Probabilistic Graphical Models

Bayesian Inference

Recap

Continuous Probability Distributions

Replace mass with density, sums with integrals

Probability of any single outcome is zero

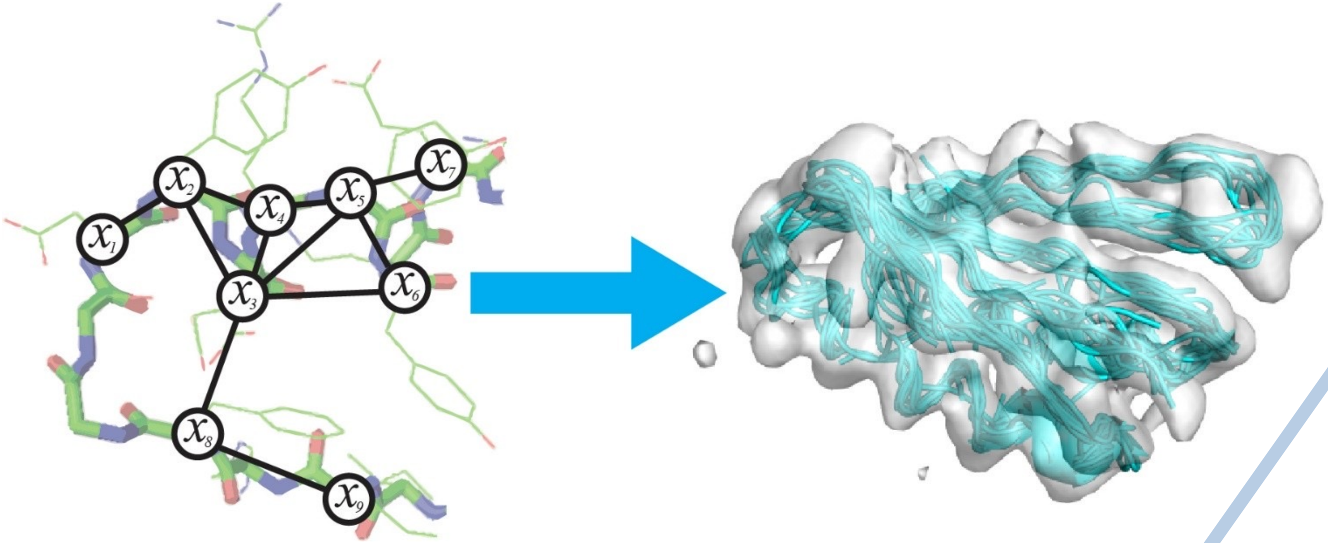
Convergence of Random Variables

WLLN: The sample mean converges to the true mean

CLT: A scaled version of the sample mean converges in distribution to the standard normal

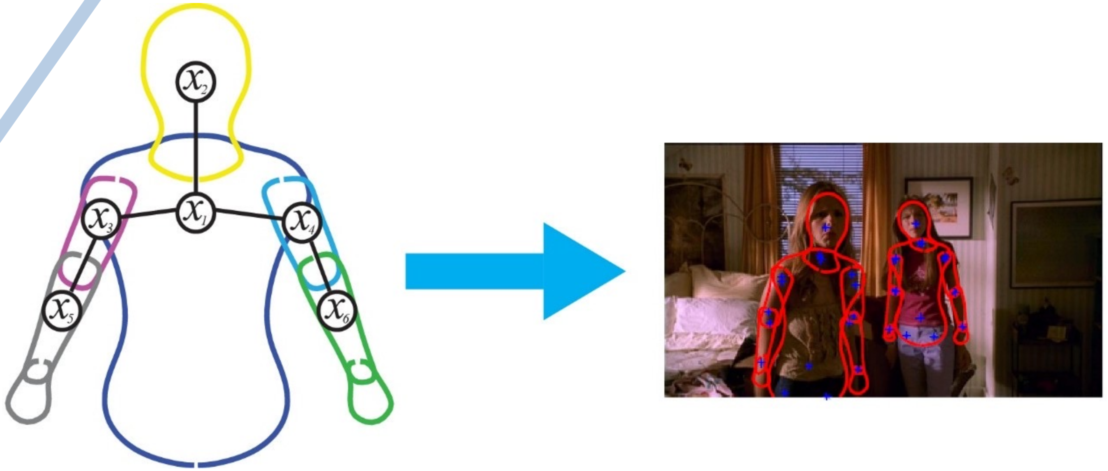
Why Graphical Models?

