$

Bayes rule

We can prove Bayes rule in just three lines!

$$p(A, B) = p(B, A)$$

$$p(A|B)P(B) = p(B|A)p(A)$$

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

Statistical inference via Bayes rule

Let $A = \theta$ be our unobserved variables of interest and $B = \mathcal{D}$ be our observed data.

Statistical inference via Bayes rule

Let $A = \theta$ be our unobserved variables of interest and $B = \mathcal{D}$ be our observed data.

The posterior distribution $p(\theta|\mathcal{D})$ represents our belief or knowledge of θ after the data has been taken into account.

Statistical inference via Bayes rule

Let $A = \theta$ be our unobserved variables of interest and $B = \mathcal{D}$ be our observed data.

The posterior distribution $p(\theta|\mathcal{D})$ represents our belief or knowledge of θ after the data has been taken into account.

θ can be as simple as a mean or proportion...

Statistical inference via Bayes rule

Let $A = \theta$ be our unobserved variables of interest and $B = \mathcal{D}$ be our observed data.

The posterior distribution $p(\theta|\mathcal{D})$ represents our belief or knowledge of θ after the data has been taken into account.

θ can be as simple as a mean or proportion...

...or it can be an arbitrarily large set of parameters for a complex model.

Statistical inference via Bayes rule

Once we have a summary of $P(\theta|\mathcal{D})$, we can ask questions like:

What is the probability that $\mu > 0$?

What is the probability that $\mu_1 > \mu_2$ for groups 1 and 2?

What is the probability that the slope of a linear regression $\beta > 0$?

Maximum a posteriori (MAP) estimation

Recall Bayes rule:

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})}$$

Maximum a posteriori (MAP) estimation

Recall Bayes rule:

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})}$$

Maximum a posteriori (MAP) estimation maximizes $P(\theta|\mathcal{D}) \propto P(\mathcal{D}|\theta)P(\theta)$ w.r.t θ.

Maximum a posteriori (MAP) estimation

Recall Bayes rule:

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})}$$

Maximum a posteriori (MAP) estimation maximizes $P(\theta|\mathcal{D}) \propto P(\mathcal{D}|\theta)P(\theta)$ w.r.t θ.

Or equivalently, minimizes:

$$\mathcal{L}_{\mathcal{D}}(\theta) = -\log P(\mathcal{D}|\theta) - \log P(\theta)$$

Maximum a posteriori (MAP) estimation

Recall Bayes rule:

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})}$$

Maximum a posteriori (MAP) estimation maximizes $P(\theta|\mathcal{D}) \propto P(\mathcal{D}|\theta)P(\theta)$ w.r.t θ.

Or equivalently, minimizes:

$$\mathcal{L}_{\mathcal{D}}(\theta) = -\log P(\mathcal{D}|\theta) - \log P(\theta)$$

Notice that maximum likelihood estimation (MLE) is just a special case of MAP.

Estimating the posterior distribution

MAP provides a point estimate of θ.

Estimating the posterior distribution

MAP provides a point estimate of θ.

As Bayesians, we are interested in estimating the full distribution $P(\theta|\mathcal{D})$.

Estimating the posterior distribution

MAP provides a point estimate of θ.

As Bayesians, we are interested in estimating the full distribution $P(\theta|\mathcal{D})$.

Unfortunately, that's pretty hard!

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{\int_\theta P(\mathcal{D}|\theta)P(\theta)d\theta}$$

The integral on the bottom is typically intractable for non-trivial models.

Markov Chain Monte Carlo

Enter... Markov Chain Monte Carlo! (MCMC)

Markov Chain Monte Carlo

Enter... Markov Chain Monte Carlo! (MCMC)

Basic idea: Sample numerically via a **random walk** in the space of θ.

Markov Chain Monte Carlo

Enter... Markov Chain Monte Carlo! (MCMC)

Basic idea: Sample numerically via a **random walk** in the space of θ.

In the limit of infinite steps, the path of the walk will follow $P(\theta|\mathcal{D})$.

Markov Chain Monte Carlo

Enter... Markov Chain Monte Carlo! (MCMC)

Basic idea: Sample numerically via a **random walk** in the space of θ.

