STASH

ML Paper Group - March 15, 2019

[Bayesianism and causality, or, why I am only half-Bayesian]

Intro / Motivation

- Arguments in favor of the Bayesian approach:
  - It is plain silly to ignore what we know
  - It is natural and useful to cast what we know in the language of probabilities
  - If our subjective probabilities are erroneous, their impact will get washed out in due time, as the number of observations increases.

- Combining human knowledge with empirical data is "the key to scientific enquiry and intelligent behavior" but in order to do this we need to be able to cast them both in a formal language

- Most human knowledge is expressed through causal relationships, not probabilistic ones
  - causal: "Rain causes mud"
  - probabilistic: "When there is rain there is mud" (or "When there is mud there is rain")

Bayesian vs Frequentist Regressions

- Frequentist data analysis evaluates the likelihood function *P(D|H)*, which represents the probability of the *data* given the *hypothesis*. Bayesian analysis evaluate the inverse of the likelihood function *P(H|D)*, which is the probability of the *hypothesis* given the data. We can transform *P(D|H)* to *P(H|D)* using the rules of conditional probability.

- In Frequentist regressions (traditional regression approach), we assume that parameters (betas) are normally distributed.
    - This view comes from a Physics experimentation view, where after enough trials of an experiment, we converge on the mean of the true parameters.

- In Bayesian regressions, we can make different assumptions on the distribution of the parameters; these beliefs are called priors.
    - Often, a good choice for econometrics, where we are definitely not dealing with repeatable experiments
    - Examples of priors: normal prior, flat prior
    - We then derive (or simulate in the case there is no analytical solution) the posterior distributions of the parameters using Bayes' Theorem.
    - Each prior type has a corresponding conjugate likelihood formula

Bayesian Inference

Define:

- x as an individual data point
- θ as a parameter in the data points distribution, e.g $x \sim p(x \mid \theta)$
- α as a hyperparameter of the parameter distribution, e.g. $\theta \sim p(\theta \mid \alpha)$
- $\mathbf{X}$ is a sample of data points

The posterior distribution can be described as:

$$p(\theta \mid \mathbf{X}, \alpha) = \frac{p(\theta, \mathbf{X}, \alpha)}{p(\mathbf{X}, \alpha)} = \frac{p(\mathbf{X} \mid \theta, \alpha)p(\theta, \alpha)}{p(\mathbf{X} \mid \alpha)p(\alpha)} = \frac{p(\mathbf{X} \mid \theta, \alpha)p(\theta \mid \alpha)}{p(\mathbf{X} \mid \alpha)} \propto p(\mathbf{X} \mid \theta, \alpha)p(\theta \mid \alpha)$$

Example Derivation

$$(\mathbf{y}-\mathbf{X}\boldsymbol\beta)^{\mathrm T}(\mathbf{y}-\mathbf{X}\boldsymbol\beta)=(\mathbf{y}-\mathbf{X}\hat{\boldsymbol\beta})^{\mathrm T}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol\beta})+(\boldsymbol\beta-\hat{\boldsymbol\beta})^{\mathrm T}(\mathbf{X}^{\mathrm T}\mathbf{X})(\boldsymbol\beta-\hat{\boldsymbol\beta}).$$

← Algebraic formation of the linear equation

$$\rho(\mathbf{y}|\mathbf{X},\boldsymbol\beta,\sigma^2)\propto(\sigma^2)^{-\frac{v}{2}}\exp\left(-\frac{vs^2}{2\sigma^2}\right)(\sigma^2)^{-\frac{n-v}{2}}\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol\beta-\hat{\boldsymbol\beta})^{\mathrm T}(\mathbf{X}^{\mathrm T}\mathbf{X})(\boldsymbol\beta-\hat{\boldsymbol\beta})\right),$$

← Prior is normal

$$vs^2=(\mathbf{y}-\mathbf{X}\hat{\boldsymbol\beta})^{\mathrm T}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol\beta})\quad\text{and}\quad v=n-k,$$

$$\rho(\boldsymbol\beta,\sigma^2)=\rho(\sigma^2)\rho(\boldsymbol\beta|\sigma^2),$$

$$\rho(\sigma^2)\propto(\sigma^2)^{-\frac{v_0}{2}-1}\exp\left(-\frac{v_0s_0^2}{2\sigma^2}\right).$$