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



Protein Structure

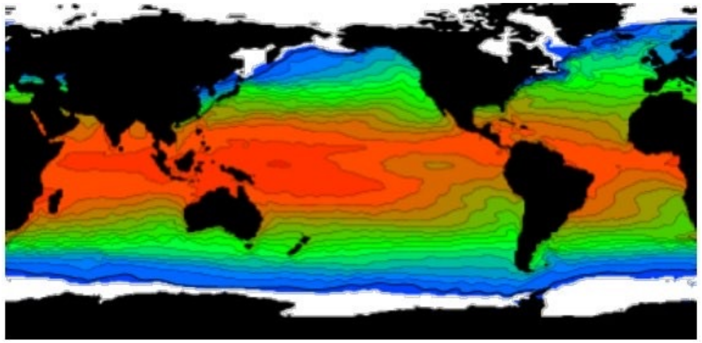
Pose Estimation



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 in structured ways

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

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

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 prior belief $p(\theta)$ about coin bias
- Coin bias θ is a random quantity
- 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 single trial

Interval Interpretation: For this trial 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.

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.

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?

Bayesian Inference

Likelihood Prior

$$\boxed{P(\theta|X_1,\ldots,X_n)} = \frac{\boxed{P(X_1,\ldots,X_n|\theta)}\boxed{P(\theta)}}{\boxed{P(X_1,\ldots,X_n)}}$$

Posterior

Normalizing Constant

Bayes Rule

Example: Bernoulli Distribution

$X_1, \dots, X_n$ follow a Bernoulli distribution
We want to estimate $P(\theta | X_1, \dots, X_n)$

$$P(\theta | X_1, \dots, X_n) = \frac{P(X_1, \dots, X_n | \theta) P(\theta)}{P(X_1, \dots, X_n)}$$

Assume we observe $X_1, \dots, X_{40}$ with $\sum_{i=1}^{40} x_i = 10$

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
- Depends on the prior!
- NP-hard to compute in general (actually #P)

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

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

Beta distribution

X has the Beta distribution with parameters $\alpha, \beta > 0$ if

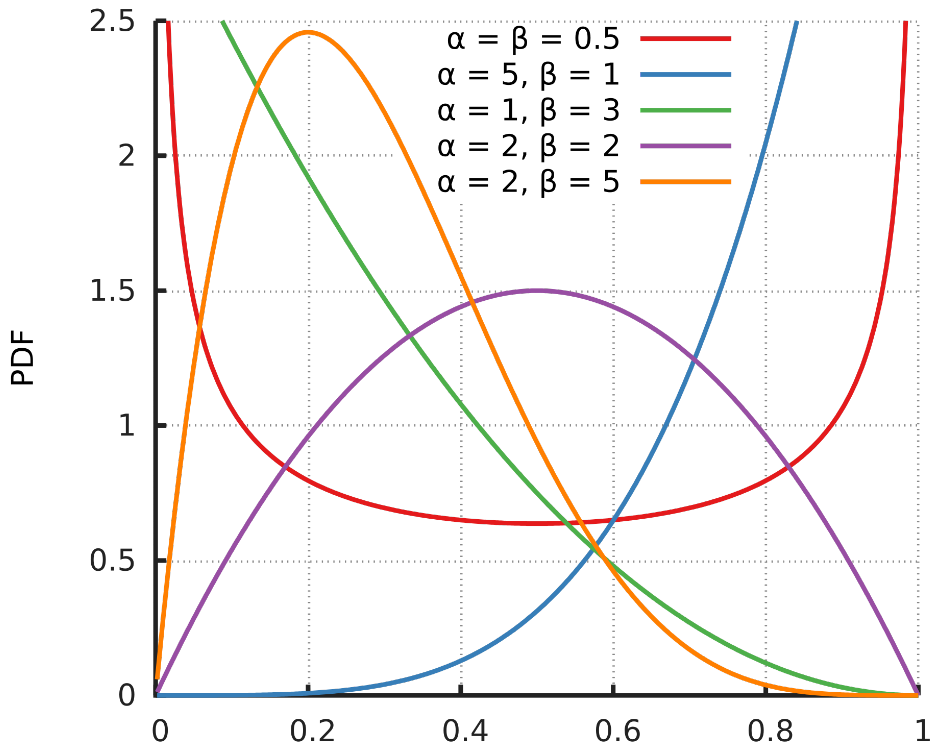
$$f(x \mid \alpha, \beta) = \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} & x \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

