Definitions

Shannon Entropy: $H(p) = -\int p(x) \ln p(x) dx$ Cross Entropy: $H(p,q) = -\int p(x) \ln q(x)$

Kullback-Leibler Divergence: $\mathrm{KL}(p\|q) = \int p(x) \ln \frac{p(x)}{q(x)} dx$ Mutual Information: $I(X,Y) = \mathrm{KL}(p(X,Y)\|p(X)p(Y))$

Exponential Family: $p(x|\theta) = h(x) \exp\left(\sum_{j=1}^{k} \zeta_j(\theta) T_j(x)\right) C(\theta) =$

 $h(x) \exp (\eta^{\top} T(x) - \Psi(\eta))$, where $\eta = \zeta(\theta)$ are called the natural parameters.

Choice of Priors

Subjective Priors

Conjugate Priors: If \mathcal{P} is in exponential family in the 2nd form in Section Definitions, the conjugate family over Θ , parameterised by μ and λ , is then

$$\mathcal{F} = \{\pi(\theta|\mu,\lambda)\} \propto \exp\left(\zeta(\theta)^{\top}\mu - \lambda\Psi(\zeta(\theta))\right) : \mu \in \mathbb{R}^k, \lambda \in \mathbb{R}^+.$$

For $\pi = \pi(\theta|\mu_0, \lambda_0)$, the posterior is $\pi(\theta|\mu_0 + T(x), \lambda_0 + 1)$.

Non-Informative Priors

- Laplace Priors: $\pi(\theta) \propto \text{const.}$ Maximising Shannon entropy $H(\pi) = \sum_{i=1}^{n} \pi(\theta_i) \ln \pi(\theta_i)$.
- Jeffreys Priors: $\pi(\theta) \propto \sqrt{\det(I(\theta))}$, $I(\theta) := -\mathbb{E}_{p(x|\theta)} \frac{\partial^2 \ln p(x|\theta)}{\partial^2 \theta}$. Derived from parameterisation invariant metric $I_2(p(\cdot|\theta_1), p(\cdot|\theta_2)) := \mathrm{KL}(p(\cdot|\theta_1), p(\cdot|\theta_2)) + \mathrm{KL}(p(\cdot|\theta_2), p(\cdot|\theta_1)) \approx (\theta_1 \theta_2)^{\top} I(\theta)(\theta_1 \theta_2)$.
- Reference Priors: Information gain is $I(\mathcal{P}, \pi) = \int p(x) \mathrm{KL}(\pi(\cdot|x) \| \pi(\cdot)) dx$, an "average" gain over all data space \mathcal{X} . By rewriting it as $I(\mathcal{P}, \pi) = H(\pi) \int p(x) H(\cdot|x) dx$, we want more information in posterior and less in prior.

Decision Theory

In additional to Bayes model, we have

- Decision (space): $d \in \mathcal{D}$. e.g. $\mathcal{D} = \Theta$ for parameter inference, $\mathcal{D} = \mathcal{X}$ for prediction, or $\mathcal{D} = \{0, 1\}$ for testing, etc.
- Decision Rule: $\delta : x \in \mathcal{X} \to d = \delta(x) \in \mathcal{D}$.
- Loss Function: $L : (\theta, d) \in \Theta \times \mathcal{D} \to L(\theta, d) \in \mathbb{R}$. Here θ is the true parameter.

Expectations of Loss

- Risk: $R(\theta, \delta) := \int L(\theta, \delta(x)) p(x|\theta) dx$
- Bayes Risk $r(\pi, \delta) := \int R(\theta, \delta) \pi(\theta) d\theta$ (further integrated over π)
- Posterior Integrated Loss $\rho(\pi, d|x) := \int L(\theta, d)\pi(\theta|x)d\theta$

Decision Rules

- Inadmissible decision rule δ : There exists another rule which is no worse than δ for all $\theta \in \Theta$, and is strictly better than δ for at least one $\theta \in \Theta$, measured by risk. (Need to be compared for every θ .)
- Bayesian decision rule δ^{π} : δ^{π} that minimises the expected risk $\delta^{\pi} := \underset{\delta}{\arg\min} \ r(\pi, \delta)$. (For each $x \in \mathcal{X}$ we can find the decision $d \in \mathcal{D}$ using $\delta^{\pi}(x) = \underset{d}{\arg\min} \rho(\pi, d|x)$).
- Minimax decision rule δ^* : δ^* that minimises the maximum risk for $\theta \in \Theta$, where δ^* is searched within the expanded randomised decision rule space.

