Machine Learning II

Bayesian inference and conjugate priors

Bayesian estimation

Bayesian statistics is about reasoning under uncertainty.

Dutch book: Rationality demands that subjective belief can be modelled as probabilities!

Example: Should you believe this coin is fair?

- Consider a coin that has been tossed 20 times
- Suppose that we observed 15 heads

Classical estimation

- Let θ denote the (unknown) bias of our coin
- Probability to observe k heads on n tosses is given by the Binomial distribution

$$P(H = k|n,\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

Maximum likelihood estimate $\hat{\theta}_{ML}$ obtained as

$$\begin{aligned}\hat{\theta}_{ML} &= \texttt{argmax}_\theta P(H = k|n,\theta) \\ &= \frac{k}{n}\end{aligned}$$

p-value that coin is unbiased:

$$\begin{aligned}P(H \geq k|n,\theta = \frac{1}{2}) &= \sum_{i=k}^{n}\binom{n}{k}2^{-n} \\ &\approx 0.02\end{aligned}$$

Classical estimation

- Let θ denote the (unknown) bias of our coin
- Probability to observe k heads on n tosses is given by the Binomial distribution

$$P(H = k|n, \theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

Maximum likelihood estimate $\hat{\theta}_{ML}$ obtained as

$$\begin{aligned}\hat{\theta}_{ML} &= \mathtt{argmax}_\theta P(H = k|n, \theta) \\ &= \frac{k}{n}\end{aligned}$$

Extreme example:

- Coin tossed three times: H H H
- *Overfitting:* Do you really believe that T is impossible?

Bayesian estimation

- Consider coin bias θ as a random variable with prior probability distribution $p(\theta)$
- Compute the posterior

$$p(\theta|Data) \propto p(\theta)p(Data|\theta)$$

Bayesian estimation

- Consider coin bias θ as a random variable with prior probability distribution $p(\theta)$
- Compute the posterior

$$p(\theta|Data) \propto p(\theta)p(Data|\theta)$$

Convenient choice for the prior is a Beta distribution

$$p(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- with the Gamma function $\Gamma(x)$ (Note: $\Gamma(x+1) = x\Gamma(x)$)
- mean $\mathbb{E}[\theta] = \frac{\alpha}{\alpha+\beta}$
- and variance $\mathbb{E}[(\theta - \mathbb{E}[\theta])^2] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Bayesian estimation

Posterior with beta prior is found to be

$$\begin{aligned} p(\theta|k,n) &\propto p(\theta|\alpha,\beta)p(H=k|n,\theta) \\ &\propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\binom{n}{k}\theta^{k}(1-\theta)^{n-k} \\ &\propto \theta^{\alpha+k-1}(1-\theta)^{\beta+(n-k)-1} \end{aligned}$$

Thus, posterior is again a Beta distribution with parameter $\alpha_k = \alpha + k, \beta_k = \beta + (n-k)$:

- This is an example of a *conjugate prior*

Definition

A prior is called *conjugate* (for the likelihood function) if the posterior $p(\theta|\mathcal{D})$ is in the same family of distributions as the prior $p(\theta)$.

- The *hyperparameters* α, β can be considered as pseudo-counts, i.e. interpreted as virtual observations

Bayesian learning

Bayesian learning: Updating information from prior to posterior
This can always be done *sequentially*:

1. Assume two independent observations D_1, D_2 (or split the one you have), e.g. coin tossed 10 time for 7 heads and then another 10 times for 8 heads
2. Then,

$$
\begin{aligned}
p(\theta|D_1, D_2) &\propto p(\theta)p(D_1, D_2|\theta) \\
&= p(\theta)p(D_2|\theta)p(D_1|\theta) \\
&\quad \text{since } D_1 \text{ and } D_2 \text{ are conditionally indepdendent given } \theta \\
&\propto p(\theta|D_1)p(D_2|\theta)
\end{aligned}
$$

Thus, posterior $p(\theta|D_1)$ serves as prior when learning from D_2

Bayesian learning

Illustration of coin toss example:

- In **Python** the density of the Beta distribution is available as

```
scipy.stats.beta
```

- Above example can be plotted as follows:

```
import numpy as np
from matplotlib import pyplot as plt
from scipy.stats import beta as Beta

n, k = 20, 15
alpha, beta = 1, 1
x = np.linspace(0, 1, num=100)
# Prior
plt.plot(x, Beta.pdf(x, alpha, beta), 'b-')
# Posterior
plt.plot(x, Beta.pdf(x, alpha + k, beta + (n-k)), 'r-')
```

Point estimates

Posterior $p(\theta|D)$ summarizes knowledge about θ after data D was observed

How can we construct a point estimate, i.e. $\hat{\theta}_{Bayes}$?