Suitable for RV in $[0,1]$

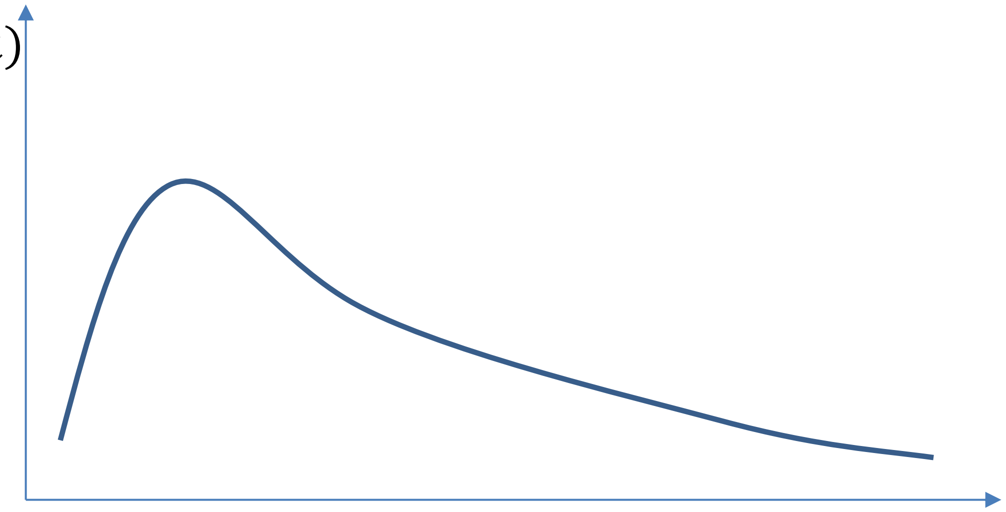
Parameter space: $\alpha, \beta > 0$.

$$E(X) = \frac{\alpha}{\alpha+\beta}, \mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

$$\mathrm{B}(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$



Computing the Posterior



- Pick a prior, e.g., Beta(2,2) :

$$f(\theta) = \frac{1}{B(2,2)}\theta(1-\theta)$$

- Compute the likelihood:

$$f(x_1, \dots, x_{40} \mid \theta) = \prod_{i=1}^{40} f(x_i \mid \theta) = \theta^{10}(1-\theta)^{30}$$

- Compute the posterior up to a constant:

$$f(\theta \mid x_1, \dots, x_{40}) = \frac{1}{\mathrm{B}(2,2) f(x_1, \dots, x_{40})} f(\theta) f(x_1, \dots, x_{40} \mid \theta) = C\theta^{10+1}(1-\theta)^{30+1}$$

- C is a constant, $f(\theta \mid x_1, \dots, x_{40})$ is a Beta(12,32) distribution.

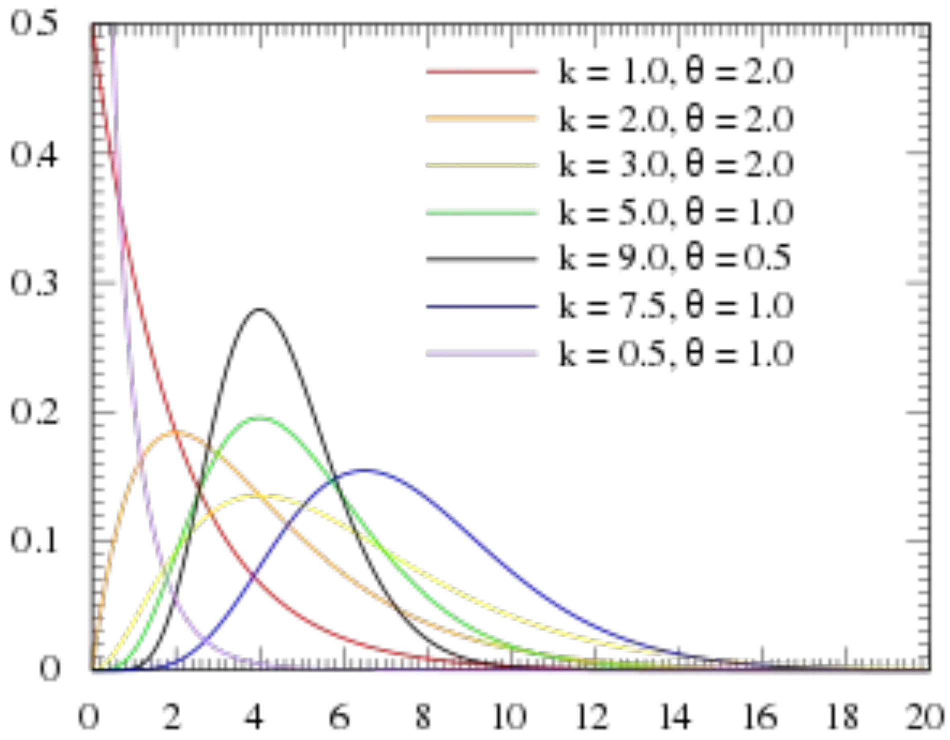
Beta Prior – Beta Posterior:
Beta is a conjugate distribution for the Bernoulli Likelihood

