1 Introduction

- measurement inherently random, variability
- Assume a latent data generating process

$$\pi(\mathcal{D})$$

- There is a distribution over the measurement.
- Space of all distribution $\mathcal{M}$ including pathological ones that may not be relevant
- For practicality, limit ourselves $\mathcal{S}$ small world. The true distribution may fall under here. But this may not be sophisticated enough
- After choosing the small world $\mathcal{S}$, $\theta \to \pi_{\mathcal{S}}(\mathcal{D}|\theta)$. likelihood.
- Measurement $\tilde{\mathcal{D}}$. Two different objects.
- Bayesian inference: probability is used to quantify consistency of the model with data. $\pi_{\mathcal{S}}(\theta|\tilde{\mathcal{D}})$. Density over $\mathcal{S}$.

$$\pi_{\mathcal{S}}(\theta|\tilde{\mathcal{D}}) = \frac{\pi_{\mathcal{S}}(\mathcal{D}|\theta)\pi_{\mathcal{S}}(\theta)}{\pi_{\mathcal{S}}(\mathcal{D})}$$

- Normalization constant. marginalized likelihood. Called evidence in physics. Not necessary.
- Prior tells us what we need before measurement. *a priori* soft weighting. some parts of the parameter space is more likely
- Likelihood is the weighting function based on the measurement.
- Posterior concentrates more. As data increase,
- Updating procedure.
- One operation: execration
- All inferential questions are answered by posterior expectations.
- Mean: centrality μ
- Variance: diffuse σ^2
- Utility $U(A)$

$$\mu = \int \mathrm{d}\theta \pi_{\mathcal{S}}(\theta|\tilde{\mathcal{D}})$$
$$\sigma^2 = \int \mathrm{d}\theta \pi_{\mathcal{S}}(\theta^2|\tilde{\mathcal{D}}) - \mu^2$$

- validation: predictive model comparison
- averaging prediction. posterior predictive distribution.

$$\pi_{\mathcal{S}}(\mathcal{D}|\tilde{\mathcal{D}}) = \int \mathrm{d}\theta \pi_{\mathcal{S}}(\mathcal{D}|\theta)\pi_{\mathcal{S}}(\theta|\tilde{\mathcal{D}})$$

- Comparing $\pi_{\mathcal{S}}(\mathcal{D}|\tilde{\mathcal{D}})$ vs $\pi(\mathcal{D})$ (truth; not possible)
- Comparing $\pi_{\mathcal{S}}(\mathcal{D}|\tilde{\mathcal{D}})$ vs $\tilde{\mathcal{D}}$ (estimator)
- misfit
- Overfitting: difference in ML vs Statisticians
- Chasing the noise. pathological behavior
- Two-step posterior predictive distribution

$$\theta \sim \pi_{\mathcal{S}}(\theta|\tilde{\mathcal{D}})$$
$$\mathcal{D} \sim \pi_{\mathcal{S}}(\tilde{\mathcal{D}}|\theta)$$

- Visual inspection: need low dimensional summary of the data that is more sensitive to abnormality

2 Stan

- facilitates Bayesian inference with a modeling language and state-of-the-art computational methods
- Modeling language
- Automatic differentiation (abstraction)
- Hamiltonian Monte Carlo

2.1 Modeling language

- Posterior requires three inputs
- Data, parameter, posterior density relating former two.
- Joint density of data and parameter $\pi(\mathcal{D}, \theta)$
- log is easier to make it summing

$$\log \pi(\theta|\mathcal{D}) = \sum_n \log \pi(\mathcal{D}_n)$$

```
data {
    int<lower=1> N;
    real x[N]
}


model {
    /* ~ adds stuff to  */
    beta ~ normal(0,1);
}
```

- `beta ~ normal(0,1)` adds $\log \pi += \log \mathcal{N}(\beta|0,1)$
- `alpha ~ normal(0,1)` adds $\log \pi += \log \mathcal{N}(\alpha|0,1)$
- Blocks: functions, data, transformed data, parameters, transformed parameters, model, generated quantities (after sampling; deterministic calculation or RNG)
- Strong static typing
- int, real
- matrix[M,N], vector[M], row_vector[N]
- bounded

