Computer
Science

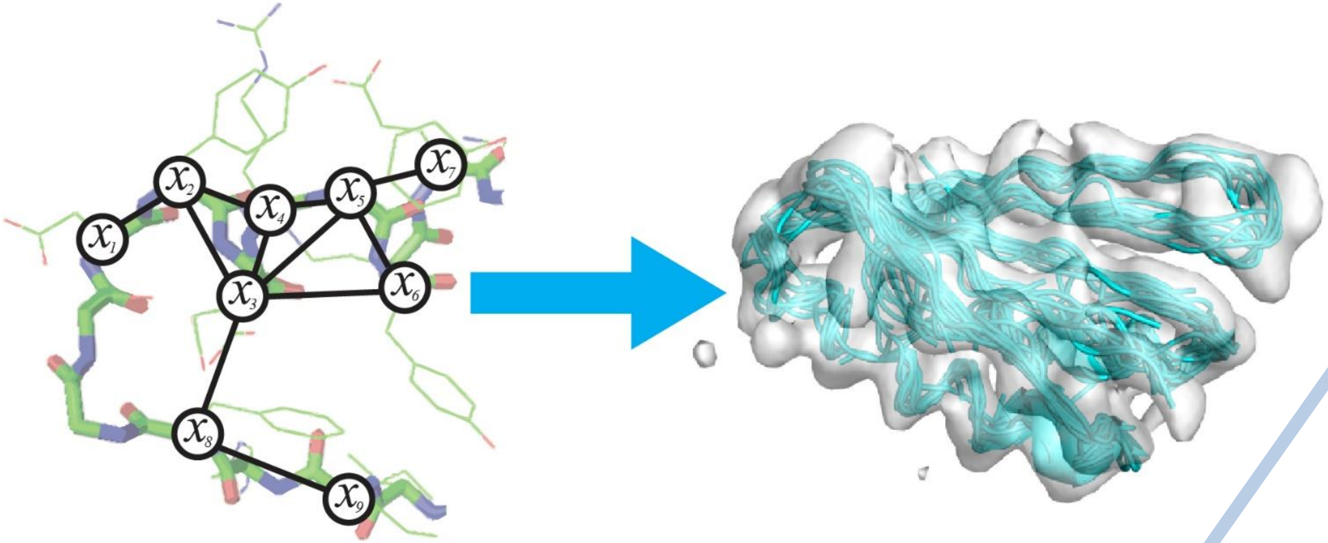
CSC 665-1: Advanced Topics in Probabilistic Graphical Models

Bayesian Inference

Instructor: Prof. Jason Pacheco

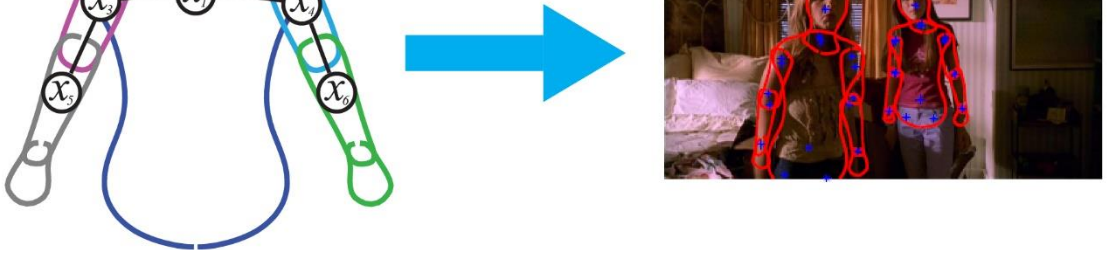
Why Graphical Models?

Data elements often have dependence arising from **structure**



Protein Structure

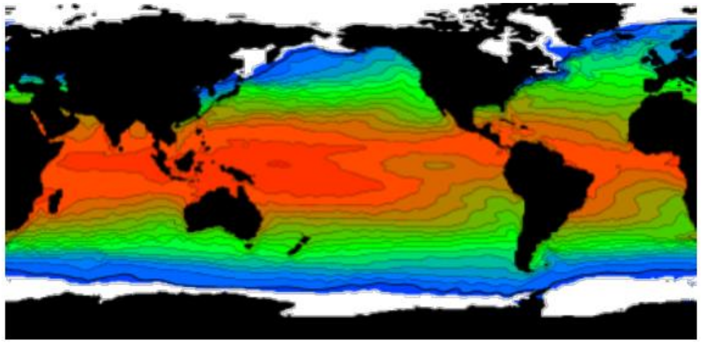
Pose Estimation



Exploit structure to simplify **representation** and **computation**

Why “Probabilistic”?

