Marginalization, Prediction, Decisions, Exponential family

PhD course in Statistical Inference

Mattias Villani

Department of Statistics
Stockholm University
and
Department of Computer and Information Science
Linköping University

- **Marginalization**
- **Prediction**
- **Decision making**
- Bayesian inference for the **exponential family**

- Models with **multiple parameters** $\theta_1, \theta_2, \ldots$.
- Examples: $x_i \stackrel{iid}{\sim} N(\theta, \sigma^2)$; multiple regression ...
- **Joint posterior distribution**

$$p(\theta_1, \theta_2, \ldots, \theta_p|y) \propto p(y|\theta_1, \theta_2, \ldots, \theta_p)p(\theta_1, \theta_2, \ldots, \theta_p).$$

$$p(\theta|y) \propto p(y|\theta)p(\theta).$$

- **Marginalize** out parameter of no direct interest (**nuisance**).
- Example: $\theta = (\theta_1, \theta_2)'$. **Marginal posterior** of θ_1

$$p(\theta_1|y) \quad = \quad \int p(\theta_1, \theta_2|y)d\theta_2 = \int p(\theta_1|\theta_2, y)p(\theta_2|y)d\theta_2.$$

- **Model**

$$x_1, \ldots, x_n | \theta, \sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2)$$

- **Prior**

$$p(\theta, \sigma^2) \propto (\sigma^2)^{-1}$$

- **Joint posterior**

$$\theta | \sigma^2, \mathbf{x} \sim N\left(\bar{x}, \frac{\sigma^2}{n}\right)$$

$$\sigma^2 | \mathbf{x} \sim \mathrm{Inv} - \chi^2(n-1, s^2),$$

  where

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

  is the usual sample variance.

- **Marginal posterior** of θ

$$\theta | \mathbf{x} \sim t_{n-1}\left(\bar{x}, \frac{\sigma^2}{n}\right)$$

- **Posterior predictive density** for future $\tilde{y}$ given observed $\mathbf{y}$

$$p(\tilde{y}|\mathbf{y}) = \int_\theta p(\tilde{y}|\theta, \mathbf{y}) p(\theta|\mathbf{y}) d\theta$$

- If $p(\tilde{y}|\theta, \mathbf{y}) = p(\tilde{y}|\theta)$ [not true for time series], then

$$p(\tilde{y}|\mathbf{y}) = \int_\theta p(\tilde{y}|\theta) p(\theta|\mathbf{y}) d\theta$$

- **Parameter uncertainty** in $p(\tilde{y}|\mathbf{y})$ by **averaging over** $p(\theta|\mathbf{y})$.

- Predictive distribution is normal (next slide).
- Remember the posterior: $\theta|\mathbf{y} \sim N(\mu_n, \tau_n^2)$.
- Law of iterated expectation:

$$E(\tilde{y}|\mathbf{y}) = E_{\theta|\mathbf{y}}[E_{\tilde{y}|\theta}(\tilde{y})] = E_{\theta|\mathbf{y}}(\theta) = \mu_n$$

- The predictive variance of $\tilde{y}$ (total variance formula):

$$\begin{aligned} V(\tilde{y}|\mathbf{y}) &= E_{\theta|\mathbf{y}}[V_{\tilde{y}|\theta}(\tilde{y})] + V_{\theta|\mathbf{y}}[E_{\tilde{y}|\theta}(\tilde{y})] \\ &= E_{\theta|\mathbf{y}}(\sigma^2) + V_{\theta|\mathbf{y}}(\theta) \\ &= \sigma^2 + \tau_n^2 \end{aligned}$$

- In summary:

$$\tilde{y}|\mathbf{y} \sim N(\mu_n, \sigma^2 + \tau_n^2).$$

Simulation algorithm:

1. Generate a **posterior draw** of θ ($\theta^{(1)}$) from $N(\mu_n, \tau_n^2)$
2. Generate a **predictive draw** of $\tilde{y}$ ($\tilde{y}^{(1)}$) from $N(\theta^{(1)}, \sigma^2)$
3. Repeat Steps 1 and 2 N times to output:
   - Sequence of posterior draws: $\theta^{(1)}, ..., \theta^{(N)}$
   - Sequence of predictive draws: $\tilde{y}^{(1)}, ..., \tilde{y}^{(N)}$.

- Note: $\tilde{y}^{(1)} = \theta^{(1)} + \sigma Z_1 = (\mu_n + \tau_n Z_2) + \sigma Z_1$ where Z_1, Z_2 are $N(0, 1)$. So $\tilde{y}^{(1)}$ is normal.