- Constrained vectors: simplex (sum to one), ordered[N] ()
- Constrained matrices: cov_matrix[K]
- ordinary differential equations
- vectorization

```
/*  */
for (n in 1:N)
    y[n] ~ normal(mu[n], sigma[n]);

/* vectorized faster */
y ~ normal(mu, sigma);
```

- Hamiltonian Monte Carlo
- $\mathbb{E}[f] = \int \mathrm{d}\theta \pi(\theta|\mathcal{D}) f(\theta)$
- Diagnostics for Hamiltonian
- HMC requires strong geometric ergodicity conditions to ensure trustworthy results

$$\frac{1}{N}\sum_{n=1}^{N} f(\theta_n)\ \mathcal{N}\left(\mathbb{E}[f],\ \right)$$

- Rhat Rubin statistics (multiple chains). Have they seen the same space. < 1.3 ok.
- stan_utility.check_rhat(fit)
- divergence if encountering a hard to explore region.
- stan_utility.check_div(sampler_parameers): prop divergence.

2.2 Good practice

- maintain reproducibility by saving the model, data, and inits in files and the Python commands in scripts
- version control on module files and scripts
- Start simple (easier to diagnose). the last attempt is the culprit if sequentially done.
- Simulate data and fit on it. Workflow validation.
- Keep an eye on those diagnostics!
- documentation tutorials and case studies.

2.3 links

- `https://betanalpha.github.io`
- Material: `https://betanalpha.github.io/workshops/computefest/material.zip`

3 Exercise 1

- Zero-inflated truncated Poisson distribution.
- Any one dimensional distribution can be truncated.

4 Regression modeling

- Model building. modular
- Likelihood and prior should be defined together.

$$\pi(\theta|\mathcal{D}) \propto \pi(\mathcal{D}|\theta)\pi(\theta)$$

- Four components to data, can be decomposed. and simplify (sequence of data of distribution). how the data was collected.
- similar for prior
- break down to smaller pieces with less dependencies
- dependent variable (more variability), covariates x (more easily measured). asymmetry.

$$\pi(y, x|\theta) = \pi(y|x, \theta)\pi(x|\theta)$$

- Assumption $\pi(x|\theta) = \pi(x)$. no selection bias. Assuming no relevance of x.

$$\pi(y, x|\theta) \propto \pi(y|x, \theta)$$

- statistical model

$$\pi(y|x, \theta) = \pi(y|f(x, \theta_1), \theta_2)$$

- separate x and y

$$\pi(y|x, \theta) = \mathcal{N}(y|f(x, \theta), \sigma)$$
$$\pi(y|x, \theta) = \text{Binomial}(y|f(x, \theta), N)$$

- two parameter family

$$\pi(y|x, \theta) = \pi(y|f_1(x, \theta_1), f_2(x, \theta_2), \theta_3)$$

- linear function of x and intercept

$$f(x, \alpha, \boldsymbol{\beta}) = \mathbf{X}^T\boldsymbol{\beta} + \alpha$$

- linear regression

$$\pi(y|\mathbf{X}, \alpha, \boldsymbol{\beta}, \sigma) = \mathcal{N}(y|\mathbf{X}^T\boldsymbol{\beta} + \alpha, \sigma)$$

- less data and parameters. likelihood will be non-identifiable. collinear model. plane of possibility.

- weakly informative priors to prevent unreasonable values. Do not set a hard bound. soft bound.
- wrong scaling. reason about the unit. Both the data and prior should be on the same unit for

$$\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Omega})$$
$$\beta_i \sim \mathcal{N}(0, \omega_i)$$

- scale is important
- half normal for σ
- generalized linear models
- g is inverse link function
- linear response and
- logistic

$$\text{Ber}(y|\text{logistic}(\mathbf{X}^T\boldsymbol{\beta} + \alpha))$$

- Poisson

$$\text{Poisson}(y|\exp(\mathbf{X}^T\boldsymbol{\beta} + \alpha))$$

-