- Could simply take posterior mean, median or mode ...

Point estimates

Posterior $p(\theta|D)$ summarizes knowledge about θ after data D was observed

How can we construct a point estimate, i.e. $\hat{\theta}_{Bayes}$?

- Could simply take posterior mean, median or mode ...
- More principled approach considers estimation as a decision problem:
  - Define loss function $L : \Theta \times \Theta \to \mathbb{R}$ that specifies cost of estimating $\hat{\theta}$ when true parameter was θ
    Loss function satisfies $L(\hat{\theta}, \theta) \geq 0$ with equality if and only if $\hat{\theta} = \theta$.
  - Point estimate is decision rule which minimizes expected loss, also called *Bayes risk*:

    $$\hat{\theta}_{Bayes} = \mathtt{argmin}_{\hat{\theta}(D)} \mathbb{E}[L(\hat{\theta}(D), \theta)]$$

    Note: Expectation is taken over joint distribution $p(\theta, D) = p(\theta)p(D|\theta)$

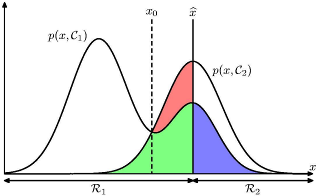
Point estimates

Common loss functions include

- 0-1-loss (for categorical values):

$$L(\hat{\theta}, \theta) = \begin{cases} 1 & \text{if } \hat{\theta} = \theta \\ 0 & \text{if } \hat{\theta} \neq \theta \end{cases}$$

Minimum Misclassification Rate



$$\begin{aligned} p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\,\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\,\mathrm{d}\mathbf{x}. \end{aligned}$$

$$\begin{aligned} \theta &\in \{\mathcal{C}_1, \mathcal{C}_2\} \\ \hat{\theta}(x) &= \begin{cases} \mathcal{C}_1, & x \in \mathcal{R}_1 \\ \mathcal{C}_2, & x \in \mathcal{R}_2 \end{cases} \\ \mathbb{E}[L(\hat{\theta}(x), \theta)] &= \sum_{\theta} \int p(x, \theta) L(\hat{\theta}(x), \theta) dx \\ &= \int_{\mathcal{R}_1} p(x, \mathcal{C}_2) dx \\ &+ \int_{\mathcal{R}_2} p(x, \mathcal{C}_1) dx \end{aligned}$$

Point estimates

Common loss functions include

- 0-1-loss (for categorical values):

$$L(\hat{\theta}, \theta) = \begin{cases} 1 & \text{if } \hat{\theta} = \theta \\ 0 & \text{if } \hat{\theta} \neq \theta \end{cases}$$

$\implies \hat{\theta}_{Bayes}$ is posterior mode

- Squared error:

$$L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$$

$\implies \hat{\theta}_{Bayes}$ is posterior mean

- Absolute error:

$$L(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$$

$\implies \hat{\theta}_{Bayes}$ is posterior median

Point estimates

Posterior mean minimizes mean squared error, i.e.

$$\texttt{argmin}_{\hat{\theta}(D)}\mathbb{E}_{p(\theta,D)}[(\hat{\theta}(D)-\theta)^2] = \mathbb{E}_{p(\theta|D)}[\theta]$$

Proof:

$$\begin{aligned}\mathbb{E}_{p(\theta,D)}[(\hat{\theta}(D)-\theta)^2] &= \int p(\theta,D)(\hat{\theta}(D)-\theta)^2 d\theta dD \\ &= \int p(D)\int p(\theta|D)(\hat{\theta}(D)-\theta)^2 d\theta\ dD\end{aligned}$$

Now, for each D:

$$\begin{aligned}\frac{\partial}{\partial\hat{\theta}}\int p(\theta|D)(\hat{\theta}-\theta)^2 d\theta &= 2\int p(\theta|D)(\hat{\theta}-\theta)d\theta \\ &= 2\left(\int p(\theta|D)\hat{\theta}d\theta - \int p(\theta|D)\theta d\theta\right) \\ &= 2\left(\hat{\theta} - \int p(\theta|D)\theta d\theta\right)\end{aligned}$$