← Posterior is gamma

$$\begin{aligned}\rho(\boldsymbol\beta,\sigma^2|\mathbf{y},\mathbf{X})&\propto\rho(\mathbf{y}|\mathbf{X},\boldsymbol\beta,\sigma^2)\rho(\boldsymbol\beta|\sigma^2)\rho(\sigma^2)\\&\propto(\sigma^2)^{-\frac{n}{2}}\exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol\beta)^{\mathrm T}(\mathbf{y}-\mathbf{X}\boldsymbol\beta)\right)(\sigma^2)^{-\frac{k}{2}}\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol\beta-\boldsymbol\mu_0)^{\mathrm T}\boldsymbol\Lambda_0(\boldsymbol\beta-\boldsymbol\mu_0)\right)(\sigma^2)^{-(a_0+1)}\exp\left(-\frac{b_0}{\sigma^2}\right)\end{aligned}$$

← Rewrite equation

$$\boldsymbol\mu_n=(\mathbf{X}^{\mathrm T}\mathbf{X}+\boldsymbol\Lambda_0)^{-1}(\mathbf{X}^{\mathrm T}\mathbf{X}\hat{\boldsymbol\beta}+\boldsymbol\Lambda_0\boldsymbol\mu_0).$$

← Mean of the posterior distribution

$$(\mathbf{y}-\mathbf{X}\boldsymbol\beta)^{\mathrm T}(\mathbf{y}-\mathbf{X}\boldsymbol\beta)+(\boldsymbol\beta-\boldsymbol\mu_0)^{\mathrm T}\boldsymbol\Lambda_0(\boldsymbol\beta-\boldsymbol\mu_0)=(\boldsymbol\beta-\boldsymbol\mu_n)^{\mathrm T}(\mathbf{X}^{\mathrm T}\mathbf{X}+\boldsymbol\Lambda_0)(\boldsymbol\beta-\boldsymbol\mu_n)+\mathbf{y}^{\mathrm T}\mathbf{y}-\boldsymbol\mu_n^{\mathrm T}(\mathbf{X}^{\mathrm T}\mathbf{X}+\boldsymbol\Lambda_0)\boldsymbol\mu_n+\boldsymbol\mu_0^{\mathrm T}\boldsymbol\Lambda_0\boldsymbol\mu_0.$$

← Rewrite in quadratic expansion

Continued Derivation

$$\rho(\boldsymbol{\beta},\sigma^2|\mathbf{y},\mathbf{X}) \propto (\sigma^2)^{-\frac{k}{2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{\mu}_n)^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X}+\boldsymbol{\Lambda}_0)(\boldsymbol{\beta}-\boldsymbol{\mu}_n)\right)(\sigma^2)^{-\frac{n+2a_0}{2}-1}\exp\left(-\frac{2b_0+\mathbf{y}^{\mathsf{T}}\mathbf{y}-\boldsymbol{\mu}_n^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X}+\boldsymbol{\Lambda}_0)\boldsymbol{\mu}_n+\boldsymbol{\mu}_0^{\mathsf{T}}\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0}{2\sigma^2}\right).$$

← Rewrite as a normal * gamma

$$\boldsymbol{\Lambda}_n = (\mathbf{X}^{\mathsf{T}}\mathbf{X}+\boldsymbol{\Lambda}_0), \quad \boldsymbol{\mu}_n = (\boldsymbol{\Lambda}_n)^{-1}(\mathbf{X}^{\mathsf{T}}\mathbf{X}\hat{\boldsymbol{\beta}}+\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0),$$

$$a_n = a_0 + \frac{n}{2}, \qquad b_n = b_0 + \frac{1}{2}(\mathbf{y}^{\mathsf{T}}\mathbf{y}+\boldsymbol{\mu}_0^{\mathsf{T}}\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0-\boldsymbol{\mu}_n^{\mathsf{T}}\boldsymbol{\Lambda}_n\boldsymbol{\mu}_n).$$

← Now, we have formulas for our parameters

That was cool, but sometimes there’s no analytical solution

- There are a whole suite of simulation and sampling methods to employ to obtain the marginal distribution of the prior and posteriors. Often, they use Gibbs Sampling with Monte Carlo Markov Chains

