III Conjugate priors and exponential family.

III (1) intro

- Often, computing the posterior density is difficult because of $D = \int_{(\Theta)} P_t(x)\,\pi(t)\,d\mu(t)$, which may not have an analytical expression.
- Estimating the density via numerical methods approaching D is not that useful because what matters is usually rather the mean or the posterior quantiles.
- Sometimes (as in our leading example) everything goes well: the prior is parameterized by hyper parameters (α, β in our example of a $\mathrm{Beta}(\alpha,\beta)$ distribution).

definition Hyper parameters

> when the prior π is chosen in a family $\mathcal{F} = \{\pi_\gamma,\ \gamma \in \Gamma\}$, with $\Gamma \subset \mathbb{R}^d$, the parameter γ characterizing π_γ is called "hyper-parameter".

In our example, $\Gamma = \mathbb{R}_+^* \times \mathbb{R}_+^*$. and
$\mathcal{F} = \{\mathrm{Beta}(\alpha,\beta),\ (\alpha,\beta) \in \Gamma\}$ is the family of all Beta distributions.
Moreover, the posterior $\pi(\theta|x)$ belongs to the same family Γ, namely
$\pi(\theta|x) = \mathrm{Beta}(\alpha + \sum_{i=1}^n x_i, \beta + n - \sum_{i=1}^n x_i)$. Such a prior is called conjugate

Definition Conjugate prior

> a family $\mathcal{F}$ of probability distributions over Θ is "conjugate" for a likelihood $p_\theta(x)$ if $\forall x \in \mathcal{X}$, $\pi(\theta|x) \in \mathcal{F}$

III (2) examples

(a) Gaussian model with known variances
$\Theta = \mathbb{R}$, $p_\theta(x) =$

$$\propto \exp\{\text{quadratic function of } \theta\}$$
as a function of θ.

if $\pi(\theta) \propto \exp\{ \text{— " —} \}$

then $\pi(\theta|x) \propto p_\theta(x)\,\pi(\theta)$
$\qquad \propto \exp\{\text{quad. fn "}\}$

but if $\pi(\theta) \propto \exp\{\text{quad. " —}\}$
then $\pi(\theta) = \mathcal{N}(\mu, \delta^2)$ for some μ, δ^2.

details $\quad \pi(\theta) \propto \exp\{-\frac{1}{2s^2}(\theta-\mu)^2\}$, $x=(x_1 \dots x_n)$

$$\pi(\theta|x) \propto \exp\left\{-\frac{1}{2}\left\{\frac{\sum_i (x_i-\theta)^2}{\sigma^2} + \frac{(\theta-\mu)^2}{s^2}\right\}\right\}$$

$$= \exp\left\{-\frac{1}{2}\left(\theta^2\left(\frac{n}{\sigma^2}+\frac{1}{s^2}\right) - 2\theta\left(\frac{\sum x_i}{\sigma^2}+\frac{\mu}{s^2}\right) + C\right)\right\}$$

$$\propto \exp\left\{-\frac{1}{2\left[\frac{n}{\sigma^2}+\frac{1}{s^2}\right]^{-1}}\left[\theta^2 - 2\theta\left(\frac{\sum x_i}{\sigma^2}+\frac{\mu}{s^2}\right)\frac{s^2\sigma^2/n}{s^2+\sigma^2/n}\right]\right\}$$

$\propto \mathcal{N}(\mu_n, \sigma_n^2)$ with

$$\begin{cases} \mu_n = \left(s^2 \, \frac{1}{n}\sum x_i + \frac{\sigma^2}{n}\mu\right) \times \frac{1}{s^2+\sigma^2/n} \\ 1/\sigma_n^2 = \frac{1}{s^2} + \frac{n}{\sigma^2} \end{cases}$$

rem : • maximum likelihood estimator :

