Graphical Models

Jacob Steinhardt

September 17, 2020

Last Time

Bayesian Inference

- Setup
- Conjugate priors
- Computing posteriors
- Inference
  - Full posterior
  - MAP, LMSE

This time: more complex models, and a (visual) language for describing them

Recall: Heights and Gender

[Jupyter demo]

Heights and Gender: Bayesian Model

- Person i: gender $z_i \in \{0, 1\}$, height $x_i \in \mathbb{R}$

Heights and Gender: Bayesian Model

- Person i: gender $z_i \in \{0, 1\}$, height $x_i \in \mathbb{R}$
- $x_i \mid z_i \sim N(\mu_{z_i}, \sigma^2)$, i.e. $p(x_i \mid z_i) \propto \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu_{z_i})^2\right)$

Heights and Gender: Bayesian Model

- Person i: gender $z_i \in \{0, 1\}$, height $x_i \in \mathbb{R}$
- $x_i \mid z_i \sim N(\mu_{z_i}, \sigma^2)$, i.e. $p(x_i \mid z_i) \propto \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu_{z_i})^2\right)$
- $p(z_i) = \pi^{z_i}(1-\pi)^{1-z_i}$ (Bernoulli with probability π)

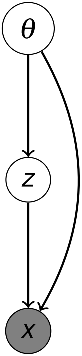
Heights and Gender: Bayesian Model

- Person i: gender $z_i \in \{0, 1\}$, height $x_i \in \mathbb{R}$
- $x_i \mid z_i \sim N(\mu_{z_i}, \sigma^2)$, i.e. $p(x_i \mid z_i) \propto \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu_{z_i})^2\right)$
- $p(z_i) = \pi^{z_i}(1-\pi)^{1-z_i}$ (Bernoulli with probability π)

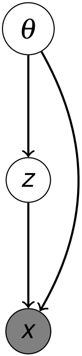
“Hyperparameters”: $\mu_0, \mu_1, \sigma^2, \pi$

[draw graphical model]

Latent Variable Model: General Form



Latent Variable Model: General Form

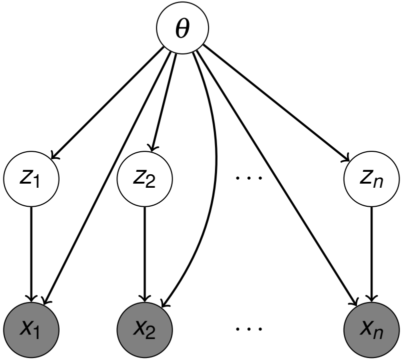


hyperparameters $(\mu_1, \mu_2, \sigma, \pi)$

latent structure (gender)

observed output (height)

Special Case: Hierarchical Model

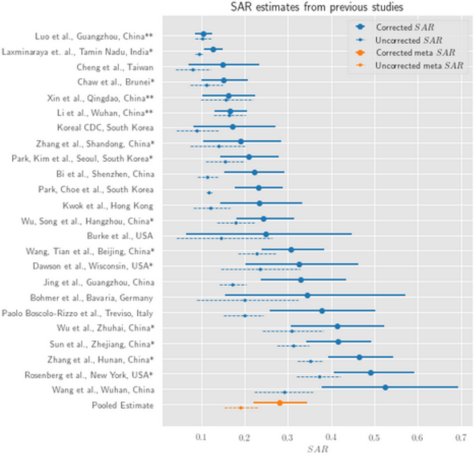


“Bayesian hierarchical model”

Example: COVID Meta-Analysis

[on board]

Example: COVID Meta-Analysis



Take-away: hierarchical models help model heterogeneity while pooling statistical strength

Another Example: HMMs

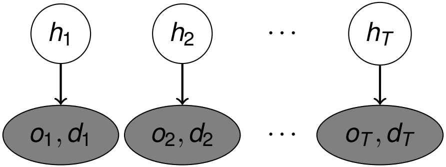
Temperature from ice cores

- Ice cores collected at times $1, \ldots, T$
- Observe ^{18}O and deuterium concentrations at each time (O_t, D_t)
- Known (noisy) relationship with temperature H_t: $O_t \approx aH_t + b$, and $D_t \approx cH_t + d$