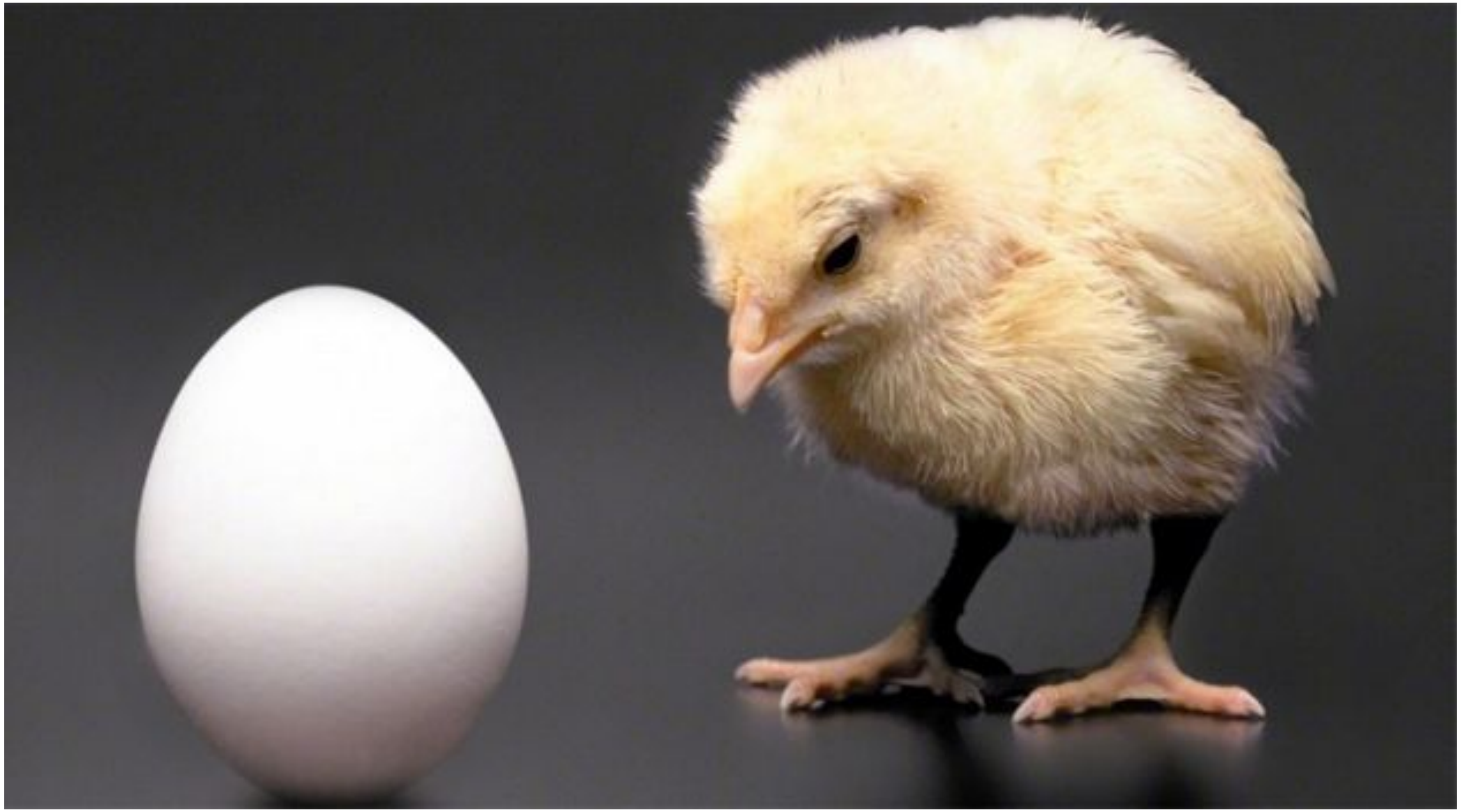
- Gibbs sampling is an iterative algorithm that produces samples from the posterior distribution of each parameter of interest.
- To use Gibbs, we need to identify the conditional posterior of each parameter:

$$p(\beta_0, \beta_1, \phi|\vec{y}) \propto \phi^{-n/2} e^{-\frac{1}{2\phi}\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1 x_i))^2} e^{-\frac{1}{2\tau_0}(\beta_0-\mu_0)^2} e^{-\frac{1}{2\tau_1}(\beta_1-\mu_1)^2} \phi^{-(\alpha+1)} e^{-\frac{\gamma}{\phi}}$$