Bridges between Frequentist and Bayesian

- Bayes estimators are admissible under (reasonable) conditions: 1. π does not exclude any θ ; 2. $r(\delta, \pi)$ for all δ are bounded; 3. $R(\theta, \delta)$ is continuous for $\theta \in \Theta$.
- Bayes estimator associated with the least favourable prior π_0 , which is defined by sup $r(\pi) = r(\pi_0)$, is a minimax estimator.
- If Bayes rule δ^{π} has a constant risk $R(\theta, \delta^{\pi}) = \text{const.}, \ \delta^{\pi}$ is minimax.

Asymptotic Theory

Strong Consistency Assume a Bayes model $\{\mathcal{P}, \pi\}$ and random variable $X_n \sim P_{\theta_0}^n$. The sequence of posteriors $\pi^n(\theta|x_n)$ is called strongly consistent at θ_0 iff for any open subset $O \subset \Theta$ with $\theta_0 \in O$ it holds that $\Pr^n(\theta \in O|x_n) \to 1$ as $n \to \infty$. (Alternatively we can show this by proving $\mathbb{E}(\theta|x_n) = \theta_0$ and $\mathbb{V}(\theta|x_n) = 0$ as $n \to \infty$)

Asymptotic Behaviours of Consistent Priors Assume $X_n \sim P_{\theta_0}^n$ and two priors π_1 and π_2 w.r.t model \mathcal{P} which both have strongly consistent posteriors, then $\sup_A |\operatorname{Pr}_n^{\pi_1}(A|x_n) - \operatorname{Pr}_n^{\pi_2}(A|x_n)| \to 0$ for $n \to \infty$. a.s. Then sup term is called total distribution distance. (Priors with consistent posteriors have identical asymptotic behaviours.)

Bayesian Linear Model

Model of the form $\mathcal{P} = \{ \mathcal{N}_n(\mathbf{X}\beta, \sigma^2 \Sigma) \colon \beta \in \mathbb{R}^p, \sigma^2 \in \mathbb{R}^+ \}$ where **X** is a matrix of known constants and Σ is known. Parameters are $\theta = \{\beta, \sigma^2\}$ or $\theta = \beta$ with σ^2 known. Also, **X** is of full rank: $r(\mathbf{X}) = p$.

Conjugate Prior π_c

Case 1: $\theta = \beta$ with σ^2 known. Formulation: $\mathbf{y}|\theta \sim \mathcal{N}_n(\mathbf{X}\beta, \Sigma), \beta \sim \mathcal{N}_p(\gamma, \Gamma)$

Results:

$$\begin{split} \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{y} \end{pmatrix} &\sim \mathcal{N}_{n+p} \left(\begin{pmatrix} \boldsymbol{\gamma} \\ \mathbf{X} \boldsymbol{\gamma} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Gamma} & \boldsymbol{\Gamma} \mathbf{X}^{\top} \\ \mathbf{X} \boldsymbol{\Gamma} & \boldsymbol{\Sigma} + \mathbf{X} \boldsymbol{\Gamma} \mathbf{X}^{\top} \end{pmatrix} \right) \\ \mathbf{y} &\sim \mathcal{N}_{n} (\mu_{y}, \boldsymbol{\Sigma}_{y}) \\ \boldsymbol{\beta} | \boldsymbol{y} &\sim \mathcal{N}_{p} (\mu_{\beta|y}, \boldsymbol{\Sigma}_{\beta|y}) \end{split}$$

where

$$\mu_{y} = \mathbf{X}\gamma \qquad \qquad \mu_{\beta|y} = \gamma + \Gamma \mathbf{X}^{\top} (\Sigma + \mathbf{X}\Gamma \mathbf{X}^{\top})^{-1} (y - \mathbf{X}\gamma)$$

$$\Sigma_{y} = \Sigma + \mathbf{X}\Gamma \mathbf{X}^{\top} \qquad \Sigma_{\beta|y} = \Gamma - \Gamma \mathbf{X}^{\top} (\Sigma + \mathbf{X}\Gamma \mathbf{X}^{\top})^{-1} \mathbf{X}\Gamma.$$