In the limit of infinite steps, the path of the walk will follow $P(\theta|\mathcal{D})$.

Unfortunately, this is often quite inefficient.

Markov Chain Monte Carlo

Enter... Markov Chain Monte Carlo! (MCMC)

Basic idea: Sample numerically via a **random walk** in the space of θ.

In the limit of infinite steps, the path of the walk will follow $P(\theta|\mathcal{D})$.

Unfortunately, this is often quite inefficient.

Modern state-of-the-art samplers are typically variants **Hamiltonian Monte Carlo** (HMC) which require gradients.

Hamiltonian Monte Carlo

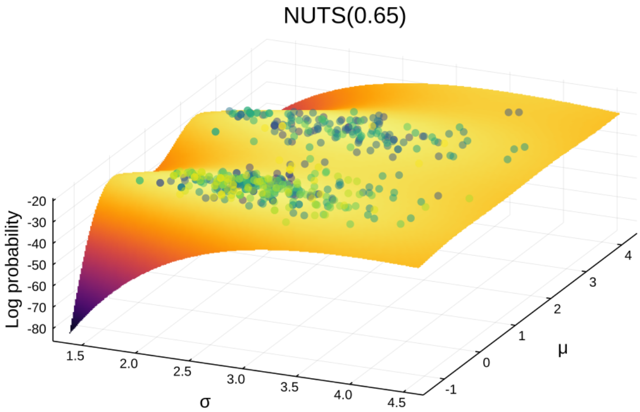


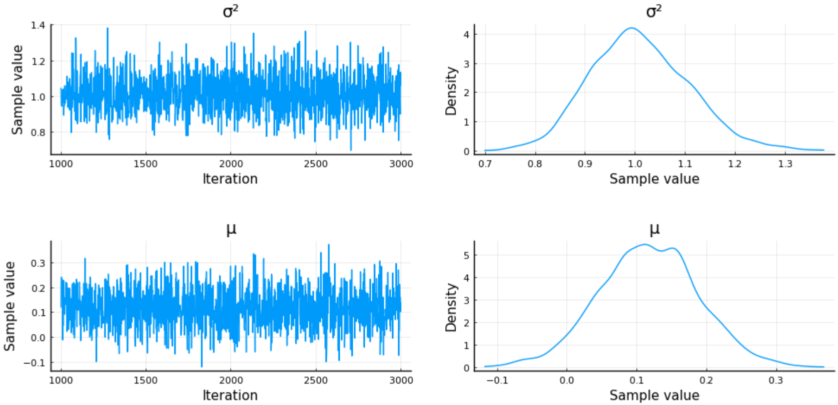
Figure 1: NUTS sampler visualization (source: Turing.jl documentation)

Interpreting the output

The output of MCMC and HMC algorithms is sequence of samples, typically referred to as a **chain** or a **trace**.

Interpreting the output

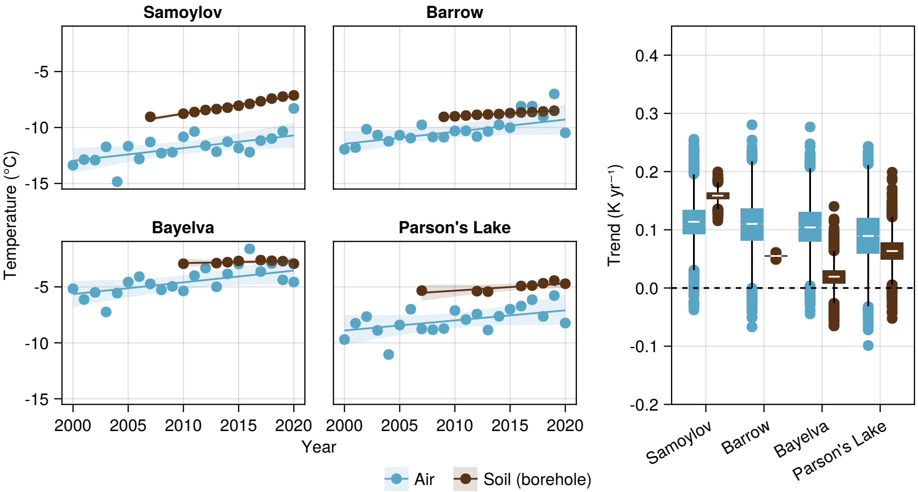
The output of MCMC and HMC algorithms is sequence of samples, typically referred to as a **chain** or a **trace**.