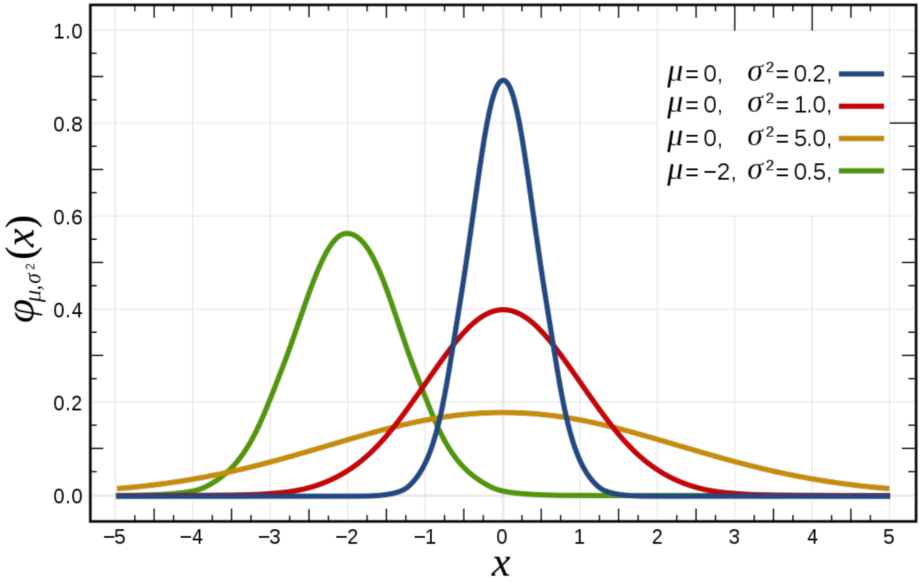
$$p(\beta_0|\phi, \beta_1, \mu_0, \tau_0, \vec{y}) \propto e^{-\frac{1}{2\phi}\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1 x_i))^2} e^{-\frac{1}{2\tau_0}(\beta_0-\mu_0)^2}$$

$$p(\beta_1|\beta_0, \phi, \mu_1, \tau_1, \vec{y}) \propto e^{-\frac{1}{2\phi}\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1 x_i))^2} e^{-\frac{1}{2\tau_1}(\beta_1-\mu_1)^2}$$

statistics & causality



statistics



goal

- infer **parameters of a distribution** from samples

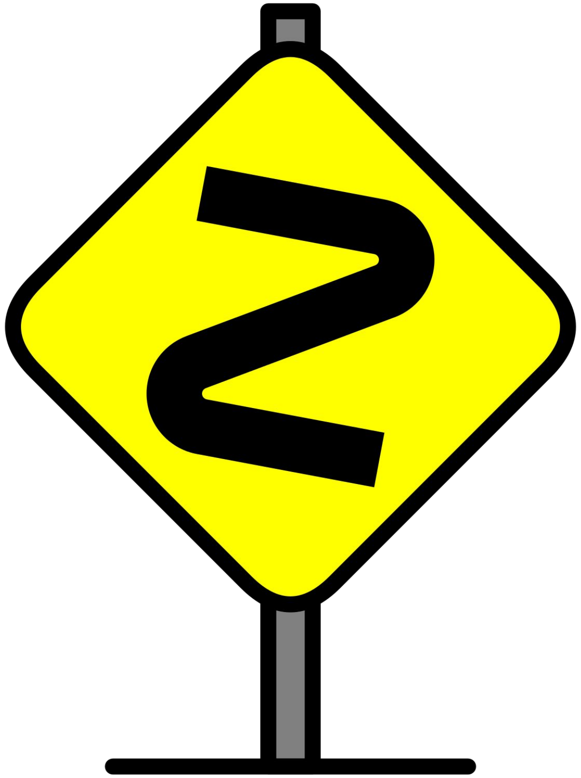
examples

- correlation, regression, distributions

notes

- deals with uncertainty under **static** conditions
- claims are **quantitative**
- describes concepts in terms of **distributions**

causality



goal

- infer aspects of **data generation process**

examples

- attribution, randomization, counterfactuals

notes

- deals with uncertainty under **changing** conditions
- claims are **qualitative**
- describes concepts **not defined** by distributions

obstacles to causal logic

A causal statement is **harder to test** than a probabilistic statement.