Another Example: HMMs

Temperature from ice cores

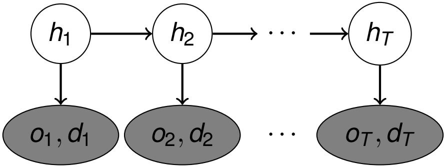
- Ice cores collected at times $1, \ldots, T$
- Observe ^{18}O and deuterium concentrations at each time (O_t, D_t)
- Known (noisy) relationship with temperature H_t: $O_t \approx aH_t + b$, and $D_t \approx cH_t + d$



Another Example: HMMs

Temperature from ice cores

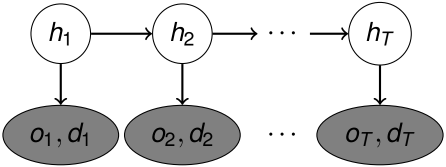
- Ice cores collected at times $1, \dots, T$
- Observe ^{18}O and deuterium concentrations at each time (O_t, D_t)
- Known (noisy) relationship with temperature H_t: $O_t \approx aH_t + b$, and $D_t \approx cH_t + d$



Another Example: HMMs

Temperature from ice cores

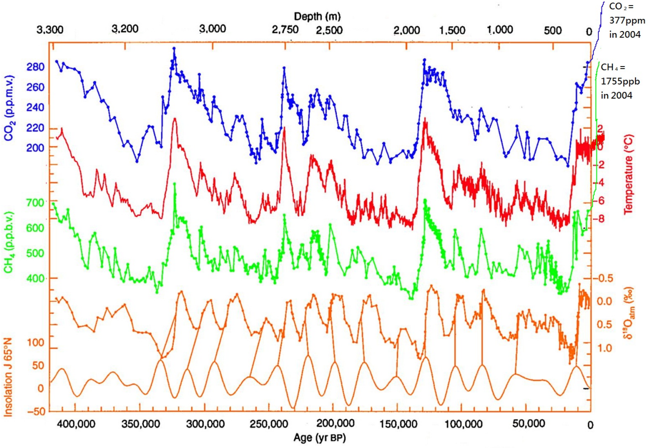
- Ice cores collected at times $1, \ldots, T$
- Observe ^{18}O and deuterium concentrations at each time (O_t, D_t)
- Known (noisy) relationship with temperature H_t: $O_t \approx aH_t + b$, and $D_t \approx cH_t + d$



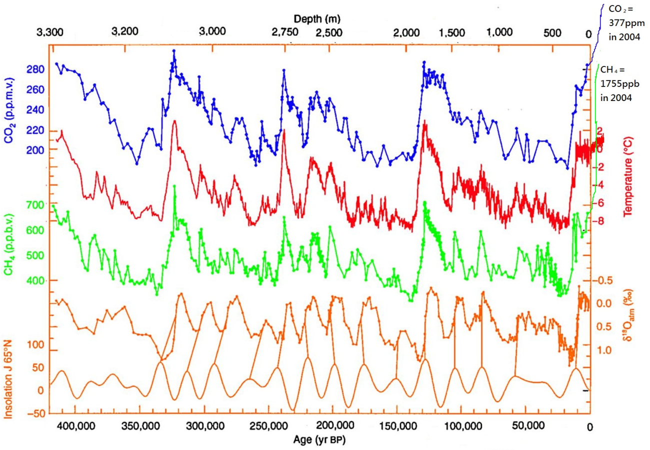
- Model:

$$d_t \sim N(ch_t + d, \sigma_d^2), \quad h_{t+1} \sim N(h_t, \sigma_h^2), \quad h_0 \sim N(\mu, \sigma_0^2)$$

Checking the assumptions



Checking the assumptions

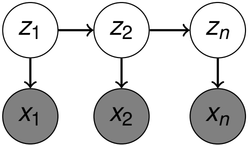


Take-away: Time series models can pool across time, but need to get dynamics right!