Conjugate Distributions

Gamma distribution

X has the Gamma distribution with parameters $\alpha, \beta > 0$ if

$$f(x \mid \alpha, \beta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$



Suitable for RVs in $(0, \infty)$

Parameter space: $\alpha, \beta > 0$.

$E(X) = \frac{\alpha}{\beta}, \mathrm{Var}(X) = \frac{\alpha}{\beta^2}$.

Conjugate Distributions

Dirichlet distribution

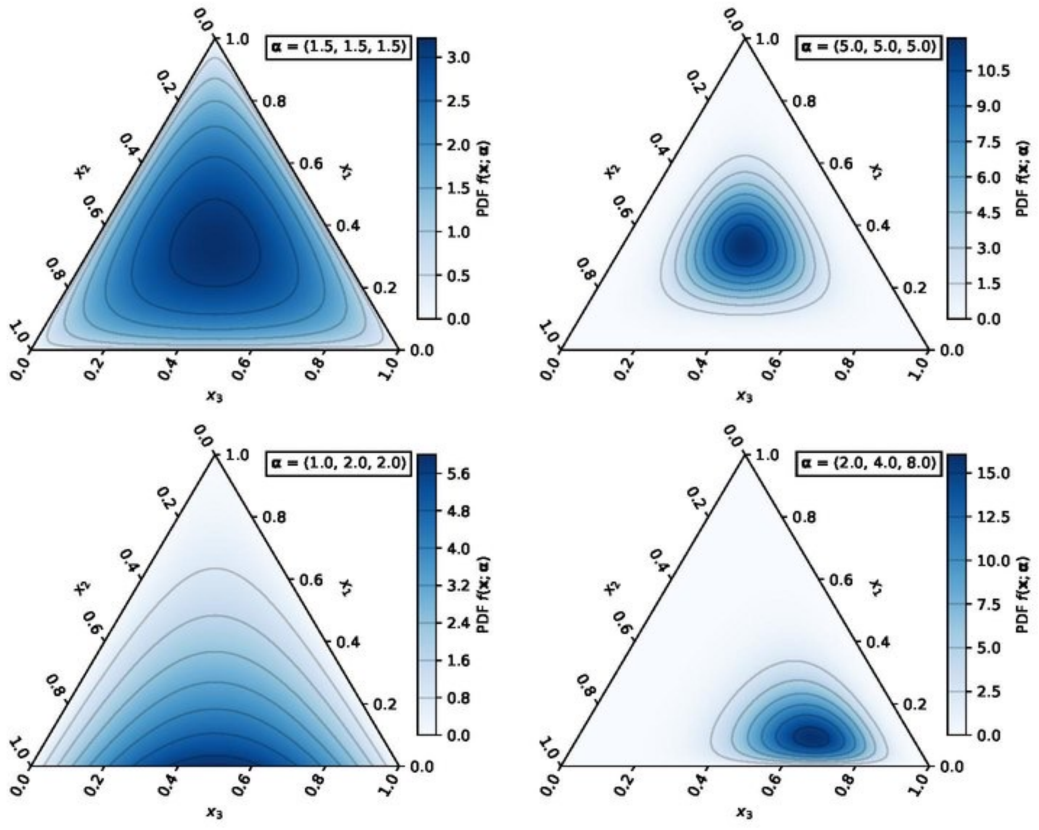
$X = X_1, \dots X_K$ have the Dirichlet distribution with parameters $\alpha_1, \dots, a_k$

$$f(x_1, \dots, x_k \mid \alpha_1, \dots, a_k) = \begin{cases} \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{a_i - 1}, x_i = 0,1 \\ 0, \quad\quad otherwise \end{cases}$$

where $B(\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\alpha_0)}, \alpha_0 = \sum_{i=1}^{K} \alpha_i$