- **Autoregressive process**

$$y_t = \mu + \phi_1(y_{t-1} - \mu) + \ldots + \phi_p(y_{t-p} - \mu) + \varepsilon_t, \ \varepsilon_t \overset{iid}{\sim} N(0, \sigma^2)$$

Simulation algorithm. Repeat N times:

1. Generate a **posterior draw** of $\theta^{(1)} = (\phi_1^{(1)}, \ldots, \phi_p^{(1)}, \mu^{(1)}, \sigma^{(1)})$ from $p(\phi_1, \ldots, \phi_p, \mu, \sigma | \mathbf{y}_{1:T})$.
2. Generate a **predictive draw** of future time series by:
   2.1 $\tilde{y}_{T+1} \sim p(y_{T+1} | y_T, y_{T-1}, \ldots, y_{T-p}, \theta^{(1)})$
   2.2 $\tilde{y}_{T+2} \sim p(y_{T+2} | \tilde{y}_{T+1}, y_T, \ldots, y_{T-p}, \theta^{(1)})$
   2.3 $\tilde{y}_{T+3} \sim p(y_{T+3} | \tilde{y}_{T+2}, \tilde{y}_{T+1}, y_T, \ldots, y_{T-p}, \theta^{(1)})$
   2.4 ...

- Let θ be an **unknown quantity**. **State of nature**. Examples: Future inflation, Global temperature, Disease.
- Let $a \in \mathcal{A}$ be an **action**. Ex: Interest rate, Energy tax, Surgery.
- Choosing action a when state of nature is θ gives **utility**

$$U(a, \theta)$$

- Example:
  - θ is the number of items demanded of a product
  - a is the number of items in stock
  - Utility

$$U(a, \theta) = \begin{cases} p \cdot \theta - c_1(a - \theta) & \text{if } a > \theta \text{ [too much stock]} \\ p \cdot a - c_2(\theta - a)^2 & \text{if } a \leq \theta \text{ [too little stock]} \end{cases}$$

Optimal Bayesian decisions

- Ad hoc decision rules: *Minimax*. *Minimax-regret* etc
- **Bayesian theory**: maximize the **posterior expected utility**:

$$a_{bayes} = \mathrm{argmax}_{a \in \mathcal{A}} \, E_{p(\theta|y)}[U(a, \theta)],$$

  where $E_{p(\theta|y)}$ denotes the posterior expectation.
- Using simulated draws $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(N)}$ from $p(\theta|y)$:

$$E_{p(\theta|y)}[U(a, \theta)] \approx N^{-1} \sum_{i=1}^{N} U(a, \theta^{(i)})$$

- **Separation principle**:

1. First obtain $p(\theta|y)$
2. then form $U(a, \theta)$ and finally
3. choose a that maximes $E_{p(\theta|y)}[U(a, \theta)]$.

- **Model**

$$y_1, ..., y_n|\theta \overset{iid}{\sim} Pois(\theta)$$

- **Poisson distribution**

$$p(y) = \frac{\theta^y e^{-\theta}}{y!}$$

- **Likelihood** from iid Poisson sample $y = (y_1, ..., y_n)$

$$p(y|\theta) = \left[\prod_{i=1}^n p(y_i|\theta)\right] \propto \theta^{(\sum_{i=1}^n y_i)} \exp(-\theta n),$$

- ***Prior***

$$p(\theta) \propto \theta^{\alpha-1} \exp(-\theta\beta) \propto Gamma(\alpha, \beta)$$

which contains the info: $\alpha - 1$ counts in β observations.

- **Posterior**

$$\begin{aligned} p(\theta|y_1,...,y_n) &\propto \left[\prod_{i=1}^n p(y_i|\theta)\right] p(\theta) \\ &\propto \theta^{\sum_{i=1}^n y_i} \exp(-\theta n)\theta^{\alpha-1}\exp(-\theta\beta) \\ &= \theta^{\alpha+\sum_{i=1}^n y_i - 1}\exp[-\theta(\beta+n)], \end{aligned}$$

proportional to the $Gamma(\alpha+\sum_{i=1}^n y_i, \beta+n)$ distribution.

- **Prior-to-Posterior mapping**

$$\text{Model: } y_1,...,y_n|\theta \overset{iid}{\sim} Pois(\theta)$$
$$\text{Prior: } \theta \sim Gamma(\alpha,\beta)$$
$$\text{Posterior: } \theta|y_1,...,y_n \sim Gamma(\alpha+\sum_{i=1}^n y_i, \beta+n).$$

Poisson example - Bomb hits in London

$n = 576$, $\sum_{i=1}^{n} y_i = 229 \cdot 0 + 211 \cdot 1 + 93 * 2 + 35 * 3 + 7 * 4 + 1 \cdot 5 = 537$.