Setting derivative to zero, it follows that

$$\hat{\theta}(D) = \int p(\theta|D)\theta d\theta = \mathbb{E}_{p(\theta|D)}[\theta]$$

Point estimates

Back to coin example:

- Posterior is Beta distribution with parameters $\alpha + k$ and $\beta + (n - k)$
- Posterior mean is

$$\frac{\alpha + k}{\alpha + k + \beta + (n - k)} = \frac{\alpha + k}{\alpha + \beta + n}$$

- Defining $m = \alpha + \beta$ posterior mean can be written as

$$\frac{m}{n+m}\frac{\alpha}{\alpha+\beta} + \frac{n}{n+m}\frac{k}{n}$$

  - Convex combination between prior mean $\theta_0 = \frac{\alpha}{\alpha+\beta}$ and ML estimate $\hat{\theta}_{ML} = \frac{k}{n}$
  - Weights correspond to relative number of (pseudo-)observations

Likelihood principle

Summary of Bayesian estimation:

- Posterior combines information from prior and likelihood of data:
$$\textit{posterior} \propto \textit{prior} \times \textit{likelihood}$$
- Prior represents (subjective) belief/information before data are obtained
Sequential learning: Prior can arise from data learned about previously

Likelihood principle:

Data enters via likelihood $p(D|\theta)$ only, i.e. inference from D_1 and D_2 is the same when $p(D_1|\theta) = p(D_2|\theta)$

Likelihood principle

Summary of Bayesian estimation:

- Posterior combines information from prior and likelihood of data:

$$\textit{posterior} \propto \textit{prior} \times \textit{likelihood}$$

- Prior represents (subjective) belief/information before data are obtained
  Sequential learning: Prior can arise from data learned about previously

Likelihood principle:

Data enters via likelihood $p(D|\theta)$ only, i.e. inference from D_1 and D_2 is the same when $p(D_1|\theta) = p(D_2|\theta)$

Q: Would your estimate change if I told you that

- coin was tossed 20 times
- or coin was tossed until 15 heads were obtained?

Binomial distribution

Recall coin tossing example:

- Binomial distribution:

$$p(k|n,\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

  - Sampling space: $\mathcal{K} = \{0,\ldots,n\}$
  - Paramter space: $\Theta = [0,1]$
  - Mean: $n\theta$
  - Variance: $n\theta(1-\theta)$

- Conjugate prior is Beta distribution:

$$p(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Poisson distribution

Used to model number of independent events occuring in unit time interval:

$$p(k|\lambda) = e^{-\lambda}\frac{\lambda^k}{k!}$$

- Sample space: $\mathcal{K} = \mathbb{N}$
- Parameter space: $\Lambda = \mathbb{R}_{>0}$
  Positive *rate* parameter λ
- Mean: λ
- Variance: λ
- Arises as limit of binomial distribution with $\lambda = \lim_{n\to\infty} n\theta$

Q: Can you find the conjugate prior?

Gamma distribution

Conjugate prior for rate parameter λ of Poisson distribution:

$$p(\lambda|\alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

- Sampling space: $\Lambda = (0, \infty)$
- Parameter space: $\alpha > 0$ (shape) and $\beta > 0$ (rate)
- Mean: $\frac{\alpha}{\beta}$
- Variance: $\frac{\alpha}{\beta^2}$

Normal distribution

$$p(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

- Sampling space: $\mathcal{X} = \mathbb{R}$
- Parameter space: $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}_{>0}$
- Mean: μ
- Variance: σ^2

Normal distribution

Often, it is more convenient to use a different parametrization:

$$p(x|\mu, \tau) = \left(\frac{\tau}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{1}{2}\tau(x-\mu)^2}$$

with *precision* $\tau = \frac{1}{\sigma^2}$
Next, we consider

- Estimating μ when τ is known
- Estimating τ when μ is known
- Estimating μ and τ together

Mean estimation

Conjugate prior for μ is again a normal distribution:

$$p(\mu|\mu_0, \tau_0) = \left(\frac{\tau_0}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{1}{2}\tau_0(\mu-\mu_0)^2}$$

