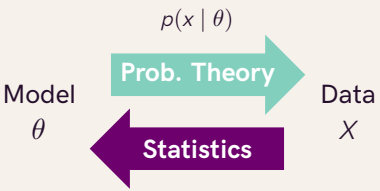
Bayesian Methods

Bayesian methods



"**Frequentist**": Bag of tools to estimate $\hat{\theta}$: MLE, regularized MLE, max entropy, moment matching, ERM, ...

"**Subjective Bayesian**": Use probabilities everywhere there is uncertainty

$$\underbrace{p(\theta \mid \text{data})}_{\text{Posterior}} \propto \underbrace{p(\text{data} \mid \theta)}_{\text{Likelihood}} \underbrace{p(\theta)}_{\text{Prior}}$$

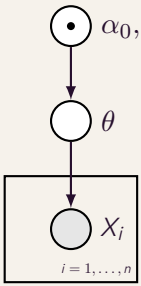
Caricature:

Bayesian is "optimist": they think you can get "good" models $\Rightarrow$ obtain a method by doing inference in a model

Frequentist is "pessimist": they use analysis tools

Example: biased coin

Recall from lecture 4



Hyperparameters for the prior

e.g., $\theta \sim \text{Uniform}([0,1]) = \text{Beta}(1,1)$

$$\theta \mid \alpha_0, \beta_0 \sim \text{Beta}(\alpha_0, \beta_0)$$

$$X_i \mid \theta \overset{iid}{\sim} \text{Bernoulli}(\theta)$$

$$p(x_i \mid \theta) = \theta^{x_i}(1-\theta)^{1-x_i}$$

Posterior:

$$p(\theta \mid x_{1:n}) \propto \left(\prod_{i=1}^{n} p(x_i \mid \theta)\right) p(\theta)$$

$$= \theta^{\overbrace{\sum_i x_i}^{\triangleq n_1}}(1-\theta)^{n-\sum_i x_i}\theta^{\alpha_0-1}(1-\theta)^{\beta_0-1}\mathbb{1}_{[0,1]}(\theta)$$

$$p(\theta \mid x_{1:n}) = \text{Beta}(\theta \mid \alpha_0 + n_1, \beta_0 + n - n_1)$$

conjugate prior to the Bernoulli likelihood model

Conjugate priors

Conjugate family

Consider a family of distributions on θ: $\mathcal{F} = \{p(\theta \mid \alpha) \mid \alpha \in \mathcal{A}\}$

We say that $\mathcal{F}$ is a **conjugate family** to the observation model $p(x \mid \theta)$ if for any $x \sim X \mid \theta$, the posterior $p(\theta \mid x, \alpha) \in \mathcal{F}$.

i.e., there exists some $\alpha'(x, \alpha) \in \mathcal{A}$ s.t. $p(\theta \mid x, \alpha) = p(\theta \mid \alpha')$

Examples

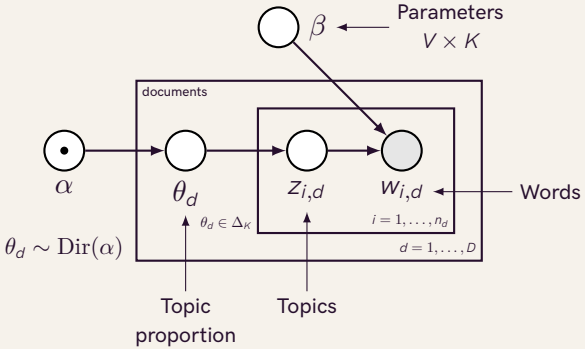
Biased coin (Lecture 4): the Beta prior is conjugate to the Bernoulli likelihood model.

Homework 1: the Dirichlet prior is conjugate to the multinomial likelihood model.

Conjugate priors

Sidenote: if you use a conjugate prior in a DGM, then Gibbs sampling can be easy

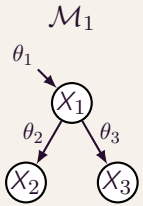
e.g., this is the case in **LDA topic model**



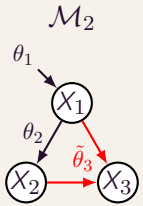
Model Selection

Model selection

Model selection: selecting different hyperparameters, models, etc...
Say we want to choose between 2 DGM



$p(x_3 \mid x_1; \theta_3)$



$p(x_3 \mid x_1, x_2; \tilde{\theta}_3)$

Note here that "$\mathcal{M}_1 \subseteq \mathcal{M}_2$"

As a frequentist

"cavalier notation"

$$\hat{\theta}^{\text{MLE}}_{\mathcal{M}_1} = \arg\max_{\theta_1,\theta_2,\theta_3} \log p(\text{data} \mid \theta_1, \theta_2, \theta_3, \text{“model} = \mathcal{M}_1\text{”})$$

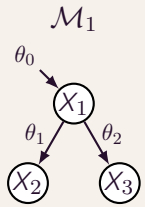
$$\hat{\theta}^{\text{MLE}}_{\mathcal{M}_2} = \arg\max_{\theta_1,\theta_2,\tilde{\theta}_3} \log p(\text{data} \mid \theta_1, \theta_2, \tilde{\theta}_3, \text{“model} = \mathcal{M}_2\text{”})$$

Different space than θ_3

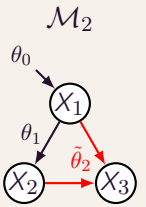
How to choose between models?