$$\hat{\mu}_{ML} = \frac{1}{n}\sum x_i$$

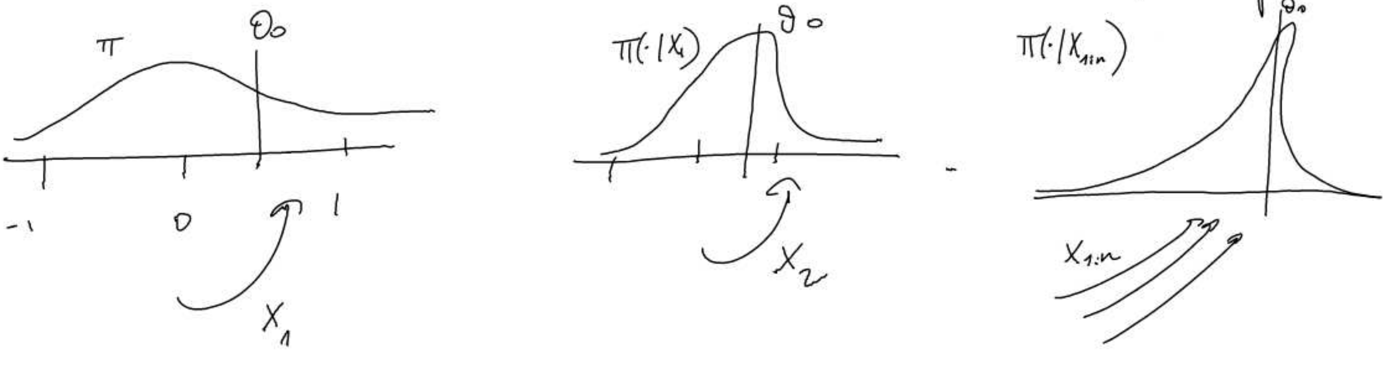
• Posterior mean, $\mu_n = \left(s^2 \hat{\mu}_{ML} + \frac{\sigma^2}{n}\mu\right)\frac{1}{s^2+\sigma^2/n}$: weighted average. (regularized version)

• $n=0$: $\mu_n = \hat{\mu}_{ML}$
$\sigma_n^2 = s^2$

• $n \to \infty$: $\mu_n \sim \hat{\mu}_{ML}$ ($\to \mu_0$ true mean)
$\sigma_n^2 \sim \frac{\sigma^2}{n} \to 0$: posterior becomes peaked around $\hat{\mu}_{ML}$

• $s^2 \to \infty$: $\sigma_n^2 \sim \frac{\sigma^2}{n}$.

• Illustration : concentration of the posterior distribution around the true parameter



(5) Gaussian model, unknown variance, known mean μ

Conjugate prior? Working with $\lambda = \frac{1}{\sigma^2}$.

$\theta = \lambda$ $\qquad p_\theta^{\otimes n}(x) \propto \lambda^{n/2} \exp\left\{ -\frac{\lambda}{2} \sum_i (x_i - \mu)^2 \right\}$

$$\propto \lambda^{n/2} \exp\left\{ -\frac{\lambda}{2} S^2 \right\} \qquad S^2 = \sum (x_i - \mu)^2.$$

remind the Gamma distribution: $(a, b > 0)$

$$\mathcal{G}_{a,b}(\lambda) = \frac{1}{\Gamma(a)} \lambda^{a-1} \exp\{-b\lambda\} \; : \; E_{ab}(\lambda) = \frac{a}{b}, \; \mathrm{Var}_{ab}(\lambda) = \frac{a}{b^2}$$

thus $P_\lambda(x) \propto \mathcal{G}\left(\frac{n}{2}+1, \frac{S^2}{2}\right)$

take $\pi(\lambda) = \mathcal{G}_{a,b}(\lambda)$

then $$\pi(\lambda|x) \propto \lambda^{a-1+\frac{n}{2}} \exp\left\{ -\lambda\left(\frac{S^2}{2} + b \right) \right\}$$
$$= \mathcal{G}_{\left\{a+\frac{n}{2},\; b+\frac{S^2}{2}\right\}}(\lambda)$$

$$\Rightarrow \begin{cases} \mathbb{E}(\lambda|x) = \dfrac{a+n/2}{b+S^2/2} & \rightarrow \hat{\lambda} = \dfrac{a+n/2}{b+S^2/2} \\ \mathrm{Var}(\lambda|x) = \dfrac{a+n/2}{(b+S^2/2)^2} \end{cases}$$