Parameter space: $\alpha_1, \dots, a_k$

$E(X_\mathrm{i}) = \frac{\alpha_\mathrm{i}}{\alpha_0}, \mathrm{Var}(X_\mathrm{i}) = \frac{\alpha_o - \alpha_i}{\alpha_o^2(\alpha_o + 1)}.$



How do we perform Bayesian Inference

- Pick a prior $P(\theta)$
- Compute the likelihood $P(X_1, \ldots, X_n|\theta)$
- Compute the normalization constant

  $P(X_1, \ldots, X_n) = \int P(X_1, \ldots, X_n|\theta) f(\theta) d\theta$

  (Also known as the marginal likelihood)

- Difficult, not always necessary

Conjugate Distributions

Likelihood	Prior	Posterior
$\mathrm{Bern}(p)$	$\mathrm{Beta}(\alpha,\beta)$	$\mathrm{Beta}(\alpha+\sum_{i=1}^{n}x_i, \beta+\mathrm{N}-\sum_{i=1}^{n}x_i)$
$\mathrm{Binom}(\mathrm{N},p)$	$\mathrm{Beta}(\alpha,\beta)$	$\mathrm{Beta}(\alpha+\sum_{i=1}^{n}x_i, \beta+\mathrm{N}-\sum_{i=1}^{n}x_i)$
$\mathrm{Pois}(\lambda)$	$\mathrm{Gamma}(\alpha,\beta)$	$\mathrm{Gamma}(\alpha+\sum_{i=1}^{n}x_i, \beta+n)$
$\mathrm{Expo}(\lambda)$	$\mathrm{Gamma}(\alpha,\beta)$	$\mathrm{Gamma}(\alpha+n, \beta+\sum_{i=1}^{n}x_i)$

Example: Gamma - Exponential

Improper priors

Sometimes it is convenient to pick a prior that does not have a proper distribution. This is called an improper prior.

Example: “Uniform” Prior for Normal Distribution

Loss Function

$L\left(\theta,\hat{\theta}\right)$: Quantifies how far your estimate $\hat{\theta}$ is from the true value θ.

Examples of loss functions:

Mean Squared Error: $\left(\hat{\theta}-\theta\right)^2$

Mean Absolute Error: $\left|\hat{\theta}-\theta\right|$

Zero-one loss: $0, if\,\hat{\theta}=\theta, 1\ otherwise$.

The loss is a random variable

We are looking for the estimate $\hat{\theta}$ that minimizes $E(L(\theta,\hat{\theta})|Data)$

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):

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

Posterior mean always minimizes squared error.

Bayesian 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,

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

Bayesian Updating

Consider two *conditionally independent* observations X_1 and X_2, their joint distribution is:

Probability chain rule

$$p(\theta, X_1, X_2) = p(\theta)p(X_1 \mid \theta)p(X_2 \mid \theta) = p(\theta \mid X_1)p(X_1)p(X_2 \mid \theta)$$

So, conditioned on X_1:

Update prior belief after seeing X_1

$$p(\theta, X_2 \mid X_1) = p(\theta \mid X_1)p(X_2 \mid \theta)$$

This is proportional to the **full posterior** by Bayes' rule:

$$p(\theta \mid X_1, X_2) \propto p(\theta \mid X_1)p(X_2 \mid \theta)$$

Normalizer is marginal likelihood $p(X_1,X_2)$

In general, given conditionally independent $X_1, \ldots, X_N$:

$$p(\theta \mid X_1, \ldots, X_N) \propto p(\theta \mid X_1, \ldots, X_{N-1})p(X_N \mid \theta)$$

Bayesian Inference for the Gaussian (2)