Model selection

How to choose between models?



$p(x_3 \mid x_1; \theta_2)$



$p(x_3 \mid x_1, x_2; \tilde{\theta}_2)$

We can't compare

$\log p(\text{data} \mid \hat{\theta}^{\text{MLE}}_{\mathcal{M}_1}, \mathcal{M} = \mathcal{M}_1)$ vs. $\log p(\text{data} \mid \hat{\theta}^{\text{MLE}}_{\mathcal{M}_2}, \mathcal{M} = \mathcal{M}_2)$

because $\text{LHS} \leq \text{RHS}$, since $\mathcal{M}_1 \subseteq \mathcal{M}_2$

i.e. you would always choose the "bigger model"

As a frequentist:

Use cross-validation

Use a validation set

i.e. $\log p(\text{test data} \mid \hat{\theta}^{\text{MLE}}_{\mathcal{M}_i}(\text{train data}), \mathcal{M} = \mathcal{M}_i)$

Bayesian model selection

True Bayesian: sum over models (integrate out uncertainty about $\mathcal{M}$)

Introduce a **prior over models** $p(\mathcal{M})$

$$p(x_{\text{new}} \mid \mathcal{D}) = \sum_{\mathcal{M}} \overbrace{p(x_{\text{new}} \mid \mathcal{D}, \mathcal{M})p(\mathcal{M} \mid \mathcal{D})}^{\text{Standard Bayesian predictive distribution for one model}}$$

$$= \sum_{\mathcal{M}} \left[\int_{\Theta_{\mathcal{M}}} p(x_{\text{new}} \mid \theta, \mathcal{M}) \underbrace{p(\theta \mid \mathcal{D}, \mathcal{M})}_{\text{posterior on } \theta \text{ given data } \mathcal{D} \text{ and model } \mathcal{M}} d\theta \right] \underbrace{p(\mathcal{M} \mid \mathcal{D})}_{\text{model averaging: sum over posterior over models}}$$

How to obtain the posterior $p(\mathcal{M} \mid \mathcal{D})$ or $p(\mathcal{M}, \theta \mid \mathcal{D})$?

variational inference, (RJ)MCMC, etc...

Marginal likelihood

Posterior over models: $p(\mathcal{M} \mid \mathcal{D}) \propto \underbrace{p(\mathcal{D} \mid \mathcal{M})}p(\mathcal{M})$

$$\int_{\Theta_{\mathcal{M}}} \underbrace{p(\mathcal{D} \mid \theta, \mathcal{M})}_{\text{Likelihood}} p(\theta \mid \mathcal{M}) d\theta$$

$p(\mathcal{D} \mid \mathcal{M}) =$ marginal likelihood

How to compute the marginal likelihood

Closed form with parametric assumptions (e.g., using conjugate priors)

Use approximations: variational inference, sampling, etc...

Simple approximation: **Bayesian information criterion**

Empirical Bayes

In model selection, we are forced to pick **one model**

Pick the model that maximizes $p(\mathcal{M} \mid \mathcal{D}) \propto p(\mathcal{D} \mid \mathcal{M})p(\mathcal{M})$

To compare models, look at

$$\frac{p(\mathcal{M} = \mathcal{M}_1 \mid \mathcal{D})}{p(\mathcal{M} = \mathcal{M}_2 \mid \mathcal{D})} = \underbrace{\frac{p(\mathcal{D} \mid \mathcal{M}_1)}{p(\mathcal{D} \mid \mathcal{M}_2)}}_{\text{Bayes factor}} \underbrace{\frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)}}_{\text{Prior ratio}}$$

If "uniform prior" over models, then **pick by** $p(\mathcal{D} \mid \mathcal{M}_i)$ among K models $\mathcal{M}_1, \ldots, \mathcal{M}_K$

When the number of models is "small", this approach is "fine" (i.e., it won't overfit)

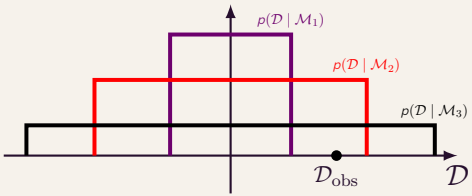
Empirical Bayes

or

Type II Maximum Likelihood

Empirical Bayes

Zoubin's cartoon: suppose "$\mathcal{M}_1 \subseteq \mathcal{M}_2 \subseteq \mathcal{M}_3$"



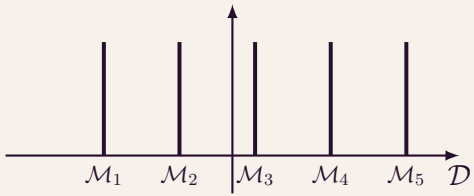
$p(\mathcal{D} \mid \mathcal{M})$ **normalized** over $\mathcal{D}$

vs.