Stochastic processes have many sources of uncertainty



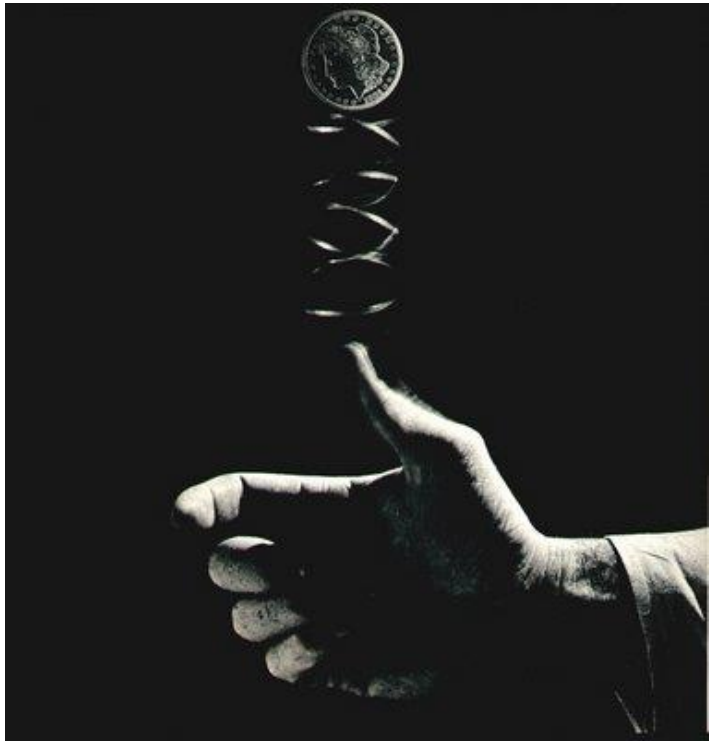
Randomness in State of Nature

Measurement Process

PGMs let us represent and reason about these

What is Probability?

What does it mean that the probability of heads is ½ ?



Two schools of thought…

Frequentist Perspective
Proportion of successes (heads) in repeated trials (coin tosses)

Bayesian Perspective

Belief of outcomes based on assumptions about nature and the physics of coin flips

Neither is better/worse, but we can compare interpretations…

Frequentist & Bayesian Modeling

We will use the following notation throughout:

θ - Unknown (e.g. coin bias)

y - Data

Frequentist

(Conditional Model)

$$p(y;\theta)$$

- θ is a non-random unknown parameter
- $p(y;\theta)$ is the *sampling / data generating distribution*

Bayesian

(Generative Model)

Prior Belief → $p(\theta)p(y \mid \theta)$ ← **Likelihood**

- θ is a random variable (latent)
- Requires specifying $p(\theta)$ the prior belief

Frequentist Inference

Example: Suppose we observe the outcome of N coin flips. $y = \{y_1, \ldots, y_N\}$. What is the probability of heads θ (coin bias)?

- Coin bias θ is <u>not random</u> (e.g. there is some *true* value)
- Uncertainty reported as <u>confidence interval</u> (typically 95%)

  Correct Interpretation: On repeated trials of N coin flips θ will fall inside the confidence interval 95% of the time (in the limit)

- Inferences are valid for multiple trials, never on single trials

  Wrong Interpretation: For *this trial* there is a 95% chance θ falls in the confidence interval

Bayesian Inference

Posterior distribution is complete representation of uncertainty

Posterior computed by **Bayes' rule:**

Prior Belief

Likelihood

$$p(\theta \mid y) = \frac{p(\theta)p(y \mid \theta)}{p(y)}$$

Marginal Likelihood
(more on this later)