Combined with a Gaussian prior over μ

$$p(\mu) = \mathcal{N}\left(\mu|\mu_0, \sigma_0^2\right).$$

this gives the posterior $p(\mu|\mathbf{x}) \propto p(\mathbf{x}|\mu)p(\mu).$

Completing the square over μ, we see that

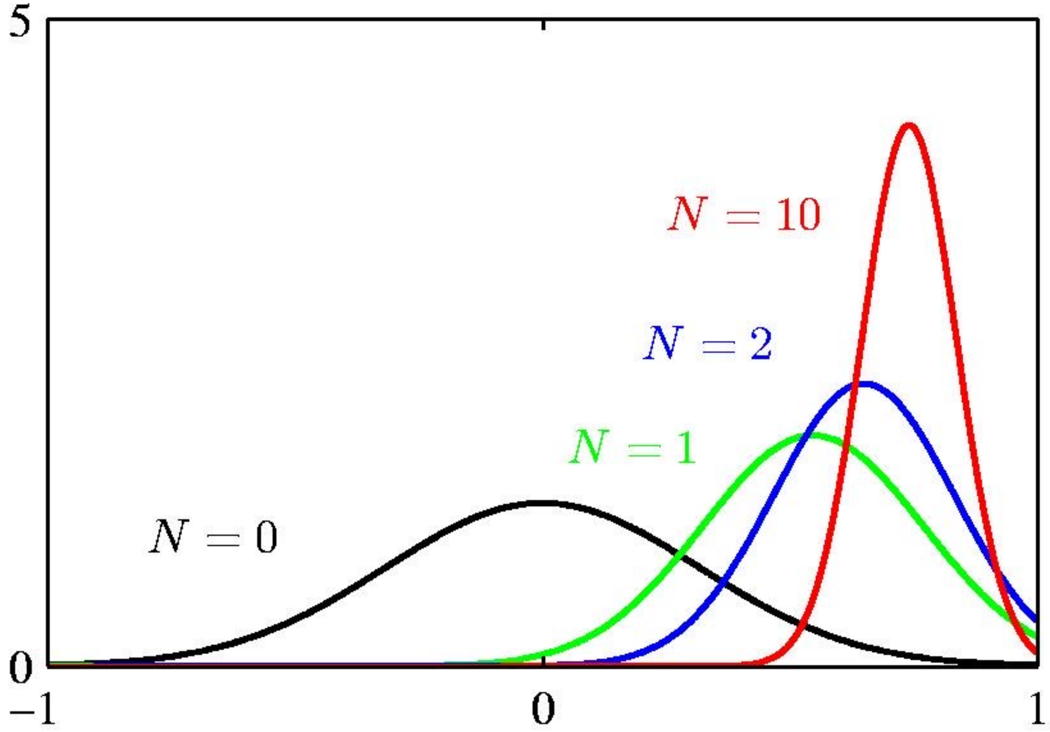
$$p(\mu|\mathbf{x}) = \mathcal{N}\left(\mu|\mu_N, \sigma_N^2\right)$$

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{\mathrm{ML}}, \qquad \mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}.$$

Bayesian Inference for the Gaussian (3)

Example: $p(\mu|\mathbf{x}) = \mathcal{N}\left(\mu|\mu_N, \sigma_N^2\right)$ for N = 0, 1, 2 and 10.



Bayesian Inference for the Gaussian (4)

Sequential Estimation

$$
\begin{aligned}
p(\mu|\mathbf{x}) &\propto p(\mu)p(\mathbf{x}|\mu) \\
&= \left[p(\mu)\prod_{n=1}^{N-1} p(x_n|\mu)\right] p(x_N|\mu) \\
&\propto \mathcal{N}\left(\mu|\mu_{N-1}, \sigma_{N-1}^2\right) p(x_N|\mu)
\end{aligned}
$$

The posterior obtained after observing N -1 data points becomes the prior when we observe the Nth data point.

Bayesian Inference for the Gaussian (5)

Now assume μ is known. The likelihood function for $\lambda = 1/\sigma^2$ is given by

$$p(\mathbf{x}|\lambda) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \lambda^{-1}) \propto \lambda^{N/2} \exp\left\{ -\frac{\lambda}{2} \sum_{n=1}^{N} (x_n - \mu)^2 \right\}.$$

This has a Gamma shape as a function of λ.

Bayesian Inference for the Gaussian (6)

Now we combine a Gamma prior, $\mathrm{Gam}(\lambda|a_0, b_0)$, with the likelihood function for ¸ to obtain

$$p(\lambda|\mathbf{x}) \propto \lambda^{a_0-1}\lambda^{N/2}\exp\left\{-b_0\lambda - \frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}$$

which we recognize as $Gam(\lambda|\alpha_N, b_N)$ with

$$\begin{aligned} a_N &= a_0 + \frac{N}{2} \\ b_N &= b_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2 \end{aligned}$$

Bayesian Inference for the Gaussian (7)

If both μ and λ are unknown, the joint likelihood function is given by

$$p(\mathbf{x}|\mu, \lambda) = \prod_{n=1}^{N} \left(\frac{\lambda}{2\pi}\right)^{1/2} \exp\left\{-\frac{\lambda}{2}(x_n - \mu)^2\right\}$$

$$\propto \left[\lambda^{1/2} \exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^N \exp\left\{\lambda\mu \sum_{n=1}^{N} x_n - \frac{\lambda}{2} \sum_{n=1}^{N} x_n^2\right\}.$$