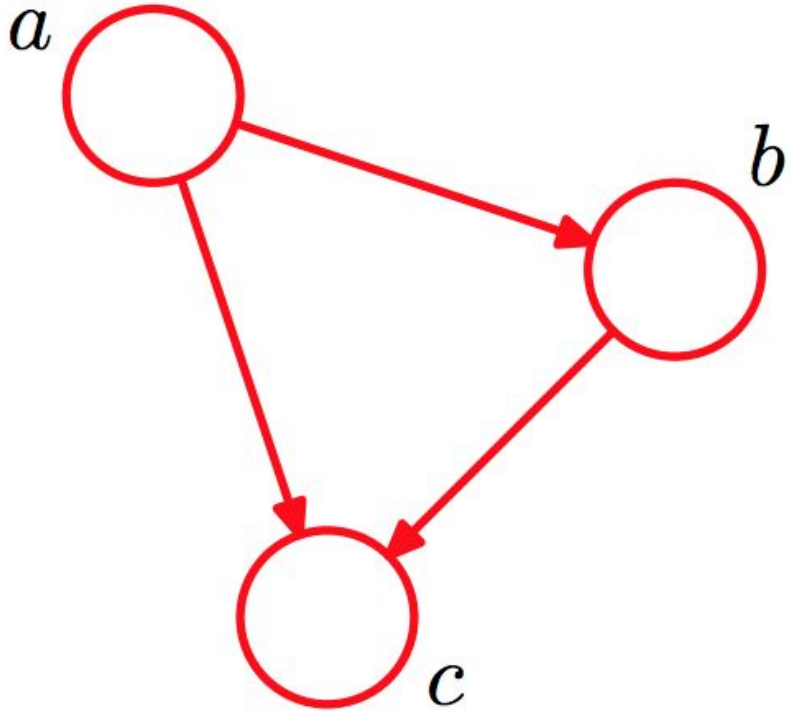
- requires an experiment

By definition, probabilistic language is **insufficient** to discuss issues of causality.

→ This requires a **new vocabulary**!

“The need to adopt a new notation, foreign to the province of probability theory, has been ***traumatic*** *to most persons trained in statistics..."*

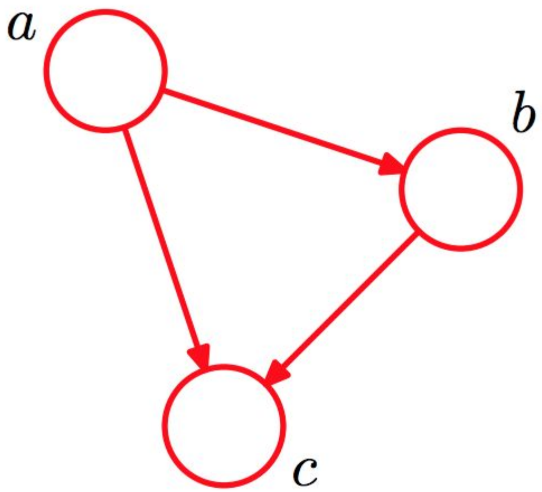
the language of causality



the language of causality

“How do you express mathematically the common understanding that symptoms do not cause disease?"

step 1: **DAG**



notes

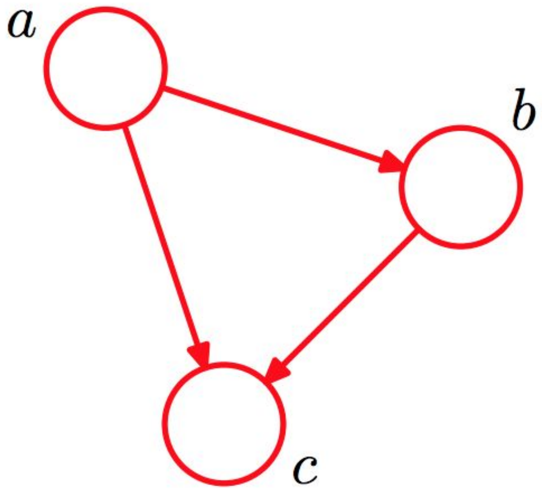
- **non-algebraic** object
- causal structure represented by graph **topology**
- **no cycles**; these would be logical fallacies

the language of causality

step 2: **DAG + structural equations**

$$\boxed{\begin{aligned} a &= f(u) \\ b &= g(a, v) \\ c &= h(a, b, w) \end{aligned}}$$

a, b, c	**endogenous** factors
u, v, w	**exogenous** factors
f, g, h	functions



notes:

- *b* is **influenced** by *a*
- *c* is **influenced** by *a, b*
- exogenous factors *u, v, w* **jointly independent**
- dependency relationships (non-causal) can be expressed through probabilities
- jumping off point for **hierarchical models**

  (→ stay tuned!)

summary

There is a **basic conflict** between the mission & practice of Bayesian methodology.

- mission: fuse judgment with data
- practice: use vocabulary of probability

*"...**too crude** a vocabulary, given the grand mission..."*

*"...**causality** deals with how probability functions change in response to influences...that originate from **outside** the probability space..."*

*"The grounds are now ready for **mission-oriented Bayesianism**."*