Posterior given data $D = (x_1, \ldots, x_N)$ is found by "completing the square":

$$\begin{aligned}
p(\mu|D) &\propto \prod_{i=1}^{N} p(x_i|\mu, \tau)p(\mu|\mu_0, \tau_0) \\
&\propto e^{-\frac{1}{2}\tau\sum_i(x_i-\mu)^2} e^{-\frac{1}{2}\tau_0(\mu-\mu_0)^2} \\
&= e^{-\frac{1}{2}(\tau\sum_i x_i^2 - 2\tau\mu\sum_i x_i + N\tau\mu^2 + \tau_0\mu^2 - 2\tau_0\mu\mu_0 + \tau_0\mu_0^2)} \\
&\propto e^{-\frac{1}{2}((N\tau+\tau_0)\mu^2 - 2(\tau\sum_i x_i + \tau_0\mu_0)\mu)} \\
&\propto e^{-\frac{1}{2}(N\tau+\tau_0)(\mu-\frac{\tau\sum_i x_i+\tau_0\mu_0}{N\tau+\tau_o})^2}
\end{aligned}$$

Mean estimation

Conjugate prior for μ is again a normal distribution:

$$p(\mu|\mu_0, \tau_0) = \left(\frac{\tau_0}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{1}{2}\tau_0(\mu-\mu_0)^2}$$

Posterior given data $D = (x_1, \ldots, x_N)$ is found by "completing the square":

$$\begin{aligned} p(\mu|D) &\propto \prod_{i=1}^{N} p(x_i|\mu,\tau)p(\mu|\mu_0,\tau_0) \\ &\propto e^{-\frac{1}{2}(N\tau+\tau_0)(\mu-\frac{\tau\sum_i x_i+\tau_0\mu_0}{N\tau+\tau_o})^2} \end{aligned}$$

Normal distribution with

- precision $\tau_D = N\tau + \tau_0$
- mean

$$\mu_D = \frac{\tau \sum_i x_i + \tau_0\mu_0}{N\tau + \tau_o} = \frac{1}{\tau_D}(N\tau\hat{\mu} + \tau_0\mu_0)$$

where $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$

Precision estimation

Now, assume that the mean μ is know and we want to infer the precision τ:

$$
\begin{aligned}
p(\tau|D,\mu) &\propto p(\tau)\prod_{i=1}^{N} p(x_i|D,\mu,\tau) \\
&= p(\tau)\left(\frac{\tau}{2\pi}\right)^{\frac{N}{2}} e^{-\frac{1}{2}\tau\sum_i (x_i-\mu)^2}
\end{aligned}
$$

Thus, τ occurs in the likelihood as $\tau^{\frac{N}{2}} e^{-\frac{1}{2}\sum_i (x_i-\mu)^2\tau}$ and the conjugate prior is a gamma distribution:

$$
\begin{aligned}
p(\tau|D,\mu) &\propto \tau^{\alpha-1} e^{-\beta\tau}\tau^{\frac{N}{2}} e^{-\frac{1}{2}\sum_i (x_i-\mu)^2\tau} \\
&= \tau^{\alpha'-1} e^{-\beta'\tau}
\end{aligned}
$$

where $\alpha' = \alpha + \frac{N}{2}$ and $\beta' = \beta + \frac{1}{2}\sum_i (x_i - \mu)^2$ are the posterior Gamma parameters

Student's t-distribution

$$p(x|\mu,\sigma^2,\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi\sigma^2}\Gamma(\frac{\nu}{2})}\left(1+\frac{1}{\nu}\frac{(x-\mu)^2}{\sigma^2}\right)^{-\frac{\nu+1}{2}}$$

- Sampling space: $\mathcal{X} = \mathbb{R}$
- Parameter space: $\mu \in \mathbb{R}$ (mean), $\sigma^2 > 0$ (variance) and $\nu > 0$ (degrees of freedom)
- Mean: μ for $\nu > 1$
- Variance: $\sigma^2 \frac{\nu}{\nu-2}$ for $\nu > 2$

Marginal distribution of Gaussian with unknown precision:

$$\begin{aligned} p(x|\mu,\alpha,\beta) &= \int_0^\infty \mathcal{N}(x|\mu,\tau)\texttt{Gamma}(\tau|\alpha,\beta)d\tau \\ &= \texttt{Stud}(x|\mu,\frac{\beta}{\alpha},2\alpha) \end{aligned}$$