$p(\mathcal{D} \mid \hat{\theta}_{\text{MLE}}(\mathcal{D}), \mathcal{M})$ can **overfit** badly

But type II ML can still overfit if we have too many models

e.g., $p(\mathcal{D} \mid \mathcal{M}) = \delta(\mathcal{D} \mid \mathcal{M})$

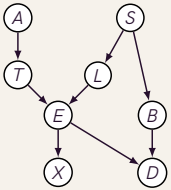


Structure Learning

Structure Learning

	A	B	D	$\dots$	T	X
Sample 1	1.22	-0.12	0.27	$\cdots$	1.09	0.99
Sample 2	-0.04	0.00	0.09	$\cdots$	0.03	-0.47
Sample 3	0.11	0.23	2.23	$\cdots$	-0.07	1.68
$\dots$			$\dots$			

Data

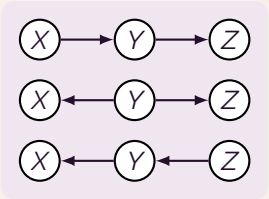


Bayesian Network

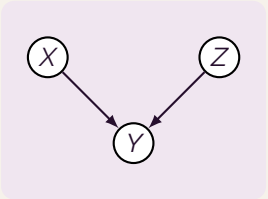
Markov Equivalence

Recall: A Directed Graphical Model encodes the Conditional Independence of a distribution.

Multiple DAGs may encode the same Conditional Independence statements.



$X \not\perp\!\!\!\perp Z$ and $X \perp\!\!\!\perp Z \mid Y$



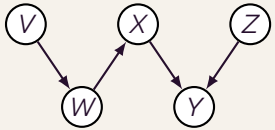
$X \perp\!\!\!\perp Z$ and $X \not\perp\!\!\!\perp Z \mid Y$

Two DAGs encoding the same Conditional Independence statements are called **Markov Equivalent**.

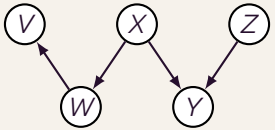
Markov Equivalence

Theorem (Verma & Pearl, 1991)

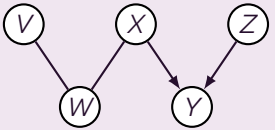
*Two DAGs G_1 and G_2 are **Markov Equivalent** if and only if they have the same skeleton and the same v-structures.*



G_1



G_2



CPDAG

Markov Equivalence Classes can be represented as a **Completed Partially Directed Acyclic Graph** (CPDAG).

Faithfulness

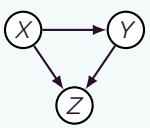
A & B are d-separated by C in $\mathcal{G}$

Global Markov Prop. ⟹

⟸ **Faithfulness**

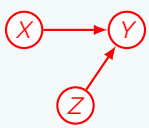
$X_A \perp\!\!\!\perp X_B \mid X_C$

Exercise: Violation of Faithfulness



$$X := N_X$$
$$Y := X + N_Y$$
$$Z := X - Y + N_Z$$
with $N_X, N_Y, N_Z \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

Structure Learning ⟹



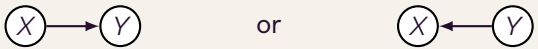
$p(X, Y, Z)$ is a Multivariate Normal distribution, where the only conditional independence statements are: $X \perp\!\!\!\perp Z$ and $X \not\perp\!\!\!\perp Z \mid Y$.

Structure Identifiability

Theorem

*If p is faithful wrt. $\mathcal{G}^0$, then the Markov Equivalence class of $\mathcal{G}^0$ is **identifiable** from p.*

Only the Markov Equivalence class is identifiable from observations, **not an individual graph**. Two Markov Equivalent graphs may lead to different causal conclusions!



Under different assumptions, an individual DAG may be identifiable

- Additive Noise Model (ANM): $X_j := f_j(X_{\mathrm{Pa}_j}) + N_j$, $N_j \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, where f_j are nonlinear.
- Using **interventional data** (i.e. data resulting from controlled experiments).