- Must specify a <u>prior belief</u> $p(\theta)$ about coin bias
- Coin bias θ is a <u>random quantity</u>
- Interval $p(l(y) < \theta < u(y) \mid y) = 0.95$ can be reported in lieu of full posterior, and takes intuitive interpretation for a <u>single trial</u>

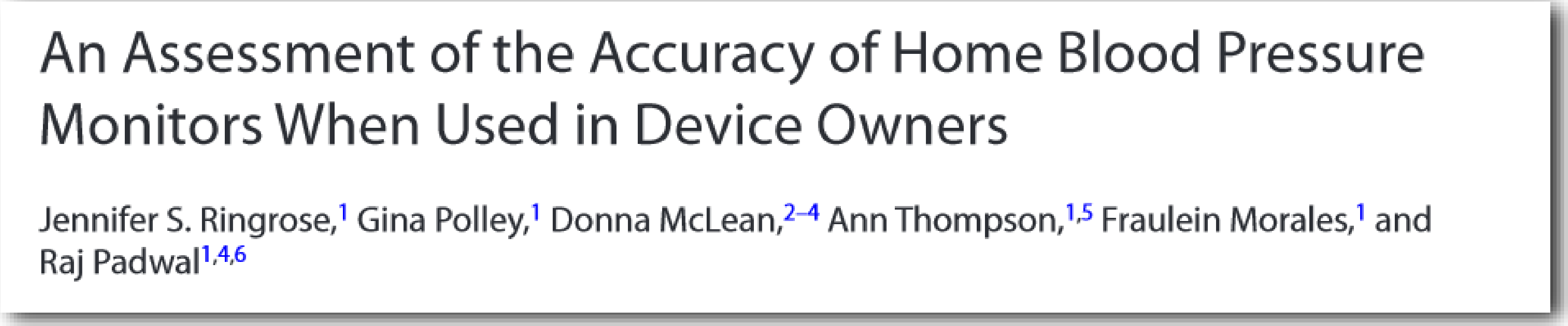
Interval Interpretation: For <u>this trial</u> there is a 95% chance that θ lies in the interval

Bayesian Inference Example

About 29% of American adults have high blood pressure (BP). Home test has 30% false positive rate and no false negative error.

Getty Images

A recent home test states that you have high BP. Should you start medication?



An Assessment of the Accuracy of Home Blood Pressure Monitors When Used in Device Owners

Jennifer S. Ringrose,[1] Gina Polley,[1] Donna McLean,[2–4] Ann Thompson,[1,5] Fraulein Morales,[1] and Raj Padwal[1,4,6]

Bayesian Inference Example

About 29% of American adults have high blood pressure (BP). Home test has 30% false positive rate and no false negative error.

- Latent quantity of interest is hypertension: $\theta \in \{true, false\}$
- Measurement of hypertension: $y \in \{true, false\}$
- Prior: $p(\theta = true) = 0.29$
- Likelihood: $p(y = true \mid \theta = false) = 0.30$

$$p(y = true \mid \theta = true) = 1.00$$

Bayesian Inference Example

About 29% of American adults have high blood pressure (BP). Home test has 30% false positive rate and no false negative error.


Getty Images

Suppose we get a positive measurement, then posterior is:

$$p(\theta = true \mid y = true) = \frac{p(\theta = true)p(y = true \mid \theta = true)}{p(y = true)}$$

$$= \frac{0.29 * 1.00}{0.29 * 1.00 + 0.71 * 0.30} \approx 0.58$$

What conclusions can be drawn from this calculation?

Marginal Likelihood

Posterior calculation requires the **marginal likelihood**,

$$p(\theta \mid y) = \frac{p(\theta)p(y \mid \theta)}{p(y)} \qquad p(y) = \int p(\theta)p(y \mid \theta)\, d\theta$$

- Also called the **partition function** or **evidence**
- Key quantity for model learning and selection
- NP-hard to compute in general (actually #P)

Example: Consider the vector $\theta = (\theta_1, \ldots, \theta_d)^T$ with binary $\theta_i \in \{0, 1\}$,

$$p(y) = \underbrace{\sum_{\theta_1=0}^{1} \sum_{\theta_2=0}^{1} \cdots \sum_{\theta_d=0}^{1}}_{\mathcal{O}(2^d)} p(\theta)p(y \mid \theta)$$