Case 2: $\theta = \{\beta, \sigma^2\}$ Formulation:

$$\sigma^{2} \sim \mathcal{IG}(a/2, b/2)$$
$$\beta | \sigma^{2} \sim \mathcal{N}_{p}(\gamma, \sigma^{2}\Gamma)$$
$$\mathbf{y} | \beta, \sigma^{2} \sim \mathcal{N}_{n}(\mathbf{X}\beta, \sigma^{2}\Sigma)$$

where $\mathcal{IG}(a/2, b/2)$ denotes inverse gamma distribution with parameters a/2 and b/2. The prior of $\theta = (\beta, \sigma^2)^{\top}$ is normal-inverse gamma $\mathcal{NIG}(a, b, \gamma, \Gamma)$.

Results:

Joint posterior: $(\beta, \sigma^2)|y \sim \mathcal{NIG}(a_1, b_1, \gamma_1, \Gamma_1)$ where

$$a_{1} = a + n b_{1} = b + (y - X\gamma)^{\top} (\Sigma + X\Gamma X^{\top})^{-1} (y - X\gamma)$$
$$\gamma_{1} = \gamma + \Gamma X^{\top} (\Sigma + X\Gamma X^{\top})^{-1} (y - X\gamma)$$
$$\Gamma_{1} = \Gamma - \Gamma X^{\top} (\Sigma + X\Gamma X^{\top})^{-1} X\Gamma.$$

Marginal posterior: For $\pi(\beta|\mathbf{y})$ or $\pi(\sigma^2|\mathbf{y})$, use properties of normal-inverse gamma distribution on the joint posterior as below. Normal-Inverse Gamma Property: Assume $(\beta, \sigma^2) \sim \mathcal{NIG}(a, b, \gamma, \Gamma)$.

Normal-Inverse Gamma Property: Assume $(\beta, \sigma^2) \sim \mathcal{NIG}(a, b, \gamma, \Gamma)$, we have $\beta \sim t_p(a, \gamma, b\Gamma/a)$ and $\sigma^2 \sim \mathcal{IG}(a/2, b/2)$, where t_p denotes student-t distribution with degree of freedom p.

Joint \mathbf{y} and σ^2 : $(\mathbf{y}, \sigma^2)^{\top} \sim \mathcal{NIG}(a, b, \mathbf{m}, \mathbf{M})$ where $\mathbf{m} = \mathbf{X}\gamma$ and $\mathbf{M} = \Sigma + \mathbf{X}\Gamma\mathbf{X}^{\top}$.

Jeffreys Prior π_J

We consider $\theta = \{\beta, \sigma^2\}.$

Derivation:

$$I(\theta) = \begin{pmatrix} \frac{1}{\sigma^2} \mathbf{X}^\top \Sigma^{-1} \mathbf{X} & 0\\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}$$

 $\pi(\theta) \propto 1/(\sigma^2)^{p/2+1}$ if we assume β and σ^2 are dependent. $\pi(\theta) \propto 1/\sigma^2$ if β and σ^2 are independent $\pi(\beta, \sigma^2) = \pi(\beta)\pi(\sigma^2)$.

Posterior: The posterior for Jeffreys Prior falls back to the conjugate family:

Under Jeffreys priors $\pi(\beta, \sigma^2) \propto (\sigma^2)^{-m}$ with 2m = p + 2 or m = 1 assuming dependency between β and σ^2 or not. The posterior is then $(\beta, \sigma^2)|\mathbf{y} \sim \mathcal{NIG}(a_m, b, \gamma, \Gamma)$ with

$$a_{m} = 2m + n - p - 2$$

$$b = (\mathbf{y} - \mathbf{X}\hat{\beta}_{\Sigma})^{\top} \Sigma^{-1} (\mathbf{y} - \mathbf{X}\hat{\beta}_{\Sigma})$$

$$\gamma = (\mathbf{X}^{\top} \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}^{\top} \Sigma^{-1} \mathbf{y} = \hat{\beta}_{\Sigma} \text{ (Estimation of } \beta)$$

$$\Gamma = (\mathbf{X}^{\top} \Sigma^{-1} \mathbf{X})^{-1}$$

Additionally, if $\Sigma = \mathbf{I}_n$, $\hat{\beta}_{\Sigma}$ is identical to the least-square solution.