Bayesian Inference for the Gaussian (8)

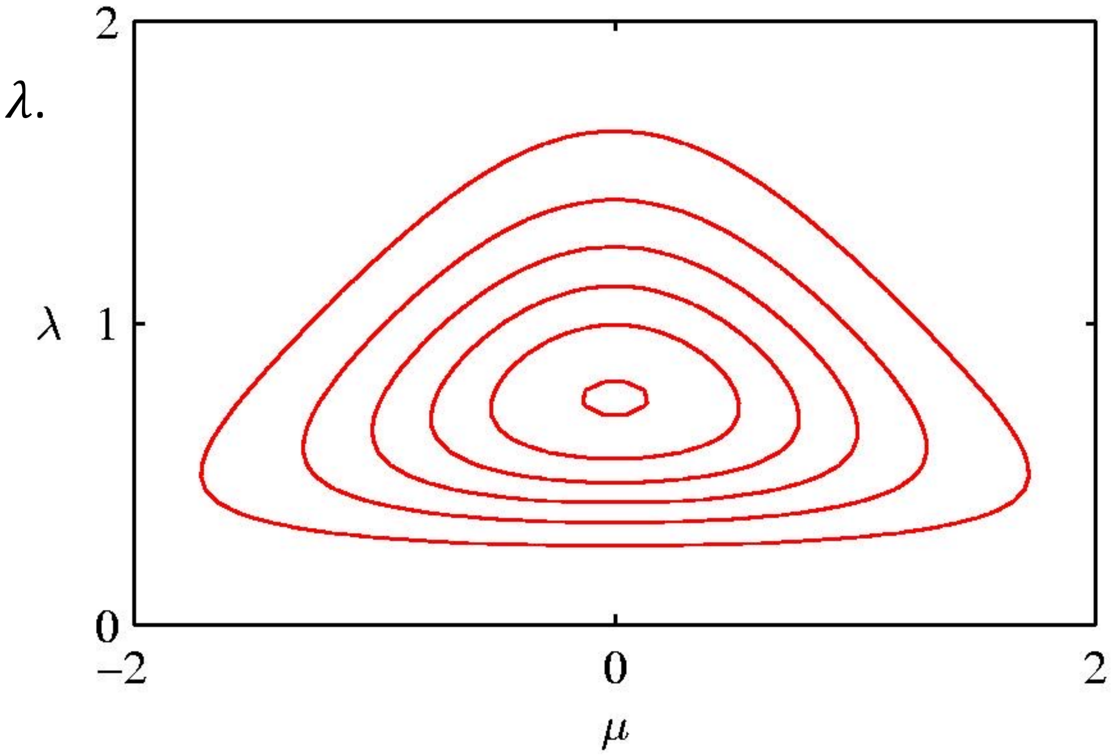
The Gaussian-gamma distribution

$$p(\mu, \lambda) = \mathcal{N}(\mu|\mu_0, (\beta\lambda)^{-1})\mathrm{Gam}(\lambda|a, b)$$

$$\propto \quad \exp\left\{-\frac{\beta\lambda}{2}(\mu - \mu_0)^2\right\} \lambda^{a-1}\exp\{-b\lambda\}$$

- Quadratic in μ.
- Linear in λ.

- Gamma distribution over λ.
- Independent of μ



Bayesian Inference for the Gaussian (9)

- $\boldsymbol{\mu}$ unknown, Λ known: $p(\boldsymbol{\mu})$ Gaussian.
- $\boldsymbol{\Lambda}$ unknown, $\boldsymbol{\mu}$ known: $p(\boldsymbol{\Lambda})$ Wishart, $\mathcal{W}(\boldsymbol{\Lambda} \mid \mathbf{W}, \nu) = B \mid \boldsymbol{\Lambda}^{(\nu-D-1)/2} \exp\left(-\frac{1}{2}\mathrm{Tr}(\mathbf{W}^{-1}\boldsymbol{\Lambda})\right)$.
- $\boldsymbol{\Lambda}$ and $\boldsymbol{\mu}$ unknown: $p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ GaussianWishart, $p(\boldsymbol{\mu}, \boldsymbol{\Lambda} \mid \boldsymbol{\mu}_0, \beta, \mathbf{W}, \nu) =$

$$\mathcal{N}(\boldsymbol{\mu} \mid \boldsymbol{\mu}_0, (\beta\boldsymbol{\Lambda})^{-1})\mathcal{W}(\boldsymbol{\Lambda} \mid \mathbf{W}, \nu)$$