Exchangeability

We often assume the model is invariant to data ordering

Def: Consider *N* random variables $\{y_i\}_{i=1}^N$ and any permutation $\rho(\cdot)$ of indices. The variables are *exchangeable* if every permutation has equal probability,

$$p\big(y_1, y_2, \ldots, y_N\big) = p\big(y_{\rho(1)}, y_{\rho(2)}, \ldots, y_{\rho(N)}\big)$$

- $\{y_i\}_{i=1}^\infty$ is *infinitely exchangeable* if every finite subsequence is exchangeable
- Independence implies exchangeability, but the converse is not true

de Finetti’s Theorem

Simple hierarchical representation for exchangeable models

Thm. (de Finetti) *For any infinitely exchangeable sequence of random variables* $\{y_i\}_{i=1}^{\infty}$ *there exists some random variable* θ *with density* $p(\theta)$ *such that the joint probability of any* N *observations has a mixture representation:*

$$p(y_1, y_2, \ldots, y_N) = \int p(\theta) \prod_{i=1}^{N} p(y_i \mid \theta) \, d\theta$$

- Observe: this is the marginal likelihood for a model with prior $p(\theta)$
- Often used as justification for Bayesian statistics
- Technically only true for *infinitely exchangeable sequences* but reasonable approximation for many finite sequences

Posterior Marginal

In hierarchical models a subset of variables may be of interest

Normal distribution with random parameters:

$$\begin{aligned} y_i \mid \mu, \tau &\sim \mathcal{N}(\mu, \tau) \;\; i.i.d. \\ \mu \mid \tau &\sim \mathcal{N}(\mu_0, n_0 \tau) \\ \tau &\sim \mathrm{Gamma}(\alpha, \beta) \end{aligned}$$

Nuisance variable ($\mu \mid \tau \sim \mathcal{N}(\mu_0, n_0 \tau)$)

Quantity of interest ($\tau \sim \mathrm{Gamma}(\alpha, \beta)$)

Marginalize out nuisance variables:

$$p(\tau \mid x) = \int \mathrm{Gamma}(\tau \mid \alpha, \beta) \mathcal{N}(\mu \mid \mu_0, n_0 \tau) \prod_i \mathcal{N}(x_i \mid \mu, \tau) \, d\mu$$

Use of conjugate prior ensures analytic posterior

$$= \mathrm{Gamma}\left(\tau \mid \alpha + \frac{n}{2}, \beta + \frac{1}{2} \sum_i (x_i - \bar{x})^2 + \frac{n n_0}{2(n + n_0)} (\bar{x} - \mu_0)^2 \right)$$

Prediction

Can make predictions of unobserved $\tilde{y}$ before seeing any data,

$$p(\tilde{y}) = \int p(\theta) p(\tilde{y} \mid \theta) \, d\theta$$

Similar calculation to marginal likelihood

*This is the **prior predictive** distribution*

When we observe y we can predict future observations $\tilde{y}$,

$$p(\tilde{y} \mid y) = \int p(\theta \mid y) p(\tilde{y} \mid \theta) \, d\theta$$

*This is the **posterior predictive** distribution*

Prediction Example

About 29% of American adults have high blood pressure (BP). Home test has 30% false positive rate and no false negative error.

What is the likelihood of *another* positive measurement?

$$p(\tilde{y} = true \mid y = true) = \sum_{\theta \in \{true, false\}} p(\theta \mid y = true) p(\tilde{y} = true \mid \theta)$$

$$= 0.42 * 0.30 + 0.58 * 1.00 \approx 0.71$$

What conclusions can be drawn from this calculation?