Constraint-based methods

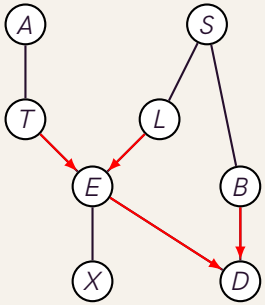
Step 1: Identify the skeleton

For each pair of nodes X & Y, and $\mathbf{A} \subseteq \mathbf{V} \backslash \{X, Y\}$, test if $X \perp\!\!\!\perp_{\mathcal{D}} Y \mid \mathbf{A}$.

If there is no set $\mathbf{A}$ s.t. $X \perp\!\!\!\perp_{\mathcal{D}} Y \mid \mathbf{A}$, then add an edge $X - Y$.

Step 2: Identify the v-structures

For each structure $X - Z - Y$ with no edge between X & Y, orient $X \rightarrow Z \leftarrow Y$ iff $Z \notin \mathbf{A}$, where $\mathbf{A}$ is such that $X \perp\!\!\!\perp_{\mathcal{D}} Y \mid \mathbf{A}$.

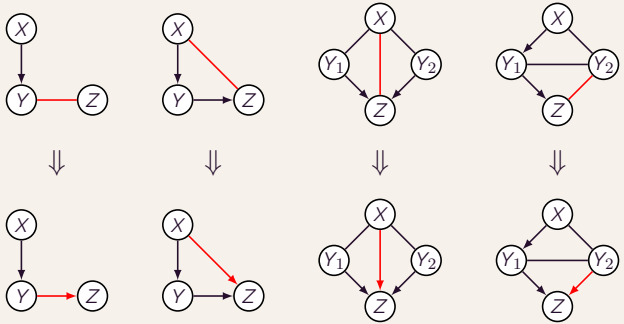


IC Algorithm

Constraint-based methods

Step 2′: Additional orientations

Use **Meek's orientation rules** to orient some of the remaining edges.



Score-based methods

Treat the problem of learning the structure of the DAG as a **model selection problem**

$$\max_{\mathcal{G} \in \text{DAG}} \text{score}(\mathcal{G} \mid \mathcal{D})$$

Recall: choices of scores

Likelihood score:

$$\text{score}_L(\mathcal{G} \mid \mathcal{D}) = \log p(\mathcal{D} \mid \hat{\theta}_{\mathcal{G}}^{\text{MLE}}, \mathcal{G})$$

Bayesian score:

$$\text{score}_B(\mathcal{G} \mid \mathcal{D}) = \log p(\mathcal{D} \mid \mathcal{G}) + \log p(\mathcal{G})$$

Bayesian Information Criterion (BIC):

$$\text{score}_{BIC}(\mathcal{G} \mid \mathcal{D}) = \log p(\mathcal{D} \mid \hat{\theta}_{\mathcal{G}}^{\text{MLE}}, \mathcal{G}) - \frac{\log N}{2} \text{Dim}[\mathcal{G}]$$

Score-based methods

$$\max_{\mathcal{G} \in \text{DAG}} \text{score}(\mathcal{G} \mid \mathcal{D})$$

How to search over the space of DAGs?

The number of DAGs over n nodes is **super-exponential** in n: $2^{\Theta(n^2)}$.

Theorem

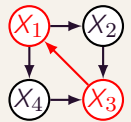
*Let $G_{\leq d} = \{\mathcal{G} \text{ a } DAG \mid \text{every node has at most } d \text{ parents}\}$. Finding a DAG in $G_{\leq d}$ that maximizes a score is **NP-hard** for $d \geq 2$.*

Heuristic solutions:

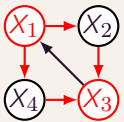
- **Greedy algorithms**: Hill climbing, GES
- **Genetic algorithms**
- **Constrained continuous optimization**: NOTEARS, Gran-DAG, DCDI, etc...

Continuous relaxation

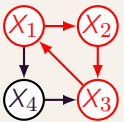
Powers of the adjacency matrix of a graph count the paths of a certain length in $\mathcal{G}$.



$$\begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$



$$\begin{bmatrix} 0 & 0 & 2 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$



$$\begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 2 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}$$

$\cdots$

$$\exp(A) = I + \frac{A}{1!} + \frac{A^2}{2!} + \frac{A^3}{3!} + \cdots$$

$\mathcal{G}$ is a DAG

$$\operatorname{Tr}\big(\exp(A)\big) = d$$

Constrained continuous optimization

$$\max_{\mathcal{G} \in \text{DAG}} \text{score}(\mathcal{G} \mid \mathcal{D})$$

$$\max_{A} \text{score}(A \mid \mathcal{D}) \quad \text{s.t. } \text{Tr}\big(\exp(A)\big) = d$$

This can be solved using **constrained optimization techniques** (e.g., Augmented Lagrangian)

Exercise

Show that

$$\frac{\partial}{\partial A} \text{Tr}\big(\exp(A)\big) = \exp(A)^\top$$