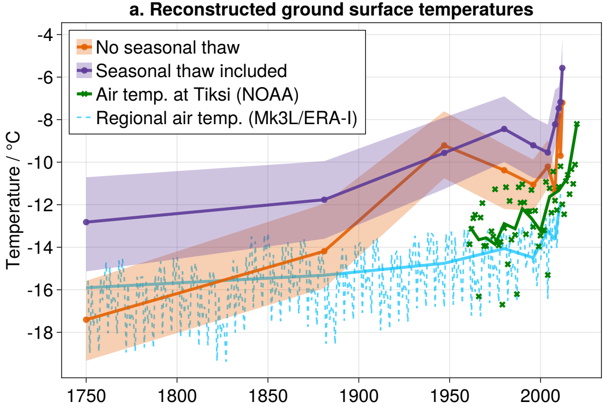
Overview of algorithms for Bayesian inference

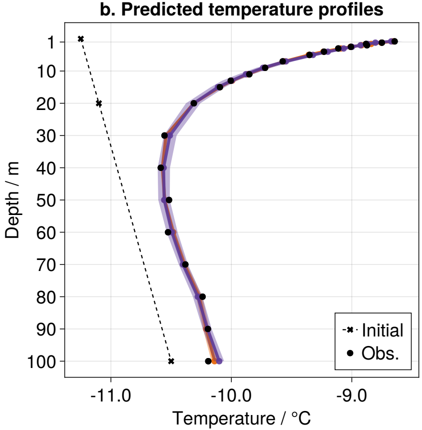
	Forward runs	Gradients required?	Parallelizable?	Asymptotic convergence
MCMC	10^4 to 10^6	No	No	Yes
HMC	10^3 to 10^5	Yes	No	Yes
VI	10^2 to 10^4	Yes	Batch	No
SMC	10^3 to 10^5	No	Yes	Yes
EKS	10^3 to 10^4	No	Yes	Gaussian
IS	10^1 to 10^3	No	Yes	No

Use case: Multi-site trend analysis



Use case: Surface temperature reconstruction





Limitations and challenges

Bayesian inference provides a flexible and intuitive framework for data analysis and generative modeling... **but**

Limitations and challenges

Bayesian inference provides a flexible and intuitive framework for data analysis and generative modeling... **but**

- Choosing good priors can be difficult.

Limitations and challenges

Bayesian inference provides a flexible and intuitive framework for data analysis and generative modeling... **but**

- Choosing good priors can be difficult.
- MCMC is computationally expensive, especially with large numbers of parameters.

Limitations and challenges

Bayesian inference provides a flexible and intuitive framework for data analysis and generative modeling… **but**

- Choosing good priors can be difficult.
- MCMC is computationally expensive, especially with large numbers of parameters.
- Practical analysis of the inference results can be time consuming.

Limitations and challenges

Bayesian inference provides a flexible and intuitive framework for data analysis and generative modeling… **but**

- Choosing good priors can be difficult.
- MCMC is computationally expensive, especially with large numbers of parameters.
- Practical analysis of the inference results can be time consuming.
- Software support is not as comprehensive as for “classical” frequentist analyses.

Coming up next. . .

- Probabilistic programming

Coming up next. . .

- Probabilistic programming
- Bayesian generalized linear models

Coming up next. . .

- Probabilistic programming
- Bayesian generalized linear models
- Model diagnostics and criticism

Coming up next. . .

- Probabilistic programming
- Bayesian generalized linear models
- Model diagnostics and criticism
- Example application to observational data from the Arctic

Table of Contents

Follow along!

Part II of the workshop consists of a series of Jupyter notebooks.

You can directly open these notebooks on Google Colab if you do not want to bother setting up a new python environment!