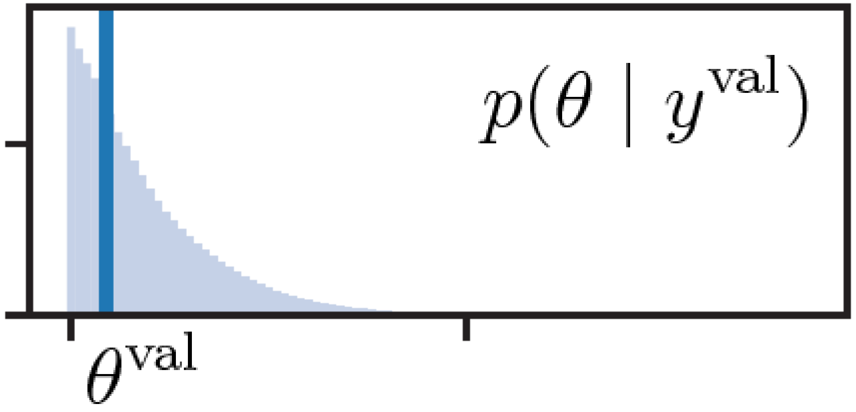
Model Validation

How do we know if the model $p(\theta, y)$ *is* <u>*good*</u>*?*

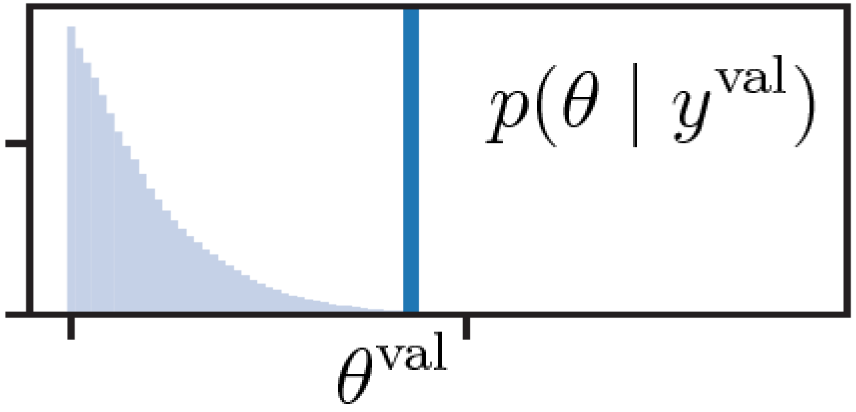
Supervised Learning

Validation set $\{(\theta^{\text{val}}, y^{\text{val}})\}$ consists of known θ^{val}. Are true values typically preferred under the posterior?

Good (maybe lucky)



Not Good (maybe unlucky)