Joint estimation

To compute the joint posterior $p(\mu, \tau|D)$ we proceed similarly and investigate the form of the likelihood:

$$\begin{aligned}
p(D|\mu,\tau) &\propto \tau^{\frac{N}{2}} e^{-\frac{1}{2}\sum_i (x_i-\mu)^2\tau} \\
&= \tau^{\frac{N}{2}} e^{-\frac{1}{2}\tau(\sum_i x_i^2 - 2\mu\sum_i x_i + N\mu^2)} \\
&= \tau^{\frac{N}{2}} e^{-\frac{1}{2}(\sum_i x_i^2 - N\hat{\mu}^2)\tau} e^{-\frac{1}{2}N\tau(\mu-\hat{\mu})^2} \\
&= \tau^{\frac{N}{2}} e^{-\frac{1}{2}(\sum_i (x_i-\hat{\mu})^2)\tau} e^{-\frac{1}{2}N\tau(\mu-\hat{\mu})^2}
\end{aligned}$$

where $\hat{\mu} = \frac{1}{N}\sum_i x_i$
Considering the above form as the product $p(\tau)p(\mu|\tau)$ we can identify a suitable prior:

$$p(\tau)p(\mu|\tau) \propto \tau^{\frac{\eta}{2}} e^{-\beta\tau} e^{-\frac{1}{2}\eta\tau(\mu-\mu_0)^2},$$

i.e. the product of a Gamma prior $p(\tau|\alpha, \beta)$ and a normal prior $p(\mu|\mu_0, \tau')$ where $\alpha = \frac{\eta}{2} + 1$ and $\tau' = \eta\tau$.

Joint estimation

Using **Python**, we can illustrate the joint inference as follows:

```
import numpy as np
from scipy.stats import gamma, norm

def normal_gamma (mu, tau, mu0, eta, beta):
  return gamma.pdf(tau, eta/2. + 1, scale=1./beta) \
       * norm.pdf(mu, loc=mu0, scale=1./sqrt(eta*tau))

N = 10
x = norm.rvs(size=N, loc=0.5, scale=0.25)
mu_hat = np.mean(x)
alpha_0, beta_0 = 1, 1
mu_0 = 0.
eta_0 = 2*(alpha_0 -1); eta_D = eta_0 + N;
mu_D = 1./eta_D*(eta_0*mu_0 + N*mu_hat)
beta_D = beta_0 + 0.5*(np.sum((x-mu_hat)**2) + eta_0/eta_D*(mu_hat-mu_0)**2)

mu, tau = np.meshgrid(np.linspace(-0.5, 1.5, num=100), np.linspace(1, 35, num=100))
plt.contour(mu, tau, normal_gamma (mu, tau, mu_D, eta_D, beta_D))
plt.scatter(np.mean(x), 1./np.var(x))
```

- Mean and standard deviation are not independent
- Normal-Gamma distribution often specified with separate parameters η and α

Multi-variate normal

Generalizes normal distribution to d dimensions:

$$p(\mathbf{x}|\mu, \Sigma) = (2\pi)^{-\frac{d}{2}} \det|\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^t \Sigma^{-1} (\mathbf{x}-\mu)}$$

- Sampling space: $\mathbf{x} \in \mathbb{R}^d$
- Parameter space: $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$
- Mean: $\mathbb{E}[\mathbf{x}] = \mu$, i.e. $\mathbb{E}[x_i] = \mu_i$
- Covariance matrix: $\mathbb{E}[(x_i - \mathbb{E}[x_i])(x_j - \mathbb{E}[x_j])] = \Sigma_{ij}$
  Matrix notation: $\Sigma = \mathbb{E}[(\mathbf{x} - \mu)(\mathbf{x} - \mu)^t]$

Multi-variate normal

Properties of the multi-variate normal distribution:

- Precision matrix: $\Lambda = \Sigma^{-1}$
- With diagonal covariance matrix, i.e. $\Sigma = \mathtt{diag}(\sigma_1^2, \ldots, \sigma_d^2)$:

$$p(\mathbf{x}|\mu, \Sigma) = \prod_{i=1}^{d} (2\pi\sigma_i^2)^{-\frac{1}{2}} e^{-\frac{1}{2}\frac{(x_i-\mu_i)^2}{\sigma_i^2}},$$

  i.e. reduces to a product of d 1-dimensional Gaussians
- Iso-probability lines are ellipses defined by the quadratic form

$$(\mathbf{x} - \mu)^t \Sigma^{-1} (\mathbf{x} - \mu) = \mathtt{const}$$

  Most easily seen in 2-dimensions (assuming $\mu = \mathbf{0}$):
  - Diagonal covariance:

$$(x_1 x_2) \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \frac{x_1^2}{\sigma_1^2} + \frac{x_2^2}{\sigma_2^2} = \mathtt{const}$$