Average number of hits per region=$\bar{y} = 537/576 \approx 0.9323$.

$$p(\theta|y) \propto \theta^{\alpha+537-1} \exp[-\theta(\beta + 576)]$$

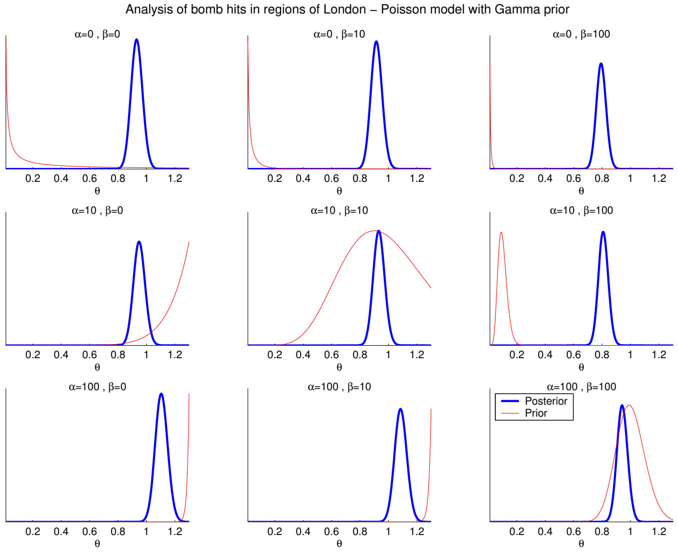
$$E(\theta|y) = \frac{\alpha + \sum_{i=1}^{n} y_i}{\beta + n} \approx \bar{y} \approx 0.9323,$$

and

$$SD(\theta|y) = \left(\frac{\alpha + \sum_{i=1}^{n} y_i}{(\beta + n)^2}\right)^{1/2} = \frac{(\alpha + \sum_{i=1}^{n} y_i)^{1/2}}{(\beta + n)} \approx \frac{(537)^{1/2}}{576} \approx 0.0402.$$

if α and β are small compared to $\sum_{i=1}^{n} y_i$ and n.

Poisson bomb hits in London

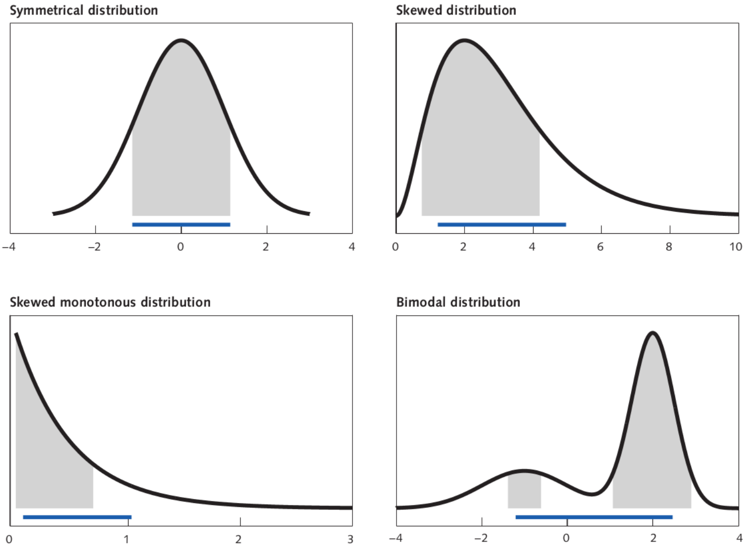


- **Bayesian 95% credible interval**: the probability that the unknown parameter θ lies in the interval is 0.95.
- Approximate 95% **credible interval** for θ (for small α and β):

$$E(\theta|y) \pm 1.96 \cdot SD(\theta|y) = [0.8535; 1.0111]$$

- An exact 95% **equal-tail interval** is $[0.8550; 1.0125]$ (assuming $\alpha = \beta = 0$)
- **Highest Posterior Density** (**HPD**) interval contains the θ values with highest pdf.
- An exact Highest Posterior Density (HPD) interval is $[0.8525; 1.0144]$. Obtained numerically, assuming $\alpha = \beta = 0$.

Illustration of different interval types



- Normal likelihood: Normal prior$\rightarrow$Normal posterior.
- Bernoulli likelihood: Beta prior$\rightarrow$Beta posterior.
- Poisson likelihood: Gamma prior$\rightarrow$Gamma posterior.

- **Conjugate priors**: A prior is conjugate to a model if the prior and posterior belong to the same distributional family.