Repeat trials over validation set for more certainty

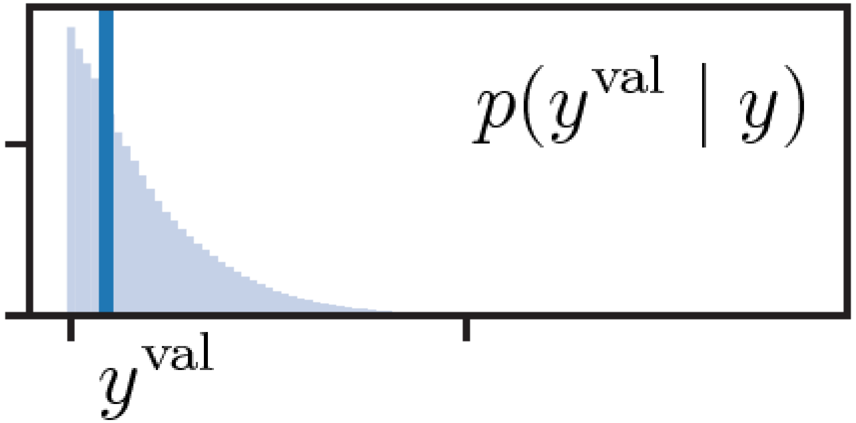
Model Validation

How do we know if the model $p(\theta, y)$ *is <u>good</u>?*

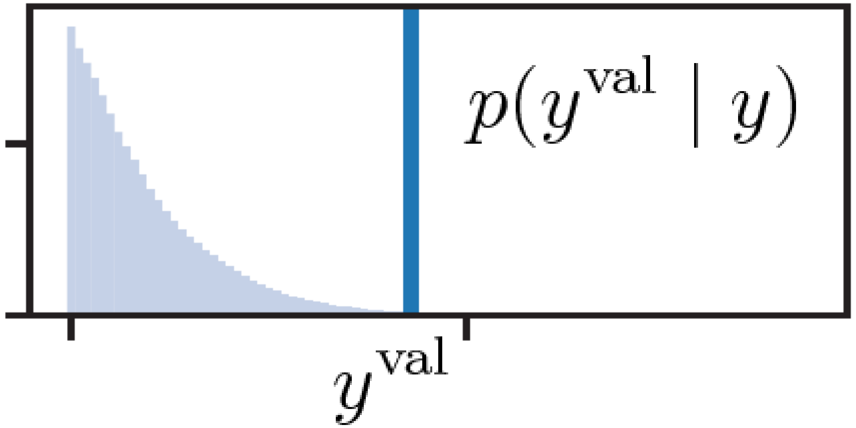
Unsupervised Learning

Validation set $\{y^{\text{val}}\}$ only contains observable data. Check validation data against posterior-predictive distribution.

Good (maybe lucky)



Not Good (maybe unlucky)



Repeat trials over validation set for more certainty

Likelihood and Odds Ratios

Which parameter value θ_1 or θ_2 is more likely to have generated the observed data y ?

The **posterior odds ratio** is:

$$\frac{p(\theta_1 \mid y)}{p(\theta_2 \mid y)} = \underbrace{\frac{p(\theta_1)}{p(\theta_2)}}_{\textbf{Prior Odds Ratio}} \underbrace{\frac{p(y \mid \theta_1)}{p(y \mid \theta_2)}}_{\textbf{Likelihood Ratio}} \cancel{\frac{p(y)}{p(y)}}$$

Observe: the marginal likelihood $p(y)$ cancels!

Bayesian Estimation

Task: *produce an estimate* $\hat{\theta}$ *of* θ *after observing data* y

Bayes estimators minimize expected **loss function**:

$$\mathbb{E}[L(\theta, \hat{\theta}) \mid y] = \int p(\theta \mid y) L(\theta, \hat{\theta}) \, d\theta$$

Example: Minimum mean squared error (MMSE):

$$\hat{\theta}^{\text{MMSE}} = \arg\min \mathbb{E}[(\hat{\theta} - \theta)^2 \mid y] = E[\theta \mid y]$$

Posterior mean always minimizes squared error.

Bayes Estimation: More Examples

Minimum absolute error:

$$\arg\min \mathbb{E}[|\hat{\theta} - \theta| \mid y] = \text{median}(\theta \mid y)$$

Note: Same answer for linear function $L(\theta, \hat{\theta}) = c|\hat{\theta} - \theta|$.

Maximum *a posteriori* (MAP):

Very common to produce maximum probability estimates,

$$\hat{\theta}^{\text{MAP}} = \arg max\, p(\theta \mid y)$$

Loss function is degenerate,

$$\lim_{c \to 0} L(\theta, \hat{\theta}) = \begin{cases} 0, & \text{if } |\hat{\theta} - \theta| < c \\ 1, & \text{otherwise} \end{cases}$$

Not a Bayes estimator!

(unless discrete)

Posterior Summarization

Ideally we would report the full posterior distribution as the result of inference…but this is not always possible

Summary of Posterior Location:

Point estimates: mean (MMSE), mode, median (min. absolute error)

Summary of Posterior Uncertainty:

Credible intervals / regions, posterior entropy, variance

Bayesian analysis should report uncertainty when possible

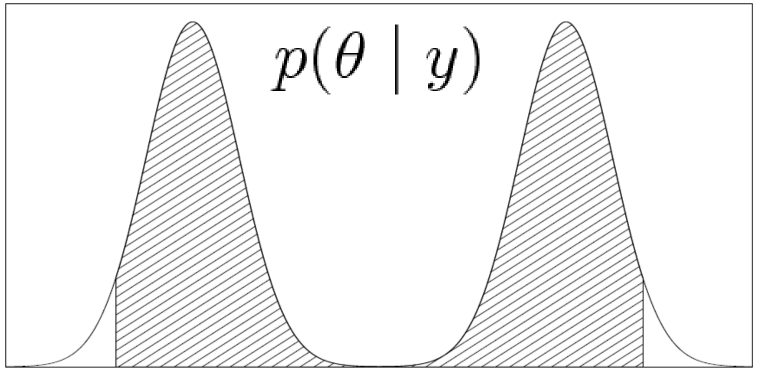
Credible Interval

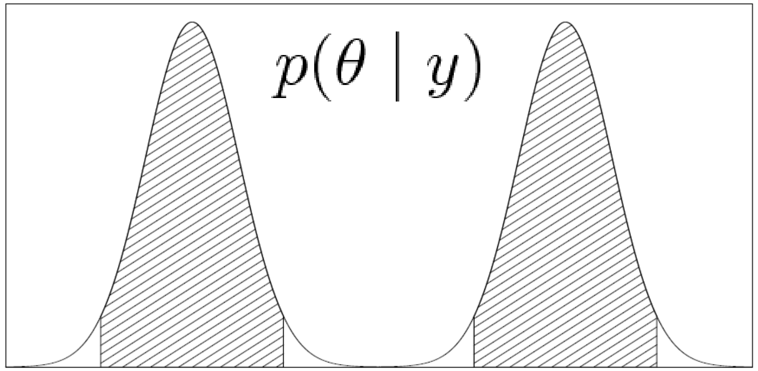
Def. For parameter $0 < \alpha < 1$ the $100(1-\alpha)\%$ a credible interval $(L(y), U(y))$ satisfies,

$$p(L(y) < \theta < U(y) \mid y) = \int_{L(y)}^{U(y)} p(\theta \mid y) = 1 - \alpha$$

Interval containing fixed percentage of posterior probability density.

Note: This is not unique -- consider the 95% intervals below:





[Source: Gelman et al., “Bayesian Data Analysis”]

Summary

- Marginal likelihood required for Bayesian inference, which can be hard:

$$p(\theta \mid y) = \frac{p(\theta)p(y \mid \theta)}{p(y)} \qquad p(y) = \int p(\theta)p(y \mid \theta)\, d\theta$$

- One exception is posterior odds (used in model selection, hypothesis testing, …)

$$\frac{p(\theta_1 \mid y)}{p(\theta_2 \mid y)} = \frac{p(\theta_1)}{p(\theta_2)} \frac{p(y \mid \theta_1)}{p(y \mid \theta_2)} \cancel{\frac{p(y)}{p(y)}}$$

- Posterior predictive can be used for model quality in unsupervised setting:

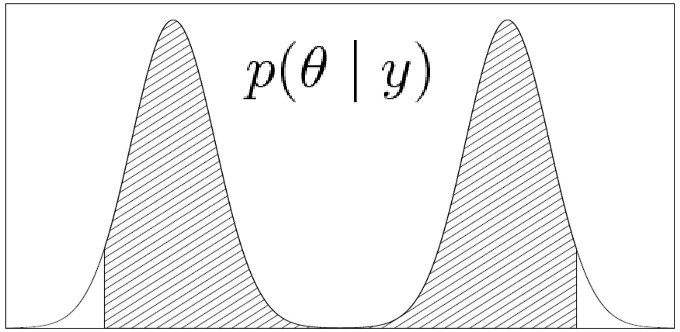
$$p(\tilde{y} \mid y) = \int p(\theta \mid y)p(\tilde{y} \mid \theta)\, d\theta$$

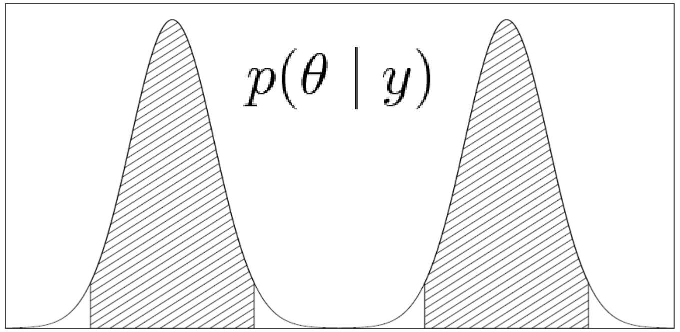
Summary

- Bayesian estimation minimizes expected loss function:

$$\mathbb{E}[L(\theta,\hat{\theta}) \mid y] = \int p(\theta \mid y) L(\theta,\hat{\theta})\, d\theta$$

- Common estimators: Posterior mean → MMSE, Median → MAE
- Posterior uncertainty can be summarized by (not necessarily unique) credible intervals:





- Interpretation: For <u>this trial</u> parameter lies in interval with specified probability (e.g. 0.95)