Interlude: Factorization

- Graphical models directly relate to **algebraic** structure of probability distribution
- HMM:

$$p(z_1, z_2, z_3, x_1, x_2, x_3) = p(z_1)p(z_2 \mid z_1)p(z_3 \mid z_2) \\ \times p(x_1 \mid z_1)p(x_2 \mid z_2)p(x_3 \mid z_3)$$



- Parents in graphical model $\leftrightarrow$ what to condition on in factorization

Final Example: Election Forecasting

2016 election forecasting

- Want to know fraction of people who will vote for Clinton in each state

Final Example: Election Forecasting

2016 election forecasting

- Want to know fraction of people who will vote for Clinton in each state
- Each of 50 states has some number of polls

Final Example: Election Forecasting

2016 election forecasting

- Want to know fraction of people who will vote for Clinton in each state
- Each of 50 states has some number of polls
- Each poll has large enough sample size that we can treat error as normal-distributed

Final Example: Election Forecasting

2016 election forecasting

- Want to know fraction of people who will vote for Clinton in each state
- Each of 50 states has some number of polls
- Each poll has large enough sample size that we can treat error as normal-distributed
- So have independent Gaussian margin of error in each state

Final Example: Election Forecasting

2016 election forecasting

- Want to know fraction of people who will vote for Clinton in each state
- Each of 50 states has some number of polls
- Each poll has large enough sample size that we can treat error as normal-distributed
- So have independent Gaussian margin of error in each state
- Sample true fraction of Clinton supporters for each state, look at how often Clinton wins

Final Example: Election Forecasting

2016 election forecasting

- Want to know fraction of people who will vote for Clinton in each state
- Each of 50 states has some number of polls
- Each poll has large enough sample size that we can treat error as normal-distributed
- So have independent Gaussian margin of error in each state
- Sample true fraction of Clinton supporters for each state, look at how often Clinton wins

Something like this predicted 90% Clinton in 2016, but Trump won.

Final Example: Election Forecasting

2016 election forecasting

- Want to know fraction of people who will vote for Clinton in each state
- Each of 50 states has some number of polls
- Each poll has large enough sample size that we can treat error as normal-distributed
- So have independent Gaussian margin of error in each state
- Sample true fraction of Clinton supporters for each state, look at how often Clinton wins

Something like this predicted 90% Clinton in 2016, but Trump won.

What is wrong with this analysis? [At least 2 things...]

Election Forecasting Model

[on board]

Election Forecasting Model

[on board]

Next: efficient algorithms

Motivation: Exponential Sums

How to do inference in latent variable models?

- Method 1: place prior on θ, sample $p(\theta, z \mid x)$ (next time)
- Method 2: maximize $\log p(x \mid \theta) = \log\left(\sum_z p(x, z \mid \theta)\right)$
  - "half-Bayesian"

Motivation: Exponential Sums

How to do inference in latent variable models?

- Method 1: place prior on θ, sample $p(\theta, z \mid x)$ (next time)
- Method 2: maximize $\log p(x \mid \theta) = \log\left(\sum_z p(x, z \mid \theta)\right)$
  - "half-Bayesian"

How many possibilities for z? Height/gender example:

Motivation: Exponential Sums

How to do inference in latent variable models?

- Method 1: place prior on θ, sample $p(\theta, z \mid x)$ (next time)
- Method 2: maximize $\log p(x \mid \theta) = \log \left(\sum_z p(x, z \mid \theta) \right)$
    - "half-Bayesian"

How many possibilities for z? Height/gender example:

- 100 people, genders $z_1, \ldots, z_{100}$
- $2^{100} \approx 10^{30}$ possibilities

Motivation: Exponential Sums

How to do inference in latent variable models?

- Method 1: place prior on θ, sample $p(\theta, z \mid x)$ (next time)
- Method 2: maximize $\log p(x \mid \theta) = \log\left(\sum_z p(x, z \mid \theta)\right)$
  - "half-Bayesian"

How many possibilities for z? Height/gender example:

- 100 people, genders $z_1, \ldots, z_{100}$
- $2^{100} \approx 10^{30}$ possibilities

Need a better strategy! (Sampling: next time)

Recap

- Many problems have unobserved structure / dependencies (hierarchical models, hidden Markov models, ...)
- Graphical models: flexible visual language for specifying this structure
- Failing to model these can lead to wrong/overconfident predictions (heterogeneity, time dynamics, independence)
- Latent variables $\implies$ exponential blow-up $\implies$ need good algorithms!