    Recall: $x_1^2 + x_2^2 = 1$ defines unit circle
  - General covariance matrix rotates axis. Elongation of ellipses is longest when precision is low

Multi-variate normal

Properties of the multi-variate normal distribution:

- Precision matrix: $\Lambda = \Sigma^{-1}$
- With diagonal covariance matrix, i.e. $\Sigma = \mathtt{diag}(\sigma_1^2, \ldots, \sigma_d^2)$:

$$p(\mathbf{x}|\mu, \Sigma) = \prod_{i=1}^{d} (2\pi\sigma_i^2)^{-\frac{1}{2}} e^{-\frac{1}{2}\frac{(x_i-\mu_i)^2}{\sigma_i^2}},$$

  i.e. reduces to a product of d 1-dimensional Gaussians
- Iso-probability lines are ellipses defined by the quadratic form

$$(\mathbf{x} - \mu)^t \Sigma^{-1} (\mathbf{x} - \mu) = \mathtt{const}$$

- Consider $\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}$. Then,
  - Marginal distribution $p(\mathbf{x}_a)$ is again Gaussian with mean μ_a and covariance matrix Σ_{aa}
  - Conditional distribution $p(\mathbf{x}_a|\mathbf{x}_b)$ is again Gaussian with mean $\mu_{a|b} = \mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(\mathbf{x}_b - \mu_b)$ and precision $\Lambda_{a|b} = \Lambda_{aa}$

Multi-variate normal

Show that the conditional distribution $p(\mathbf{x}_a|\mathbf{x}_b)$ is Gaussian:

$$
\begin{aligned}
p(\mathbf{x}_a|\mathbf{x}_b) &\propto p(\mathbf{x}_a, \mathbf{x}_b) \\
&\propto \exp\left\{-\frac{1}{2}\begin{pmatrix} \mathbf{x}_a - \mu_a \\ \mathbf{x}_b - \mu_b \end{pmatrix}^t \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix} \begin{pmatrix} \mathbf{x}_a - \mu_a \\ \mathbf{x}_b - \mu_b \end{pmatrix}\right\} \\
&\propto \exp\left\{-\frac{1}{2}(\mathbf{x}_a^t \Lambda_{aa} \mathbf{x}_a - 2\mathbf{x}_a^t(\Lambda_{aa}\mu_a - \Lambda_{ab}(\mathbf{x}_b - \mu_b)))\right\} \\
&\propto \exp\left\{-\frac{1}{2}(\mathbf{x}_a - \mu_{a|b})^t \Lambda_{aa}(\mathbf{x}_a - \mu_{a|b})\right\}
\end{aligned}
$$

Wishart distribution

$$p(\mathbf{X}|\mathbf{V}, n) = \frac{1}{2^{\frac{np}{2}} \det|\mathbf{V}|^{\frac{n}{2}} \Gamma_p(\frac{n}{2})} \det|\mathbf{X}|^{\frac{n-p-1}{2}} e^{-\frac{1}{2}\mathtt{tr}(\mathbf{V}^{-1}\mathbf{X})}$$

where $\Gamma_p(\frac{n}{2}) = \pi^{\frac{p(p-1)}{4}} \prod_{j=0}^{p-1} \Gamma(\frac{n-j}{2})$

- Sampling space: $\mathbf{X} \in \mathbb{R}^{p \times p}$
- Parameter space: $\mathbf{V} \in \mathbb{R}^{p \times p}$ and degrees of freedom $n > p + 1$
- Mean: $n\mathbf{V}$
- Variance: $n(V_{ij}^2 + V_{ii}V_{jj})$

Wishart distribution

- Conjugate prior for the precision Λ of multi-variate Gaussian with known mean μ
- Sampling distribution of empirical covariance matrix, i.e. $\widehat{\Sigma} = \sum_{i=1}^{N} \mathbf{x}\mathbf{x}^t \sim \texttt{Wishart}(\Sigma, n)$ when $x_i \sim \texttt{Normal}(\mathbf{0}, \Sigma)$
- Posterior $p(\Lambda|D)$ is Wishart distribution with

$$\Lambda_D = (\Lambda_0^{-1} + \widehat{\Sigma})^{-1} \text{ and } n_D = n + n_0$$

- Unknown mean and precision:

$$p(\mu, \Lambda) = \texttt{Normal}(\mu|\mu_0, (\eta\Lambda)^{-1})\texttt{Wishart}(\Lambda|\Lambda_0, n_0)$$

Normal-Wishart distribution as conjugate prior

Chosing priors

How should we choose a suitable prior?