Slides by Christopher Bishop

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:

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

Prior Odds Ratio

Likelihood Ratio

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

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*

Posterior Marginal

In hierarchical models a subset of variables may be of interest

Normal distribution with random parameters:

$$y_i \mid \mu, \tau \sim \mathcal{N}(\mu, \tau) \;\; i.i.d.$$

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

$$\tau \sim \text{Gamma}(\alpha, \beta)$$ ← **Quantity of interest**

Marginalize out nuisance variables:

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

Use of conjugate prior ensures analytic posterior

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

Posterior Summarization

Ideally we would report the <u>full posterior distribution</u> 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

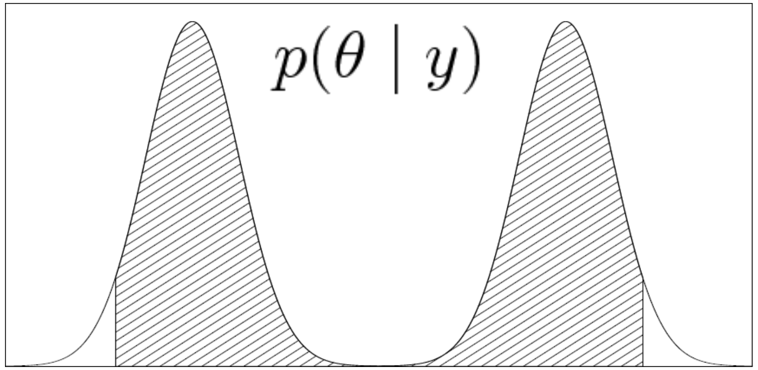
Credible Interval

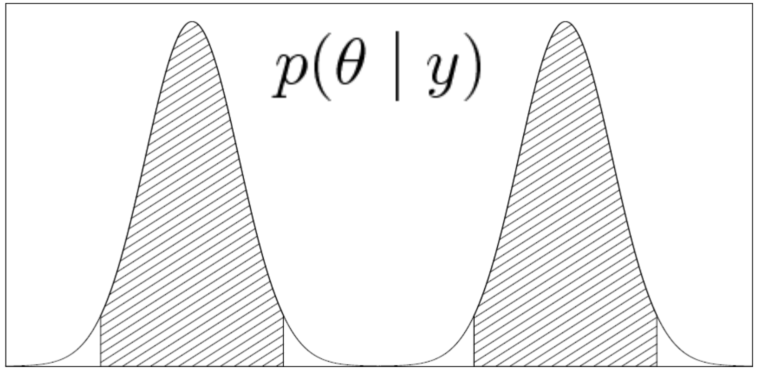
Def. For parameter $0 < \alpha < 1$ the $100(1-\alpha)\%$ a credible interval $(L(y), U(y))$ satisfies,

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

Interval containing fixed percentage of posterior probability density.

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





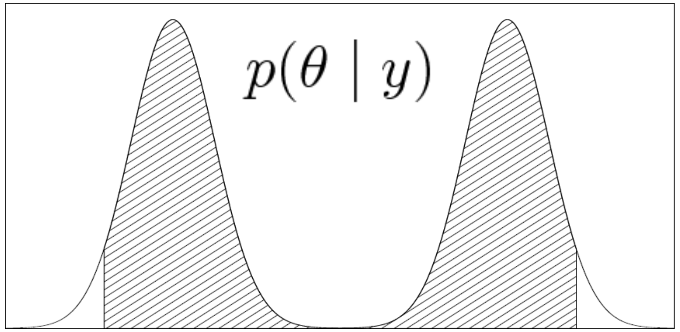
[Source: Gelman et al., "Bayesian Data Analysis"]

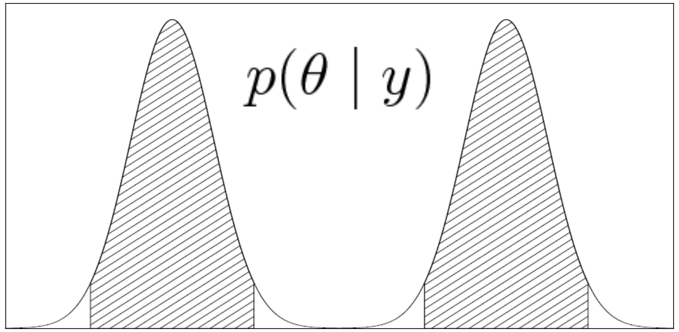
Summary

- 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)

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, …)

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

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

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