Parameter Estimation

Maximum Likelihood Estimator (MLE): $\hat{\theta}_{\text{MLE}}(x) = \underset{\theta \in \Theta}{\arg \max} \ell(\theta|x)$.

Maximum a Posteriori (MAP) Estimator

 $\hat{\theta}_{\text{MAP}} = \underset{\theta \in \Theta}{\arg \max} \pi(\theta|x)$

- Make estimation with the mode of the posterior
- Enough to know the kernel only $\pi(\theta|x) \propto \pi(\theta)\ell(\theta|x)$

Connections between MAP and MLE For a linear model $y = X\beta + \epsilon$ with parameter $\theta = \beta \in \mathbb{R}^p$, by rearranging $\hat{\theta}_{\text{MAP}} = \arg\max_{\theta \in \Theta} [\ln \ell(\theta|x) + \ln \pi(\theta)]$, we have $\hat{\theta}_{\text{MAP}}$ is equivalent to the regularised MLE: $\ln \pi(\theta) = \text{Pen}(\theta) + \text{const.}$. If $\pi(\theta)$ is flat/constant, $\hat{\theta}_{\text{MAP}} = \hat{\theta}_{\text{MLE}}$.

Bayes Rules Estimator

- $\hat{\theta}_{L2} = \mathbb{E}_{\pi(\theta|x)}\theta$: Minimise L_2 loss.
- $\hat{\theta}_{L_1} = \text{Median}_{\pi(\theta|x)}(\theta)$: Minimise L_1 loss.

For symmetric single mode posterior, $\hat{\theta}_{L2} = \hat{\theta}_{L1}$.

Credible Sets

Plays the similar rule as the confidence interval, to quantify the precision of the estimation.

- A set C_x is an α -credible region iff $\Pr(\theta_0 \in C_x | x) \ge 1 \alpha, \alpha \in [0, 1]$.
- The region is Highest Posterior Density α -credible region iff if can be written as $\{\theta \colon \pi(\theta|x) > k_{\alpha}\} \subset C_x \subset \{\theta \colon \pi(\theta|x) \geq k_{\alpha}\}$ where k_{α} is the largest bound. (Intuitively, HPD: The α -credible region with the shortest interval)

Predictions

We assume a Bayes model $\{\mathcal{P}, \pi\}$ which has the posterior $\pi(\theta|x)$. The future data x_f is generated by distribution Q_θ with probability function $q(x_f|\theta,x)$. Note generally $q(x_f|\theta,x) = q(x_f|\theta)$ unless it is autoregressive, future data depending on history data.

- 1. Define the prediction error $L_{\text{pred}}(x_f, d)$ for a pair of future data point x_f and prediction/decision d. Note: It is not the loss function.
- 2. Loss function is given by $L(\theta, d) = \int L_{\text{pred}}(x_f, d) q(x_f | \theta, x) dx_f$.
- 3. We get standard risk $R(\theta, \delta)$, Bayes risk $r(\delta, \pi)$, integrated posterior loss $\rho(d, \pi|x)$.

4. By minimising Bayes risk we obtain the Bayes predictor, which can be practically obtained from the integrated posterior loss.

Interestingly, we have

$$\rho(d, \pi | x) = \iint L_{\text{pred}}(x_f, d) q(x_f | \theta, x) \pi(\theta | x) d\theta dx_f \tag{1}$$

$$\int L_{\text{pred}}(x_f, d) \underbrace{\int q(x_f | \theta, x) \pi(\theta | x) d\theta}_{\text{predictive distribution } \pi(x_f | x)} dx_f \qquad (2)$$

- Predictive distribution $\pi(x_f|x)$ is the main tool for predictions: Point estimation and prediction region (similar to HPD credible region) are both base on $\pi(x_f|x)$.
- Similarly, using L_1 error for L_{pred} we have $\operatorname{Median}_{\pi(x_f|x)}(x_f)$ as the bayes estimator, and using L_2 error leads to $\mathbb{E}_{\pi(x_f|x)}x_f$.