- Mathematical convenience: Conjugate prior
- Let the data speak: Uninformative priors
- Invariance principle: Jeffreys prior

Conjugate priors

Conjugate priors are often a good choice as

- Posterior is analytically tractable
- Parameters often interpretable as pseudo-observations

But, can correspond to strong assumptions that are hard to justify

Uninformative priors

What to do if we have little information a-priori?

- Natural guess: Uniform prior
  But: Choice depends on parametrization:
  - Consider parameter θ with prior $p(\theta)$
  - Reparametrize as $\eta = h(\theta)$:

$$
\begin{aligned}
p(\eta) &= p(\theta)|\frac{d\theta}{d\eta}| \\
&= p(h^{-1}(\eta))|\frac{d}{d\eta}h^{-1}(\eta)|
\end{aligned}
$$

Example: $\sigma = e^{\theta}$ with $p(\theta) \propto 1$:

$$
p(\sigma) \propto 1|\frac{d}{d\sigma}\log\sigma| = \frac{1}{\sigma}
$$

Uninformative priors

What to do if we have little information a-priori?

- Natural guess: Uniform prior
  But: Choice depends on parametrization:
- Better: Invariance principle
  - Prior for *location parameter* μ

    $$p(x|\mu) = f(x - \mu) = p(x + c|\mu + c) \quad \forall c$$

    should be *translation invariant*, i.e.

    $$\int_a^b p(\mu) d\mu = \int_{a+c}^{b+c} p(\mu) d\mu = \int_a^b p(\mu + c) d\mu \quad \forall a, b, c$$

    Thus, $p(\mu) \propto 1$ uniform

Uninformative priors

What to do if we have little information a-priori?

- Natural guess: Uniform prior
  But: Choice depends on parametrization:
- Better: Invariance principle
  - Prior for *scale parameter* σ

    $$p(x|\sigma) = \frac{1}{\sigma} f(\frac{x}{\sigma})$$

    should be *scale invariant*, i.e.

    $$\int_a^b p(\sigma) d\sigma = \int_{ca}^{cb} p(\sigma) d\sigma = \int_a^b \frac{1}{c} p(\frac{\sigma}{c}) d\sigma \quad \forall a, b, c$$

    Thus, $p(\sigma) \propto \frac{1}{\sigma}$
    Note: Previous slide shows that $p(\log \sigma)$ uniform!

Improper priors

Prior $p(\theta)$ is called *improper* if it cannot be normalized, i.e.

$$\int_{\mathbb{R}} p(\theta) d\theta = \infty$$

- Uninformative priors $p(\mu) \propto 1, p(\sigma) \propto \frac{1}{\sigma}$ are improper
- Bayesian inference is still valid if posterior can be normalized:
  - Estimate mean of Gaussian sample $p(\mu|D)$
  - Conjugate prior $p(\mu|\mu_0, \tau_0)$ uniform for $\tau_0 \to 0$
  - Posterior $p(\mu|\mu_D, \tau_D)$ well defined in this limit:

$$\tau_D = N\tau \text{ and } \mu_D = \hat{\mu}$$

Note: Posterior depends on data only

Jeffreys prior

Idea: Prior density should be unchanged under reparametrization:

$$p(\theta) \propto \sqrt{I(\theta)}$$

where $I(\theta) = \mathbb{E}[(\frac{d}{d\theta} \log p(x|\theta))^2]$ denotes the Fisher information

- Consider $\eta = h(\theta)$

$$\begin{aligned} p(\eta) &= p(\theta)\left|\frac{d\theta}{d\eta}\right| \\ &\propto \sqrt{\mathbb{E}[(\frac{d}{d\theta} \log p(x|\theta))^2](\frac{d\theta}{d\eta})^2} \\ &= \sqrt{\mathbb{E}[(\frac{d}{d\theta} \log p(x|\theta)\frac{d\theta}{d\eta})^2]} \\ &= \sqrt{I(\eta)} \end{aligned}$$

Jeffreys prior

Examples:

- Gaussian distribution $p(x|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$:

$$
\begin{aligned}
p(\mu) &\propto \sqrt{\mathbb{E}[(\frac{d}{d\mu}\log p(x|\mu))^2]} \\
&= \sqrt{\mathbb{E}[(\frac{x-\mu}{\sigma^2})^2]} \\
&= \sqrt{\frac{\sigma^2}{\sigma^4}} \propto 1
\end{aligned}
$$