. Remark: $\sigma^2_{ML} = \frac{S^2}{n}$ if we take

$\hat{\sigma} = \frac{1}{\hat{\lambda}_{mean}}$, we get $\hat{\sigma} = \dfrac{b + S^2/2}{a + n/2}$ $\rightarrow$ regularized version of $\hat{\sigma}_{ML}$

(c) Gaussian model, mean and variance unknown.

$\theta = (\mu, \lambda) \qquad \lambda = \frac{1}{\sigma^2}$.

$$P_{(\mu,\lambda)}(x) = \prod_{i=1}^{n} \left(\frac{\lambda}{2\pi}\right)^{\frac{1}{2}} \exp\left\{-\frac{\lambda}{2}(x_i-\mu)^2\right\}$$

$$\propto \left[\lambda^{\frac{1}{2}} \exp\left\{-\frac{\lambda\mu^2}{2}\right\}\right]^n \exp\left\{\lambda\mu \sum_{i=1}^{n} x_i - \frac{\lambda}{2}\sum_{i=1}^{n} x_i^2\right\}$$

We need a prior $\pi(\mu,\lambda)$ with a similar form

$$\pi(\mu,\lambda) \propto \left[\lambda^{\frac{1}{2}} \exp\left\{-\frac{\lambda\mu^2}{2}\right\}\right]^{\beta} \exp\left\{c\lambda\mu - d\lambda\right\}$$

$$= \lambda^{\beta/2} \exp\left\{-\frac{\lambda\beta}{2}\left(\mu - \frac{c}{\beta}\right)^2\right\} \exp\left\{-\lambda\left(d - \frac{c^2}{2\beta}\right)\right\}$$

writes $\pi(\mu,\lambda) = \pi_1(\mu|\lambda)\, \pi_2(\lambda)$

with $\pi_1(\mu|\lambda)$ gaussian $\left(\mu_0 = \frac{c}{\beta}\right.$, $\left.1/\sigma_0^2 = \beta\lambda\right)$

$$\pi_2(\lambda) = \mathcal{G}amma\left(\underbrace{\frac{\beta}{2}+1}_{a},\ \underbrace{d-\frac{c^2}{2\beta}}_{b}\right)$$

"Normal-Gamma model":

$$\begin{cases} \lambda \sim \mathcal{G}(a,b) \\ \mu/\lambda \sim \mathcal{N}\left(\mu_0,\ \sigma_0^2 = \frac{1}{\lambda\beta}\right) \end{cases}$$

$\to$ "Normal Gamma" (μ_0, β, a, b)

<u>Posterior</u>: $a_N = a + \frac{N}{2}$; $b_N = b + \frac{1}{2}\left\{ s^2 + \beta\frac{n(\bar{x}-\mu_0)^2}{\beta+n}\right\}$; $\mu_n = \frac{\beta\mu_0+n\bar{x}}{\beta+n}$; $\beta_n = \beta+n$.

$s^2 = \sum(x_i-\bar{x})^2$; $\bar{x} = \frac{1}{n}\sum x_i$.

(d) Multivariate case: conjugate priors $\quad X \sim \mathcal{N}(\mu,\Sigma) \in \mathbb{R}^D$

(i) unknown mean: $\longrightarrow$ $\pi(\mu)$: multivariate gaussian is conjugate

(ii) unknown covariance:

$$\Lambda = \Sigma^{-1} \qquad \mathcal{W}(\Lambda) = B|\Lambda|^{(\nu-D-1)/2} \exp\left\{-\frac{1}{2}\mathrm{Tr}(W^{-1}\Lambda)\right\}$$

"Wishart distribution with ν degrees of freedom $\in \mathbb{R}^+$

$W \in \mathbb{R}^{D\times D}$; B = normalization constant