Model Testing

Combine two models together with the indicator parameter k. Consider Bayes models $\mathcal{P}_i = \{P_{\theta_i}^i(x) : \theta_i \in \Theta_i\}$ with prior π_i for $i \in \{0, 1\}$. The common model is $\mathcal{P}_m = \{(1-k)P_{\theta_0}^0(x) + kP_{\theta_1}^1(x) : \theta_m := (k, \theta_0, \theta_1) \in \{0\} \times \Theta_0 \times \emptyset \cup \{1\} \times \emptyset \times \Theta_1\}$ with mixed prior $\pi_m(\theta_m) = \pi_k(k = 0)\pi_0(\theta_0) + (1-\pi_k(k = 1))\pi_1(\theta_1)$.

Interested in k, we have the posterior $\Pr(k=i|x) = \frac{\Pr(k=i)p(x=i|k=)}{\Pr(k=0)p(x|k=0)+\Pr(k=1)p(x|k=1)}$ for $i \in \{0,1\}$. Essentially we are interested in the evidence of each model $p(x|k=i) = \int \ell(\theta_i|x)\pi_i(\theta_i)d\theta_i$ for $i \in \{0,1\}$. We call the ratio of the evidences *Bayes Factor* p(x|k=0)/p(x|k=1).

Lazy Mathematicians' Methods

All you need is posterior.

Integration

Independent Monte Carlo Integration To integrate $\mathbb{E}_{p(\theta)}m(\theta) = \int m(\theta)p(\theta)d\theta$.

- 1. Draw samples from $p(\theta)$ as $\{\theta_{(1)}, \dots, \theta_{(N)}\}$.
- 2. $\hat{\mathbb{E}}_{p(\theta)}m(\theta) = \frac{1}{N} \sum_{i=1}^{N} m(\theta_i)$.

Sampling

We want to sample $p(\theta)$ with only access to its kernel $p(\theta) \propto k(\theta)$. **Importance Sampling** Rewrite the integration $p(\theta) = \text{Normalise}(g(\theta)\frac{k(\theta)}{g(\theta)})$.

- 1. Sample form $g(\theta)$ as $\{\theta_1, \ldots, \theta_N\}$.
- 2. Calculate associated importance weights $w_i = \frac{k(\theta_i)}{g(\theta_i)}$ for $i \in [N]$.
- 3. Standardise the importance weights $w_i^s = \frac{w_i}{\sum_{i=1}^N w_i}$.
- 4. Obtain weighted samples with weights: $\{(\theta_i, w_i^s)\}_{i=1}^N$.

With weighted samples, one can

- integrate $\int m(\theta)p(\theta) \approx \sum_{i=1}^{N} m(\theta_i)w_i^s$.
- resample from $\{\theta_i\}_{i=1}^N$ with corresponding probabilities $\{w_i^s\}_{i=1}^N$, resulting unweighted samples from $p(\theta)$ directly as $\{\nu_i, \dots, \nu_M\}$.

Note: Brute-force is a special case of importance sampling, with $g(\theta) \propto {\rm const.}.$

Rejection Algorithm

- 1. For $g(\theta)$, pick a constant M s.t. $Mg(\theta) \geq k(\theta)$ for all $\theta \in \Theta$.
- 2. Sample from $g(\theta)$ as $\{\theta_1, \dots, \theta_N\}$.
- 3. Accept θ_i with probability $\Pr_i(\text{accept}) = \frac{k(\theta_i)}{Ma(\theta_i)}$.

MCMC (Metropolis-Hastings) For $i \in [N]$:

- 1. Draw ν from $T(\theta_{i+1}|\theta_i)$.
- 2. Compute acceptance criteria $r(\nu, \theta_i) = \frac{k(\nu)T(\nu|\theta_i)}{k(\theta_i)T(\theta_i|\nu)}$.
- 3. $\theta_{i+1} = \nu$ with $Pr_i(accept) = r(\nu, \theta_i)$.

Approximate Bayesian Computation (ABC) Not even kernel! We can only generate data from $p(x|\theta)$ for $\theta \in \Theta$.

- 1. Generate θ from π and generate x_{new} .
- 2. Accept it if $d(S(x_{\text{new}}), S(x))$ is small enough, where $S(\cdot)$ denotes sufficient statistics and d denotes a metric.