- Binomial distribution $p(k|\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$.
  Jeffreys prior is Beta distribution with $\alpha = \beta = \frac{1}{2}$

Some remarks

- Uninformative priors are often employed
  - Often limits of conjugate priors with analytic posterior
  - Inference resembles maximum likelihood estimates
- Not a good idea in high-dimensions
  - Where is the probability mass of a standard Gaussian in $\mathbb{R}^d$?
    - Thin shell around origin at radius $\sqrt{d}$.
    - Thus, uniformative prior puts infinite mass on infinity!
  - Inference in high-dimensions profits from informed priors:
    - Recall James-Stein phenomenon
    - Power of shrinkage and penalized maximum likelihood

Exchangeability

More on priors:

- Can existence be motivated from first principles?
- Structural assumptions expressed by priors?

Exchangeability

More on priors:

- Can existence be motivated from first principles?
- Structural assumptions expressed by priors?

Exchangeability:

- Expresses symmetries of probabilistic models
- Derives representation in terms of latent variables

Exchangeability

Consider an (infinite) sequence $X_1, X_2, \ldots$ of random variables.

Definition

A sequence $X_1, X_2, \ldots$ of random variables is called (infinitely) exchangeable *if*

$$(X_1, X_2, \ldots, X_n) \stackrel{d}{=} (X_{\pi(1)}, X_{\pi(2)}, \ldots, X_{\pi(n)})$$

for all $n \in \mathcal{N}$ and all permutations π of $1, \ldots, n$.

Intuition: Order of the sequence does not matter ...

- Sequences of independent and identically distributed random variables (i.i.d.) are exchangeable,
- but **not** every exchangeable sequence is i.i.d.

Pólya urn

Consider an urn containing r red and b blue balls. Now repeat the following process:

- Draw a ball at random and note its color
- Place it back *together with an additional ball* of the same color

Obviously, consecutive draws are **not** independent, but the process is exchangeable:

$$\begin{aligned} p(\mathtt{b}, \mathtt{b}, \mathtt{r}) &= \frac{b}{r+b}\frac{b+1}{r+b+1}\frac{r}{r+b+2} \\ p(\mathtt{b}, \mathtt{r}, \mathtt{b}) &= \frac{b}{r+b}\frac{r}{r+b+1}\frac{b+1}{r+b+2} \end{aligned}$$

De Finetti theorem

De Finetti representation theorem:

Theorem

A binary sequence $X_1, X_2, \ldots$ is exchangeable if and only if there exists a measure μ on $[0,1]$ such that for all n

$$p(x_1, \ldots, x_n) = \int_\theta \theta^{t_n}(1-\theta)^{n-t_n} d\mu(\theta)$$

where $t_n = \sum_{i=1}^n x_i$.

Further,

- Given θ the sequence is i.i.d. Bernoulli distributed, i.e.

$$p(x_1, \ldots, x_n|\theta) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}$$

  Think of θ as the bias of a coin.

De Finetti theorem

De Finetti representation theorem:

Theorem

A binary sequence $X_1, X_2, \ldots$ is exchangeable if and only if there exists a measure μ on $[0,1]$ such that for all n

$$p(x_1, \ldots, x_n) = \int_\theta \theta^{t_n}(1-\theta)^{n-t_n} d\mu(\theta)$$

where $t_n = \sum_{i=1}^n x_i$.

Further,

- Given θ the sequence is i.i.d. Bernoulli distributed.
- A law of large numbers holds, i.e.

$$\lim_{n\to\infty} \sum_{i=1}^n \frac{X_i}{n} \sim \mu$$

De Finetti theorem

De Finetti's representation theorem can be interpreted in several ways:

- *Frequentist:* There is an unknown θ such that $X_1, X_2, \ldots$ are i.i.d. $\texttt{Bernoulli}(\theta)$ distributed. θ is the limiting frequency of observed 1's (heads).
- *Bayesian:* Exchangeable distribution P expresses beliefs/assumptions about $X_1, X_2, \ldots$. Observed distribution is permutation invariant and $\mu(\theta)$ is the subjective *prior* about the coin bias θ.
- *Preferred:* Model observations $X_1, X_2, \ldots$ as a-priori alike
  - Implies a decomposition into structure θ (coin bias) and randomness $p(X_i|\theta)$ (coin flip)
  - Structural model is an (implicit) assumption $p(\theta)$.