- Formal definition: Let $\mathcal{F} = \{p(y|\theta), \theta \in \Theta\}$ be a class of sampling distributions. A family of distributions $\mathcal{P}$ is **conjugate** for $\mathcal{F}$ if

$$p(\theta) \in \mathcal{P} \Rightarrow p(\theta|\mathbf{x}) \in \mathcal{P}$$

holds for all $p(y|\theta) \in \mathcal{F}$.

- **Exponential family** in the canonical parametrization

$$p(x|\theta) = h(x) \exp\left(\theta^T \mathbf{t}(x) - A(\theta)\right)$$

where $A(\theta) = -\ln a(\theta)$ in Rolf's notation.

- **Likelihood**

$$p(x_1, \ldots, x_n|\theta) = \left[\prod_{i=1}^{n} h(x_i)\right] \exp\left(\theta^T \sum_{i=1}^{n} \mathbf{t}(x_i) - nA(\theta)\right)$$

- **Conjugate prior**

$$p(\theta) = H(\tau_0, n_0) \exp\left(\theta^T \tau_0 - n_0 A(\theta)\right),$$

where τ_0 and n_0 are prior hyperparameters and $H(\tau_0, n_0)$ is the normalizing constant which is known to exist if $n_0 > 0$.

- **Conjugate prior**

$$p(\theta) = H(\tau_0, n_0) \exp\left(\theta^T \tau_0 - n_0 A(\theta)\right)$$

- **Posterior**

$$p(\theta|x_1, ..., x_n) \propto \exp\left[\theta^T \left(\tau_0 + \sum_{i=1}^{n} \mathbf{t}(x_i)\right) - (n_0 + n)A(\theta)\right]$$

- **Prior-to-posterior updating**

$$\tau_0 \Longrightarrow \tau_n = \tau_0 + \sum_{i=1}^{n} \mathbf{t}(x_i)$$

$$n_0 \Longrightarrow n_0 + n$$

- **Exponential family** in the non-canonical parametrization

$$p(x|\theta) = h(x)\exp\left(\phi(\theta)^T\mathbf{t}(x) - A(\theta)\right)$$

- **Conjugate prior**

$$p(\theta) = H(\tau_0, n_0)\exp\left(\phi(\theta)^T\tau_0 - n_0A(\theta)\right)$$

- **Bernoulli likelihood**

$$\begin{aligned} p(x_1, \ldots, x_n|\theta) &= \prod_{i=1}^{n}\theta^{x_i}(1-\theta)^{1-x_i} \\ &= \exp\left(\log\left(\frac{\theta}{1-\theta}\right)\sum_{i=1}^{n}x_i - n\log\left(\frac{1}{1-\theta}\right)\right) \\ &= \exp\left(\phi(\theta)\sum_{i=1}^{n}x_i - nA(\theta)\right) \end{aligned}$$

where $\phi = \log\left(\frac{\theta}{1-\theta}\right)$ and $A(\theta) = \log\left(\frac{1}{1-\theta}\right)$.

- **Conjugate prior $p(\phi)$**

$$\exp\left(\phi(\theta)\tau_0 - n_0A(\theta)\right) = \exp\left(\log\left(\frac{\theta}{1-\theta}\right)\tau_0 - n_0\log\left(\frac{1}{1-\theta}\right)\right) = \theta^{\tau_0}(1-\theta)^{n_0-\tau_0}$$

- **Prior mean**

$$\mathrm{E}(\mu) = \tau_0 / n_0$$

- **Posterior mean**

$$\mathrm{E}(\mu | x_1, ..., x_n) = \frac{\tau_0 + \sum_{i=1}^n \mathbf{t}(x_i)}{n_0 + n} = w\tau_0 + (1 - w)\hat{\mu}_{ML},$$

where $\hat{\mu}_{ML} = \frac{1}{n} \sum_{i=1}^n \mathbf{t}(x_i)$ and $w = n_0 / (n_0 + n)$.

- **Predictive distribution** of x_{n+1}

$$\begin{aligned} p(x_{n+1} | x_{1:n}) &= \int p(x_{n+1}|\theta) p(\theta | x_{1:n}) d\theta \\ &= \int h(x_{n+1}) \exp\left(\theta^T \mathbf{t}(x_{n+1}) - A(\theta)\right) H(\tau_n, n_0 + n) \exp\left(\theta^T \tau_n - (n_0 + n) A(\theta)\right) d \\ &= h(x_{n+1}) H(\tau_n, n_0 + n) \int \exp\left(\theta^T (\tau_n + \mathbf{t}(x_{n+1})) - (n_0 + n + 1) A(\theta)\right) d\theta \\ &= h(x_{n+1}) H(\tau_n, n_0 + n) / H(\mathbf{t}(x_{n+1}) + \tau_n, n_0 + n + 1) \end{aligned}$$